Overall, this is a very well written paper, and a solid contribution to the field. I would recommend it for publication following revisions.
There are two major issues that need strengthening before publishing.

The accuracy of the reconstruction method, which is treated as "truth" for the region's snowpack, needs to be more carefully assessed. In particular, all of the citations demonstrating the method works refer to the Sierra Nevada. Given what is known about the sources of weakness in the reconstruction method (e.g., cloud cover on the actual date of disappearance, precipitation occurring after the date of peak SWE, errors in the atmospheric forcing terms), which issues are likely to be more or less problematic in Afghanistan compared to California and how are these errors likely to propagate to accuracy of the reconstructed SWE? (I realize that some issues may be hard to pin down, but the authors should be able to make some statements about the degree to which weather models are poorer in this region and the degree to which this will impact the reconstruction.)

We thank Referee #1 for her comments. Some of the comments regarding the accuracy of reconstruction in Afghanistan and ground truth in general were shared by Referee #2. Frankly, similar comments have been made by previous reviewers of this study also, which has made it difficult to get the study through peer review. We believe we've finally added a validation source that will address some of these concerns.

At the end of 2017 Oct, we received a dataset of in situ snow depth and other meteorological measurements from Afghanistan, Tajikistan, and Pakistan that we had been waiting on for about a year. Please see the revised Section 3.4 on Validation in Afghanistan. Although these measurements are all manual and do not include SWE, only snow depth, they are far better than any other ground truth accessible to us, or likely available to anyone. For instance, through CRREL, we have access to US military weather station measurements in Afghanistan, which are sadly all going offline as they abandoned. Yet these weather stations provide no snow measurements. We have also never seen manual snow pit measurements from Afghanistan.

We find that using these in situ snow measurements, albeit with point to area extrapolation problems, geolocational uncertainty in MODIS pixel locations, and uncertainty in the density model, provide a better picture of the accuracy of our SWE reconstructions than speculation about cloud cover, quality of forcings, and precipitation after peak SWE.

I would also like to see more analysis regarding potential issues with the daily-reset cold content of the snowpack in particularly cold regions. This could be examined by running a model in the traditional forward sense (with meteorological data from this region, fully accounting for multi-day accumulated "cold content") and comparing it to a model run in reconstruction mode.

Likewise, we find that a full evaluation of the daily cold content scheme not possible with the data available to us in Afghanistan. To fully evaluate whether or not the pack is ripe, snow pit, bulk temperature, or lysimeter measurements are needed. At the least, to run a model like SNOWPACK, hourly energy and mass balance forcings with snow depth or SWE are needed. We still do not have in situ measurements of these forcings available anywhere in the watersheds of Afghanistan.

In our view, the cold content scheme has been shown to work well in the Sierra Nevada and Rocky Mountains (Jepsen et al) and at predicting the onset of lysimeter discharge (this study). We now know from the FOCUS station measurements that air temperatures are quite warm, which is why all the stations classified as warm snow types: alpine, maritime, or prairie, We

conclude that most areas of Afghanistan have a snow climate similar to the Rockies (alpine) and the Sierra (maritime). Thus, we suggest that the cold content scheme can be justifiably applied to most of Afghanistan's watersheds. We agree that the daily cold content scheme should be tested in cold regions (e.g. taiga and tundra snowpacks). There are probably areas like this in Afghanistan above 6000 m, but do not have the in situ measurements to test this claim. In fact, we are not aware of any areas in all of High Mountain Asia with full energy/mass balance instrumentation at these altitudes.

2. The introduction nicely makes the connection that Afghanistan's water supply is susceptible to year-to-year variations in snowfall and that some way of making seasonal predictions of the snow available for runoff is very important. This paper demonstrates a way of doing this. However, the paper needs to clearly make the connection of how the errors inherent in the proposed method (order of 20%) compare to the errors in the current system. For example, what is the interannual variability in snowpack? How wrong would a water manager be if he/she just presumed mean runoff from snow? What methods are currently used for such a forecast, and what are their errors? (Are there any citations on this?) I'm guessing that 20% error is better than the current situation, but the actual numbers (or a best guess to the actual numbers) should be presented in the discussion and conclusions.

This is a good point and we agree that our errors should be put into context in terms of operational utility. There is little in the way of water management in Afghanistan so we'll have to use hypothetical examples. More common is that snow and glacier melt fed streams run dry in the fall without warning (e.g. Introduction).

Please see our addition of Nash-Sutcliffe efficiencies on p 10 l 1-7, which show improvement over a mean forecast for all years.

Note that Figure 2b already shows the interannual variability in absolute and relative terms. We've added a few sentences describing these figures (p 5 l 28-31)

Some more minor issues include:
1. Given that only fSCA and mean reconstructed SWE had predictive power, why not test a simpler model with just those terms? How does that compare with the full set?

A model with just those predictors should work almost as well. Building such a model is certainly less effort and something we will try in the future based on our results. However, we fail to see why a simpler model needs to be employed here given that the machine learning techniques employed should be very robust against overfitting. Moreover, the passive microwave data, while not correctly estimating deep snow, help with the shallow snow and in areas with more cloud cover could conceivably help with estimating snow-covered area.

Bagged trees and subsequently random forests were invented to prevent overfitting. In our case, the hyperparameters for the bagged trees–which include the number of variables sampled for each tree, the minimum leaf size, and the maximum number of split–were optimized by minimizing a cross-fold validation error. The end result should be that unimportant variables do not affect the model.

The neural network model is more of a black box with more potential for overfitting, however since the results are nearly identical to the bagged trees, we conclude that overfitting is not an issue.

Also, given the conclusion that only those variables mattered, why does the conclusion say that an operational system would need to ingest Passive Microwave data? Does that make a difference that warrants the effort?

Ok, we've removed the PM reference part from the conclusion.

2. The discussion should also address the implications of combined error from the forward prediction (which was trained on reconstructed SWE) and the errors in the reconstructed SWE (which the one point check suggests may be biased low 20%).

With added section 3.4, we find it unlikely that our reconstructed SWE results are biased and given the uncertainty in our density model, they have very low errors.

How large might these combined errors be, and combined, are the expected errors still better than a baseline assumption of an average year?

We've added NS statistics from Section 4 to the conclusion.

Note, I am also providing an annotated manuscript to the editorial office and the authors, which marks in the text where the issues summarized here arise.

Needhamn, J.: Water Balance and Regulation Alternative Analysis for Kajakai Reservoir using HEC-ResSim, US Army Corps of Engineers, 58, 2007.
Vuyovich, C., and Jacobs, J. M.: Snowpack and runoff generation using AMSR-E passive microwave observations in the Upper Helmand Watershed, Afghanistan, Remote Sensing of Environment, 115, 3313-3321, doi 10.1016/j.rse.2011.07.014, 2011.