Bennett et al., perform a detailed experiment at a set of study sites in Alaska to answer the increasingly important question of “how much snow exists here”. Their work examines a suite of regression models of varying sophistication to model SWE based on a set of environmental predictors like NDVI, elevation and wind. The paper was detailed, well written with a novel methodology and promising resulting model performance from the RF (suggesting followup work in this area of research). While portions of the paper are a bit verbose, after some edits I believe that this paper would be an important scientific contribution for the readers of The Cryosphere.

Response: thank you for your review and for your positive comments on our work.

Major Comments/Revisions:

1. While the paper by Bennet et al., is generally well written, it can be overly detailed in certain places. With some restructuring, I believe the paper can be much more concise and effective. For instance, the Introduction from lines 50-95 is likely unnecessary content. I would much rather get right into the meat of the problem at hand starting on line 96, as information about the importance of snow etc. can probably be a sentence or two with details left in references to previous literature. I have similar comments for Section 4 (specifically 4.1, 4.2, 4.3 and 4.4) which should not be in the results section and likely could be summarized in either the methodology or introduction in a paragraph or two. The beginning of 4.4 was answering questions I had about the model setup described in the methodology. These should absolutely be grouped together and the structure revised for clarity. Finally on this point, the discussion section 5 is again far too verbose and should likely be restructured with some of the details moved to the results section or moved to the Appendix. I would recommend limiting the discussion to a summary of uncertainties, sources of error and questions left unanswered from the results.

Response: To address the above comments, we have made the following changes to the paper:

- Removed lines 50-95 (four paragraphs from the introduction), starting with paragraph “Within the Arctic hydrologic cycle…” to final paragraph starting with “Many studies have”...
- We have replaced these paragraphs with a short paragraph referring to the importance of snow and references to previous work.
- For Section 4.0, we have not moved these paragraphs into the methods section. The methods section details how these tests were set up and then the results section details the results of all the testing and how we arrived at the final model configuration.
- We have gone through the text and shortened it up, reorganized the text, and focused the text in the Discussion section, and added a new section on Future Work.

2. Line 462, I am interested/potentially concerned about the extreme importance of Year on the accuracy of your RF. The authors mention that this Year variable is in some ways a proxy for

temperature/precipitation differences between years, and I would ask why not explicitly test for this? While I agree that this conclusion is probably correct, incorporating temperature and precipitation data from a well-validated reanalysis product like ERA5 or MERRA2 could help evaluate this hypothesis. It would also help explain which of these two variables is the most important. Furthermore, a Year variable really limits the robustness of this product for applications outside of your current study and removing it would help in predictions elsewhere. For instance, what if you want to apply your model to data retrieved last year? Would the RF understand the year value of 2020 if fed into the model? However, it could, in theory, incorporate precipitation/temperature data from 2020 without issue.

Response: To test this we ran several new tests to understand the impact of year in our analysis. We ran the following tests within the random forest modeling framework: a) using Year resorted (3 tests), b) using Daymet precipitation for Year (1 test), c) using Nome Airport precipitation for impurity (1 test), and d) using ERA5 precipitation for Year (1 test), e) using ERA5 temperature for Year (1 test), and f) using ERA5 precipitation and temperature for Year (1 test). When we tested precipitation and temperature, we used total winter precipitation and average winter temperature (Oct - Mar) for each of the years. In Table R1-1 below we show all the prediction and importance metrics for each test.

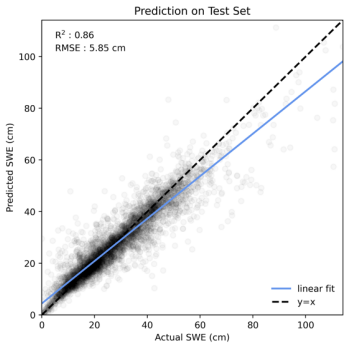
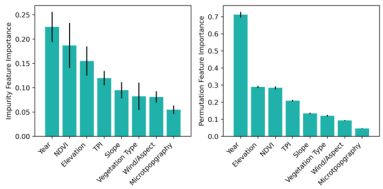
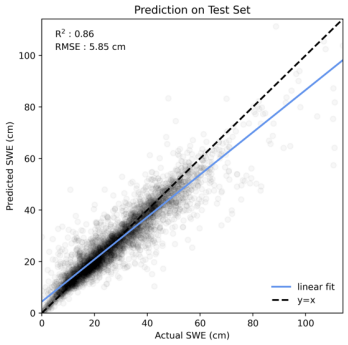
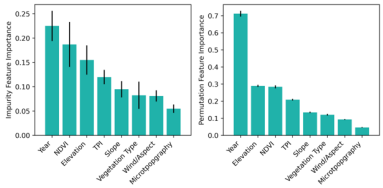
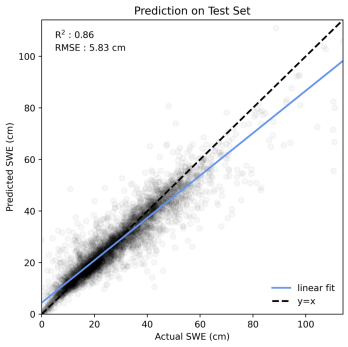
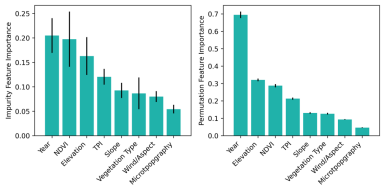
As you can see, the overall fits (R^2 and RMSE values for actual SWE vs predicted SWE) for each test did not vary that much across all of the tests. Factor importance did change slightly, with the overall four top factors remaining the same in almost every case except for ERA5 precipitation and temperature. Year and NDVI swapped in importance for the third test for impurity, with NDVI becoming the most important factor when 2018 and 2019 were interchanged (while permutation importance did not change).

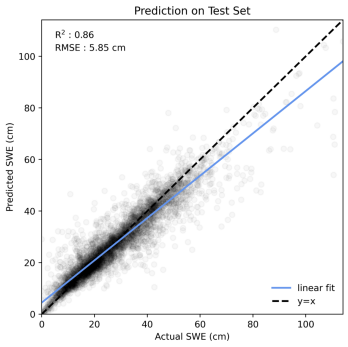
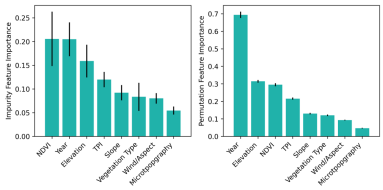
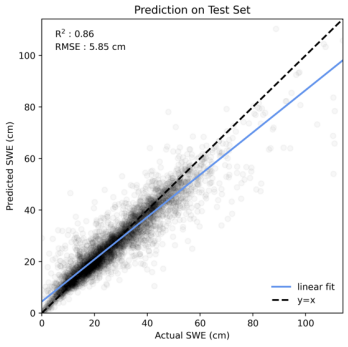
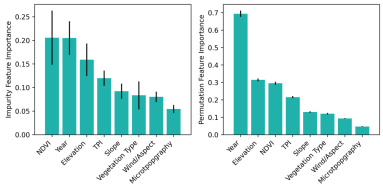
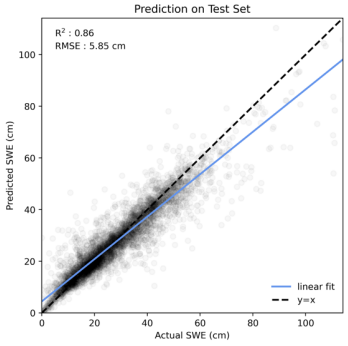
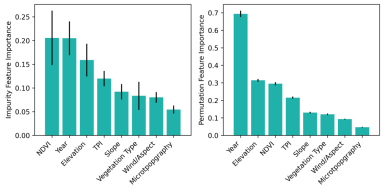
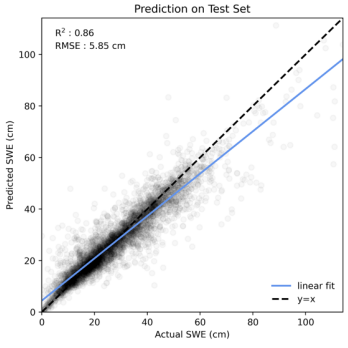
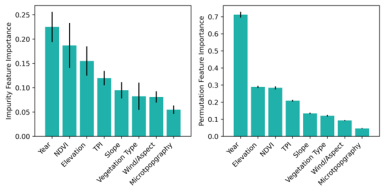
We replaced Year with three different data sets, including Daymet (Thornton et al. 2020, Figure R1-1), Nome Airport climate station, and ERA5 (Muñoz, 2019). For each of these, we also show the histograms of the SWE prediction for train, test, and whole area for each year. As you can see, the histograms do not change much between Daymet and Nome Airport station as Daymet is likely using the Airport station to estimate precipitation values. ERA5 is also quite similar with a few differences. Factor importance does change between these, Precipitation and NDVI swapped in importance when using Daymet and Nome airport data in tests, with NDVI becoming the most important factor for impurity importance. For ERA5, precipitation was the most important factor for each importance metric.

We also did two tests using ERA5 temperature, and using ERA5 precipitation and temperature. We found these results were the most different in terms of changing our modeling predictions, and also factor importances. In particular, for ERA5 precipitation and temperature tests, the top four important factors were NDVI, elevation, TPI, and Precipitation (elevation, NDVI, Year, TPI) for impurity (permutation), respectively, with temperature as either the third-to-last (impurity) or fifth (permutation) most important factor. Thus, we decided to use only precipitation as we feel that it is the most important factor and likely our Year factor represents precipitation over temperature in our study.

Due to the similarity in responses, but the wider availability of ERA5, and the fact that it does not rely on a particular station proximity, we decided to replace ERA5 precipitation in our final model and have rerun all of our analysis and output new results for the paper. We have adjusted the text of the paper as well to reflect this update. Using ERA5 will also allow us to generate the model for different years outside of our field period, and additionally, allow us to generate data for the entire Seward Peninsula region, which will be useful as we begin to compare these results and consider how to apply them within subgrid-scale earth system modeling.

Table R1-1. Tests for Year factor in random forest models.

Original year order		
a) Year resorted test 1 (2017 <- 2018, 2018 <- 2019, 2019 <- 2017)		
a) Year resorted test 2 (2017 <- 2019, 2018 <- 2017, 2019 <- 2018)		

a) Year resorted test 3 (2017 <- 2017, 2018 <- 2019, 2019 <- 2018)		
b) Year replaced with Daymet precipitation values where year=precipitation.		
c) Year replaced with Nome airport precipitation values year=precipitation.		
d) Year replaced with ERA5 precipitation values year=precipitation.		

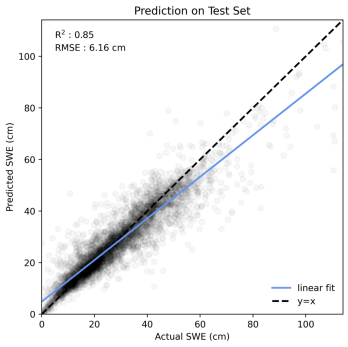
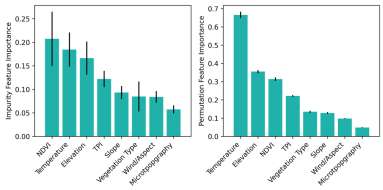
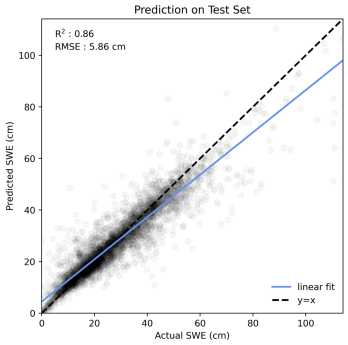
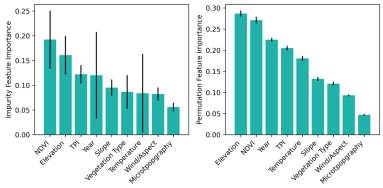
e) Year replaced with ERA5 precipitation values. Year=precipitation.		
f) Year replaced with ERA5 precipitation and temperature values.		

Figure R1-1. Daymet annual total precipitation histograms applied in random forest simulations for predicted SWE for each year.

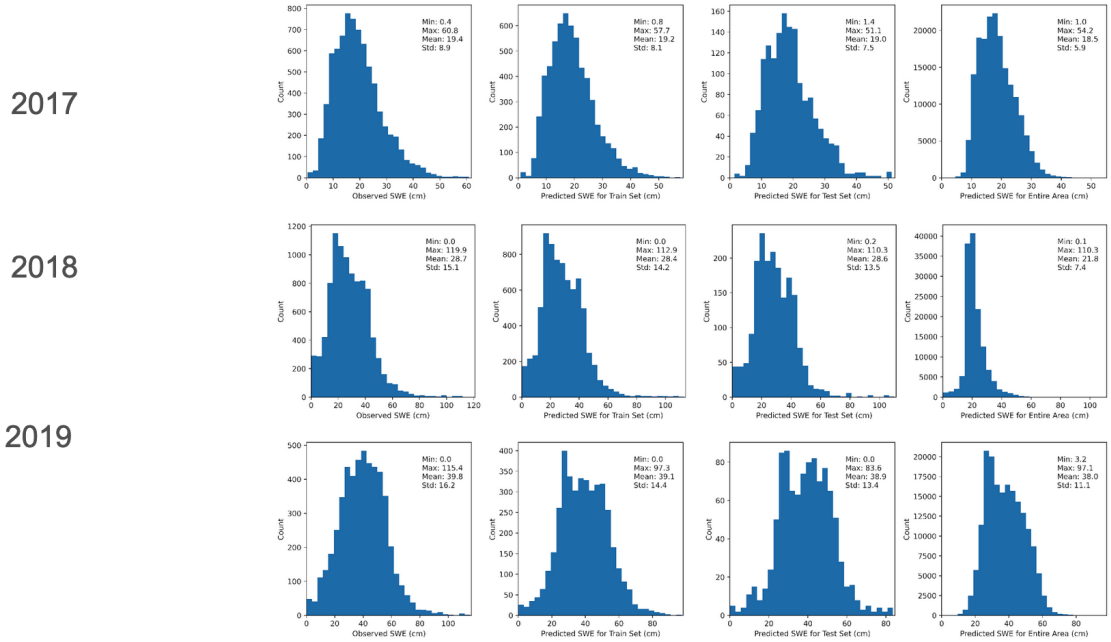


Figure R1-2. Nome Airport annual total precipitation applied in random forest simulations for predicted SWE for each year.

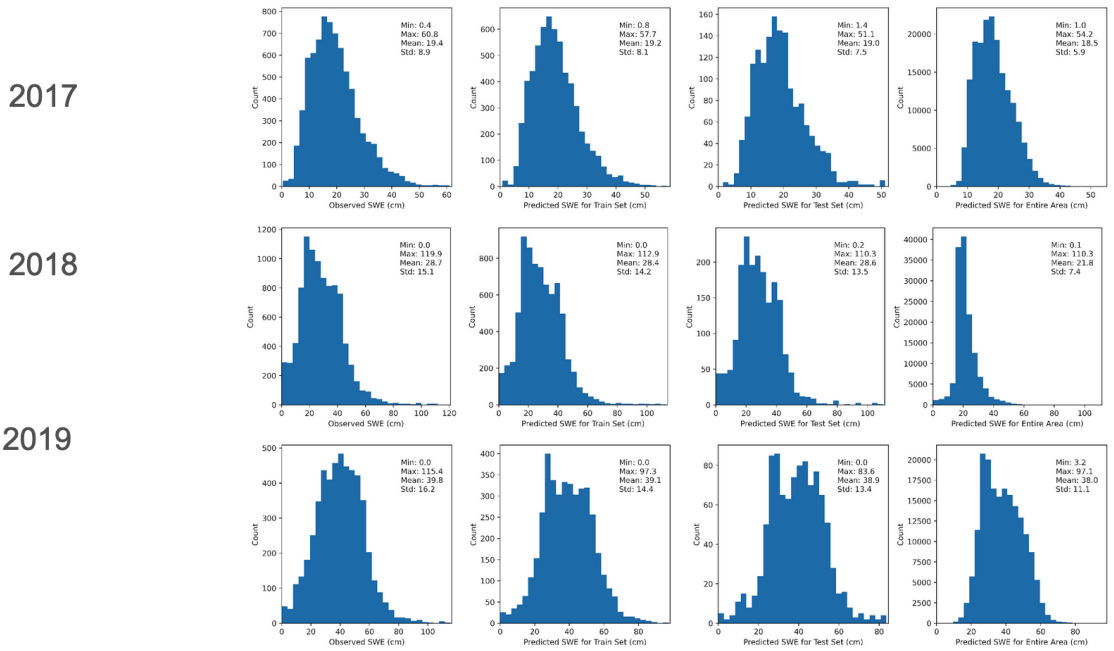
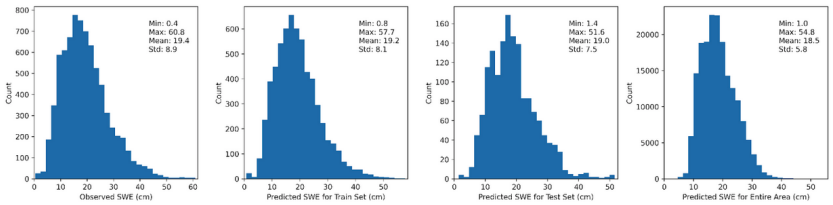
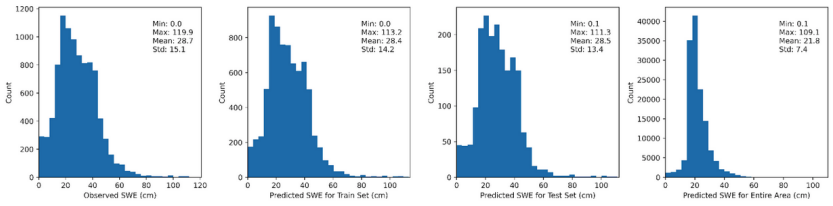


Figure R1-3. ERA5 annual total precipitation applied in random forest simulations for predicted SWE for each year.

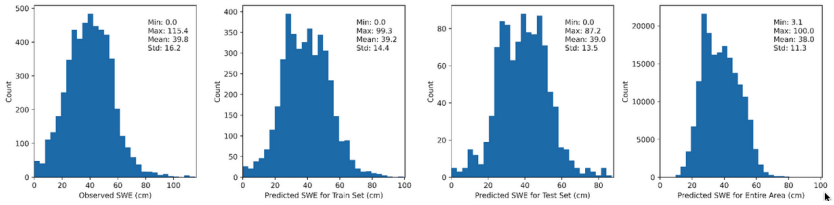
2017



2018



2019



Minor Comments/Revisions:

3. Can you speak to the different sampling distributions in Fig. 2? While I realize the Kougarok site is much larger than Teller, the spatial coverage of the samples display much more structure and consistency at Teller than Kougarok and I am curious if this sampling discrepancy impacts your results and why the sampling was so different.

Response: Teller watershed was initially the site for the snow survey, thus only Teller was surveyed in 2016 and 2017. The 2016 survey data was not used because the equipment to produce a quality survey (i.e., Snow Hydro probes and tubes) were not used in this first year. In 2018, both sites were surveyed. In 2019, the Kougarok survey was planned but sea ice in the Bering Strait went out early leading to anomalous weather conditions that meant that the field crew was limited to surveying Teller. The field team will return to Kougarok in 2022.

Although we recognize that the sampling is somewhat uneven between the sites and years, we do not think that this affects our results. And, having data in the different locations strengthened our model, rather than reduced its robustness.

We have now added maps to the Appendix that show the difference in data collection across the survey years in Teller.

4. Regarding model training I had a few questions. First, what sort of hyperparameterization are you using? It appears to be a RandomSearchCV but this isn't explicitly mentioned. Why not use

something a little more sophisticated like a Bayesian Search? Furthermore, why do you use an 80/20 split for train/test instead of a kFold CV like you do in the hyperparameterization step? This way you can operate on the full datasets.

Response: Thank you for the suggestion. We updated our hyperparameter selection method to Bayesian Search with 8-fold CV on the training set, and clarified our revised modeling approach in the paper. We updated all figures and results included in the paper as well. We chose to keep the 20% evaluation set following a similar method to past papers (e.g., Revuelto et al., 2020). In addition, the model performance does not change significantly when the training set is increased beyond 80%.

5. A table in the appendix showing the different final models (like the RF) and all incorporated predictors would be helpful for clarity.

Response: We considered this but the table would be very basic, which we do not think it would add anything to the paper. Furthermore, adding acronyms to the models (i.e., calling it T17 or something like that) would add confusion for the reader. As we describe, we use precipitation (in our updated model) for the final model, and then the individual years do not include precipitation (since it would only be one value). All the other tests we performed (100s of them) are described separately in a limited way to not overwhelm the reader (i.e., testing NDVI vs vegetation). We do not think it would make sense to introduce all of those tests and describe them since, again, it would add confusion and unnecessary detail to the paper.

6. Regarding the title, is this truly an Arctic analysis? The sites are all straddling the Arctic circle and some may consider this to be in the sub-arctic region.

Response: We have changed the title to reflect that the region is sub-Arctic rather than Arctic.

7. I commend the authors for speaking to the topic of complex terrain in different areas of the manuscript, however I am curious how your model accuracy would change as a function of the complexity of the terrain in mountainous regions. For instance, could this be applied to an alpine area? These regions don't typically have much plant life so I would expect a predictor like NDVI would be much less useful here and furthermore, the distribution of snow is extremely heterogeneous across these locations.

Response: Overall, we feel that factors might change in mountainous regions, thus the model would need to be rerun to select which factors would be most important for that particular region. But we do think the approach can be used to develop similar models and estimate snow distribution in any other area. Thus, while the exact model configuration would change, we think that the method would be transferable.

8. I am curious why you selected the RF over a method like a neural network? With such a large sample, I would expect a deep learning method like a multilayer perceptron would perform as well or better than the RF. This may be outside the scope of your paper, but something to consider.

Response: Different modeling methods in terms of neural networks or other machine learning-type approaches is something we are considering for future work, but we think it is outside the scope of the current work.

9. Fig. 3 caption should not have the definition of SWE in it

Response: We have removed the second line of the figure caption.

10. Section 4.2 heading isn't capitalized while the same words in 5.2 are? Just wondering for consistency.

Response: Right, we went through and corrected this. The capitalization was inconsistent throughout. Thanks for catching this!

References

Thornton, M.M., R. Shrestha, Y. Wei, P.E. Thornton, S. Kao, and B.E. Wilson. 2020. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/1840.

Muñoz Sabater, J., (2019): ERA5-Land hourly data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (<date of access>), doi:10.24381/cds.e2161bac