Author response to reviewer comments

I. Anonymous Referee #1

Received and published: 9 October 2015

In the manuscript “Semi-automated calibration method for modelling of mountain permafrost evolution”, Marmy et al. present simulations of the future ground thermal regime at six instrumented sites in the Swiss Alps. In its scope and effort, this work is virtually unparalleled and deserves publication in The Cryosphere. However, I am not at all convinced that this work can re-define the state-of-the-art for such studies, and I recommend major revisions before publication. From the material presented, it does not become clear to me that the method can increase the confidence in future predictions compared to much simpler methods.

I.1 Major Comment:

In a certain way, the authors treat the COUP model as a “black box” for which the calibration procedure produces an optimal set of parameters. However, at least in some cases, these parameter sets are not really physical realizations, i.e. some parameters would most likely not be confirmed if independent measurements (e.g. of surface or ground thermal properties) were available. In previous studies dedicated to future projections of the ground thermal state, the authors have chosen a sufficiently simple model (e.g. based only on heat conduction), estimated the parameters according to field knowledge/physical constraints and then compared the results to measurements e.g. in boreholes for validation. While the match with measured data is in general not as good as in this study, it will generate the right results for the right reasons, or at least the limitations will become more obvious. In particular for future simulations, which cannot be validated, the “black box calibration” approach chosen for this study has the potential to produce artifacts in the future simulation. The authors should therefore make the link between the fitted parameters and observable/observed processes much clearer (wherever this link exists). If a parameter set is clearly unphysical (see below for more specific comments), I suggest not to show the future simulations since artifacts are highly likely.

Response to the reviewer: we are thankful to the reviewer for this critical comment, which made us re-formulate our rationale behind the chosen approach in a more explicit way. As indicated in the reviewer’s comment, this semi-automated calibration approach targets the challenge of permafrost modelling at sites, where no independent measurements of ground thermal properties, porosity, snow parameters etc exist. This is in fact usually (almost always) the case, even for monitoring sites with a high number of observed parameters (such as Schilthorn or Murtèl rock glacier, which were simulated in the preceding paper by Scherler et al. 2013). Also in this case, ground thermal properties had to be assumed and were only calibrated against borehole temperatures such as in this paper.

The differences in the approaches lie in the way this calibration is performed: (1) manually, as in Scherler et al. (2013), or (2) automatically, as in the present study. Whereas in the former case, the user can manually make sure that no unphysical or implausible values are used for the unknown soil and snow parameters, this has to be assured by plausible parameter ranges in the latter procedure. However, as the reviewer correctly pointed out, a bad combination of (physically plausible) values could still lead to a correct calibration but for the wrong reasons, which might yield unrealistic projections for the future. If for example a too high thermal conductivity of the snow cover in winter is balanced by an unrealistic soil albedo in summer, this combination of parameters may lead to good calibration results during a 10-year calibration phase, but it will be wrong as soon as the snow cover will diminish in a future (warmer) climate, as the soil albedo will not change in a compensating way.

However, to avoid that, we made sure in our original manuscript that (1) all parameter used for semi-automated calibration stayed in a physically plausible range; and (2) that we identify those parameter with a large influence on calibration (Fig. 5) in order to spot potential combinations of unrealistic parameters, which may lead to implausible projections.

Author changes in revised manuscript: In our revised manuscript we tried to explain this rationale in a much clearer way in order to point the reader to the potential disadvantages of this calibration approach (e.g. in new paragraphs in section 7.1, 7.3 and 5.1). We also followed the suggestion of the reviewer and make the explicit link between the obtained parameter values during calibration and real observed processes (see section 5.1); we also enlarged our analysis of the importance of the various parameters for calibration to determine if there are sites, where implausible combinations of parameter values might lead to erroneous projections for the future. We identified one site (Ritigraben) where this is most probably the case, and consequently removed the long-term simulation of this site from the paper (revised Figures 7 and 8). For two other sites, where (a) missing processes (Murtèl-Corvatsch) or (b) missing input data (Muot da Barba Peider) influence the long-term results, we enlarged the discussion within chapters 5-7, and included additional Figures (new Figure 11 and supplementary material).

Finally, we tried to be more clear and detailed throughout the manuscript with respect to the above mentioned points by the reviewer and hope that the manuscript improved significantly. In the following we will address each minor comment of the reviewer in detail:

I.2 Minor Points:

-p.4788, l. 27: Explain what is meant by “GCM-RCM chain” and “ENSEMBLES data set”.

Response and changes in manuscript: changed to *“…from the different data sets produced by coupled simulations of Global and Regional Climate Models (GCM-RCM) within the EU-ENSEMBLES project.”*

-p. 4789, l. 2: I don’t think that “Langer et al. (2013)” is not a good reference in this context. It would be much more appropriate in l. 20ff.

Response and changes in manuscript: *changed accordingly* (*Langer et al. 2013 in the second paragraph of the Introduction, and Romanovsky et al. 2010 in the first part instead of the original Langer reference*)

-p. 4789, l. 20 ff: The classification in 1D-2D- and 3D models does not follow strict and logical criteria, or at least it does not become clear which variable or process are 1D, 2D or 3D. To me, the mentioned 1D and 3D approaches have a lot in common, since they explicitly account for energy exchange processes in a more or less physically-based way. The mentioned 2D approaches, however, are more (semi-)empirical schemes which are aimed at estimating averages of the target variables of the 1D/3D schemes in a simplified way (except Hartikainen et al., which stands out in that it focuses on much longer timescales than the other studies). In addition, there is the class of spatially distributed 1D-models, sometimes referred to as 2.5D. Examples are the later mentioned Westermann et al. (2013), but also Jafarov et al. (2012) and Zhang et al. (2012).

Response and changes in manuscript: *that is correct, thanks for this clarifying comment (see also reviewer 2). We rewrote the entire paragraph and included the mentioned and additional new references.*

p. 4792, l. 1: the statement "potential scenarios of possible. . ." contains some redundancy. "scenarios of. . ." is enough, in my opinion.

Response and changes in manuscript: *Changed accordingly*

p.4795, l. 19: Wicky (2015) refers to a master thesis (in German) which has not undergone the normal review process. While this is generally problematic, the statement seems to be sufficiently backed up by another reference. I would therefore leave it up to the authors to decide whether to remove this reference or not.

Response and changes in manuscript: *We deleted the reference accordingly and rewrote sentence as follows: "These 2-dimensional (or potentially 3-dimensional) processes cannot be explicitly simulated with the COUP-model, however, their effect on the thermal regime has been indirectly confirmed by specific 1-d distributed COUP simulations at this site (Staub et al. 2015)."*

p. 4798, l. 10, p. 4800, l. 14: it is not directly clear to me why wind speeds play only a minor role in the modeling. From my experience, there are some cases (e.g. high global radiation, but cold air temperatures), where wind speeds have a pronounced effect on active layer thickness and ground temperatures using similar model approaches. It is quite possible that this is not the case for the investigated sites, but it should become clear whether this was checked, and to what extent the role is "minor".

Response and changes in manuscript (see also comment by reviewer 2): *We apologise for the misunderstanding! The reviewers are correct in pointing out that wind speed may have a pronounced effect, what we wanted to say is that the wind speed scenarios of the different GCM-RCM chains are very similar so that the choice of the specific chain does not play a major role in this case. The sentence: "As the wind plays only a minor role in the long-term trend of soil, it was satisfactory to use the median series of the seven other chains" was accordingly changed to: "As the wind speed scenarios of all available GCM-RCM chains are very similar, we consider it acceptable to use the median of these as a substitute for the seven chains with missing wind speed scenarios".*

p. 4799, l. 2: What is meant by "virtually all"?

Response and changes in manuscript: *we changed the sentence to: "While MAAT is predominantly negative in present-day climate, all six sites are subject to a significant increase in temperature and the majority of climate models indicate at four of the six sites positive mean annual temperatures by the end of the 21st century"*

p. 4800, l. 8ff: This is obviously a huge limitation for MBP, and the effect on the results is not clear at all. If there is a strong bias in the global radiation forcing the model, the optimization procedure would tend to correct this by adjusting parameters in a potentially unphysical way, making the simulated future ground temperatures more or less useless. How well can the model estimate global radiation based on latitude/air temperature? Has this been checked for the other sites, where global radiation was available?

Response: *We are thankful for this comment – and the reviewer is of course right, that this absence of radiation data is a huge limitation for the modelling of MBP. We checked the model estimates for the example of COR, which is the nearest station to MBP, and added this comparison as supplementary data. The results show an overestimation of global radiation values by the CoupModel, which may lead (in this case) to maximum surface temperature biases of up to 10°C in summer, which would of course be a serious overestimation. We further checked how the model is potentially compensating this overestimation (as we have no measured data we cannot be sure that the bias is similar for MBP) and found a very low value for the critical snow height parameter ΔS_{crit} which decouples the soil already for very low snow height parameter and could be a sign for model compensation as it affects the albedo calculation. The albedo values for dry and wet soil, however, were calibrated with average values (α_{dry} = 24.1%, α_{wet} = 19.1%), which rather points to no large radiation bias. Finally, the calibrated thermal conductivity values for the near-surface layer were about average as well (around 2-4 W/m*K) and do not indicate a large bias towards too warm radiation based surface temperatures.*
In conclusion, we see no real evidence for a large problematic compensation of a potential radiation bias, and therefore think that it will not induce a comparatively large uncertainty for future scenarios at MBP. If yes, then it would be a slight underestimation of the estimated warming.

Author changes in revised manuscript: The text in section 3.1.2 was modified accordingly: *"Global radiation for MBP has therefore been estimated by CoupModel based on potential global radiation (depending on latitude and declination) and atmospheric turbidity (Jansson 2012). Independent comparison between measured, reconstructed and CoupModel estimated global radiation values for COR showed an overestimation of global radiation by the CoupModel leading to near-surface maximum temperature biases of up to 10°C in summer (cf. supplementary material). However, the calibration technique applied (see sections 4 and 5) would compensate potential biases in the temperature simulations by adjusting related parameters in the model, e.g. snow cover parameters or the albedo. Corresponding uncertainties arising from a potential compensation in the MBP results will be further discussed below."*

In section 5.1. a corresponding paragraph was added: *"As mentioned above, a potential radiation bias could be present in the case of MBP due to the absence of on-site measured global radiation. A compensation of a potential bias would be expected either in the near-surface thermal conductivities or in the albedo values. Although the critical snow height parameter ΔS_{crit} for MBP is very low which could be a sign for model compensation as it affects the albedo calculation, the albedo values themselves were calibrated with average values (α_{dry} = 24.1%, α_{wet} = 19.1%), which rather points to no large radiation bias. Similarly, the calibrated thermal conductivity values for the near-surface layer are about average (around 2-4 W/m*K) compared to the other sites and do not indicate a large bias towards too warm radiation based surface temperatures."*

p. 4801, l. 16: at what depth is the lower boundary? Is it below the depth of zero annual amplitude? In this case, the amplitude should be negligibly small. Does this treatment take the geothermal gradient into account at all, i.e. that ground temperatures become warmer at depth? In this case, how can ground temperatures at depths of up to 80m be modeled?? How is the lower boundary condition for the future runs?

Response and changes in manuscript: *The depth of the lower boundary is different for the various sites (due to different maximum borehole depths) but at least 30 meter and therefore always well below the depth of zero annual amplitude. Therefore, the prescribed heat flux at the lower boundary condition is indeed negligible. However, this enables comparatively stable conditions at the lower boundary and accounts for the often isothermal conditions found in Alpine permafrost at this depth. Of course a potential change in ground temperatures at this depth cannot be simulated and are not interpreted within this study. It is only used as lower boundary condition. See also our response to the comment regarding the spin-up time. A corresponding paragraph was added to section 3.3 and the depth of the lower boundary for each site was added in Figure 3.*

p. 4801: I don't understand how Eqs. 1 and 2 are related. There are four fluxes, qh, qv, qh and qin, are they related somehow? In Eq. 2, the units don't match, the first term has the unit K/m, not J/m2d.

Response and changes in manuscript: *We apologise for the mistake in Equation 2, partly because of two different nomenclatures – q_h is the surface heat flux, and q_v and q_w are the vapour and water fluxes (in general in Eq. 1, and at the surface (z=0) in Eq. 2). Infiltration (as mentioned in the original manuscript) would be a different expression for q_w. We corrected this mistake.*

p. 4802, l. 7ff: This treatment seems to be mainly focused on reproducing spatial averages of the ground thermal regime. Is this appropriate for reproducing temperatures close to the surface, which would either "see" snow or now snow, not patchy snow conditions?

Response and changes in manuscript: *The critical snow depth parameter is in our case less focused on the spatial average but more on the surface conditions, i.e. the roughness. It is one of the main tuning parameter concerning snow cover conditions and their temporal variation. In theory it should reflect amongst others the roughness of the surface, e.g. if it consists of large blocks of 1-2m height (e.g. rock glacier Murtèl) 10cm snow cover has a different thermal insulation effect as at Schilthorn, where the surface is covered by sandy material. Corresponding explanatory sentences were added to this part.*

p. 4802, l. 11: What is Eq. 9?

Response and changes in manuscript: *This equation is included in Table 2. We changed the reference to "see Table 2".*

p. 4802, l. 25: I can't believe that this procedure will create anything realistic for the "deep" ground temperatures. At a depth of almost 80 m, like in Stockhorn, the ground temperatures should still be influenced (if not completely determined) by times before 2000, and even 1981. The modeled ground temperatures at depth would then only reflect unphysical steady-state conditions for the applied forcing data. If this is the case, the deep ground temperatures should be completely removed from the optimization routine and analysis, and the limitations on the future simulations resulting hereof should be mentioned.

<u>Response</u>: *This is of course true, but as mentioned above we are not aiming at simulating deep ground temperatures or very long time-scales. We consider these depths only as lower boundary condition for our near-surface simulations, and therefore include them in the calibration procedure to (semi-automatically) produce data-consistent conditions. Tests with changed spin-up procedures or lower boundary conditions did not show significant changes for the results of the time-scales considered in this study (similar results were obtained by Ekici et al. 2015 and Scherler et al. 2013).*

<u>Author changes in revised manuscript:</u> *We added the following sentences to this section : "Model tests with longer spin-up times showed only negligible differences with respect to the procedure described above. However, this approach clearly neglects all long-term effects of past climatic conditions on the ground thermal regime at larger depths. Therefore, simulation results at larger depths should not be interpreted in a climate context."*

p. 4803, l. 26ff: This is unproblematic if it is done for periods when observations are available. It is extremely problematic as a basis for future simulations, as it is done here. It is not clear at all, if the parameter set is also an optimal one for the future forcing, or if it creates a fake model reality. See also major comment.

<u>Response and changes in manuscript</u>: *See our response to the major comment. We changed the respective paragraph as follows: "In addition, a model set-up which is consistent with present day conditions may not be optimal for future climatic conditions. This well-known problem is inherent to most long-term transient simulations with a high number of parameterised and calibrated processes. One possibility to avoid compensation of two or several parameters, which all show unphysical or unrealistic values is to (1) constrain the parameter range to physical plausible values and (2) verify if the obtained calibration values for all parameters contain any outliers, which cannot be explained by site-specific conditions. However, it has to be kept in mind that the aim of the calibration procedure is not to get a physical determination of the parameter values themselves, but to get a model that is thermally most representative of for the ground thermal regime at a given site. Keeping the above constraints in mind, for long-term simulations, where no observations are available, it has to be assumed that the parameters governing the ground thermal regime do not change significantly over the duration of the simulation."*

p. 4804. L. 4: How about the parameters that do have on-site measured values. Is the uncertainty/spatial variability/changes over time taken into account? This could play a major role for parameters to which the model is highly sensitive.

<u>Response and changes in manuscript</u>: *we apologise for the incorrect wording and the misunderstanding – usually, and also in our case, no subsurface data except ground temperature is available for high mountain permafrost site. Even if one would analyse the rock type in detail, e.g. by collecting of rock samples from the surface, the physical properties of the subsurface could differ strongly, due to the strong weathering, fracturing and heterogeneity at most sites. Because of this also porosity distribution with depth is not available for the sites. Similarly, no in-situ data of snow characteristics are available (thermal conductivity, density, etc). We rephrased the sentence by omitting the term "no on-site measured values" which was misleading in the original version of the manuscript.*

p. 4804: In the fitting routine, only ground temperatures are used to determine the model performance. This could be problematic as the freezing point of water is not exceptional in this procedure. It is the same if the model is wrong by 0.2 degree C at +10 degrees or at -0.1degree C. For the latter, the differences in ground ice and thus the energy content of the ground would be substantial, with pronounced impacts on the future simulations. This would in particular influence deeper ground temperatures. The effect could potentially be moderated by using the energy content of the ground (e.g. as in Jafarov et al., 2012, calculated with the freeze curve also assumed in the model) instead of temperature. The authors should at least comment and discuss this limitation. Note that this comment does not refer to using additional measured parameters, such as water contents.

Response and changes in manuscript: *we thank the reviewer for this important comment, which we are fully aware of – as it complicated the calibration routine for us in the first place. This problem was also the reason that we used a joint bias (mean error, ME) and correlation (r2) based statistical approach to evaluate the calibration performance. While minimising the ME ensures that the absolute values are near the observed ones, the correct simulation of the timing of freeze/thaw events can be improved by maximising r2. It is clear that in the case of long freeze/thaw events a good correlation will always be difficult to achieve, but we found a reasonable good match at least for the near-surface layers by optimising the correlation. In a future step, other quantities such as the energy content of the ground as cited by the reviewer, but also the water content (see our Fig. 9) can be used to enhance the calibration. Regarding its influence on deeper temperatures the reviewer is of course right. However, as mentioned above, the temporal change of deeper ground temperatures near the lower boundary is not subject of the present study. A corresponding discussion was added in section 5.1.*

p. 4805, l. 7: see comment above on initialization/deep ground temperatures.

Response and changes in manuscript: *see our response to the respective comment above: the deep temperatures are only used to get the lower boundary right for the different sites. In addition, their influence on the overall calibration procedure is small compared to the near-surface levels (cf Fig. 5). On the other hand, the benefit of including them into the calibration routine is not only to improve the lower boundary, but also to identify possible differences between the sites regarding the sensitivity of model processes/parameters a these model levels. Therefore, we propose to keep these levels within the calibration routine, but omit their discussion and any interpretation about the temporal change of deep temperatures in the manuscript. We modified/added the following sentence to the final part of section 5.1: "When considering the absolute importance (% in Figure 5, left), we notice that deep boreholes (COR, RIT and STO) have low percentages, which is not surprising as the temperatures at those depths vary on much longer time-scales and depend primarily on the structural set-up of the model. As their future evolution is influenced by past climates, which are not included in the present study, simulated temperature changes at large depths will not be discussed within this study. However, their correct representation for present day climate is important as lower boundary condition for shallower levels."*

p. 4808, l. 5: I suggest leaving the deep temperatures out, see above.

<u>Response and changes in manuscript</u>: *see our response above; we included a clarification regarding the deep temperature calibration as proposed, and changed the text of the deep temperatures accordingly.*

p. 4809, l. 14ff: Which criteria is the statement on equifinality based upon?

<u>Response and changes in manuscript</u>: *In his PhD thesis the main author A. Marmy (2015) addressed the problem of equifinality, i.e. the fact that several model set-ups may result in different but equally well-calibrated models, which may consequently lead to different projections into the future. Sensitivity tests showed that in the context of our study (and using the parameter and parameter ranges indicated), this effect seems to play only a minor role, as model set-ups with similar calibration performances lead also to similar simulation results. However, as a thorough discussion of this topic would be beyond the scope of the present paper, we propose to omit this paragraph in the revised version.*

p. 4817, l. 20: So what's the value of the simulations in these cases then? The authors clearly state that the model cannot represent site-specific conditions and the associated ground thermal regime, so simulations of the future ground thermal regime could feature a strong bias. Wouldn't it be better to clearly state that the procedure is inappropriate for such conditions, and that it is not meaningful to conduct future simulations with the scheme in this case?

<u>Response and changes in manuscript</u>: *We thank the reviewer for this comment and apologise for the apparent misunderstanding: With saying that the structure was not manually adapted we meant that we did not hard-code a specific porosity structure or specific additional heat sources/sinks into the model, as was being done in the paper by Scherler et al. (2013). On the contrary, we gave realistic ranges of parameters (especially porosity) and let the model calibration find the optimal values – and by this giving an indication of a physically-based estimate of the parameters which we have to guess anyway (also in Scherler et al. 2013 real observed values were not available and values were obtained by "educated guesses"). If an automatic calibration leads to a mismatch in temperature or in unrealistic parameter combinations at a certain depth, we can deduce in a more scientific way that an additional heat sink would be necessary to get realistic results and, therefore, that a process is really missing. Because of this, we think that the present approach is more objective, than starting with a (potentially wrong) hypothesis already during model set-up. We tried to explain this concept and the difference to the Scherler et al. paper more clearly in the revised version and rephrased the corresponding paragraph. For the specific case of rock glacier Murtel you are right in saying that the original long-term simulation of Scherler et al. 2013 is probably more correct, even though we do not know whether this process will be unchanged during the next hundred years. Because of this we included a comparison of our and the Scherler et al. 2013 simulation results for rock glacier Murtèl in the revised version (Figure 11).*

p. 4813, l. 17ff: It would be nice if the authors could directly provide some statements on how these two points affect the outcome of this study.

<u>Response and changes in manuscript</u>: *We added a corresponding paragraph to the discussion, as mentioned by the reviewer: "The calibration with GLUE depends on several subjective initial assumptions: a) choice of tested parameters and their range:*

this choice has to be made by the modeler prior to the calibration and is a result of previous tests to identify relevant and sensitive parameters and, b) the choice of criteria of acceptance. For the former, we tried to include a representative set of parameters for surface processes (snow, albedo, evaporation), subsurface processes (thermal and hydraulic conductivity) and properties which are characteristic for the specific geomorphological sites (porosity) in order to provide enough degrees of freedom for a satisfactory calibration. In addition we used our prior experience with CoupModel (cf. Engelhardt et al. 2010, Scherler et al. 2010, 2013, 2014, Marmy et al. 2014, Staub et al. 2015) to identify the most sensitive parameters. We tried to fix the allowed parameter range to physically plausible ranges and verified that the obtained values during calibration were not distributed at the limits of these ranges. Regarding the choice of criteria of acceptance, we gave priority to good correlation coefficients near the surface and at intermediate levels while making sure that absolute values (via the RMSE) were acceptable at all depths. Whereas a different set of calibration parameters would not change Here, different simulation results would have been obtained by e.g. giving more weight to intermediate levels, however, due to the uncertainties regarding the influence of past climates at the lower boundary and regarding the exact representation of temperature evolution near the freezing point, the results would be less certain than in the case of a well-calibrated model at the upper boundary."

p. 4815, l. 28: remove "provide again . . ."?

<u>Response and changes in manuscript</u>: *Sorry for this editing error! The missing link was inserted.*

p. 4816, l. 8: How is the partitioning between transpiration and evaporation controlled in COUP? Are there really plants on Stockhorn, and is it realistic to assume that the latent heat flux is strongly controlled by transpiration rather than evaporation? If no, this would be a good example where the effects of incomplete or even flawed model physics, input data and/or other biased model parameters is compensated by tuning the model in an unphysical way. In my opinion, this does not result in a model that can describe reality in a better way (although it can describe the training data sets).

<u>Response and changes in manuscript</u>: *We are thankful to the reviewer for spotting this error. Indeed no vegetation is present at all our modelled sites, and the corresponding vegetation module in CoupModel was switched off for all simulations. The wilting point parameterisation mentioned in our old Figure 9 (and corresponding text) is used in Coup as a soil physical property and is part of the water retention parametrisation (corresponding to a pF of 4.2 or 150 m of water). This applies also in the absence of plants. We apologise for the incorrect usage of the term wilting point. The corresponding sentences were changed to:*

"In a second step, we manually calibrated the soil physical parameter used in the water retention curve to define the minimal residual water, which has also notable influence on the freezing-point depression." and

"Figure 9. comparison of the simulated (red) and measured (black) soil moisture data at 12 cm (left panels) and 60 cm (right panels) at SCH. (a) and (b) are the results for soil moisture of the best thermal calibration while (c) and (d) are the results after a further calibration of the soil physical parameter of the water retention curve, showing that the calibration can be further improved with additional data sets."

p. 4817, l. 23+25: I would interpret these two statements in the way that the results are not meaningful for this case. The optimization procedure yielded unphysical model parameters to compensate for the incomplete model physics, and then the model is run in this configuration with the future forcing, not knowing anything about the effect of the biased configuration under the warmer future conditions.

Response and changes in manuscript: *see our response to the same comment above. The aim of this study was to introduce and evaluate a new calibration approach, and show the advantages and disadvantages for various (and different) mountain permafrost sites. The long-term simulations (and their comparison with the earlier published estimates of Scherler et al. 2013) show the effects of incomplete calibration if specific processes (here convection within the coarse blocky surface layer) are not included. We added a paragraph discussing these effects, and also included the comparison between the Scherler et al. (2013) results and the results of the present study as new Figure 11. We find a thorough discussion of the advantages/disadvantages better than just leaving out the long-term simulations for rock glacier Murtèl. A corresponding paragraph was also added in the Discussion chapter.*

p. 4820, Conclusions: I expect a clear statement from the authors if the considerable efforts involved in this method can bring a performance gain over much simpler methods (e.g. the comparatively primitive approach to simulate future scenarios for borehole temperatures in Etzelmüller et al. 2011, Hipp et al. 2012, or the spatial modeling of Jafarov et al., 2012).

Response and changes in manuscript: *We added a corresponding statement to chapter 8, Conclusion: "The method was generally suitable for large-scale or long-term modelling but is not recommended for site-specific process analysis, if there are existing dominant processes which are not included in the CoupModel formulation. In these cases, manual calibration and parameterization of the missing processes have to be added. In comparison to other, simpler approaches to simulate future scenarios for borehole temperatures (as e.g. in Etzelmüller et al. 2011, Hipp et al. 2012 or, regarding spatial modelling, in Jafarov et al. 2012) the approach of this study focuses more on the site-specific processes understanding, while the long-term simulation results will not necessarily be better than results from simpler approaches as in the above cited studies. But we believe that the considerably higher efforts of our approach are well justified by the knowledge gain regarding the effect of the dominant processes of the different sites. Of course, future work has to be directed into including the already identified missing processes into the model formulation (i.e. convection)."*

II. Anonymous Referee #2

Received and published: 1 February 2016

The manuscript by Marmy et al. is concisely written and dealing with an important question in permafrost modeling, trying to bridge the gap between site specific calibration and its practical adaptation to spatially distributed modeling on larger scales. In my opinion it is a valuable contribution to permafrost research using a new and time saving approach to calibrate complex models at a set of sites in the Swiss Alps and is as such suited to be published in this journal after some minor revisions.

II.1 General comments

The English should be improved and the manuscript should be shortened. The site descriptions are quite extensive and could in part be moved to a table instead. Also the abstract should be shortened.

Response and changes in manuscript*: Thank you very much for your constructive comments! In general we improved the English, shortened the abstract and tried to shortened the paper wherever possible. We shortened the site description but did not include an additional table, as we feel that it would not shorten the paper with respect to space, and would not improve the readability, due to the many specific characteristics of the sites. As response to reviewer 1, certain paragraphs especially regarding the discussion of the calibration procedure had to be slightly enlarged.*

- The distinction between 1D, 2D and 3D models is not consistent, as the cited 2D models partly refer to empirical statistical models whereas the 1D and 3D approaches refer to process based numerical models.

Response and changes in manuscript (see also reviewer 1): *that is correct, thanks for this clarifying comment (see also reviewer 1). We rewrote the entire paragraph and included new references*

- It is stated that wind speed has virtually no influence on the ground thermal regime. I would doubt this, so either a reference should be given or this simplification has to be addressed in the discussion.

Response and changes in manuscript (see also reviewer 1): *We apologise for the misunderstanding! The reviewers are correct in pointing out that wind speed may have a pronounced effect, what we wanted to say is that the wind speed scenarios of the different GCM-RCM chains are very similar so that the choice of the specific chain does not play a major role in this case. The sentence: "As the wind plays only a minor role in the long-term trend of soil, it was satisfactory to use the median series of the seven other chains" was accordingly changed to: "As the wind speed scenarios of all available GCM-RCM chains are very similar, we consider it acceptable to use the median of these as a substitute for the seven chains with missing wind speed scenarios".*

- The improved calibration using soil moisture data (Figure 9) is very interesting – it would be of great interest for the reader to see how the future projections for this specific site would change by applying the new calibration.

Response and changes in manuscript: We supplied a new figure (Fig. 10) and explanations in the text, which shows the changes (~0.3K colder and later degradation) for Schilthorn by applying the new calibration.

II.2 Technical corrections

P4788
L1: Permafrost also exists in other regions than the European Alps. Be more precise.

Response and changes in manuscript: *changed to "Permafrost is a widespread phenomenon in mountainous regions of the world such as the European Alps".*

L5: ..which allow for . . .

Response and changes in manuscript: *changed accordingly*

L15: Is the calibration method "automated" or "semi-automated"?

Response and changes in manuscript: "semi-automated" → *changed accordingly*

L21: . . . by the end of . . .
L26: climate input data

Response and changes in manuscript: *changed accordingly*

P4789
L1-2: Sentence is unclear and should be rewritten.

Response and changes in manuscript: *changed to: "Permafrost is the thermal state of a soil or rock subsurface with a temperature that remains below 0°C for two or more consecutive years"*

L12: . . . has had notable effects on permafrost that are apparent..

Response and changes in manuscript: *changed accordingly*

P4791
L15: .. multiple sites . . .
L20: .. interpolating in between by the help of . . .

Response and changes in manuscript: *changed accordingly*

P4792
L14: ..massif site ..

Response and changes in manuscript: *changed accordingly*

L16: The expression "non-vegetated lithology" sounds unfamiliar to me.

Response and changes in manuscript: *changed to "The lithology of this non-vegetated site is dominated by…"*

P4797
L23: . . . complete on-site meteorological . . .

Response and changes in manuscript: *changed accordingly*

P4798
L7: ..resolution for the . . .
L9: . . . that some . . .

Response and changes in manuscript: *changed accordingly*

P4800
L13-15: See "General comments"

Response and changes in manuscript: *see response to the general comments*

L17: For calibration, we used . . .
L19: ..errors are possible due to . . .

Response and changes in manuscript: *changed accordingly*

P4803
L6: ..them based on/using . . .
L27: ..for the ground thermal . . .

Response and changes in manuscript: *changed accordingly*

P4804
L4-6: Give references to why these parameters are important in reality. It would also be interesting to read what kind of preliminary analysis is mentioned.

Response and changes in manuscript: *the references were added. With "preliminary analysis" we meant the previous CoupModel permafrost studies by our research group. These references were also added and the sentence was changed accordingly..*

L13: ..tested as well as their range . . .

Response and changes in manuscript: *changed accordingly*

P4805
L15: "As expected": Why was this expected? Give a reference or a reasonable for this expectation.

Response and changes in manuscript: *the respective references were added.*

L24: ..thick snow cover. . .

<u>Response and changes in manuscript</u>: *changed accordingly*

L24-27: Sentence is unclear.

<u>Response and changes in manuscript</u>: *If the snow cover for a specific field site is very thick (e.g. maximum snow cover thickness of >2m) the ground is decoupled from atmosphere during most part of the year (often October/November-July). During this time, small inaccuracies in the snow cover parameterisation will not influence the model results as they do not affect the ground thermal regime. The site is therefore less sensitive. On the contrary, if a site has generally only a thin snow cover (e.g. maximum snow cover <1m), small uncertainties in the snow cover parameterisation will directly affect the simulated ground thermal regime, so the site is more sensitive. From a modellers point of view these sites are more easy to calibrate, as the snow parameters present an easy possibility for an automatic calibration. As we see that this might be misleading in the sentence we omitted the sentence in brackets (“*and therefore more difficult to calibrate”*) in the revised version.*

P4807
L9: How broad is the range of thermal conductivities needed?

<u>Response and changes in manuscript</u>: *we added the respective values to the text (between 0.3 and 2.5 W/m*K).*

L24: ..should be.. Or should it read "has to be"?

<u>Response and changes in manuscript</u>: *see the changes to this sentence shown in the comment below.*

L24-26: This is somehow contradictory to the findings that evaporation is not important for the ground thermal regime at most sites, as it seems to be a tuning parameter for one site. Can you explain why?

<u>Response and changes in manuscript</u>: In the beginning *we were not completely sure of the reasons, because the obtained value for the parameter which was used for evaporation tuning was not very different for RIT as for the other sites (in fact the value for RIT is in the middle of the spread of all sites). Meanwhile we looked more in detail into all parameters from the calibration (see also response to reviewer 1) and found the probable reason for this behaviour. We therefore included the following paragraph into the discussion of the calibration results:*

“When analyzing the specific values obtained for the different calibration parameters at all sites, it was noted that even though the parameter related to evaporation (water tension Ψ_{eg}, cf. Table 2) did not show specifically high or low values for RIT, the parameterized values for T_{snow} (minimum temperature at which precipitation falls only as snow) and ΔS_{crit} (critical snow depth, at which the whole surface is considered to be covered by snow) are very low (T_{snow} = -4.86°C) and very high (ΔS_{crit} = 1.9m), respectively. Whereas the former leads to comparatively large precipitation input as rain, the latter leads to an almost never completely snow-covered surface. In addition, the wet soil albedo for RIT is calibrated with the lowest value of all sites (α_{wet} = 7.0) whereas its dry albedo is comparatively high (α_{dry} = 34.6). In total, this parameter combination enables additional energy input by liquid water into the subsurface, which of course also explains the high sensitivity to evaporation. Even though this parameter combination may lead to an unrealistic process representation in the model, it is still in good accordance with observations, as the effect of 3-d advective water flow from the melting snow cover has been observed in the borehole temperatures (see also Luethi et al. 2016), which is probably the reason for this specific calibration outcome. Of course, the real 3D-process of melt water infiltration is not included in the model."

P4808
L6: boreholes
L17: ..which is because..

Response and changes in manuscript: *changed accordingly*

P4811
L3:. . . slightly positive . . .
L11:. . . set of..

Response and changes in manuscript: *changed accordingly*

P4812
L4:. . . artifact is due to . . .
L5:. . . propagated when run..
L6:. . . by 2080..

Response and changes in manuscript: *changed accordingly*

P4813
L9-15: Does the model have structural errors (L14) or is it realistic (L10)? Be more consistent.

Response and changes in manuscript: The paragraph was changed as follows: "The GLUE calibration method is not meant to determine the physical values of a parameter. The model is physically-based regarding its underlying equations, but has to rely on parameterisations for many of the complex processes in the subsurface and at the soil-snow-atmosphere boundary. The values for all model parameters at all depths cannot be known exactly, especially as almost no direct measurements of these properties are available., The GLUE method gives the ability of finding the value which gives the best fit with the observations within the number of tested runs. But as the system is complex, with sometimes highly uncertain initial and boundary conditions, non-linear processes and model structural simplifications make an optimum calibration impossible (Beven, 2002). It is therefore more meaningful to analyze the residuals and the sensitivity to parameters than the values of the parameter themselves."

P4814
L8: What is meant by "semi-infinite"?

Response and changes in manuscript: *changed to "extremely large number…"*

L14: parameters

Response and changes in manuscript: *changed accordingly*

P4815
L28-29: What is meant by "provide again the P3-link"?

Response and changes in manuscript: *Sorry for this editing error! The missing link was inserted.*

L29: .. as a validation. . .

Response and changes in manuscript: *changed accordingly*

P4816
L9: Give a proper definition of the wilting point.

Response and changes in manuscript (see also reviewer 1): *See our answer to the respective comment of reviewer 1 – we apologise for the error. Indeed no vegetation is present at all our modelled sites, and the corresponding vegetation module in CoupModel was switched off for all simulations. The wilting point parameterisation mentioned in our old Figure 9 (and corresponding text) is used in Coup as a soil physical property and is part of the water retention parametrisation (corresponding to a pF of 4.2 or 150 m of water). This applies also in the absence of plants. We apologise for the incorrect usage of the term wilting point. The corresponding sentences were changed to:*

*"*In a second step, we manually calibrated the soil physical parameter used in the water retention curve to define the minimal residual water, which has also notable influence on the freezing-point depression." *and*

"Figure 9. comparison of the simulated (red) and measured (black) soil moisture data at 12 cm (left panels) and 60 cm (right panels) at SCH. (a) and (b) are the results for soil moisture of the best thermal calibration while (c) and (d) are the results after a further calibration of the soil physical parameter of the water retention curve, showing that the calibration can be further improved with additional data sets."

L13: ..further improved. . .

Response and changes in manuscript: *changed accordingly*

P4819
L24:. . .analyze the. . .
L27:. . .can be drawn. . .

Response and changes in manuscript: *changed accordingly*

P4820
L4: ..precisely whereas. . .

Response and changes in manuscript: *changed accordingly*

L25-26: Is this sentence complete?

<u>Response and changes in manuscript</u>: *changed to:* “The degradation is primarily driven by the change in air temperature during the snow-free period and the change in snow cover duration.”

P4821
L1: . . . decrease from . .

<u>Response and changes in manuscript</u>: *changed to:* “to decrease by values between 20 % and 37 %”