I only focus on changes asked by the reviewers and that authors chose not to take into account. My comments are in bold.

Reviewer#1

P. 5, L. 30; P. 6, L. 1: I think the results from RACMO2.1 forced by HadGEM2 is not necessary in this paper, because it can confuse readers who do not know much about RACMO2. If the authors think this part is really important for this paper, they should at least indicate key differences between RACMO2.1 and RACMO2.3p2 briefly in Sect. 2.2. Also, brief introduction of HadGEM2 would be needed as well. P. 6, L. 15-16: Same as the above comment. We deem that the comparison is valuable and shows how ESMs climate forcing has improved in time. To keep the manuscript concise and because the differences between the various RACMO2 model versions and associated forcing have been previously discussed in Van Angelen et al. (2013a,b) and Noël et al. (2015; 2018), we prefer to directly refer the reader to those publications. We included the following sentence in P6 L11-13: "For additional information about the HadGEM2-forced RACMO2.1 simulation and settings, we refer the reader to Van Angelen et al. (2013a); key differences between RACMO2.1, RACMO2.3p1 and p2 are discussed in Noël et al. (2015; 2018)."

I agree that RACMO2.1 forced by HadGEM2 is not necessary in this paper. We cannot know if the difference is coming from RACMO2.1, the downscaling method to 1 km, of HadGEM2 vs. CESM2, whereas only the later is of interest for this article. I think that Fig. 2 is not useful and should be removed, on focus only on the difference between 11 km and 1 km, as SMB components are shown in Fig. 3.

Reviewer#2

2) The authors focus their analysis only on the GrIS surface mass balance If this study should become a standalone piece of work without the promised future projections, then the authors should be highly encouraged to consider at least a subset of other parameters to validate their single-simulation analysis to exclude the likelihood of compensating biases leading to a "correct" SMB due to "false" physical reasons - (a) Surface energy budget vs. observations (b) Albedo vs. observations (c) Temperature and/or cloud properties vs. observations. We decided to limit the evaluation to SMB measurements, as the ability of CESM2 to represent key surface processes (including the near surface climate and the surface energy budget, SEB) has been addressed in other recent publications that emerged from the CESM2 development phase, e.g. Van Kampenhout et al. (2019) and Sellevold et al. (2019). In addition, direct comparison to daily in situ measurements (e.g. PROMICE, GC-NET) of (a) S EB components, (b) snow albedo, (c) near-surface temperature and cloud properties is not appropriate since ESMs, as opposed to reanalysis, do not assimilate observations and hence cannot reproduce the actual weather and exact timing of extremes (as in e.g. 2010 and 2012). See also our response to Reviewer #1 on Figure 3. We therefore deem the good agreement with in situ SMB measurements in different regions of the GrIS, characterized by very different climate conditions, to be a solid model evaluation, especially in view of the excellent agreement with temporal mass loss from GRACE.

I agree with reviewer#2 that given that the aim of the article is to validate CESM2, the authors have to show that CESM2-RACMO2 SMB is right for the right reason. The authors could focus on evaluating the forcing fields of RACMO2 compared to ERA-40 and ERA-Interim in a climatological point of view, for all ensemble members, similarly as done in Fig. 4, adding more variables such as T850, Z500. In particular, is the increase in temperature in

CESM2 ensemble related to circulation change after the 90's as observed (Hanna et al. 2018, cited by the authors in the introduction), or to high climate sensitivity?

Looking at separate surface energy balance components as suggested by reviewer#2, in a climatological point of view, can also help answer if RACMO2-CESM2 and RACMO2-ERA are increasing melt after the 90s for the same reason.

P4.L25ff: But what about other parameters such as the surface energy budget, temperature and clouds? How does it compare to recent circulation and cloud anomalies over Greenland which have been shown to be important for future projections? Upper atmospheric temperature (T700) in the CESM2 forcing is now evaluated using ECMWF reanalyses in Fig. 4a. See also our response to scientific assessment #2. Addressing circulation and cloud anomalies is beyond the scope of this study: this work assesses the ability of the CESM2 climate forcing to reconstruct the present-day SMB of the GrIS.

I disagree with the authors and agree with reviewer#2. I can't see why assessing clouds and circulation changes is beyond the scope of this study. It is the core of this study to assess whereas CESM2 is doing the right SMB for the right reason.