We thank the reviewer for taking the time to read through our paper so carefully, for their detailed review and for their insightful comments. Please find our replies below as inserted in red text.

General comments

The paper by T. Malle et al. addresses the effects of spatial resolution, atmospheric forcing data, and land surface characterisation on snow depth, gross primary production (GPP) and evapotranspiration (ET) simulated by the Community Land Model version 5 (CLM5) over Switzerland. Factorial model experiments with different combinations of resolution and input data were used to isolate the effects of the three sources of variability in the model output. The authors conclude that deficiencies in all three aspects contribute to the current uncertainty in the results of land surface models (LSMs), particularly in heterogeneous regions like the Swiss alps. In this context, they call for the use of more fine-grained input data and for evaluating LSM simulations at high resolution.

The study contains several elements that are interesting for the audience of ESD, including modelers and users of land/climate model output over heterogeneous regions. It probably involved huge efforts, considering the generation of various input datasets, model experiments, the analysis of three simulated quantities (snow depth, GPP, ET), and their comparison against observational or modelled benchmarks. For a high-quality modelling paper, I think the technical foundations need to be improved (e.g. what controls simulated snow depth, GPP, and ET in CLM5-SP offline?). Also, the study addresses several complex issues in land/climate modelling (e.g. ecophysiology, carbon cycle dynamics, feedbacks, etc.) that are poorly captured in the model setup chosen for this study (i.e. CLM5 with prescribed vegetation phenology, largely inactive carbon and nitrogen cycles, and prescribed atmospheric conditions). I think this distracts from the relevant key findings and confuses the overall storyline. Below, I list suggestions how the study and its findings could be improved to become more useful for the audience of ESD. I think this involves substantial revisions. My comments refer to main text only, not the Appendix.

Thank you for this comment, and for the generally positive assessment of our work. Our goal is to maximize the paper's usefulness for the modelling community, so we very much welcome your suggestions for improvement.

Initially, we opted to use CLM5 with prescribed vegetation phenology to represent the vegetation of Switzerland today in the most accurate manner, which would be very difficult to achieve when running in prognostic biogeochemical mode, at least without data assimilation. We also did/do not have the computational resources to perform CLM5 simulations at high resolution in bgc mode, especially not the required spin-up. We will therefore follow your suggestion of re-focusing the manuscript on snow-related analysis and will not focus on the link between snow and GPP/ET.

We will elaborate in our response to the specific comments below.

Specific comments

(1) Relevance of higher resolution and implications of the results. It is obvious that higher resolution and more specific input data improve model performance for a specific location, i.e. when CLM5 output is evaluated against high-resolution simulations or point/site-scale observations. This is known by the modelling community and it is thus not surprising that the 1km simulations with high native-resolution input data perform best. What additional insights can the study provide, e.g. should we focus on (1) high resolution atmospheric forcing data in offline land surface modelling, (2) high resolution land surface data, or (3) high LSM resolution to capture non-linear effects between atmospheric forcing and land surface characterization? Or should model evaluation and benchmarking be done at high resolution, to rule out deficiencies due to insufficient resolution and focus on process uncertainties? What do users get if they run CLM5 at high (<0.5°) resolution with CRU as the atmospheric forcing dataset, is CRU 0.5° simply interpolated and could "smart downscaling" considering the temperature lapse rate be useful as a (built-in) solution? (this would be very interesting for the CLM community!)

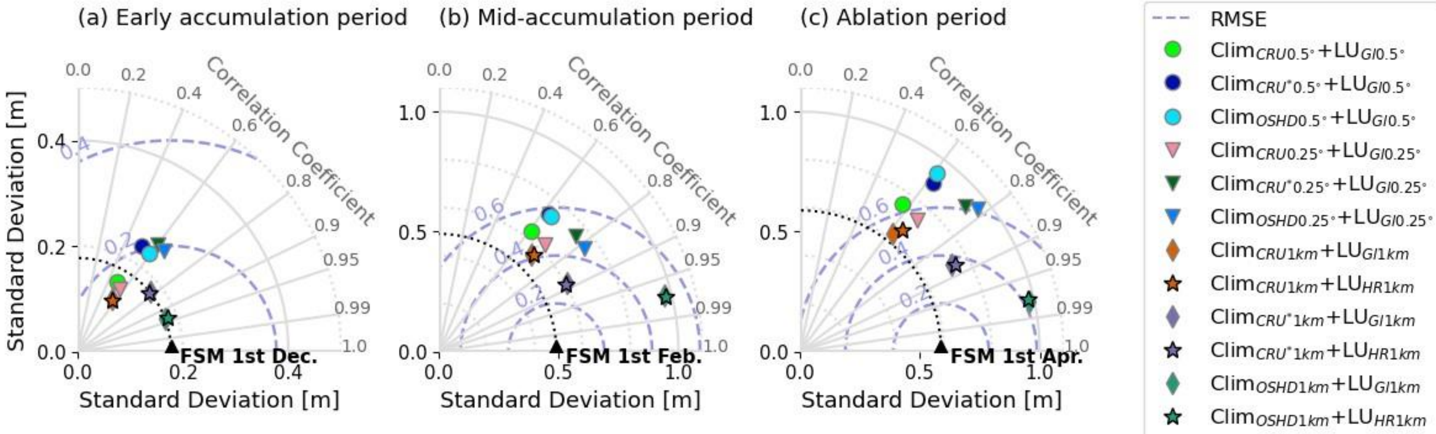
We will discuss our results in this context in greater detail in a revised version of the manuscript.

(2) Benchmarking and implications. The purpose of a land surface model like CLM5 is to simulate larger spatial scales ranging from one grid cell to regional domains or the globe, in spatially representative grid cells rather than point fashion (only at very high resolution, grid cells start to resemble point/site-scale conditions…). Therefore, it might be useful to evaluate the output of various simulations (1km, 0.25°, 0.5°) at the lowest resolution of 0.5°, to assess if there are major differences due to input data quality and non-linearities. Such results could inform the modelling community if there is a need to account for small-scale heterogeneity to obtain accurate fluxes and pools at larger spatial scales of one to several kilometres. This might ultimately be more relevant for the purpose of CLM5 (or LSMs in general) than mimicking point/site-scale conditions.

This is a very valid point. We have redone evaluations of gridded snow simulations at 0.25°, as we believe that given the complexity of the topography across our modelling domain and its relatively small size, and considering today's ever-increasing computational resources, 0.25° should be a fair target for the main analysis.

We have upscaled results from the 1km simulations to 0.25°, using a conservative upscaling approach which preserves areal averages. For this purpose, we had to decrease our evaluation domain slightly, as we performed the 1km simulations with a mask running exactly along the edges of our modelling domain, making it impossible to upscale these areas to 0.25° without crude assumptions. The 0.5° simulations were downscaled to 0.25°, and all simulations were evaluated across the same domain.

As seen from the updated version of the Taylor's diagram below (Figure 3 in original manuscript), the difference between different land-use datasets with regards to simulated snow depth remains small and increasing spatial resolution in isolation only has a marginal effect on accuracy of simulated seasonal snow cover. Upscaled 1km simulations with highest quality meteorological forcing datasets ($Clim_{OSHD1km}$) perform best during the early accumulation period but performance is matched/exceeded by the lapse-rate corrected global dataset ($Clim_{CRU*1km}$) for mid-accumulation and ablation period, underlining the effect of a relatively simple lapse-rate based downscaled temperature input to better account for sub grid variability. We will discuss these results in greater detail in an updated version of the manuscript.



(3) Land surface dataset methodology and evaluation. The generation of a high-resolution land surface dataset based on national land cover data and satellite-based LAI (not sure I interpreted the brief description in section 2.3.2 correctly) seems quite innovative and interesting. I think the procedures should be described in more detail, so that other could potentially follow a similar approach. Some summary/evaluation/validation beyond what is shown in Figure 1 (e.g. how did land cover fractions and LAI change across Switzerland, regionally, or for the point locations?).
We will describe how the land surface dataset was created in a more detailed manner. Additionally, we will add additional figures to the appendix demonstrating how different percentile fractions (e.g. crop, vegetation) and LAI/SAI changed spatiotemporally across the model domain.

(4) Overall scope including snow depth and GPP/ET. The manuscript might benefit from limiting the scope and analysis to snow depth, and in that area developing the causes for improved model performance more thoroughly (e.g. how can spatial resolution, atmospheric forcing data, and land surface characterization influence snow depth considering forced precipitation, rain/snow partitioning inside CLM5, land cover, LAI, slope, etc.). A caveat of this is, of course, that snow depth is likely closely linked to the forced precipitation and temperature fields, which could make the results appear trivial. Yet, the authors might identify interesting aspects related to, e.g., land cover, LAI or slope. In any case, there are multiple reasons why the analysis of effects on GPP and ET should be excluded, or included only after substantial revisions:

a) The link between Hypothesis 2 and the chosen methods is currently very weak. Most importantly, the modelling setup and the correlation analysis do not allow to isolate the effects of differences in snow depth from differences in its drivers (i.e. spatial resolution, atmospheric forcing data, and land surface characterization) on GPP and ET. The framing of GPP and ET as "snow-cover dependent ecophysiological variables" is confusing, because likely most of the GPP and ET differences are driven by the resolution-dependent forcing fields directly (i.e. atmospheric variables or land surface characterization) and not indirectly via snow-cover changes. This joint independent driver could lead exactly to the correlation between snow cover and GPP/ET found by the authors, without any effect of snow cover itself on GPP/ET.

b) In CLM5 in SP mode, vegetation phenology is prescribed as a climatological seasonal cycle of LAI, and LAI controls the leaf to canopy scaling of all fluxes including carbon (GPP) and water (ET). Therefore, in SP mode the model has very limited capabilities to show an ecophysiological response to snow cover change, in the sense of seasonally shifting growth. As a minimum, I recommend to discuss the precise implications (e.g. is the response we see solely due to changes in temperature and water availability, or what can affect GPP in SP mode at all?). However, I strongly doubt that an ecophysiological response can be quantified in SP mode. I imagine it works as follows, although I am not 100% sure: if the snow season is shorter than in the climatology, suitable temperatures for growth coincide with zero LAI in the model and nothing happens until the climatological growth begins; if the snow season is longer than in the climatology, suitable temperatures for growth coincide with high LAI in the model, leading to a jump start and no compensation later in the season.

c) Ecophysiology refers specifically to (plant) organisms and not to land as a whole. So, for ecophysiological effects on ET I suggest analyzing plant transpiration and canopy evaporation fields of CLM5, or total ET in the vegetated parts of grid cells excluding the bare soil PFT. However, considering that the experiments include changes in land surface characterisation, the best option might be to remove the term "ecophysiological" and to refer to responses in land ET, which include e.g. the effects of varying bare soil fraction.

d) Simulated GPP and ET are compared against the "best-effort" configuration of CLM5 at 1km resolution as a benchmark, and deviations from this benchmark are considered model uncertainty. An objective benchmark (e.g. observational data or output of a dedicated model, like for snow depth) is lacking. Also, I am not sure the term "model uncertainty" is appropriate in this context. Is it model uncertainty if lower resolution models perform poorer at producing high-resolution-model-like outputs? The comparison could potentially be done the other direction (see point 2).

Thanks for the suggestions and elaborate explanations. We will follow the suggestion of re-focusing the manuscript on snow depth and will not focus on the link between snow and GPP/ET. We will discuss rain/snow partitioning in more detail and include an analysis of snow-depth accuracy per elevational band. Figure A1 will be moved to the main paper and Appendix B will be removed entirely.

We would like to keep Figure 4 in the paper, however, not as a main result, but in order to demonstrate how different GPP/ET results can be in our various simulation setups. It will be integrated as part of the discussion and help pointing at the magnitude of uncertainties, which are related to questions like resolution, input data, and resolving vs. not resolving sub grid variability.

(5) Introduction: The background could be more technical (e.g. what controls simulated snow depth, GPP, and ET in CLM5-SP offline?) and focused on aspects that matter for this study (i.e. strengthen the main story and cut out things that are interesting but not directly relevant).
We will include a more technical paragraph in the introduction of a revised manuscript.

(6) Discussion: The findings need to be contextualised, considering the capabilities of CLM5 in SP mode offline (see above). For the comparison with other studies (e.g. Birch et al.) to be valid and useful for the reader, it would be good to understand what these studies did or how they explain their biases (e.g. also land cover specification, atmospheric forcing and model resolution, or something completely different like effects of grazing animals or artic plant types?). Certain

links made between the findings of this study (with a very specific setup and research focus) and model uncertainties (with sometimes known different root causes) are not appropriate (e.g. L371). I recommend reconsidering these and focusing the discussion more on new insights gained through this study (see points 1 and 2).
We will revise the discussion and focus it more on the highlights of this study.

(7) Language and figures: The manuscript is very well written. There are a couple of “empty” phrases that highlight something but do not actually deliver new insights (e.g. the results have profound implications, the study highlights the importance of model development, the study highlights the utility of multi-resolution modelling, etc.). I think those could be filled with content or removed. The figures have high quality and are visually appealing.
Thank you. We will remove empty phrases from the paper.

Technical and line-by-line comments
L14 ff: “Earth’s systems” sounds a bit unconventional and far-reaching; maybe use a more concrete/narrow term?
We will rephrase this sentence.

L19: add water ok

L23, L24: maybe use “influence” instead of “control/determine” ok

L24, L35, L65: there is no feedback among the mentioned dependencies/effects
We will use ‘exchange’ instead.

L29, L31: check logical link for “thus” and “as” ok

L49: offer literature for multi-resolution modelling?
We will cite Singh et al. (2021, https://doi.org/10.1002/2014WR015686) and Meissner et al. (2009, https://doi.org/10.1127/0941-2948/2009/0400) here.

L50: they allow evaluating ok

L51: I guess any gridded LSM can be evaluated against point/site-scale observations by taking the individual grid cells that match the locations best; there is no need for dedicated point simulations for this
In principle yes, but for this study we used downscaled meteorological forcing data to the exact location of the point observation for our ‘1km’ simulations, which in case of the OSHD dataset also included station observations and should represent a best-case scenario. Also, for the lapse-rate based temperature correction we used the exact coordinate and a high-quality GPS elevation measurement for the downscaling. We will make this clearer in the methods section of a revised version of the manuscript. We also acknowledge that our choice of naming of the point simulations might have fueled this confusion, which we will therefore update in a revised version of this work (e.g. simply $Clim_{OSHD}$ and $Clim_{CRU*}$ instead of $Clim_{OSHD1km}$ and $Clim_{CRU*1km}$) to make this clearer.

L56: what is meant by snow cover "dynamics"? the temporal evolution?
Here we mean depth, duration as well as variability of snow cover across space and time. We will make this sentence clearer.

L73: consider limitations in SP mode ok

L80: in my understanding, "process representation" refers literally to how processes are represented in the model, i.e. the equations used to calculate snow depth, GPP and ET; I think this term is not appropriate for the modifications in forcing/input data and resolution made in this study
We will avoid using the expression 'process representations' in a revised version of the manuscript.

L87: remove "heat fluxes" if not addressed in the results
Heat fluxes will be removed here.

L88-90: consider reformulating, the sentence sounds nice but it does not deliver any content (or at least I cannot understand it)
We will reformulate the sentence.

L98-102: the methods need some technical precision here: which state variables, which datasets (e.g. what time period/ past conditions do they represent), not only "natural" vegetation (crop PFTs), how does GPP work in SP mode (if the LUNA model active, also cite Ali et al. 2016 for photosynthesis)
We will cite Ali et al. 2016 here (LUNA model is active in our simulations) and give more technical background on our simulation setup.

L100: "is approximated": I think this a used choice, not done by the model.
We will rewrite this sentence.

L103: revise components of ET, e.g. soil sublimation does not make sense; maybe "ice" is missing?
Yes, soil should read ice here, we will update it accordingly.

L105-111: snow cover is the focus of this study, so I think the foundations of snow cover calculations should be provided for convenience (and understanding), including rain/snow partitioning in CLM5
We will document and explain CLM5 snow cover calculations including rain/snow partitioning in greater detail here.

L115, L197: PFTs are patch level, mention that prescribed LAI etc. is in SP mode.
We will update this sentence accordingly.

L116-L119: is there any difference between taking out individual grid cells from the regional simulations and running dedicated point simulations (i.e. a 1x1 regional grid) in your setup? they should be identical, provided that (1) the grid is anchored and specified identically, and (2) there are no lateral exchanges between grid cells (which depends on the CLM5 compset, if river routing is off there is no lateral exchange)

There is no lateral exchange in our model setup, so in principle there is no difference in the simulation setup. However, as mentioned above we used downscaled meteorological forcing data to the exact location of the point observation, resulting in different forcing datasets than for the 1km simulations. We will make this clearer in the text.

L124: for comparing between resolutions, it would have made sense to subdivide the 0.5° and 0.25° grids, e.g. by using 4x the number of cells for 2x resolution, i.e. 20x12 grid cells for 0.25°; that way you could preserve the grid anchoring and preclude differences due to a new "positioning" of the grid; maybe motivate your choice and/or mention potential effects of different grids on the results
The 1km grid was pre-determined by the OSHD grid of snow-simulations and meteorological forcing data, which we used as a starting point. The 0.5° and 0.25° grid were then determined to closely match the extent of this initial grid. We will discuss potential effects of this choice on the results in a revised version of the manuscript.

L128: a "the" is missing ok

L137 ff: are the grid cells for point simulations centered on the station coordinates? using the "nearest neighbor" grid cell for land surface characterization seems like a very simplified approach compared to all the other sophisticated things done in this study; why not use conservative regridding so you would get something more realistic? ideally, one would generate surface data from the raw PFT fraction data at 0.05° resolution (CLM's own methods) or the raw Swiss national data (your methods); in contrast, taking the nearest neighbor effectively shifts surface information and pairs it with the correctly positioned atmosphere; if this is acceptable, one might as well take the nearest (or interpolated/regridded) results of the gridded simulation? (see comment on L116-L119)
Yes, the grid cells for point simulations are centered on the station coordinates. We did actually first calculate the domain and surface dataset for each point location separately via CLM's own methods, hence using the raw PFT fraction data at 0.05° resolution. We then updated this dataset with the 1km HighRes dataset taking the nearest neighbor. Given that some of the underlying datasets are higher resolution than 1km we acknowledge that it would be beneficial to also run our own methods at the effective point location. We will prepare this dataset, and since all station locations are open, non-forested sites we will ensure that we have a 100% vegetated non-forested grid cells for all LU_{HR} simulations. We will include these new simulations in a revised version of the manuscript and describe it accordingly in the text. We also acknowledge that we should re-name the point simulations (e.g. simply LU_{HR} and LU_{Gl} instead of LU_{HR1km} and LU_{G1kml}) to avoid confusion and make this point clearer.

L154: is "accelerated decomposition" valid/applicable for SP mode? it sounds like BGC; by "cycling" (remove "re-")
You are right, "accelerated decomposition" is not necessary for SP mode, this part of the sentence will be removed. "Re-cycling" will be replaced by "cycling".

Figure 1 caption: is "percentage vegetation cover" the natural vegetation landunit including bare soil, or the sum of vegetation PFTs and CFTs? (the latter would be good)
Currently it is the natural vegetation landunit including bare soil, but we will update this accordingly to represent the sum of vegetation PFTs and CFTs.

L159, L177: is CRU a station-based interpolated dataset and the OSHD based on the COSMO model? for OSHD, it is also a bit unclear if the dataset was produced or re-used for this study
Yes, we will make this clearer in the text of a revised manuscript.

L187: was the native 0.05° PFT and LAI data reprojected and regridded, or was this done based on an existing surface dataset at e.g. 0.5° resolution? depending on the data used, there might be several regridding steps involved (with every step further degenerating the final product) and the native input might be 0.05 or 0.25°; was this done with the CLM5 tools, with which regridding algorithm (bilinear, conservative)?
CLM5 tools were used separately for the 1km, 0.25° and 0.5° resolution to generate each of the 'global' datasets of this study ($LU_{Gl0.5}$, $LU_{Gl0.25}$, LU_{Gl1km}), using the conservative regridding algorithm and all underlying raw input data files (e.g. 0.05° raw PFT fraction data).

L200: FSM2 output is not "observational"
We will clarify this.
L216-217: remove "were", "ground truth" is usually used in remote sensing?, does "upscaled" mean regridded and if so, with which algorithm (bilinear, conservative)?
Ok - "were" will be removed, and "ground truth" will be replaced by "best-case reference". Here upscaled means regridded with the conservative algorithm. We will add this information to the sentence.

L230-234: consider doing this at 0.5° resolution, see point 2
See our answer to point 2 earlier.

L238: to be honest I am a bit lost by now – was 3.1.1 done with the point simulation results? maybe for each Results sub-section this could be highlighted in the title or mentioned in the first sentence.
Thanks for this suggestion, we will make the sub-section titles more descriptive.

L247: for this section it would be really good to understand how land cover and LAI and potentially affect snow depth in CLM5 (see point 4)
From our analysis of snow-depth simulations we saw that the choice of surface dataset only had marginal effects on simulated snow cover, which is why we have not included an analysis of the effects of land cover and LAI on snow cover here.

L283: replace "parameters" by "variables" ok

L284: why is peak GPP assessed and not total GPP? I see the "motivation" later in L328, but because there is a bigger effect does not mean it is more relevant? I think this is related to limitations in SP mode (see point 4b); the effects described in section 3.2 go way beyond ecophysiology (see point 4c)
We will update Figure 4 to use show total annual GPP instead of peak GPP. This part of the results section will be rewritten.

Figure 4 labels: replace "climatological" by "meteorological" ok

L322: snow cover and ET are negatively correlated, but I doubt this is driven by snow but rather by cold temperatures and energy (not water) limitation (see point 4a); feedbacks to the atmosphere are missing in CLM5 offline by construction

Following your suggestion in the specific comments we will refrain from discussing these links in the paper. The section 'Seasonal snow cover development and ecophysiological variables' will be omitted from the paper. Instead, we will prepare a more detailed assessment of snow-related simulation results.

L348 ff: for the calculation of variations in (monthly) total GPP across Switzerland, it would be useful to have an observational benchmark and to relate the amounts to total GPP (i.e. % variation of total GPP). Is there a good reason for not calculating variation in total annual GPP? I would find this quantity more informative

We will update Figure 4 to show variations in the total annual GPP. As we shifted this paper's focus, we will not include observational benchmarks but discuss relative differences.

L368: ET can also be water limited in Switzerland, at least in some regions seasonally

We will add this to a revised version of the manuscript.