Application of machine learning techniques for regional bias correction of SWE estimates in Ontario, Canada

Response letter to all reviewers

Fraser King, Andre E. Erler, Steven K. Frey, Christopher G. Fletcher - May 2020

Reviewer 1

1) The results indicate that there is a clear difference in SNOWDAS agreement (against in situ SWE) in the period 2011-2013 and 2014-onwards. It will be interesting to see/understand why? Is it the change in assimilation frequency, sources used in as similation, their accuracy? I think such understanding can then support the selection of approach used for bias correction. It has some implications also for the design of this study. If there is a step change in SNOWDAS, then it is not surprising that simple mean subtraction method is not working well for the entire period. It will be interesting to see why does the random forest outperform the other methods in such case and what factors are controlling its efficiency? (Is it because using year of observation?) Will it be not more fair in this case to compare the methods in two separate periods?

We agree with the reviewer that the change in bias post-2014 is of interest, and we mention on lines 27-30 of section 4.2 that newly assimilated datasets are likely the dominant contributing factors to the reduction in the intensity of the SNODAS bias during this period. We argue that while the bias is reduced post-2014, it is still non-zero and the approaches explored in our work continue to provide improvements to SNODAS estimates during this time. The decision tree and random forest approaches outperform traditional methods like SLR and mean bias subtraction due to this nonlinearity in the bias and the ability for the machine learning techniques to recognize these patterns and better correct for them. As shown in the predictor importance scores of table 2, year does play a somewhat important factor along with other climatic variables like temperature and total precipitation.

We agree with the reviewer that further descriptions of bias correction model performance (with respect to bias and RMSE) when trained/tested over these

two separate periods (before and after 2014) would be beneficial, and therefore additional text describing the results of these comparisons has been added to the manuscript in section 3.3 (pages 9-10).

2) I think that the referencing (used in the Introduction and Discussion) can be improved. There are some relevant papers which are not addressed: e.g. Zahmatkesh et al. (2019) evaluating bias correction of SNODAS in Canadian basins or some studies cited in Lv et al. (2019) focusing on the accuracy assessment of SNODAS. Please consider to formulate how does this study compare to these studies (in Intro and Discussion sections).

We thank the reviewer for recommending these relevant papers from Zahmatkesh et al. (2019) and Lv et al. (2019). These references have been added in the manuscript as additional motivation to our work in section 1 (page 2).

3) I have to say that the part related to evaluation of the impacts of different bias corrected SWE estimates on snowmelt is not clear to me. Using monthly estimates without accounting for evapotranspiration and other processes is somewhat less robust. Comparison of observed daily discharge with daily simulations driven by a hydrologic model will be more representative example.

We thank the reviewer for this comment, as this point may not be immediately obvious: a direct comparison between SWE estimates and streamflow is not straight forward and presents a major methodological challenge, as outlined below. The primary purpose of this section (and Figure 7) is to demonstrate that SNODAS SWE values are clearly too high and unphysical, especially during the time period before 2015, where estimated snowmelt exceeds total spring runoff in several cases. After bias-correction this is not the case anymore, suggesting that the bias-corrected values are at least plausible.

The methodological challenge preventing direct validation of SWE estimates against streamflow gauges is the fact that runoff is generated by snowmelt and snowmelt has to be estimated from SWE changes. However, SWE also changes due to snow fall (and sublimation); snow fall, sublimation and melt occurring during the same time period cannot be separated easily (and can cancel each other). A better estimate of melt and runoff therefore would require additional data on precipitation, precipitation phase and/or temperature at high temporal frequency and a series of non-obvious judgements (such as estimating sublimation) would be required. This could be a topic of a potential follow-up study, but is beyond the scope of this manuscript.

A hydrologic or land surface model, which would be necessary to properly account for sublimation and evapotranspiration, would not be helpful for this purpose, as these models compute snowpack internally and one would be left with a comparison against modeled snowpack (SWE). Furthermore, if SWE values from SNODAS were to be assimilated into the model, melt and runoff values would potentially be worse, since data assimilation violates mass conservation.

As a case in point, we note that SNODAS also computes snowmelt internally, however, these values suffer from biases even larger than the biases in SWE. The reason for this is likely that snowmelt is not assimilated and at the same time artifacts are introduced by the assimilation of other variables (mass conservation is violated). Unfortunately, direct observation of snowmelt is not possible.

4) How to account for scale gap between SNODAS and in situ observations?

In our analysis, we compare gridded estimates of SWE from SNODAS (1 km resolution) to snow survey estimates (which is essentially point data taken over 10 m). Due to the relatively high spatial resolution of SNODAS, along with the fact that the in situ measurement sites are taken at distances > 1 km from each other, we do not compare multiple in situ points to a single grid cell. This allows us to complete a simple point to grid cell comparison where we assume the snow survey SWE estimate is representative of the wider, containing grid cell.

This assumption of representativeness across the grid cell introduces additional uncertainty, as SWE is highly variable at even small spatial scales, and we have therefore included additional details in the paper to make these uncertainties clearer to the reader in section 4.2 (page 12).

5) Fig.1b. What do the lines represent? Mean over 383 stations?

The reviewer is correct, the lines in Figure 1.b represent the daily mean SWE on ground for all survey locations (383 sites) across the full study period.

6) Fig.2,3,4,5. Please explain the meaning of abbreviations MBS, SLR, etc. in figure caption.

We thank the reviewer for this comment, and we have included an additional description of the abbreviations for MBS, SLR, DT and RF in the caption of Figure 2.

Reviewer 2

1) First of all and most important the applied machine learning methods are not described at all and references are missing. I don't think that all readers of this journal are familiar with Decision Trees (DT) and Random Forest (RF) methods. Therefore a short description should be included, especially explaining the RF model in more detail, which shows the best results, and what's the difference to the DT models. Related to that comment, it doesn't make too much sense to mention on page 5 (line 30) that you run the model with a forest size of 100 trees and tree depth of 15, when you don't explain what that parameter mean.

We thank the reviewer for this comment, and agree that additional details should be included in the text which further describe the methodology behind the decision tree (DT) and random forest (RF) techniques we employ in this work.

We have updated the document to include further references/details regarding what these techniques are and how they operate, along with further descriptions of what parameters like forest size and tree depth mean with respect to the RF model in section 2.3 of the manuscript (page 6).

2) Additionally, there are some points which are not clear to me and which should be comment clarified before publishing the paper: You didn't explain how you handled the scaling issue when you compare point data and gridded data (up- or downscaling?). Since you could identify a change in the bias between the first and the second half of the period, it would be reasonable to split the analysis into these two periods and fit different models and take 2 different means separately for each period.

In our analysis, we compare gridded estimates of SWE from SNODAS (1 km resolution) to snow survey estimates (which is essentially point data taken over 10 m). Due to the relatively high spatial resolution of SNODAS, along with the fact that the in situ measurement sites are taken at distances > 1 km from each other, we do not compare multiple in situ points to a single grid cell. This allows us to complete a simple point to grid cell comparison where we assume the snow survey SWE estimate is representative of the wider, containing grid cell. The snow survey sites are selected to generally be representative of the area around it and are not just random point measurements which would contain higher variability in their estimates.

However, this assumption of representativeness across the grid cell introduces additional uncertainty, as SWE is highly variable at even small spatial scales, and we have therefore included additional details in the paper (section 4.2, page 12) to make these uncertainties clearer to the reader. Furthermore, we agree with the reviewer that due to the change in the intensity of the bias post-2014, a description of how the bias correction models perform over these separate two

periods would be interesting and complimentary to our analysis. Therefore, we have also updated the results section 3.3 (pages 9-10) of the paper with the results of this test.

3) On page 3 you specify the 383 locations with in situ measurements. In line 14-15 you write that an average SWE is estimated taken from 10 fixed sampling stations. What does this mean? Is this the average SWE for Ontario estimated from 10 stations, or is this the average for each of the 383 stations taken from the 10 surrounding stations??

What we are referring to on page 3 is the method by which in situ measurements are retrieved (snow survey), where a sampling location is selected and then 10 point measurements are taken using a snow coring device over approximately 10 meters at that location. These 10 SWE measurements along the snow survey are then averaged together to provide a single SWE estimate for that location. This is the technique used at all 383 in situ measurement sites.

We now include additional details on how these measurements are retrieved to add further clarity to the reader in section 2.1 (page 3) of the manuscript.

4) Page 5: You should mention that the period of 1981-2010 is used for calculating the climatology, which is not clear.

We thank the reviewer for noticing this detail, and we now make the temporal period used for the calculation clear in the paper on page 5 (section 2.2.2).

5) Also, you should explain why you have used the difference between the precipitation estimates from NRCAN and the SNODAS! It would be interesting to see the results if you would include actual meteorological observations as predictors (for example available at: https://data.noaa.gov/dataset/dataset/globalsurface-summary-of-the-day-gsod, provided by the National Centers for Environment Information). I could imagine that in that case the importance of these variables would not be neglectable and could further improve the bias correction.

We thank the reviewer for the comment; we have also considered this option; however, we have chosen to limit meteorological data to basic monthly climate normals. Analogous to the choice of model complexity, there is always a trade-off between accuracy, complexity and the risk of over-fitting. Using a large set of predictors requires a more complex model, which increases the risk of over-fitting. Therefore we have chosen to only include monthly normal surface temperature and precipitation, as these two variables are usually readily available and characterize the type of climate reasonably well. The rational behind including climate variables was that, on average, snow characteristics (like density, albedo, ice content) vary between different climates. It is true that these characteristics would be predictable (to some extend) from the actual evolution of these meteorological forcings; however, the processes that govern such characteristics are very complex and involve long-term memory effects, which would

require a much more complex model (like an LSTM), which would approach the complexity of physical snow models. Considering the data requirements and complexity of this approach, we believe that the use of monthly normals represents the best compromise.

As for the reason, the difference between SNODAS average precipitation and NRCan normals was used, rather than total precipitation from NRCan (or SNODAS): this choice was made because notable biases in the precipitation fields used by SNODAS over Canada were found early on in the analysis, and it appears obvious that the size of these biases would have a first-order effect on the resulting SWE bias in SNODAS. At the same time, in order to reduce the number of input variables, we did not want to include multiple, possibly redundant, precipitation variables.

6) Page 7: When you write in 3.2.1 about mean bias, I suppose that this mean bias is calculated as the average of the mean bias of all stations? Similar to that I'm a bit confused about what you write on page 8 regarding SLR. I was assuming that you fit a regression model for each station individually. But that seems to be not the case, otherwise I could not understand why there should be a bias overcorrection. It would be nice if you could clarify this, whether you fitted separate models for each station or not.

The reviewer is correct in that the mean bias is calculated as the difference between the average SWE across the full temporal period for SNODAS minus the average SWE for all 383 survey sites (ie. the two dashed lines in the time-series Figure 1.b). The reviewer is also correct in that we did not fit a SLR model to each station, but instead trained a single model across all survey sites for our full temporal period (as well as the partitioned upper and lower regions in Figure 2 to see if multiple models showed improvement; and we found they did not). The bias overcorrection in the linear techniques like SLR stems from the fact that the SLR is attempting to model a linear relationship across all years which is problematic due to the nonlinearity in the bias introduced post-2014. This results in an overcorrection in some periods and an undercorrection in others.

7) Although you wrote in the beginning that you took 75% for training, you didn't mention if all the calculated verification measures refer to the remaining 25% testing period.

The reviewer is correct that the calculated verification measurements on model performance when performing validation testing refer to the remaining 25% of the dataset, we now make this clearer in the text in section 3.3 (page 10).

8) In the legend of Figure 2 you write Lower and Upper. Shouldn't it be southern and northern?

We thank the reviewer for noticing this naming discrepancy and we have updated the Figure 2 legend to show Southern and Northern instead of Lower and Upper.

Reviewer 3

1) The potential strengths of machine learning are highlighted but a justification for the selection of random forests (RF) is not particularly apparent. The authors mention applications of support vector machines and neural networks in geosciences detailed in Lary et al., 2009, a study of aerosol optical depth, but neglect to review specific literature around machine learning applications in SWE estimation (e.g. Wrzesien et al., 2017, Snauffer et al., 2018, Xue et al., 2018). A review of such advances is warranted.

We thank the reviewer for their suggestion to include additional motivation behind our selection of the random forest technique for bias correction. As mentioned by the reviewer, this choice primarily stems from the strengths this technique has shown in previous literature for bias correcting data in the geosciences (Reichstein et al., 2019; Shen, 2018; Lary et al., 2009).

However, we agree that additional motivation with respect to bias correcting SWE would be beneficial, and we have now included additional literature focusing on the application of random forest bias correction towards SWE datasets from Wrzesien et al. (2017), Snauffer et al. (2018), Xue et al. (2018), Zahmatkesh et al. (2019) and Lv et al. (2019) in section 1 (page 2) and section 4.2 (page 12) of the manuscript.

2) RF model structure and hyperparameter descriptions should be moved to the methods section. The authors mention RF is run with a forest size of 100 and maximum tree depth of 15, but it is unclear how these hyperparameters were selected beyond a mention of "sensitivity tuning experiments". Generally hyperparameters should be tuned using a standard method (e.g. grid search, particle swarm optimization, evolutionary strategy, etc.) on each test split and reported accordingly. Is the maximum number of terminal nodes for a given tree specified or are the trees allowed to grow to full extent?

We thank the reviewer for this comment and question. During the model training phase of our analysis, we experimented with a variety of values for forest size and maximum tree depth to find a balance between model accuracy and run time efficiency. This sensitivity experiment was performed through a brute-force grid search approach of nudging each parameter value to find a set of parameters which exhibit both high general performance (low RMSE and bias), and an efficient RF model run time. This resulted in the selection of a forest size of 100, along with a max tree depth of 15. As per the maximum number of leaf nodes for each tree, this was left to allow each tree to grow to its full extent. We have now included further details on how hyperparamaterization was performed in the same section (2.3) on page 6 to add further clarity to the reader.

3) RF and DT are stated to be trained on 75% of the data and evaluated on the remaining 25% test set, but are also evaluated using

a 10-fold cross-validation, resulting in an average RMSE reduction of 4.7 mm. The change to bias is unclear, as is the motivation for using both a 75-25 and 10-fold split structure. Since you've appropriately gone to the effort to run a full 10-fold cross-validation, why aren't you just using these results?

When training and running our RF model, we used a 75/25 split (75% training and 25% testing) of our dataset to help mitigate against potential model overfitting while maintaining good model performance (low bias and RMSE). We experimented with a variety of values for the training and testing set and found the 75/25 ratio provided a balance between strong model performance, and a large test set of data to compare against. This train/test ratio also aligns with standard RF test sizes as mentioned in the Scikit-learn documentation (Pedregosa et al., 2011). After calculating our results, in order to further mitigate against potential model overfitting and to evaluate model performance on unseen data, we then went ahead and employed an additional 10-fold cross validation which resulted in an average RMSE reduction which was complimentary to our 75/25 structured model. Our 75/25 model was therefore used as the primary structure for our results since it was the original model developed, reported similar results ($<$ 1.5 mm SWE difference) to our followup CV experiments, and was overall much more efficient to run.

4) The manuscript would be strengthened with a description of the efforts you've undertaken to mitigate temporal and spatial auto-correlation in your training and test sets. The manuscript would be strengthened with further descriptions of the efforts you've undertaken to mitigate overfitting. A comparison of training and validation errors would be an appropriate way to do this.

We thank the reviewer for this comment. In order to mitigate against spatial auto-correlation, we broke the training and testing datasets spatially as seen in Fig. 2 of the manuscript into northern and southern regions, to evaluate model performance in areas with differing magnitudes of bias and station densities. With respect to mitigating against temporal autocorrelation, we use monthly averaging of the biweekly station data which does help to some extent, however in order to fully avoid issues with auto-correlation, we would need to employ a strategy of removing stations/periods which are consistently correlated, and this would introduce new biases in the training dataset for our model. Overall, stations are usually selected in a representative manner by the Conservation Authorities throughout the region and we trust in the integrity of the station network to help mitigate this issue.

5) In Table 2, what are Year Id and Month Id? Are you using straight numerical values, cyclical temporal sin-cos pairs, 1-of-c indicators (Bishop, 1995)?

The Year Id and Month Id predictors are 1-of-c indicators (numerical values of 0 or 1) with 0 representing the absence of either a month/year and 1

representing the presence of a month/year.

6) The water balance analysis averages melt over a watershed associated with a given stream gauge, asserting the stream gauge provides a reasonable estimate of snowmelt while at the same time neglecting evapotransportation and rainfall (actually any precipitation). Such an assertion requires that evapotransportation and subsequent precipitation are not as significant a signal as snowmelt to runoff. This may be true, but it should be backed up by analysis and references, or minimally one of these. Baseflow should also be at a minimum mentioned.

These are fair comments and we agree that the argument can be strengthened by quantitative data. We have conducted an analysis of the dominant hydrological components across all three catchment areas, based on climate normals obtained from NRCan/CFS for the period of 1980-2010. The figure is included in this response and could be included in supplementary material. It shows that in all cases average liquid precipitation (rain) during the spring freshet season exceeds potential evapotranspiration, so that it can be argued that snowmelt places a lower bound on the spring freshet volume. A nuance here is that the snowmelt peak estimated following the method of Erler et al. (2019) can (and does) exceed the streamflow peak due to routing delay within the catchment area. The peak of negative SWE differences (which is shown in Fig. 7 of the manuscript) is shown in the Figure for comparison: it is evident that the value is significantly lower than the former snowmelt estimate and does not exceed the streamflow peak. The reason is that negative SWE differences do not include water from additional snowfall during the melt period. Comparing SNODAS SWE differences with those estimated from NRCan climate normals and streamflow, it is clear that the uncorrected SNODAS values are unphysical, while the bias-corrected values appear reasonable. For a detailed discussion of the variables shown in the Figure and how they were processed, see section 3.2 and S2 of Erler et al. (2019)); the Figure is analogous to their Fig. 2 and the datasets and methods employed are the same. The reason that this figure was not included initially is that it is based on climate normals for a period before our analysis period. Unfortunately the PET and snow depth data used in the figure are not available past 2010, so that it was not possible to update the figure. Curation of a new PET dataset (for just this figure) would be beyond the scope of this study.

7) You conclude that MBS and SLR exhibit an inability to capture year-to-year variability present in the bias, but interannual correlations are not present in the analysis. The ability of bias-correction methods particularly of the non-linear flavor to capture changes over time is arguably one of their greatest strengths, as simple offsets are more easily calculated, as you have done. A simple correlation calculation may serve as further evidence of the utility of the nonlinear method.

We thank the reviewer for this comment and agree that the inclusion of interannual correlations between in situ the bias corrected SWE datasets would further highlight the utility of nonlinear techniques. These results have been included in section 3.3 of the manuscript (page 10).

8) Fig 5 is hard to read with the scales and lines used, especially the in situ values, which are key to the plot. No description of shading used is given in the figure caption. Suggest changing line thicknesses/colors and/or adjusting scales, orientation, or paneling to make better use of available space.

The Fig. 5 caption has been updated to include a description of the shaded regions (95% sampling confidence intervals). We have also updated line thickness for the in situ data to improve visibility for the reader.

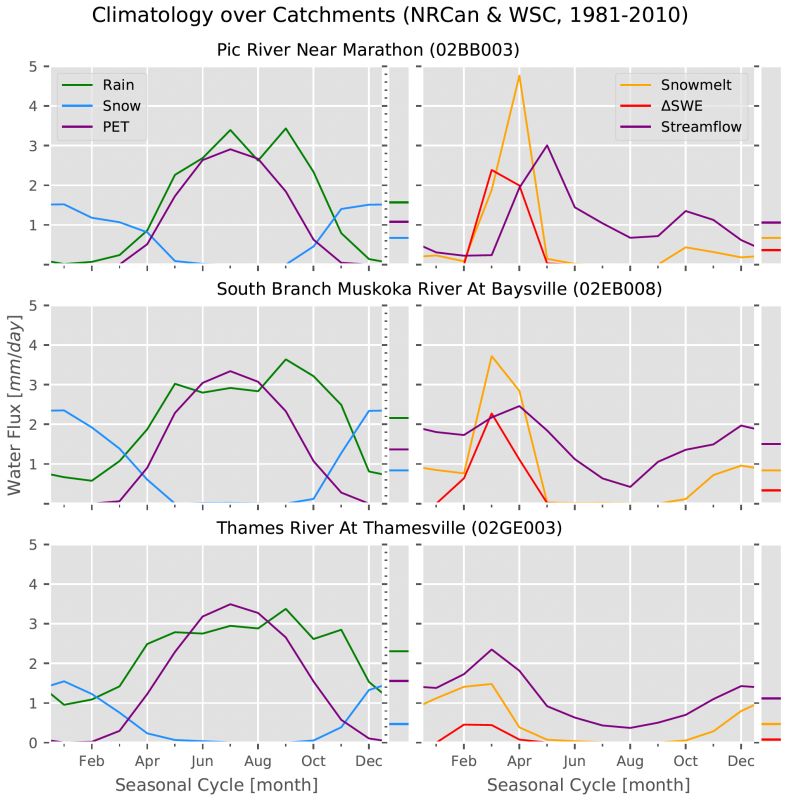

Figure 1: Catchment water flux climatology (1981-2010) for NRCan data and stream gauge data from the Water Survey of Canada.

References

Erler, A. R., Frey, S. K., Khader, O., d'Orgeville, M., Park, Y.-J., Hwang, H.-T., Lapen, D. R., Peltier, W. R., and Sudicky, E. A. (2019). Simulating Climate Change Impacts on Surface Water Resources Within a Lake-Affected Region Using Regional Climate Projections. *Water Resources Research*, 55(1):130–155.

Lary, D. J., Remer, L. A., MacNeill, D., Roscoe, B., and Paradise, S. (2009). Machine Learning and Bias Correction of MODIS Aerosol Optical Depth. *IEEE Geoscience and Remote Sensing Letters*, 6(4):694–698.

Lv, Z., Pomeroy, J. W., and Fang, X. (2019). Evaluation of SNODAS Snow Water Equivalent in Western Canada and Assimilation Into a Cold Region Hydrological Model. *Water Resources Research*, 55(12):11166–11187. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019WR025333.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,

Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195–204.

Shen, C. (2018). A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. *Water Resources Research*, 54(11):8558–8593.

Snauffer, A. M., Hsieh, W. W., Cannon, A. J., and Schnorbus, M. A. (2018). Improving gridded snow water equivalent products in British Columbia, Canada: multi-source data fusion by neural network models. *The Cryosphere*, 12(3):891–905. Publisher: Copernicus GmbH.

Wrzesien, M. L., Durand, M. T., Pavelsky, T. M., Howat, I. M., Margulis, S. A., and Huning, L. S. (2017). Comparison of Methods to Estimate Snow Water Equivalent at the Mountain Range Scale: A Case Study of the California Sierra Nevada. *Journal of Hydrometeorology*, 18(4):1101–1119. Publisher: American Meteorological Society.

Xue, Y., Forman, B. A., and Reichle, R. H. (2018). Estimating Snow Mass in North America Through Assimilation of Advanced Microwave Scanning Radiometer Brightness Temperature Observations Using the Catchment Land Surface Model and Support Vector Machines. *Water Resources Research*, 54(9):6488–6509. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2017WR022219.

Zahmatkesh, Z., Tapsoba, D., Leach, J., and Coulibaly, P. (2019). Evaluation and bias correction of SNODAS snow water equivalent (SWE) for streamflow simulation in eastern Canadian basins. *Hydrological Sciences Journal*, 64(13):1541–1555. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/02626667.2019.1660780.

Application of machine learning techniques for regional bias correction of SWE estimates in Ontario, Canada

Fraser King[1], Andre R. Erler[2], Steven K. Frey[2,3], and Christopher G. Fletcher[1]

[1]Dept. of Geography & Environmental Management, University of Waterloo, Ontario, Canada
[2]Aquanty, Waterloo, Ontario, Canada
[3]Dept. of Earth & Environmental Sciences, University of Waterloo, Ontario, Canada

Correspondence: Fraser King (fdmking@uwaterloo.ca)

Abstract. Snow is a critical contributor to Ontario's water-energy budget with impacts on water resource management and flood forecasting. Snow water equivalent (SWE) describes the amount of water stored in a snowpack and is important in deriving estimates of snowmelt. However, only a limited number of sparsely distributed snow survey sites (n=383) exist throughout Ontario. The SNOw Data Assimilation System (SNODAS) is a daily, 1 km gridded SWE product that provides uniform spatial coverage across this region; however, we show here that SWE estimates from SNODAS display a strong positive mean bias of 50% (16 mm SWE) when compared to in situ observations from 2011 to 2018. This study evaluates multiple statistical techniques of varying complexity, including simple subtraction, linear regression and machine learning methods to bias-correct SNODAS SWE estimates using absolute mean bias and RMSE as evaluation criteria. Results show that the random forest (RF) algorithm is most effective at reducing bias in SNODAS SWE, with an absolute mean bias of 0.2 mm and RMSE of 3.64 mm when compared with in situ observations. Other methods, such as mean bias subtraction and linear regression, are somewhat effective at bias reduction however, only the RF method captures the nonlinearity in the bias, and its interannual variability. Applying the RF model to the full spatio-temporal domain shows that the SWE bias is largest before 2015, during the spring melt period, north of 44.5° N and East (downwind) of the Great Lakes. As an independent validation, we also compare estimated snowmelt volumes with observed hydrographs, and demonstrate that uncorrected SNODAS SWE is associated with unrealistically large volumes at the time of the spring freshet, while bias-corrected SWE values are highly consistent with observed discharge volumes.

1 Introduction

Snow melt is an important factor for determining flood risk in many regions within both Northern latitudes and higher elevation across Europe and North America (Berghuijs et al., 2019, 2016; Buttle et al., 2016). Accordingly, predicting the impact of snowmelt on flooding is contingent on having reasonable spatially distributed estimates of the snowpack snow water equivalent (SWE). SWE is the amount of liquid water that is produced from completely and instantly melting a snowpack and is defined in terms of snowpack depth and bulk density (or, equivalently, mass per unit area). Traditionally, ground-based observations have been used to assess and quantify SWE; however, such an approach does not always capture the full spatial variability.

In Canada, large scale snowmelt is often a key driver of flooding across much of the southern, and more populated, parts of the country (Buttle et al., 2016), and one can posit that an improved ability to characterize snowpack SWE would allow better characterization of flood risk, propagation, and duration. Particularly, within the Canadian provinces of Ontario and Quebec, snowmelt and rain-on-snow events are the most frequent initiators of flooding (Buttle et al., 2016; Irvine and Drake, 1987). Regional flood danger was realized in a 2017 flood across Southern Quebec which damaged over four thousand homes and lead to approximately 200 million dollars worth of insured damages (Davies, 2017). Additional serious snowmelt and rain-on-snow induced flooding has occurred in Southern Quebec and Ontario as recently as spring 2019 (Floodlist, 2019). These recent events indicate that even though future SWE is projected to decline in south-central Canada on account of warmer winters, snowmelt will be a major flood factor in this region for the foreseeable future, with a strong likelihood of an increase in the frequency of rain-on-snow events (Byun et al., 2019).

Simulation and operational forecasting of flood risk necessitates insight on the contribution of snowmelt to the active component of the terrestrial hydrologic cycle. This is particularly important if snowmelt is anticipated to influence flood behavior (Li and Simonovic, 2002; Bokhorst et al., 2016), and the modelling tools employed for such applications include the capability to simulate snowpack dynamics (Anderson, 1976; Jordan, 1991). However, due to the high spatial and temporal variability of the snowpack, combined with the sparse distribution of in situ observations, it is difficult to properly initialize and validate forecast models. For this reason, data assimilation products such as SNODAS, which operates at 1 km spatial resolution (Carroll et al., 2001) with a daily update cycle, offer an attractive option for characterizing SWE. For example, Shen and Anagnostou (2017) employed SNODAS data aggregated to 8 km x 8 km grid cells as a validation dataset for their hydrologic model of the $20,000\ km^2$ Connecticut River Basin, wherein SWE was simulated using an energy balance approach. In CONUS scale work, Vuyovich et al. (2014) utilized SNODAS SWE as a comparative benchmark in their evaluation of SWE derived from passive microwave satellite sensors, wherein the data were aggregated at the scale of watersheds with an average size of $3,700\ km^2$.

An inherent challenge with using SNODAS as either a validation target or as direct forcing data for hydrologic modelling is that SNODAS itself may contain biases or errors that will in turn propagate through to the model outputs. In this context, the motivation for this work derives from an initial comparison between SNODAS and an independent set of in situ SWE surveys throughout Ontario (section 3.1), which suggested a positive bias of approximately 50% in the SNODAS estimates. ~~If a hydrologic model is being used to simulate flooding, deviation between actual and simulated SWE may lead to notable differences between actual and simulated flood severity, which could in turn diminish the operational value of the model/forecast. Yet~~Wrzesien et al. (2017) performed a comparison between SNODAS and in situ SWE over alpine regions in North America and found SNODAS performed best in areas with a high density of in situ measurements, however SNODAS still exhibited a general overestimation of SWE. Additional recent studies by Leach et al. (2018), Lv et al. (2019) and Zahmatkesh et al. (2019) also suggest similar positive biases in SNODAS SWE estimates throughout other North American regions. This work builds on the comparison methods outlined in previous bias correction studies by Li et al. (2010), Themeßl et al. (2011) a Teutschbein and Seibert (2012), to examine an ensemble of bias correction techniques, quantify the skill of each model, and apply the model over a larger spatio-temporal domain to produce a gridded bias corrected SWE product. Biases in initial SWE estimates constitute a major source of uncertainty in hydrologic modelling (Islam and Déry, 2017); yet, at present this impor-

tant influence of SNODAS biases on simulated hydrologic behavior and flood magnitude is not well understood. Accordingly, the primary objectives of this work are to evaluate:

1. Biases in SNODAS across flood prone regions of Ontario, Canada.
2. The effectiveness of SNODAS bias correction from simple subtraction methods to more sophisticated machine learning techniques.
3. The relationship between the regional water balance and snowmelt estimates from SNODAS SWE and bias-corrected SWE.

Section 3.1 quantifies current biases between SNODAS and in situ SWE estimates throughout Ontario. Sections 3.2 and 3.3 present evaluations of several statistical methods for bias correction, to determine whether machine learning techniques offer improved performance compared to more traditional linear methods. In Section 4, the best-performing bias correction model is then applied across the full spatio-temporal domain to create a daily, bias corrected SWE dataset which can be compared with the uncorrected SNODAS record. Differences between these two products can provide insights into where and when the bias is strongest. Finally, in Section 4.1 the impact of these difference on snowmelt volume and the influence on the regional water balance is evaluated in three representative catchment areas.

2 Data and methods

2.1 In situ data

In situ snow survey data is retrieved from a dataset created by the Climate Research Division of Environment and Climate Change Canada (ECCC), which has been updated to include observations up to the end of 2017 (ECCC, 2000). This dataset includes snow survey measurements from approximately 33 Conservation Authorities (CAs) at 383 locations throughout Ontario between 41° N and 49.5° N and -87.875° E and -73.375° E. The locations of these survey sites are marked in Fig. 1 a along with an outline of our Ontario study region. The snow surveys provide an ~~average~~ estimate of SWE calculated ~~from~~ as an average of 10 ~~fixed sampling locations taken~~ individual snow core SWE samples taken over 10 meters at each survey site (CA, 1985). These obtservations are recorded bi-weekly around the 1^{st} and 15^{th} of each month from November to May ~~(CA, 1985). These observations~~ and have been recorded since 1933, but ~~but~~ for this study only data from January 2011 until December 2017 have been considered. Snow survey density is higher in the southernmost portion of Ontario (below 44.5° N) with 189 survey locations closely grouped near the United States (US) border between Lake Huron and Lake Ontario. A similar survey count ($n = 194$) exists above 44.5° N but with sparser spatial coverage due to the region's larger size and lower population density.

The measurement tools used for retrieving in situ observations vary between locations over the time span of our study (CA, 1985; Sturm et al., 2010). Common strategies for collecting SWE measurements by hand include the use of snow corers which are portable, handheld tubes that are inserted into the snowpack down to the soil layer and weighed to retrieve a SWE estimate

at that point within the snowpack (López-Moreno et al., 2013). An Ontario snow inventory summary completed by Metcalfe (2018) provided a questionnaire to 265 of the snow survey sites (with a 67% response rate) and found that the Federal Snow Sampler (also known as the Mount Rose sampler) was used at 94% of sites and the ESC-30 was used at the remaining 6% of locations. Methods of SWE measurement also varied with 62% using a calibrated spring balance in the field, 30% using a digital balance in the field, 6% grouping snow samples into a container and weighing in the field, and 2% bagging the samples and weighing them later (Metcalfe, 2018). Although an important observational metric, in situ measurements of SWE take, on average, 20 times longer than snow depth measurements, and due to the additional time investment this often results in poor spatial and temporal data coverage of SWE measurements across large regions (Sturm et al., 2010).

2.2 Gridded SWE products

2.2.1 The Snow Data Assimilation System (SNODAS)

SNODAS is a gridded modeling and data assimilation dataset produced by the National Oceanic and Atmospheric Administration (NOAA) National Weather Service's Operational Hydrologic Remote Sensing Center (NOHRSC) (Barrett, 2003). SNODAS provides a physically consistent framework for assimilating snow data from nearly all available North American airborne, satellite and ground station sources with a Numerical Weather Prediction (NWP) snow model (Dawson et al., 2016). Produced at 1 km resolution, SNODAS covers the continental US from approximately 25.95° N to 52.87° N and overlaps with portions of Canada including our study region (Azar et al., 2008). Daily estimates are provided from September 2003 to January 2018, however Ontario was only included within the assimilation domain starting in January 2011, providing seven years of overlapping data with the in situ SWE measurements. Additional SNODAS product details are described in Table 1.

SNODAS is composed of three primary components: the data ingestion pipeline which handles data quality control and downscaling from the NWP model forecasts, the snow mass and energy-balance model which calculates hourly snowpack property estimates, and the data assimilation scheme which updates the model state with observational snow data (Carroll et al., 2001). In order to prescribe forcing data for the snow model, SNODAS makes use of the Rapid Refresh (RAP) and High-resolution Rapid Refresh (HRRR) NWP systems, deployed by the National Centers for Environmental Prediction (NCEP) to produce high accuracy, hourly numerical weather forecasts (Benjamin et al., 2016). RAP/HRRR produces analyses and short-term forecasts of precipitation, pressure, temperature, wind and relative humidity which are corrected using station and radar data, downscaled, assessed for quality and then used to force the SNODAS snow model (Barrett, 2003). SNODAS uses a spatially distributed multi-layer mass and energy-balance snow model with 3 snow layers and 2 soil layers (Carroll et al., 2001). The snow model calculates snowpack SWE, temperature, thickness and liquid water fraction within each snow layer and produces an estimate of total SWE, runoff melt (from the base of the snowpack), as well as estimates of exchange fluxes with the atmosphere. Thermal properties of the snowpack are simulated using similar techniques to SNTHERM89 as described in Jordan (1991). After applying the surface and atmospheric forecasts from RAP/HRRR, the snow model is run at an hourly timestep, with mass and energy balance calculated at each grid cell (Barrett, 2003).

A simple nudging method (Newtonian Relaxation Procedure) is then used to update model SWE estimates with assimilated ground-based, airborne and satellite snow observations (Boniface et al., 2015). This technique examines differences between numerical model SWE estimates and assimilated observations to identify regions with significant differences (Clow and Nanus, 2011). Although many existing snow cover and SWE datasets are assimilated by SNODAS, we note that the in situ snow survey dataset employed in this study is not assimilated by SNODAS. Differences between the model estimates and observations are then interpolated to produce nudging fields (an increment used to *nudge* model estimates closer to observations) and the model is re-run for the previous 6 hours. Each hourly increment during this period is nudged using the previously computed nudging fields to produce the final SWE estimate for each grid cell, updated using assimilated observational datasets (Barrett, 2003). Previous studies by Frankenstein et al. (2008) and Rutter et al. (2008) have suggested that SNODAS strongly benefits from this data assimilation step with densely observed locations displaying high quality SWE estimates in SNODAS when compared with in situ measurements.

2.2.2 NRCan ANUSPLIN data

During the development of the bias-correction methods, a gridded, monthly climatology (spanning 1981-2010) of 2 meter air temperature and total precipitation was employed. This dataset was developed by the Canadian Forestry Service (CFS), which is a division of Natural Resources Canada (NRCan); it will be henceforth referred to as the NRCan dataset. The NRCan dataset is generated through the use of thin-plate (Laplacian) smoothing splines which interpolates point observations over a grid as implemented in the ANUSPLIN (Australian National University SPLINe) climate modeling package (Hutchinson et al., 1991; McKenney et al., 2011). The NRCan product provides additional gridded estimates of snowpack height, 2 meter air temperature and total precipitation throughout Ontario (Table 1). This product has a spatial resolution of approximately 10 km and provides monthly normal estimates of surface parameters from January 1981 to December 2010. This observational time frame overlaps with in situ survey measurements, however NRCan data ends (December 2010) just before SNODAS becomes available in this region (January 2011) which is an additional source of uncertainty (see section 4.2). The datasets used in the generation of this product are independent from both the SNODAS and the snow survey datasets.

2.3 Statistical methods for bias correction

A set of statistical methods that have previously been applied to bias correction in different contexts are analysed in this study to identify the method which displays the highest performance in reducing the bias between SNODAS SWE and in situ observations over our study period. The methods examined include: mean bias subtraction (MBS), simple linear regression (SLR), decision trees (DT) and random forest (RF). All models (excluding MBS) are implemented using the scikit-learn Python package which includes built in linear regression and machine learning modules (Pedregosa et al., 2011). For MBS, the average difference in SWE between SNODAS and in situ is calculated and then subtracted from each SNODAS estimate to produce a bias corrected dataset. More formally, mean bias (MB) is defined as: $MB = \frac{1}{n}\sum_{i=1}^{n}(x_i - z_i)$ where x_i and z_i are the respective daily SNODAS and in situ SWE measurements, and n is the number of measurements over the study period. The linear regression techniques used in this study conform to the least squares general regression model which relates a response

variable y to a linear combination of n explanatory x-variable predictors $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n$ where b_0 is the model intercept and b_1 to b_n are predictor coefficients. Both, simple linear regression (SLR) using a single explanatory variable, and multiple linear regression (MLR) with numerous explanatory variables are considered in this study.

SLR is applied using daily SWE from SNODAS as the sole predictor and in situ snow survey SWE estimates as the response variable. MLR, DT and RF methods all use the full list of predictors outlined in Table 2 to predict in situ SWE. ~~Additionally, the RF method is~~

A decision tree is a flowchart-like data structure, wherein the decision making process begins at the root node (the top of the tree) followed by a series of cascading decisions based on the included model predictors until the terminal leaves are reached at the bottom of the tree which represent the regression estimate of the response variable. As implied by its name, the random forest regression model is an ensemble of decision trees that are generated during model training (Azar et al., 2008). Each tree included in the RF model ensemble is generated from a randomized subset of the available training data, coupled with a randomized subset of the predictor variables (Breiman, 2001; Grömping, 2009). This inherent randomness improves the learning process of this technique, but also contributes to uncertainty in the accuracy of an individual tree (Barnett et al., 1988). The ensemble approach used in RF accounts for and minimizes the uncertainty present in an individual decision tree by calculating the mean prediction from all trees in the ensemble. Our RF model is run with a forest size of 100 ~~, and~~ trees in its ensemble, both RF and DT methods use a maximum tree depth of 15 (~~these values were obtained from sensitivity tuning experiments which~~ the maximum number of decisions before determining an estimate for the response variable) and each tree was allowed to grow to its full extent, with no set number of maximal terminal nodes (no set maximum number of leaf nodes). These model parameters were obtained through a brute-force grid search hyperparamaterization of the RF model, which nudged each parameter value and examined changes in model accuracy and impacts on computational efficiency ~~while controlling for model overfitting)~~to select optimal values for running the model. A variety of regression predictors were considered for use in this study, including land use parameters, elevation, and indicators of general climate. The final set of predictors for all methods (shown in Table 2) was selected based on the (non-zero) model importance score for each variable ~~from Random Forest~~ in the RF model summary output.

A range of metrics have been considered in order to determine whether additional performance can be gained from using more sophisticated statistical methods over traditional approaches. To assess model skill, we have selected absolute mean bias (accuracy) and RMSE (precision) as our model performance criteria, as these properties have demonstrated effectiveness in previous studies for assessing the capabilities of competing bias correction methods in the geosciences (Cannon et al., 2015; Grossi et al., 2017; Li et al., 2010).

In addition to applying each model to the full set of in situ survey sites for the full time span, we also run spatially and temporally partitioned models to assess changes in performance over specific regions and periods. Partitioning is applied spatially by separating the set of in situ measurement locations into northern and southern regions at 44.5° N to help account for snow survey density differences between the two regions of Ontario as described in section 2.1. Model performance is also analysed temporally with training restricted to different portions of the snow season: December to February (DJF), March to

May (MAM) and the combined period: December to May (DJFMAM). All models and partitioned datasets were trained on 75% of the data and tested on the remaining 25% (excluding MBS which does not include a model training step).

3 Bias correction results

3.1 Quantifying biases in SNODAS SWE

Initial comparisons between current SNODAS SWE estimates and in situ observations throughout Ontario describe, on average, a positive absolute mean bias of 50% in the SNODAS estimates from 2011 to 2017. Additionally, the snow survey sites in Fig. 1 a display a pattern of strong relative mean bias present in the SNODAS estimates at the majority of survey locations. Relative mean bias (RMB) is defined as $RMB = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - z_i}{z_i} * 100$ where x_i and z_i are the respective daily SNODAS and in situ SWE measurements, and n is the observation count. This relative bias is positive at 212 of the 383 measurement sites and rises above +100% relative bias at 67 locations. These sites with a strong relative bias also generally exhibit a strongly positive absolute mean bias, with SNODAS overestimating SWE by over 100 mm SWE at many of these locations. The sites with the strongest relative and absolute mean biases are typically grouped together in the northern portion of the study region above 44.5° N, as well as in areas East of both Lake Huron and Lake Superior.

There also exists a strong temporal bias in the bi-weekly SWE estimates from SNODAS when compared with in situ (Fig. 1 b). This bias is strongest during the first half of our study period until the beginning of 2015 where, although SNODAS estimates are generally still higher on average (by approximately 5 mm SWE), the overall absolute mean bias is reduced. If we consider the full temporal domain, the absolute mean bias in the SNODAS estimates is approximately 16 mm SWE, which corresponds to a 50% increase compared to that of the in situ SWE observations. The change in bias between the first and second half of our study period implies a change in the data assimilation system used by SNODAS, wherein new datasets are assimilated into the system to further reduce model error.

3.2 Simple subtraction and regression techniques

In the following section, the performance of four bias correction techniques will be discussed. The progression of mean bias and RMSE over the two study regions and three time periods is summarized in Fig 2, timeseries summary metrics for the full region are shown in Fig. 3, and the spatial pattern of (remaining) absolute mean biases at snow survey sites is shown in Fig. 4; the timeseries of corrected and uncorrected, domain-averaged SWE (along with the 95% confidence intervals based on each sample) is shown in Fig. 5 for our full study period.

3.2.1 Mean bias subtraction

We begin by quantifying how well SNODAS SWE biases can be reduced through a simple subtraction of its mean bias. Since this method is constructed through the removal of the mean bias in the SNODAS data record, MBS reduces the absolute mean bias between SNODAS and in situ to zero when averaged over all regions across all seasons as shown in Fig 2. Although zero

absolute mean bias gives the appearance of strong performance, residual biases still remain at individual days and months of the MBS corrected dataset.

The RMSE of the resulting bias corrected dataset is only slightly reduced compared to the default RMSE between SNODAS and in situ SWE. The largest decreases in RMSE of approximately 15 mm SWE (30%) occur in the northern region, with a more muted reduction throughout the southern region (approximately 2 mm SWE (10%) on average). Similar reductions in RMSE follow when this technique is applied over all years for both regions as shown in Fig. 3. RMSE is reduced from the default SNODAS value of 27.45 mm SWE to 21.9 mm SWE, which is an improvement of approximately 20%.

As was previously noted, this technique is able to reduce absolute mean bias across the full region to zero, however this is achieved at the cost of introducing strong negative biases which cancel the remaining positive biases, as shown in the change in spatial bias from Fig. 4 a to Fig. 4 b. Since MBS uniformly subtracts bias from all sites across the region, areas of low positive bias in SNODAS (eg. along the US border) have their SWE estimates reduced too aggressively and now exhibit a strong negative bias. This subtraction process can lead to unphysical, negative estimates of SWE which should be discarded if this bias correction technique is to be used in practice. Additionally, areas with the strongest positive bias in SNODAS throughout the northern region have their SWE estimates reduced by too little and continue to display a strong positive bias.

MBS results in the creation of a SWE product that has been overcorrected in some areas and undercorrected in others, and leads to high RMSE in the final corrected dataset. Similar issues are also apparent temporally in the MBS corrected timeseries of Fig. 5, with an undercorrection of SWE in the years before 2015, and an overcorrection during 2015 and the years that follow. This residual error suggests that MBS is unable to fully capture spatio-temporal differences in the SNODAS bias and that more sophisticated techniques should be investigated.

3.2.2 Linear regression

A limitation of MBS is that it is unable to benefit from predictor relationships between the snow bias and climate variables. Using daily SNODAS SWE as a predictor, SLR displays skill in significantly reducing absolute mean bias. However, this technique seems to overcompensate in the correction of the strong bias in the northern region of the study area, especially during MAM where the absolute bias drops below zero to -3.2 mm SWE (Fig. 2). RMSE is reduced from the uncorrected SNODAS values to 15-20 mm SWE on average. Similar to MBS, the SLR corrected dataset also exhibits a RMSE difference between the northern and southern regions. We note the largest decreases in RMSE in the northern portion of the study area across all time periods with improvements of approximately 50% over the uncorrected SNODAS values. However, only slight reductions in RMSE are observed throughout the southern region.

SLR performance across the full spatio-temporal study range exhibits similar results to that of its partitioned comparison, with absolute mean bias reduced to approximately -1.25 mm SWE, and overall RMSE lowered by 45% (to 14.9 mm SWE) compared to that of the default SNODAS bias (Fig. 3). In order to determine whether the inclusion of additional predictors improves the performance of linear regression, MLR was also examined. When run with the predictor set described in Table 2, MLR exhibits similar performance to SLR with approximately the same reductions in absolute bias and only slightly lower RMSE ($RMSE_{MLR} = 13.66$ mm SWE vs. $RMSE_{SLR} = 14.9$ mm SWE).

SLR continues to improve upon the results of MBS by further reducing absolute mean bias and RMSE at individual locations as shown in Fig. 4 c. The results of this technique show significant reductions in the spatial bias present in SNODAS. However, this technique also suffers from bias overcorrection. Since the SNODAS bias is not homogeneous across all snow survey sites, areas of negative bias in SNODAS are corrected by the SLR model to be even more negative (as is seen at a set of survey sites in Fig. 4 c along the coasts of Lake Huron, Lake Ontario and Lake Superior). SLR improves the overall positive bias across the majority of the northern region sites, however a strong positive bias persists at many locations after SLR is applied, suggesting undercorrection at some locations. Similar to MBS, we note both an overcorrection and undercorrection of SWE in the timeseries of Fig. 5 (with a transition occurring again in 2014), confirming our assumptions that the linear regression methods are unable to account for heterogeneity and nonstationarity in the bias between years.

3.3 Nonlinear methods

The DT method displays further improvements over MBS and SLR in terms of model skill, with the second lowest overall RMSE between 3-8 mm SWE on average, coupled with near-zero absolute mean bias when partitioned spatially and temporally (Fig. 2). Differences in RMSE are quite small between each region and time period, and the resulting RMSE between DT and the in situ observations is substantially lower, on average, than that of uncorrected SNODAS (an 80% improvement). We note similar large improvements in model performance using the DT method across the full region for all months, with an overall RMSE of 4.03 mm SWE and absolute mean bias of 0.6 mm SWE.

Building on the improvements from DT, we find that RF displays the best overall skill of all tested models by producing SWE estimates with low absolute mean bias and the lowest overall RMSE when compared with in situ SWE. As noted in the predictor importance scores of Table 2, RF incorporates information from a suite of predictor variables which allows the model to better understand how SWE biases change in both time and space. RF was found to consistently outperform the other models for all time periods for both northern and southern regions of our study area, as shown in the partitioned model run summary statistics in Fig. 2, with absolute bias reduced below 1 mm SWE and RMSE between 3 and 5 mm SWE. Furthermore, RF continues to outperform other methods of bias correction when the model is trained and run over the full spatio-temporal domain, resulting in an RMSE of 3.64 mm SWE and absolute mean bias of only 0.2 mm SWE as shown in Fig. 3. This is an 86% reduction in RMSE compared to the uncorrected SNODAS RMSE and a significant improvement over the 45% reduction achieved by SLR and the 20% reduction in RMSE from MBS. Since the the RF is composed of an ensemble of DT models, it is not surprising that both methods perform similarly when run with the same predictor set, with RF slightly outperforming a single DT, because the ensemble is more robust and reduces systematic model error caused by overfitting.

As the bias in SNODAS is nonstationary (Fig.1b), we next evaluate the bias correction methods separately for a sub-period of high bias (2011-13), and one of low bias (2014-17). This test is performed using the same predictor variables in Table 2, excluding Year Id; i.e., we implicitly assume stationarity within each sub-period. During the high bias period (with a default bias of 27.9 mm SWE and default RMSE of 38.5), we find similar results to the full period. MBS, SLR and RF all reduce the absolute mean bias down to less than 1 mm SWE, and RF reduces RMSE to the lowest value of 2.7 mm SWE, compared to 5.7 and 26.6 for SLR and MBS, respectively. The low bias period (with a default bias of 9.28 mm SWE and default RMSE

of 16.5), again exhibits a similar pattern in model performance to the full period, with all models reducing the absolute mean bias to less than 1 mm SWE and RF again showing the lowest RMSE of 4.9 mm SWE, compared to 14.0 and 13.6 for SLR and MBS. In summary, the sub-period analysis shows consistent performance from the RF model, but improved performance of the SLR model during the high bias period, when the bias in SNODAS appears more uniform from year-to-year (Fig.1b).

To further mitigate against model overfitting, the data for the RF and DT models ~~were~~ are randomly split into a training set composed of 75% of the ~~dataset, and evaluated on~~ values, and a testing set which comprises the remaining 25%. Additionally, a separate 10-fold cross validation (CV) resampling procedure was applied to further evaluate model performance on unseen data. The CV K-fold splits the full dataset in time into 10 consecutive groups of samples which are held constant for the full CV procedure. We then train the model on each combination of $k-1$ folds and their performance is calculated as the average of all training and testing scores for each K-fold split. When applied to the full spatio-temporal dataset, this technique results in an average reduction in RMSE to 4.7 mm SWE between the RF CV models and in situ observations in the 25% remaining testing portion of the dataset. The fold value of $k = 10$ was selected as a compromise between the size of the training sample and the computational overhead~~. Furthermore, it has been suggested in the literature that 10 folds empirically yield testing error rates that suffer less from high variance datasets compared to other values of k~~ , and is a typical choice for similar applications. (James et al., 2013).

The RF model displays the best overall performance in terms of reducing bias and RMSE, and this skill is demonstrated spatially in Fig. 4 d. Compared to the other bias correction methods, the RF model is the most effective at reducing the spatial bias in SNODAS, with only small differences between model corrected SWE values and in situ SWE across the majority of the region. This accuracy is also evident in the timeseries of domain-averaged SWE values shown in Fig. 5, with the RF corrected SWE estimates closely tracking the in situ observations across all years. ~~However, no model is perfect, and it is clear that even with its strong general performance, the~~ Comparisons of interannual correlations further emphasize the strengths of the nonlinear techniques over traditional bias correction methods at capturing changes in bias over time. Interannual correlations between RF corrected SWE ~~still exhibits areas of non-zero absolute mean bias (most notably in the high-bias northern portion of the study region and at a few coastal sites which display a slight negative bias)~~and in situ are the highest at approximately 0.99, with correlations of approximately 0.93 for linear regression and of approximately 0.90 between the default SNODAS and in situ SWE. The RF model is therefore selected as the best-performing candidate model to perform bias correction on the Ontario-wide SNODAS data.

4 Application of the random forest model

~~Running~~ In this section, we apply the trained RF model ~~over the full Ontario grid (~~to the full 1 km SNODAS grid for all of Ontario (approx 1.5 million grid cells, Fig. 1 a)~~allows us to~~ , and derive a gridded estimate of corrected SWE throughout the entire region. ~~When running on a standard, 4-core desktop predicting against approximately 1.5 million grid cells (at 1 km resolution), this~~ This operation takes around 30 seconds per day of SNODAS observations (approximately 1.5 megabytes per day in storage space) on a modern, 4-core desktop computer. After running the RF model at 1 km resolution, we plot

the resulting average monthly SWE bias between SNODAS and the RF corrected grid in Fig. 6 for December through May ($SNODAS - RF$). From these plots we note a strong positive monthly bias from January through April with the largest bias in SNODAS SWE estimates in March and April (averaging 57.7 mm SWE and 55.8 mm SWE, respectively), when the amount of snow on the ground is generally at its highest in Ontario. We also note a strong bias East (downwind) of Lake Superior and Lake Huron where SNODAS may be producing too much lake-effect snow.

4.1 ~~Water balance analysis~~

~~The difference in snow water volume between uncorrected and bias-corrected SNODAS SWE has important implications for understanding the regional water balance of Ontario.~~ Through the application of the RF bias correction, estimated mean SWE during December to May in the study region outlined in Fig. 1 a is reduced by approximately 33 mm (Fig. 6). ~~Naturally, these SWE reductions have implications for regional~~

4.1 Water balance analysis

The difference in snow water volume between uncorrected and bias-corrected SNODAS SWE has important implications for understanding the regional water balance of Ontario. The reduction in mean SWE resulting from the bias correction should reduce the regional melt estimates, ~~and these are estimated~~ which we estimate using hydrographs of area-normalized discharge from three Ontario river gauges ~~from~~ for the period 2011 to 2018: Pic River (48.77° N) near Marathon, the South Branch Muskoka River (45.14° N) at Baysville, and the Thames River (42.54° N) at Thamesville (basins outlined in Fig. 1 a). Monthly melt water amounts are estimated as the ~~negative~~ (negative) SWE differences between consecutive monthly means from the SNODAS and RF SWE datasets (Erler et al., 2019). To compare the melt volumes with normalized discharge values, the melt ~~has been~~ is averaged over the drainage area associated with each stream gauge. ~~It can be argued~~ We argue that this provides a reasonable estimate of the amount of water being released from the snowpack during the spring freshet period in each watershed. Note, however, that this does not include losses of water due to evapotranspiration or additional water input from rainfall.

Figure 7 shows the timeseries of area-normalized discharge and estimated melt rate over the study period for the three ~~catchment areaslisted above~~drainage areas. The timing of observed peak streamflow closely aligns with melt rate peaks during the spring freshet at the northern gauges of Pic River and the South Muskoka River. Since the melt water estimates do not include rainfall, they should be considered a conservative estimate of potential spring discharge.~~It is evident that the~~ The melt volumes derived from the corrected dataset are close to, but ~~usually~~ mostly below, observed discharge values in the two northern catchments, while the estimates based on the uncorrected SNODAS SWE data significantly exceed the observed discharge, and can thus be considered unphysical. This serves as an independent validation of the physical plausibility of the bias correction method proposed here.

We further note that the differences between the corrected and uncorrected melt estimates are most apparent during the period of high bias prior to ~~2015, as was reported in section 3.1.~~ 2015. In this context it is also interesting to note that the accumulation of SWE in the uncorrected SNODAS dataset exceeds the total amount of precipitation (based on the NRCan dataset) for most

winter months prior to 2015, and for isolated winter months after 2015. In the southern Thames River watershed, on the other hand, there exists a much lower bias between SNODAS and RF-corrected melt compared to the two northern watersheds, which is consistent with the previously discussed spatial pattern of biases in Fig 1 a. In addition, the Thames River watershed is not snowmelt dominated, so the biases do not affect streamflow in the same way as they do in the two northern watersheds. The changes in the magnitude of snowmelt shown in Fig. 7 suggest that the RF bias-corrected SWE constitutes an improvement over the uncorrected SNODAS-derived melt estimates throughout the study region, and that the RF-corrected dataset could provide a valuable new resource for hydrologic modeling and flood risk forecasting.

4.2 Discussion

Linear regression and machine learning techniques have previously been used effectively across the geosciences for bias correction of global and regional climate model output (Teutschbein and Seibert, 2012; Li et al., 2010; Lary et al., 2009; Reichstein et al., 2019; Shen, 2018). Previous studies on the estimation of North American SWE using artificial neural networks and support vector machines also exhibit similar results, with machine learning techniques outperforming general linear models (Snauffer et al., 2018; Xue et al., 2018). However, recent work by Dixon et al. (2016) and Ehret et al. (2012) suggests that bias correction methods have their own associated uncertainties which must be considered when applied to datasets like SNODAS. These studies suggest that potential inconsistencies can exist between real-world and model dynamics, and their interactions with bias correction techniques. This can lead to unphysical changes in the relationships between variables and model dynamics, and even violate basic physical principles. This last point is relevant to this study, as some models (like MBS) over-correct SWE on the ground to negative values, which are physically meaningless. Our research has found that more sophisticated nonlinear statistical techniques like DT and RF produce bias-corrected SWE values that adhere more closely to these physical principles.

We must also consider uncertainties in the in situ snow survey data record. Hand measured SWE observations are generally considered to be of high accuracy; however, measurement error can still occur. Common issues include snow sticking to the inside of the measurement device or falling out of the bottom of the device due to improper soil capping (López-Moreno et al., 2013; ECCC, 2000). Issues like these can lead to underestimations in SWE when measurements are being recorded. Furthermore, from the available documentation by Metcalfe (2018), not all CAs use the same snow coring device and measurement techniques when retrieving SWE samples and this may result in systematic differences in their reported SWE estimates. Errors in the reference dataset can propagate through into the bias correction model during training and negatively impact the reliability of the model, even away from the snow survey locations. Additional error also arises in our comparison of point to grid data since our analysis assumes that the snow survey data is generally representative of the surrounding area in the containing 1 km SNODAS grid cell. While snow survey locations are selected to be representative of their surrounding landscapes CA (1985), snow density varies drastically over small spatial scales, and this assumption of homogeneity contributes to further uncertainty in our analysis (Molotch and Bales, 2005).

Any additional uncertainties that exist in SNODAS and ~~NRCan~~ the NRCan gridded precipitation product (which are ingested as predictors into the bias correction models) will further contribute to the overall error in the bias corrected SWE dataset (Hay

et al., 2006). Uncertainties in the SNODAS numerical forecast model along with measurement error from the datasets being assimilated by each product add to the total uncertainty of the system. Furthermore, we note again that the reference period of the climate normals that have been used to characterize the climate in the RF model, is 1981 to 2010, while the study period is 2011 to 2018. This may introduce additional uncertainty due to decadal variability and transient shifts in climate; however, since only long-term averages (monthly normals) have been employed for this purpose, the error is likely small.

While there is no clearly documented reason behind the ~~shift in the magnitude of the~~ SNODAS SWE bias that ~~we note in~~ occurred after 2015, we believe this may be the result of ~~a change in~~ new datasets being inserted into the data assimilation scheme ~~being~~ used by SNODAS~~wherein new datasets are being assimilated into the system~~. Although one may argue that since the general magnitude of the SNODAS bias is reduced post 2014, a bias correction of SNODAS SWE in this region is unnecessary. We suggest that the bias correction is still a valuable contribution, since the SNODAS bias remains non-zero (approximately 5 mm SWE on average, and even higher throughout the northern region) during this period when compared with in situ, and the extended bias corrected data record allows us to better calibrate current hydrologic models. Another area of potential interest for other groups using SNODAS in Canada exists in the latitudinal gradient of bias we note in Fig. 4 a, which suggests that the mean SNODAS bias increases in magnitude as we move further away from the US border (SNODAS is a US product which mainly ingests US data).

~~Accurate and robust estimates of SWE are critical for predicting the effects of snowmelt on flooding in cold regions. Gridded SWE products like SNODAS aid in filling observational gaps between in situ measurements, however model biases and uncertainties in these products can impair the reliability of their estimates. This was apparent in SNODAS throughout our Ontario study region, with daily SWE estimates approximately 50% higher, on average, than corresponding in situ measurements. This work builds on the comparison methods outlined in previous bias correction studies by Li et al. (2010), Themeßl et al. (2011) and Teutschbein and Seibert (2012), to examine an ensemble of bias correction techniques, quantify the skill of each model, and apply the model over a larger spatio-temporal domain to produce a gridded bias corrected SWE product .~~

Each of the bias correction methods examined here ~~show~~ shows skill in reducing the absolute bias present between SNODAS and in situ SWE observations, from the default 16 mm SWE in SNODAS to less than 1.5 mm SWE across all techniques. MBS and SLR exhibit an inability to capture year-to-year variability present in the bias and often overcorrect or undercorrect the amount of SWE on ground, resulting in high RMSE between their corrected estimates and in situ observations. The more sophisticated machine learning techniques display further improvements in skill, with RF reducing RMSE by approximately 86% compared to that of the uncorrected SNODAS RMSE, and a reduction in absolute bias throughout the region to 0.2 mm SWE. The additional predictors combined with the ability of the model to capture nonlinear behavior, allows the RF model to closely reproduce observed SWE values and remain within physically plausible limits. The RF model also provides insights into the strengths of the relationships between biases and various model predictors, suggesting a connection between SNODAS biases and elevation, total precipitation and air temperature. Furthermore, it is also evident that the bias diminishes over time, even though this may not be adequately reflected in the predictor importance ranking of the calendar year variable. Unfortunately, due to lack of documentation regarding changes in the assimilation system of SNODAS, it is not possible to identify the reasons behind these changes.

In this study we have only employed simple linear regression and decision tree-based methods of bias correction. Nevertheless, we have demonstrated that nonlinear techniques can be used very effectively for bias correction, and are far superior to linear methods. ~~Support vector machines and neural networks~~ Neural networks and support vector machines have also been effectively implemented for the purpose of bias correction in the geosciences ~~, and it~~ (Lary et al., 2009). A paper by Xue et al. (2018) also found that machine learning methods can act as effective operators at estimating North American snow mass. It is possible that these other machine learning techniques may offer further improvements to the methods examined here~~(Lary et al., 2009)~~, and should be considered in additional followup work. Furthermore, it has been suggested by Reichstein et al. (2019) and Shen (2018) that deep learning methods can provide powerful new perspectives in addressing common challenges in information extraction for water resource research. However, the region that was considered in this study is relatively small and climatologically homogeneous, and the number of in situ observations is likely insufficient to justify the use of more complex techniques that typically require very large training data sets. If, on the other hand, bias correction were to be attempted on a larger scale, for example the entire SNODAS domain, a more complex technique should be considered: likely a deep neural network, potentially with recurrent properties or convolutional layers, so as to account for memory effects and spatial structure. In this scenario, it would also be possible to make use of significantly more in situ observations across North America (e.g., SNOTEL sites), that could be used to train such a model.

Data availability. SNODAS SWE data is publicly available for download via National Snow and Ice Data Center (https://doi.org/10.7265/N5TB14TC). NRCan ANUSPLIN gridded products can be downloaded from Natural Resources Canada (https://cfs.nrcan.gc.ca/projects/3). ECCC snow survey records are available for public download on GitHub (https://github.com/frasertheking/ontario_snow_surveys).

Author contributions. AE and SF identified the bias and conceived the project. All authors developed the methodology and provided interpretations of the results. CF and AE supervised this work. AE obtained the data, FK organized the data, trained and tuned the models, developed the bias corrected dataset, and produced Figures 1-6. AE performed the water balance calculations and produced Figure 7. The introduction was written by SF, with the remainder of the document written by FK and reviewed by AE and CF.

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

Acknowledgements. This study was supported by an Engage grant from the Natural Sciences and Engineering Research Council of Canada.

References

Anderson, E.: A point energy and mass balance model of a snow cover, Technical Report 19, NOAA, https://repository.library.noaa.gov/view/noaa/6392, 1976.

Azar, A. E., Ghedira, H., Romanov, P., Mahani, S., Tedesco, M., and Khanbilvardi, R.: Application of Satellite Microwave Images in Estimating Snow Water Equivalent [1], JAWRA Journal of the American Water Resources Association, 44, 1347–1362, https://doi.org/10.1111/j.1752-1688.2008.00227.x, http://doi.wiley.com/10.1111/j.1752-1688.2008.00227.x, 2008.

Barnett, T. P., Dümenil, L., Schlese, U., Roeckner, E., and Latif, M.: The Effect of Eurasian Snow Cover on Regional and Global Climate Variations, Journal of the Atmospheric Sciences, 46, 661–686, https://doi.org/10.1175/1520-0469(1989)046<0661:TEOESC>2.0.CO;2, https://journals.ametsoc.org/doi/abs/10.1175/1520-0469%281989%29046%3C0661%3ATEOESC%3E2.0.CO%3B2, publisher: American Meteorological Society, 1988.

Barrett, P. A.: National Operational Hydrologic Remote Sensing Center SNOw Data Assimilation System (SNODAS) Products at NSIDC, https://nsidc.org/sites/nsidc.org/files/files/nsidc_special_report_11.pdf, 2003.

Benjamin, S. G., Weygandt, S. S., Brown, J. M., Hu, M., Alexander, C. R., Smirnova, T. G., Olson, J. B., James, E. P., Dowell, D. C., Grell, G. A., et al.: A North American hourly assimilation and model forecast cycle: The Rapid Refresh, Monthly Weather Review, 144, 1669–1694, 2016.

Berghuijs, W. R., Woods, R. A., Hutton, C. J., and Sivapalan, M.: Dominant flood generating mechanisms across the United States, Geophysical Research Letters, 43, 4382–4390, https://doi.org/10.1002/2016GL068070, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016GL068070, 2016.

Berghuijs, W. R., Harrigan, S., Molnar, P., Slater, L. J., and Kirchner, J. W.: The Relative Importance of Different Flood-Generating Mechanisms Across Europe, Water Resources Research, 55, 4582–4593, https://doi.org/10.1029/2019WR024841, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019WR024841, 2019.

Bokhorst, S., Pedersen, S. H., Brucker, L., Anisimov, O., Bjerke, J. W., Brown, R. D., Ehrich, D., Essery, R. L. H., Heilig, A., Ingvander, S., Johansson, C., Johansson, M., Jónsdóttir, I. S., Inga, N., Luojus, K., Macelloni, G., Mariash, H., McLennan, D., Rosqvist, G. N., Sato, A., Savela, H., Schneebeli, M., Sokolov, A., Sokratov, S. A., Terzago, S., Vikhamar-Schuler, D., Williamson, S., Qiu, Y., and Callaghan, T. V.: Changing Arctic snow cover: A review of recent developments and assessment of future needs for observations, modelling, and impacts, Ambio, 45, 516–537, https://doi.org/10.1007/s13280-016-0770-0, http://link.springer.com/10.1007/s13280-016-0770-0, 2016.

Boniface, K., Braun, J. J., McCreight, J. L., and Nievinski, F. G.: Comparison of Snow Data Assimilation System with GPS reflectometry snow depth in the Western United States, Hydrological Processes, 29, 2425–2437, https://doi.org/10.1002/hyp.10346, https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.10346, 2015.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, https://doi.org/10.1023/A:1010933404324, 2001.

Buttle, J. M., Allen, D. M., Caissie, D., Davison, B., Hayashi, M., Peters, D. L., Pomeroy, J. W., Simonovic, S., St-Hilaire, A., and Whitfield, P. H.: Flood processes in Canada: Regional and special aspects, Canadian Water Resources Journal / Revue canadienne des ressources hydriques, 41, 7–30, https://doi.org/10.1080/07011784.2015.1131629, https://doi.org/10.1080/07011784.2015.1131629, 2016.

Byun, K., Chiu, C.-M., and Hamlet, A. F.: Effects of 21st century climate change on seasonal flow regimes and hydrologic extremes over the Midwest and Great Lakes region of the US, Science of The Total Environment, 650, 1261–1277, https://doi.org/10.1016/j.scitotenv.2018.09.063, http://www.sciencedirect.com/science/article/pii/S0048969718334995, 2019.

CA: Snow Surveying Manual, Standards and Procedures, Conservation Authorities and Water Management Branch, 1985.

Cannon, A. J., Sobie, S. R., and Murdock, T. Q.: Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes?, Journal of Climate, 28, 6938–6959, https://doi.org/10.1175/JCLI-D-14-00754.1, https://journals.ametsoc.org/doi/full/10.1175/JCLI-D-14-00754.1, 2015.

Carroll, T., Cline, D., Fall, G., Nilsson, A., Li, L., and Rost, A.: NOHRSC OPERATIONS AND THE SIMULATION OF SNOW COVER PROPERTIES FOR THE CONTERMINOUS U.S., in: Western Snow Conference, p. 14, 2001.

Clow, D. and Nanus, L.: Evaluation of SNODAS snow depth and snow water equivalent estimates for the Colorado Rocky Mountains, AGU Fall Meeting Abstracts, p. 0675, 2011.

Davies, R.: Canada – Over 4,400 Homes Flooded in Quebec – FloodList, http://floodlist.com/america/canada-flood-quebec-may-2017, 2017.

Dawson, N., Broxton, P., Zeng, X., Leuthold, M., Barlage, M., and Holbrook, P.: An Evaluation of Snow Initializations in NCEP Global and Regional Forecasting Models, Journal of Hydrometeorology, 17, 1885–1901, https://doi.org/10.1175/JHM-D-15-0227.1, https://journals.ametsoc.org/doi/full/10.1175/JHM-D-15-0227.1, 2016.

Dixon, K. W., Lanzante, J. R., Nath, M. J., Hayhoe, K., Stoner, A., Radhakrishnan, A., Balaji, V., and Gaitán, C. F.: Evaluating the stationarity assumption in statistically downscaled climate projections: is past performance an indicator of future results?, Climatic Change, 135, 395–408, https://doi.org/10.1007/s10584-016-1598-0, https://doi.org/10.1007/s10584-016-1598-0, 2016.

ECCC: Canadian Snow Data, CD-ROM, Climate Research Branch, Environment and Climate Change Canada, 2000.

Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., and Liebert, J.: HESS Opinions "Should we apply bias correction to global and regional climate model data?", Hydrology and Earth System Sciences, 16, 3391–3404, https://doi.org/https://doi.org/10.5194/hess-16-3391-2012, https://www.hydrol-earth-syst-sci.net/16/3391/2012/, 2012.

Erler, A. R., Frey, S. K., Khader, O., d'Orgeville, M., Park, Y.-J., Hwang, H.-T., Lapen, D. R., Peltier, W. R., and Sudicky, E. A.: Simulating Climate Change Impacts on Surface Water Resources Within a Lake-Affected Region Using Regional Climate Projections, Water Resources Research, 55, 130–155, https://doi.org/10.1029/2018WR024381, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR024381, 2019.

Floodlist: Canada – Floods Damage Over 2,000 Homes in Québec – FloodList, http://floodlist.com/america/canada-floods-quebec-april-2019, 2019.

Frankenstein, S., Sawyer, A., and Koeberle, J.: Comparison of FASST and SNTHERM in Three Snow Accumulation Regimes, Journal of Hydrometeorology, 9, 1443–1463, https://doi.org/10.1175/2008JHM865.1, https://journals.ametsoc.org/doi/full/10.1175/2008JHM865.1, 2008.

Grossi, G., Lendvai, A., Peretti, G., and Ranzi, R.: Snow Precipitation Measured by Gauges: Systematic Error Estimation and Data Series Correction in the Central Italian Alps, Water, 9, 461, https://doi.org/10.3390/w9070461, wOS:000406681700012, 2017.

Grömping, U.: Variable Importance Assessment in Regression: Linear Regression versus Random Forest, The American Statistician, 63, 308–319, https://doi.org/10.1198/tast.2009.08199, https://doi.org/10.1198/tast.2009.08199, publisher: Taylor & Francis _eprint: https://doi.org/10.1198/tast.2009.08199, 2009.

Hay, L. E., Leavesley, G. H., Clark, M. P., Markstrom, S. L., Viger, R. J., and Umemoto, M.: Step Wise, Multiple Objective Calibration of a Hydrologic Model for a Snowmelt Dominated Basin1, JAWRA Journal of the American Water Resources Association, 42, 877–890, https://doi.org/10.1111/j.1752-1688.2006.tb04501.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1752-1688.2006.tb04501.x, 2006.

Hutchinson, M. F., McKenney, D. W., Lawrence, K., Pedlar, J. H., Hopkinson, R. F., Milewska, E., and Papadopol, P.: The application of thin plate smoothing splines to continent-wide data assimilation, Data assimilation systems: Papers presented at the Second BMRC Modelling

Workshop, J.D. Jasper, Ed., Bureau of Meteorology Research Centre, pp. 104–113, https://www.researchgate.net/publication/284058675_The_application_of_thin_plate_splines_to_continent_wide_data_assimilation_Data_Assimilation_Systems, 1991.

Irvine, K. N. and Drake, J. J.: Spatial Analysis of Snow-and Rain-Generated Highflows In Southern Ontario, The Canadian Geographer / Le Géographe canadien, 31, 140–149, https://doi.org/10.1111/j.1541-0064.1987.tb01634.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0064.1987.tb01634.x, 1987.

Islam, S. U. and Déry, S. J.: Evaluating uncertainties in modelling the snow hydrology of the Fraser River Basin, British Columbia, Canada, Hydrology and Earth System Sciences, 21, 1827–1847, https://doi.org/10.5194/hess-21-1827-2017, https://www.hydrol-earth-syst-sci.net/21/1827/2017/, 2017.

James, G., Witten, D., Hastie, T., and Tibshirani, R.: An introduction to statistical learning, vol. 112, Springer, 2013.

Jordan, R. E.: A one-dimensional temperature model for a snow cover, https://pdfs.semanticscholar.org/ae2d/518793624a2d5b9d5395a5dfdf2055c2b970.pdf, 1991.

Lary, D. J., Remer, L. A., MacNeill, D., Roscoe, B., and Paradise, S.: Machine Learning and Bias Correction of MODIS Aerosol Optical Depth, IEEE Geoscience and Remote Sensing Letters, 6, 694–698, https://doi.org/10.1109/LGRS.2009.2023605, 2009.

Leach, J. M., Kornelsen, K. C., and Coulibaly, P.: Assimilation of near-real time data products into models of an urban basin, Journal of Hydrology, 563, 51–64, https://doi.org/10.1016/j.jhydrol.2018.05.064, http://www.sciencedirect.com/science/article/pii/S0022169418303925, 2018.

Li, H., Sheffield, J., and Wood, E. F.: Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching, Journal of Geophysical Research: Atmospheres, 115, https://doi.org/10.1029/2009JD012882, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009JD012882, 2010.

Li, L. and Simonovic, S. P.: System dynamics model for predicting floods from snowmelt in North American prairie watersheds, Hydrological Processes, 16, 2645–2666, https://doi.org/10.1002/hyp.1064, https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.1064, 2002.

Lv, Z., Pomeroy, J. W., and Fang, X.: Evaluation of SNODAS Snow Water Equivalent in Western Canada and Assimilation Into a Cold Region Hydrological Model, Water Resources Research, 55, 11 166–11 187, https://doi.org/10.1029/2019WR025333, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019WR025333, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019WR025333, 2019.

López-Moreno, J., Fassnacht, S., Heath, J., Musselman, K., Revuelto, J., Latron, J., Morán-Tejeda, E., and Jonas, T.: Small scale spatial variability of snow density and depth over complex alpine terrain: Implications for estimating snow water equivalent, Advances in Water Resources, 55, 40–52, https://doi.org/10.1016/j.advwatres.2012.08.010, https://linkinghub.elsevier.com/retrieve/pii/S0309170812002382, 2013.

McKenney, D. W., Hutchinson, M. F., Papadopol, P., Lawrence, K., Pedlar, J., Campbell, K., Milewska, E., Hopkinson, R. F., Price, D., and Owen, T.: Customized Spatial Climate Models for North America, Bulletin of the American Meteorological Society, 92, 1611–1622, https://doi.org/10.1175/2011BAMS3132.1, https://journals.ametsoc.org/doi/abs/10.1175/2011BAMS3132.1, 2011.

Metcalfe, R. A.: Understanding and Improving Estimates of Snowpack Conditions in Ontario, 2018.

MNRF: Provincial Digital elevation Model (PDEM), https://www.sse.gov.on.ca/sites/MNR-PublicDocs/EN/CMID/PDEM_UserGuide.pdf, PDEM Dataset Documentation, Ministry of Natural Resources and Forestry (Ontario), 2019.

Molotch, N. P. and Bales, R. C.: Scaling snow observations from the point to the grid element: Implications for observation network design, Water Resources Research, 41, https://doi.org/10.1029/2005WR004229, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005WR004229, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2005WR004229, 2005.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, http://www.nature.com/articles/s41586-019-0912-1, 2019.

Rutter, N., Cline, D., and Li, L.: Evaluation of the NOHRSC Snow Model (NSM) in a One-Dimensional Mode, Journal of Hydrometeorology, 9, 695–711, https://doi.org/10.1175/2008JHM861.1, https://journals.ametsoc.org/doi/full/10.1175/2008JHM861.1, 2008.

Shen, C.: A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists, Water Resources Research, 54, 8558–8593, https://doi.org/10.1029/2018WR022643, https://onlinelibrary.wiley.com/doi/abs/10.1029/2018WR022643, 2018.

Shen, X. and Anagnostou, E. N.: A framework to improve hyper-resolution hydrological simulation in snow-affected regions, Journal of Hydrology, 552, 1–12, https://doi.org/10.1016/j.jhydrol.2017.05.048, http://www.sciencedirect.com/science/article/pii/S0022169417303414, 2017.

Snauffer, A. M., Hsieh, W. W., Cannon, A. J., and Schnorbus, M. A.: Improving gridded snow water equivalent products in British Columbia, Canada: multi-source data fusion by neural network models, The Cryosphere, 12, 891–905, https://doi.org/https://doi.org/10.5194/tc-12-891-2018, https://www.the-cryosphere.net/12/891/2018/, publisher: Copernicus GmbH, 2018.

Sturm, M., Taras, B., Liston, G. E., Derksen, C., Jonas, T., and Lea, J.: Estimating Snow Water Equivalent Using Snow Depth Data and Climate Classes, Journal of Hydrometeorology, 11, 1380–1394, https://doi.org/10.1175/2010JHM1202.1, https://journals.ametsoc.org/doi/full/10.1175/2010JHM1202.1, 2010.

Teutschbein, C. and Seibert, J.: Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods, Journal of Hydrology, 456-457, 12–29, https://doi.org/10.1016/j.jhydrol.2012.05.052, http://www.sciencedirect.com/science/article/pii/S0022169412004556, 2012.

Themeßl, M. J., Gobiet, A., and Leuprecht, A.: Empirical-statistical downscaling and error correction of daily precipitation from regional climate models, International Journal of Climatology, 31, 1530–1544, https://doi.org/10.1002/joc.2168, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.2168, 2011.

Vuyovich, C. M., Jacobs, J. M., and Daly, S. F.: Comparison of passive microwave and modeled estimates of total watershed SWE in the continental United States, Water Resources Research, 50, 9088–9102, https://doi.org/10.1002/2013WR014734, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013WR014734, 2014.

Wrzesien, M. L., Durand, M. T., Pavelsky, T. M., Howat, I. M., Margulis, S. A., and Huning, L. S.: Comparison of Methods to Estimate Snow Water Equivalent at the Mountain Range Scale: A Case Study of the California Sierra Nevada, Journal of Hydrometeorology, 18, 1101–1119, https://doi.org/10.1175/JHM-D-16-0246.1, https://journals.ametsoc.org/doi/10.1175/JHM-D-16-0246.1, publisher: American Meteorological Society, 2017.

Xue, Y., Forman, B. A., and Reichle, R. H.: Estimating Snow Mass in North America Through Assimilation of Advanced Microwave Scanning Radiometer Brightness Temperature Observations Using the Catchment Land Surface Model and Support Vector Machines, Water Resources Research, 54, 6488–6509, https://doi.org/10.1029/2017WR022219, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2017WR022219, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2017WR022219, 2018.

Zahmatkesh, Z., Tapsoba, D., Leach, J., and Coulibaly, P.: Evaluation and bias correction of SNODAS snow water equivalent (SWE) for streamflow simulation in eastern Canadian basins, Hydrological Sciences Journal, 64, 1541–1555,

https://doi.org/10.1080/02626667.2019.1660780, https://doi.org/10.1080/02626667.2019.1660780, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/02626667.2019.1660780, 2019.

Table 1. Descriptions of the primary datasets used in our bias correction methods including relevant regression variables, their resolution, observational record coverage and data references.

Dataset	**Variable(s)**	**Horizontal resolution**	**Data period**	**Reference**
Snow Surveys	Snow water equivalent	383 points	Jan 1933–May 2018	ECCC (2000)
SNODAS	Snow water equivalent, Total precip.	1 km	Jan 2010–Dec 2018	Carroll et al. (2001)
NRCan	2-Meter temperature, Total precip.	10 km	Jan 1979–Jan 2010	McKenney et al. (2011)
Provincial DEM	Elevation	30 m	May 1978–Mar 2018	MNRF (2019)

Table 2. Predictor names and details used in the decision tree, multiple linear regression and random forest bias correction models. Also included are their respective variable units, measurement timescales, data sources and variable importance scores produced by the random forest model.

Predictor	Description	Units	Time scale	Data source(s)	RF Importance
SWE	SWE on ground	Millimeter	Daily	SNODAS	0.68
T2	2 meter air temperature	Celsius	Monthly	NRCan	0.08
TP Difference	NRCan - SNODAS total precipitation	Millimeter	Monthly	NRCan, SNODAS	0.08
Year Id	Year of observation	Year Indicator	–	–	0.07
Elevation	Height relative to sea level	Meter	–	Ontario Government	0.06
Month Id	Month of observation	Month Indicator	–	–	0.01

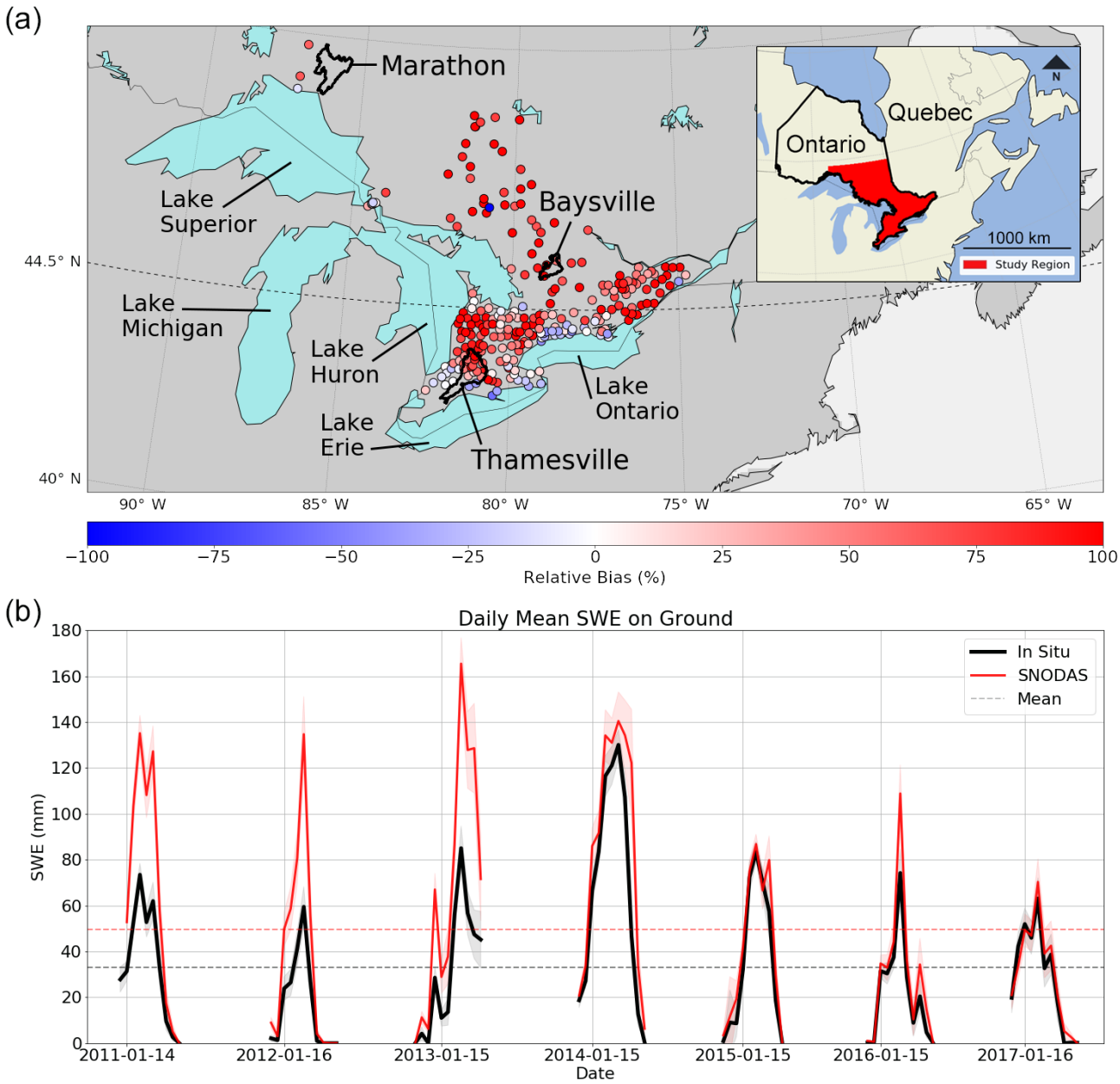

Figure 1. (a) Relative mean bias between SNODAS and in situ SWE aggregated for each snow survey site (colored points). Thicker black contours show the boundaries of the three drainage basins in the water balance analysis (Section 4.1). (b) Daily mean SWE on ground estimates from ~~SNODAS and~~ all in situ ~~SWE~~ survey sites and SNODAS, taken biweekly from November to May [2011-2017] at 383 locations across Ontario.

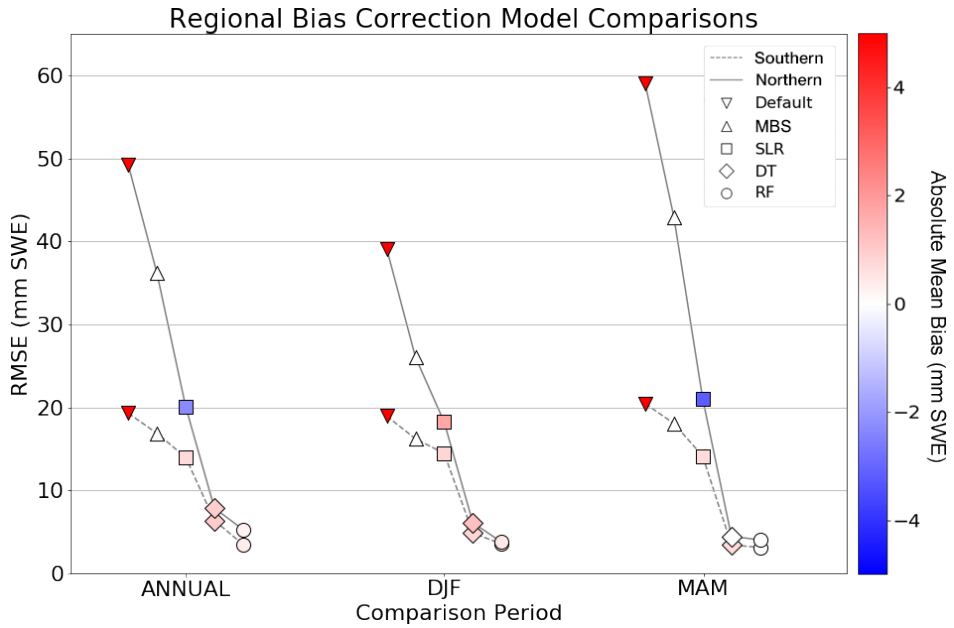

Figure 2. Performance results of regional bias correction methods (mean bias subtraction (MBS), simple linear regression (SLR), decision tree regression (DT) and random forest regression (RF)) for northern and southern geographic regions across DJF, MAM and the combined annual snow season (DJFMAM).

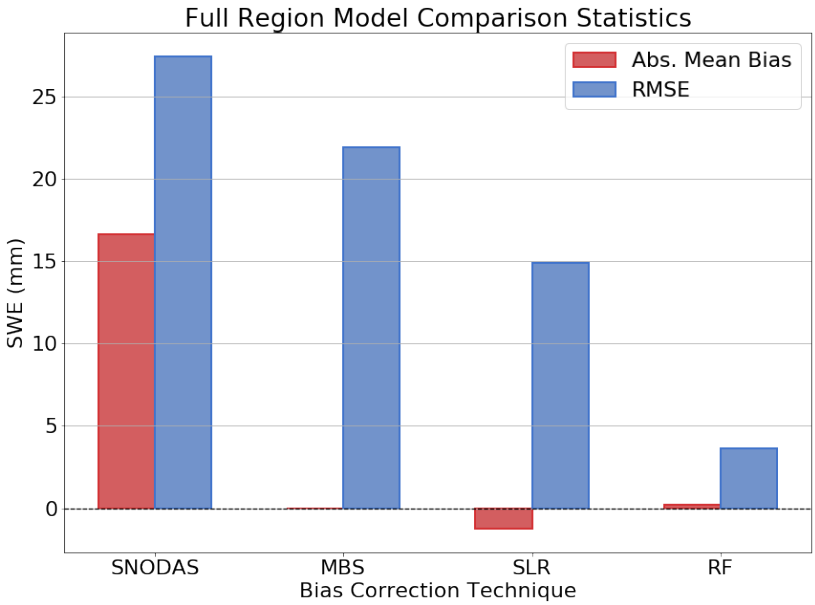

Figure 3. Bias correction model performance results for each technique across the full spatio-temporal domain.

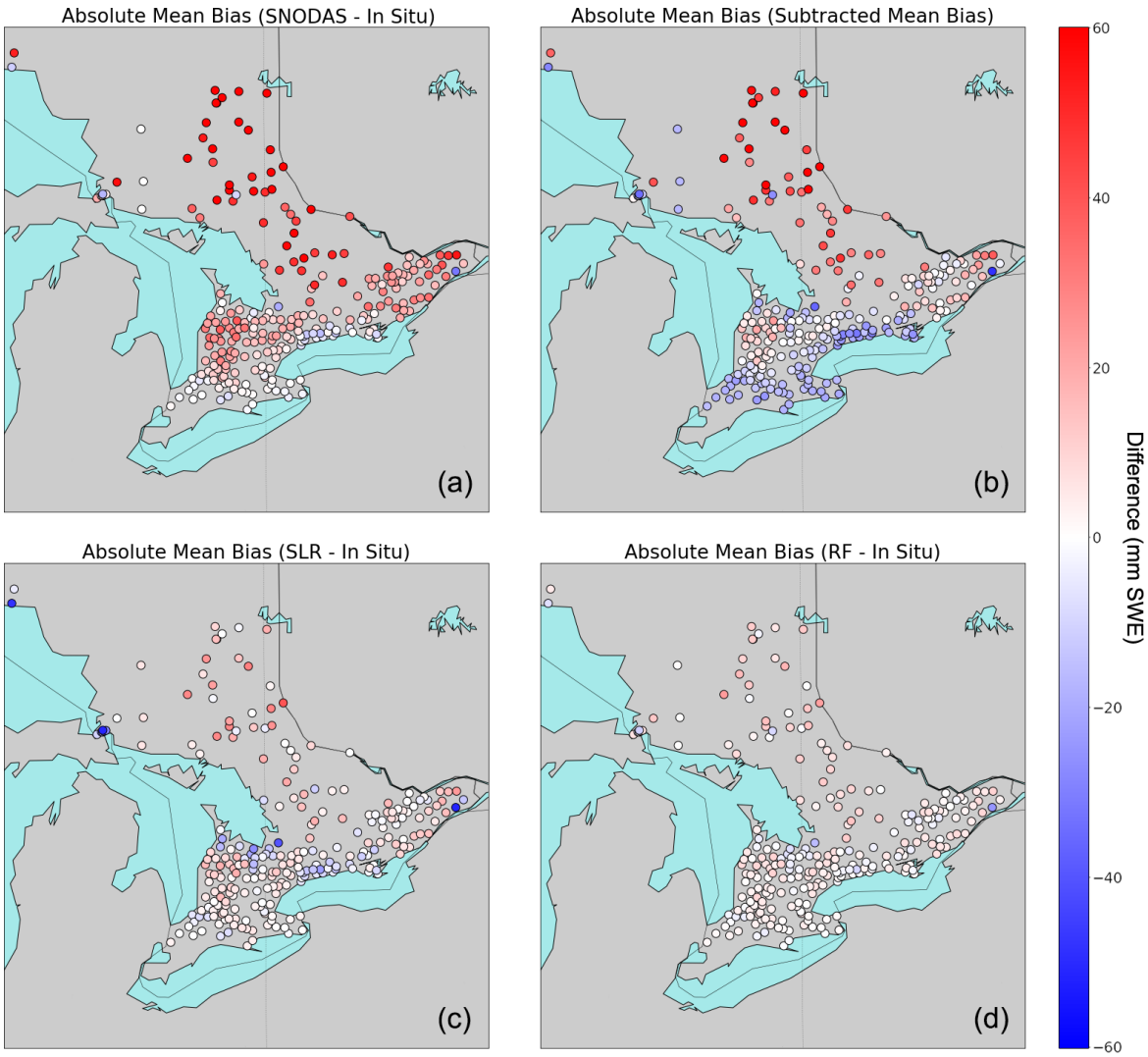

Figure 4. Absolute mean bias comparisons between in situ SWE and (a) SNODAS, (b) MBS, (c) SLR, and (d) RF, averaged at each snow survey site over the full study period.

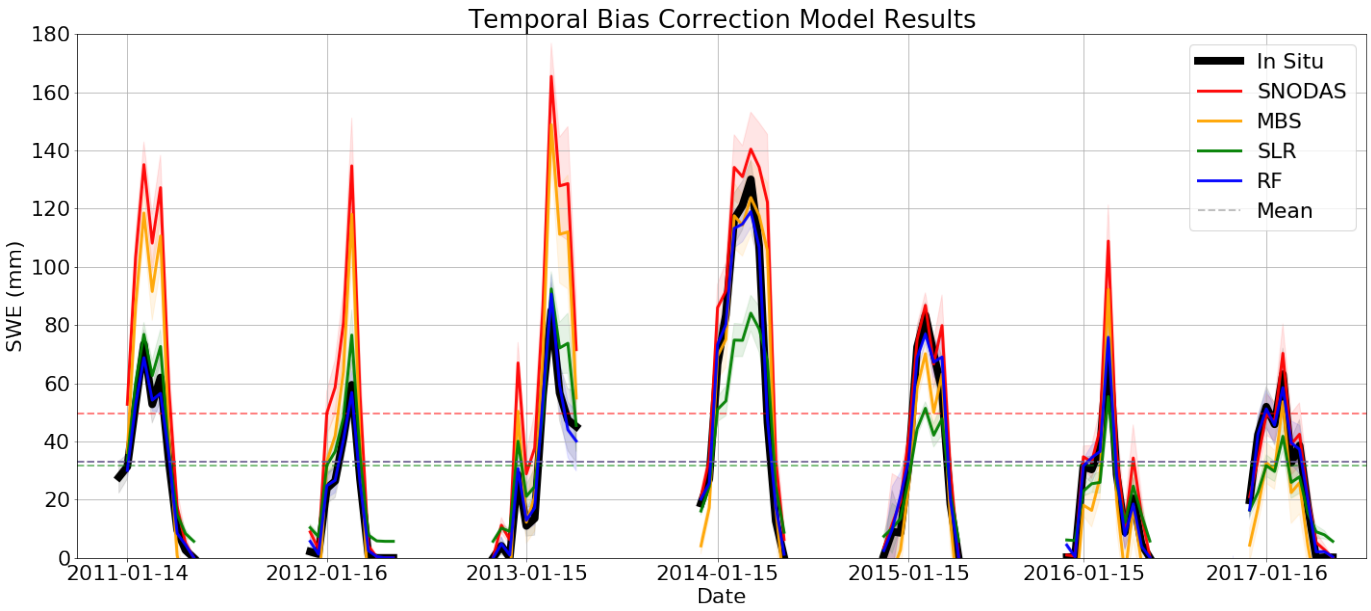

Figure 5. Daily mean SWE on ground for the MBS, SLR and RF bias corrected datasets, the default SNODAS SWE dataset and in situ SWE records. Shaded areas represent 95% confidence intervals based on the region data sample.

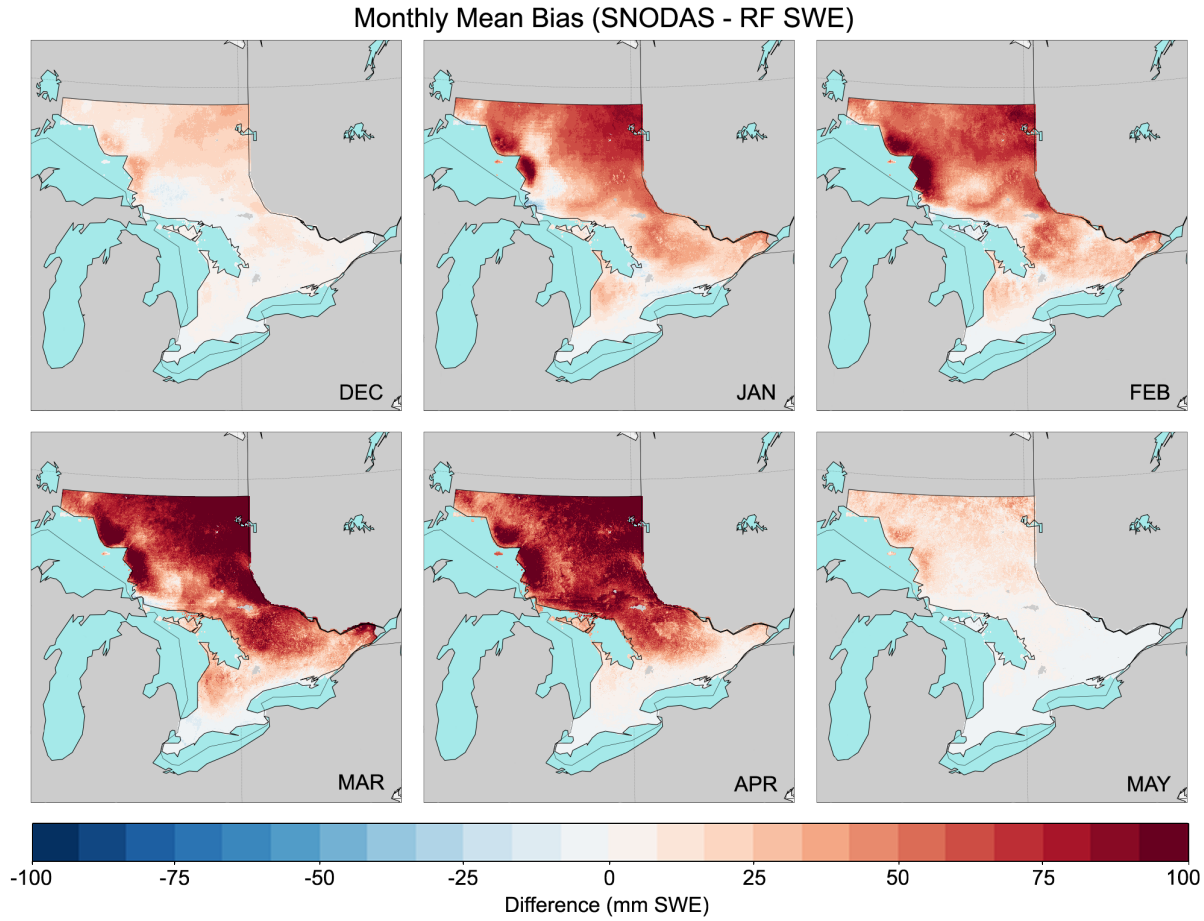

Figure 6. Monthly mean bias of SWE on ground between SNODAS and the RF bias corrected SWE dataset over December, January, February, March, April and May across the full study region at 1 km resolution.

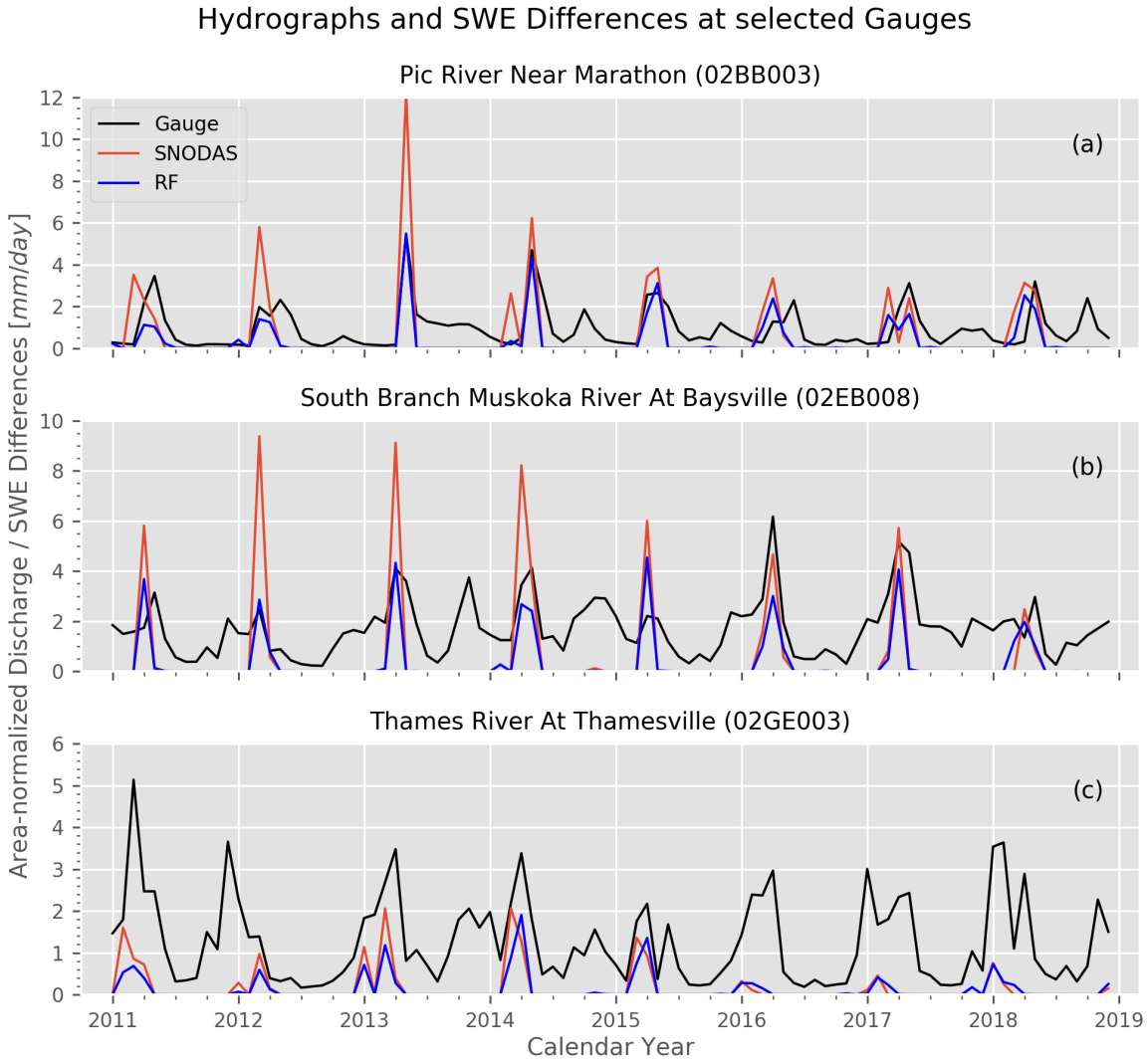

Figure 7. Monthly timeseries of area-normalized discharge from three river gauges in Ontario, along with corresponding melt estimates calculated from the SNODAS and RF corrected SWE datasets. Melt estimates are negative monthly SWE differences averaged over the drainage area of the corresponding gauge (see section 4.1).