Smith et al. present an important study using ICESat-2 altimetry data (ATL11 data) to evaluate three surface mass balance and firn models (MAR and two GSFC models) over the Greenland Ice Sheet. These two processes have been important to be considered when using altimetry data to estimate ice mass balance, in particular, to separate ice mass changes from firn compaction based on volumetric changes. Seasonally repeated surface elevation measurements from ICESat-2 provide an excellent opportunity to evaluate the firn models over an ice-sheet-wide scale which had been validated using sparsely distributed firn cores. The authors thoroughly compared ICESat-2-derived height changes with model-estimated height changes caused by surface mass accumulation+ablation and firn compaction. There are several points that need to be clarified and/or discussed. See comments below:

This study considers the surface height change anomalies and SMB/FAC anomalies over the areas with little variability of flow velocities. Although the ice-dynamic induced height changes (anomalies) can be neglected, how would the variations of local topography/roughness with (fast) ice flows affect the evaluation? This may have little impact for large-scale evaluation when the data are aggregated to a coarse resolution grid, but it would be good if the authors can comment/clarify on this point.

Correction of firn compaction has been a critical step when using altimetry data to estimate the ice mass changes. RACMO has been more widely used in literature to correct for this effect. Although it may fall out of the scope of this study, it would be very helpful for the community if the authors can comment/discuss the RACMO firn estimates as well.

Line 19. Specify the names of the three FD/SMB models evaluated in this study.
Line 22. Specify the names of the two models mentioned here.

Line 25. Specify the name of the third model here.

Line 186. Why did the authors use MARv3.5.2 for this step? How would the difference between MARv3.11.5 and MARv3.5.2 affect the evaluation? The reasons and potential biases should be clarified.

Section 2.3.1. This part (especially the first two paragraphs) is difficult to follow. Could the authors use some equations to explain the regression analysis done here?

Section 3.2.3. This part is hard to follow too, with those scaling parameters and standard deviations. It would be helpful if the authors wrote some summary/topic sentences at the beginning of this section.

Line 465. “...but the melt for GSFCv1.1 was based on a degree-day parametrization of the MARv3.11.5 melt...”. Here is confusing. Did the authors use MARv3.11.5 or MARv3.5.2 to calibrate the degree-day model? MARv3.5.2 was mentioned in the methods part.