This study explores the implementation of dry snow and firn densification formulations in the Land Model of the Energy Exascale Earth System Model. Firstly, the snowpack domain is extended to greater depths and simulated at a higher vertical resolution. Secondly, the authors compare the performance of four different densification formulations with empirical strain rates generated from the analytical firn model of Herron and Langway (1980). Finally, they evaluate results of two of the four formulations with some firn cores from the Greenland ice sheet and from Siple Dome in Antarctica.

Improving firn densification schemes in Earth System models is a valuable objective. However, many aspects of the study need to be significantly improved to possibly meet the standards of Geoscientific Model Development (GMD). I list below my concerns about the contents and the structure of the manuscript. I relate my remarks to some review criteria, which are available on the website of GMD. I also try to outline possible avenues for the authors to improve their work. My comments are separated in Major comments, which address general shortcomings of the study, and Specific comments.

Major Comments

1) The lack of information about the optimisation

Here are two review criteria of GMD:

> *Are the methods and assumptions valid and clearly outlined?*
> *Is the description sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)?*

I focus first on the optimisation method applied (Section 3.2). The entire optimisation method is described in a single sentence (lines 184-186):

"*From our estimated empirical strain rate-versus-depth data, we optimized the previously described densification model coefficients (from A'76, vK'17, and A'10) by applying a regularized least squares algorithm for two stages of densification (above and below* ρ *= 550 kg* m^{-3}*).*"

The A'76 model includes 7 different coefficients ($c_1, c_2, c_3, c_4, c_5, \rho_{dm}, \eta_0$), the vK'17 includes 10 ($c_3, c_4, c_5, \rho_{dm}, \eta_0, c_\eta, f_1, f_2, a_\eta, b_\eta$) and the A'10 includes 7 ($c_3, c_4, c_5, \rho_{dm}, k_c^{\rho<550}, k_c^{\rho>550}, E_c$). Additionally, in the Results section, the authors mention "*adding a constant compaction term*" (line 277), which does not appear in any equation and is not explained in the Methods section. Throughout the manuscript, the authors never state which coefficients are subject to the optimisation. Moreover, they mention in Section 4.2 that " *we have yet to test in ELM an optimized version of the semi-empirical model*". From my understanding, the "*semi-empirical model*" is A'10 and they decided to compute the ELM simulations with the original version of A'10 and not the optimised version. In contrast, for some reason, the authors did the ELM simulations with the optimised version (vK'17+) of vK'17. They still provide speculative avenues for a re-parameterisation of A'10 at lines 322-328. These statements are not supported by any quantitative information about a better fit of the optimised A'10 either to observed data or to the strain rates generated from the model of Herron and Langway (1980) (referred to as HL hereafter). Finally, they assert that they optimise A'76 ("*we optimized the previously described densification model coefficients (from A'76, vK'17, and A'10)*"). However, the only information to be found about the modifications brought to the model is the change of the value of c_5, but nothing about other parameters and nothing about a performance comparison between the original A'76 and the optimised version.

Coming back to the optimisation methodology itself, the authors decide to select annual mean temperatures only below -25 °C, but they then proceed to model simulations for grid cells where the annual mean temperature is as high as -20 °C (Section 3.3, Figures 2, 3 and 4). It is puzzling that the authors themselves suggest a better approach to selecting mean annual temperatures and accumulation rates, which would be easy to implement (lines 357-361). They also claim to calibrate the models by matching the computed strain rates to the HL strain rates. The issue is that several models use dynamic variables in the strain rate equations: T, σ, P and r_e in Eqs 1, 2, 3, 4, 8, 9 and 10. The HL can provide analytic solutions of strain rates for steady state annual mean temperature and accumulation. The dynamic models require values for the dynamic variables at each time step and for each layer of the firn column. A steady state annual mean temperature does not correspond to all firn layers having the same temperature year-round (temperature still varies seasonally when in steady state). Similarly, accumulation rate still varies seasonally, which means that σ also varies in time for any firn layer (again, even in steady state). Finally, the reader has no information about how r_e is calculated in the computations of these steady state strain rates.

Furthermore, the "*regularized least square algorithm*" is not described. Why not proceed to a simple least square? What is the penalty term? What are the penalised factors? And, most importantly, which coefficients are subject to the optimisation and what is the range of possible values covered by the optimisation?

I think that the authors can easily understand that the issues I raise here are concerning with respect to the GMD review criteria.

2) The ELM firn density simulations

My first concern relates to the data that is used for model evaluation. Why use the cores of Mosley-Thompson et al. (2001) and Lamorey (2003)? And why do the authors average the Greenland cores? They highlight themselves that "*variability can be large, particularly across the GrIS*" (line 337). Why not compare an observed firn core to the model simulations for the grid cell of the corresponding location? Averaging observed and modelled firn depth-density profiles makes little sense. The authors themselves seem to point out this shortcoming of the study (lines 318-320): "*though our analysis with ELM thus far is limited to a generalized comparison with a broad (climate) perspective rather than to a more site-specific comparison against direct observations*". So why was a site-specific comparison not performed?

Moreover, the data selected is not in line with the objective of the study: improving dry densification schemes. Most of the Greenland grid cells are likely affected by melt, and Siple Dome is an area of Antarctica with relatively high melt rates for the continent. Why don't the authors select data only from dry snow areas (higher accumulation zone of Greenland and more inland regions of Antarctica)? Do the authors know about the extensive SUMup dataset (Koenig and Montgomery, 2019) that includes many more firn cores? The occurrence of melt is clear because there is "*formation of ice lenses*" (line 241). But no information is provided about the model schemes for simulation of meltwater percolation and refreezing. Moreover, the simulations are performed with atmospheric forcing at very coarse resolution, which is underlined at lines 342-347 (I mention here that the resolution is not provided in the manuscript). This forces the authors to artificially adjust their evaluation: "*this large grid cell remapping lead to a cold bias, resulting in too-slow densification. Therefore, we adjusted our Siple Dome comparisons to include grid-cells away from the coast that better represent atmospheric conditions and result in a more realistic density simulation*". Firstly, I would think that grid cells away from the coast should be even colder and thus enhance the cold bias. Secondly, this further underlines the question of why choosing Siple Dome and Greenland firn cores to evaluate the models. This choice brings in problems related to the adequacy of the model forcing, which makes it very difficult to disentangle firn model deficiencies from errors due to inadequate forcing.

The spin-up period is taken to be 260 years. This is likely too short for low-accumulation grid cells (thus most of the dry snow zone) to build a full firn layer (i.e. until ice density is reached). The authors should thus support their statement (lines 226-227) that the profiles "*averaged from the final 100 years of simulation results*" (thus starting the averaging only after 160 years) are "*steady-state density profiles*". For example, after 160 years of an accumulation rate of 0.07 m w.e. yr^{-1} (the limit assumed for warm dry snow zones in Section 3.2), only 11 m w.e. have accumulated, which corresponds roughly to a firn column of 20 m. I doubt that this represents a steady state. As shown in Figure 4 (T=-34°C), firn that is 100 years old is only at 600 kg m^{-3} density, showing that the firn layer is most likely not in steady state after 160 years and not even after 260 years. Concerning the fresh snow density, the A'76 model calculates surface densities by itself, while vK'17 and vK'17+ use a fresh snow parameterisation (not given in the manuscript…). But how is fresh snow density calculated for the A'10 model? The prognostic equation for r_e should also be given or referred to.

The approximation of vertical strain rates (line 244) is also unclear. This raises the same questions that I mentioned above about the dynamic variables when assuming a steady state. In my view, the authors should include a detailed explanation about how the steady state strain rates of A'76, A'10, vK'17 and vK'17+ are calculated. This holds for both the calibration step as for the ELM simulations (i.e. the values appearing in Figure 4).

When analysing and evaluating the results, there is a severe lack of quantification. This holds for both the Equilibrium climate simulations and the Twentieth century climate simulations. I give some examples:

- "*the semi-empirical model improves the density profile*" (line 260): improves with respect to what (I guess that the authors mean A'10 improves the density profile with respect to A'76)? And the statement of "improvement" should not be based on a mere visual comparison of Figures 2 and 3. Moreover, it should be clarified that the authors evaluate the model results against results from HL, which is not a guarantee of model accuracy.
-"*Densification tapers-off at lower densities (around 450 kg* m^{-3}*) for colder climates, a temperature-dependent effect enhanced with the model from Arthern et al. (2010).*" (lines248-250): from Figures 3 and 4, it is not obvious that this effect is tronger in A'10 than in vK'17+. The enhancement of the effect should thus be quantified.
- "*A lower model variance occurs when it does not covary with the empirical model. This effectively reduces a model's prediction risk if it does not also result in an increased bias.*" (lines 303-304): If the authors discuss about the bias of models, they can simply add a "Bias" column in Table 2.
- "*both models show improvement compared to their original counterparts (ELM v1 and CLM).*" (lines 337-338): this is impossible to evaluate for the reader because (1) only the results of A'10 and vK'17+ are shown in Figure 6, and not

the ones of their so-called "*original counterparts*" (which are A'76 and vK'17 I suppose), and (2) there is no quantitative evaluation of the models' performance with respect to the observed data (e.g. RMSE, bias, etc.).
- "*Encouragingly, our simulation results compare well with firn density measurements and indicate an improved capability in the ELM.*" (lines 348-349): same remarks as for the previous point.

3) The novelty and objective of the study
GMD review criterion:

Does the paper present novel concepts, ideas, tools, or data?

Firn model optimisation has been addressed in numerous studies over the recent years (e.g. Ligtenberg et al., 2011; Kuipers Munneke et al., 2015; van Kampenhout et al., 2017; Smith et al., 2020; Verjans et al., 2020). Four of the studies mentioned have already investigated the optimisation of the model of Arthern et al. (2010), for Antarctica, Greenland or both. An easy and straightforward way to improve the ELM would be to implement the parameterisations developed in these studies. If the authors want to address the same problem, they should propose a new, original method. However, in contrast to the existing literature, they do not calibrate the model of Arthern et al. (2010) with observations but with HL-computed strain rates. And, as mentioned above, it is unclear to me how they calibrate a dynamic model to steady state strain rates. They should justify why their methodology is better suited to their objective than using what other researchers have already accomplished. Moreover, the objective stated in the conclusion of improving the capability of the ELM to better simulate refreezing rates in firn is not in line with the study itself. The focus of the calibration is on dry firn densification and does not support the statements at lines (377-381): "*With an evaluation of the simulation of dry firn densification, we have optimized the ELM firn model for future studies of the impacts of liquid water on firn density and SMB. Ultimately, this study seeks to enable better predictions of SLR as a direct result of surface melt and mass loss from the GrIS.*"
If the authors do want to better capture liquid water effects, they should focus on this very challenging topic by studying the mechanisms of wet firn compaction, meltwater percolation and refreezing.

4) The clarity of the manuscript
GMD review criterion:

Is the overall presentation well structured and clear?

It is very difficult for the reader to understand the different steps of the study. The authors alternate between different ways to refer to a same thing. For example, the A'10 model is sometimes referred to as "*A'10*" and sometimes as "*the semi-empirical model*", the vK'17 model is sometimes referred to as "*vK'17*", sometimes as "*the CLM*" and sometimes as the "*Snowpack model*" (see Figure 5). Similarly, it is never clearly stated that the "*empirical strain rates*" correspond to the ones computed with the HL. A first, simple way to improve the clarity would be to consistently use the terms A'76, A'10, vK'17 and vK'17+ throughout the manuscript, including in the captions of the Figures. In line with this, the Section 3.1 should be split in four subsections that clearly detail each of these four models instead of subsections presenting equations which are subsequently assigned to the models in a confusing way.
The Figures and Tables also lack clarity. In Figures 2 and 3, why are high values of accumulation only shown at depths greater than 15 m? Even in a high-accumulation climate, there will always be a shallow and a deep part of the firn column. And how were the surface density values chosen for the HL-computed profiles? The caption of Figure 1 mentions that the new firn model "*can extend as deep as 80 m*", whereas it is always presented as a "*semi-infinite*" grid in the text. Which of these two statements is true? In Figure 4, why are there points without a vertical error bar? And why do some points have a horizontal error bar (age should be well-determined for any firn layer of any model run)? In Table 1, equation numbers could be provided for each model to know which equations apply to which model. In Table 2, the model names should be used in the column "*Densification model*" instead of the mechanisms applied and the variable for which the statistics are calculated should be specified in the caption (presumably strain rate values). Improving the structure of the manuscript could possibly help decrease the degree of confusion for the reader when trying to understand the study.

Specific comments:
line 2:
Change "*consist*" to *consists*.
line 15:
I doubt that any paleoclimate study uses Earth System Models to determine pore close off depth and timing.
line 25:
Repetition of "*coupled*".
lines 32-35:

As far as I understand, there is a contradiction between "*does not yet exist*" and explaining the implementation of the advanced firn model in the CLM.

lines 47-48:

Change "*those predicted by Herron and Langway (1980)*" to *those predicted by the model of Herron and Langway (1980)*. Moreover, I suggest using a consistent way to refer to this model (e.g. HL'80).

lines 52-55:

Add an explanatory sentence about the fact that snowpack models and firn models also have a different vertical scale of application.

Section 2.1:

Provide units of all the variables and quantities presented. This will make clear that there are some unit inconsistencies in some of the equations (which I give below).

Equations 1, 2, 3 and 4:

The variables $\dot{\varepsilon}$ and $\left(\frac{1}{\Delta z}\frac{\partial \Delta z}{\partial z}\right)$ are equivalent to each other as far as I understand. Use either one of the two notations.

line 78:

Specify if $\left(\frac{1}{\Delta z}\frac{\partial \Delta z}{\partial z}\right)_{dm}$ is also considered in CLM (v5).

line 82:

In CLM(v5), c_η = 358 kg m^{-3} and f_2 was set to 4. Only f_1 accounts for the effects of liquid water and not $c_\eta/(f_1 f_2)$.

line 89:

The characteristic depth is not "*a single valued proxy for a given site's full density profile*" but only for the upper density profile.

line 89:

No s at *stages*.

line 89:

Add here the explanatory sentence about why models assume a two-stage densification process.

line 92:

The sentence "*Empirical firn densification models typically employ analytic functions that assume a steady-state density profile*" is not true. Only the model of Herron and Langway (1980) and a few others provide analytic functions but almost all of the recent firn models are dynamic models.

line 93:

Rephrase.

Equation 5:

This equation is erroneous. The units of the left- and right-hand sides do not match. The correction is: $w(z) = \frac{A}{\rho(z)/\rho_w}$, where ρ_w is the density of water (1000 kg m^{-3}).

Equation 7:

Again, this equation is erroneous and there is a unit inconsistency. The correction suggested above, fixes the error.

Equation 8:

The variable *P* is defined here as the "*overburden pressure*", which makes it equivalent to σ. I suspect that this variable corresponds to the *P* as defined in Equation 9, which should be called the *grain-load stress*. I underline here that in the model of Arthern et al. (2010), it is really σ that is used and not the grain-load stress. The authors should explain why they differ from the original model of Arthern et al. (2010) on this point.

lines 176-177:

Specify that the "*plausible firn density-versus-depth profiles*" were computed with the different models and the HL.

Section 3.2:

Why do the authors decide to draw annual temperatures from a distribution representative of the global Earth climates instead of the polar climates? Is the objective to have much more values close to T=-25°C? This should be clarified.

How are all these values decided:

- -25°C as a threshold (whereas ELM simulations involve warmer sites)
- -51°C as limit between low- and high-accumulation sites (many sites can have T>-51°C and A<0.07 m SWE yr^{-1})
- surface density values between 300 and 380 kg m^{-3}

line 196/

Specify the resolution of "*coarse-resolution*".

line 197:

What does "*(an "I-compset" at ne11 resolution)*" mean?

line 199:

Change *January 1st* to *January 1st 1901*.
line 206:
Define "*restart runs*". In general, it is good practice to define any term used that may not be straightforward to everyone reading the study.
Table 1:
Add a column for ρ_{dm} values. Add another column that indicates which equations apply to which model, with the corresponding equation numbers (see Major Comment 4).
lines 209-215:
All the details provided about the observational data should be given in a separate subsection. Are the sites of measurements affected by surface melt? And are these firn core measurements open access?
Section 4:
There are a lot of speculative statements in this section. I suggest splitting it into a section Results and a section Discussion, so that the reader can distinguish between model results and the thoughts of the authors.
line 219:
Change *accumulations* to *accumulation rates*.
lines 217-222:
Here, the entire methodology is again defined. This can be confusing for the readers. For example, I suggest rephrasing the sentence "*To improve the accuracy of our firn model simulations, we optimize compaction terms against empirical strain rates using statistical modeling*" because this was the point of a previous section.
line 229:
Does the statement "*the mean annual temperature is within a couple degrees of -25 °C*" refer to the results of Figure 2 for T=-27°C? If so, it would be clearer to give the exact mean annual temperature value.
lines 236-237:
"*These simulations demonstrate a stronger effect of temperature on densification rates, resulting in more variation in density with depth (Fig 3).*" I think that this sentence means more variable depth-density profiles according to the mean annual temperature. If so, it should be rephrased.
lines 242-243:
Specify which models use the "*better fresh snow density parameterization*".
lines 260-261:
Strange use of commas.
line 264:
What is "*over-densification*"?
line 265:
The notion of "*density profiles that vary too weakly with depth*" should be replaced with the one of density that increases too weakly in depth.
line 265:
What does "*Their*" refer to?
lines 274-277:
Are these the only coefficients included in the optimisation? Or the only ones that were decided to be changed? See Major Comment 1.
line 276:
The coefficient f_2 is related to the grain radius (see Vionnet et al., 2012). If the ELM calculates grain size, f_2 can be calculated accordingly. It is crucial that the authors clarify why they decide not to follow the original formulation of f_2 (as a function of grain size) but to consider it as a pure tuning factor. This is all the more relevant since it is emphasised throughout the manuscript that firn models need to account for microphysical features.
line 286:
Specify that the density model coefficients are calibrated to the HL-computed strain rates.
lines 288-290:
Specify the variable for which the statistics are calculated. See Major Comment 4.
lines 293-298:
How do the authors explain these results?
Table 2:
The mention to eq. (5) is an error because it is not the one for destructive metamorphism and c_5 does not appear in this equation.
line 300:
The value of R^2=0.67 is valid only for strain rates in the second stage. Note also that Table 2 shows R^2=0.66.

line 303:
"*A lower model variance occurs when it does not covary with the empirical model*". This statement sounds like a general statement, but I believe that it is applicable only to the results of this specific study.
line 306:
What does "*these results*" refer to? The paragraph above is about variances in the compaction rates.
lines 306-307:
"*negative correlations between overburden pressure and empirical strain rates*": this is explained by higher overburden being applied to deeper firn, which is at higher density and thus compacts less. A simple explanation could be provided to the reader.
lines 311-312:
Note that Equations 3, 4 and 8 are directly dependent on density.
lines 320-328:
Are these results or speculations? Where do these values come from? And did the "*statistical computing*" focus only on these specific coefficients of A'10? Such conclusions should not be stated without quantitative results to support them. I emphasise again the need to clarify the optimisation method and its results.
lines 331-332:
Why did the authors decide to use the optimised vK'17 but not the optimised model of A'10?
line 333:
Typo "*the the*"
line 341:
"*ne11*" is not defined.
line 350:
The statement "*we should focus on the near surface layers, as they contain the primary SMB components*" should be clarified.
line 351:
"*it could be necessary to model the upper most 20 m*": Is this figure of 20 m supported by the results? If not, references should be provided.
lines 352-353:
"*the optimized version (vK'17+) is likely the better choice for implementation into the next major release of the E3SM*": Again, it is unclear if this is speculative or if the results provide strong evidence for this. And how would an optimised version of A'10 compare to vK'17+?
lines 353-355:
A low bias is usually preferable to a high bias. What is meant by this sentence?
line 364:
"*In the near future, this could be the entire GrIS*": What is the "*near future*"? And references should be provided.
Conclusions:
See Major Comment 3.
line 370:
Change "*with steady-state empirical models*" to *with a steady-state empirical model*.
lines 370-371:
I do not believe that the analysis is "*similar to that by van Kampenhout et al. (2017) for CLM*".
Code and data availability:
The availability of the firn core measurements should be provided in this section.
Figures:
See Major Comment 4. Also, in all Figure captions and legends, the models should be consistently referred to as A'76, A'10, vK'17 and vK'17+.
Figure 1:
The vertical extension of "*80 m deep*" (caption) contradicts the "*semi-infinite*" stated in the text.
Figures 2 and 3:
The vertical/horizontal lines at ρ=550 kg m^{-3} and ρ=550 kg m^{-3} can be removed to improve the clarity of the Figures. Why do high accumulation rate values appear only at great depth?
Figure 4:
"*for various plausible accumulation rates (0.11–0.50 m SWE yr.$^{-1}$)*": Why is 0.11 m SWE yr.$^{-1}$ chosen as lower limit? The construction of the horizontal error bars is not clear to me.
Figure 5:
Do these plots show the results of A'10 (top) and vK'17 (bottom)?