We first would like to thank reviewers for their constructive comments, which led additional useful analyses. Our responses to the formulated remarks are listed below.

Answer to referee #1

- *There are two recurring statements in the paper, I have trouble with.*

  1) Several times the authors consider permafrost to exist "in equilibrium" with the current climate or to have existed in "equilibrium" during the LGM. Either during preindustrial times, today or during the LGM did we have equilibrium? All we can say is that there was and is a quasi-steady state of some sort. Example: the climate signal of the Little Ice Age has certainly not yet reached in some places the base of permafrost, let alone is equilibrated into the overall system. Likewise the LGM was preceded by a long episode of steady cooling. It would take thousands of years of temperatures (especially in Siberia) during the LGM to reach equilibrium. Those thousands of years of constant temperatures during the LGM never occurred.

  It would be better to formulate the assumption that the "NEAR-SURFACE permafrost" has reached "equilibrium" or "quasi-equilibrium" with the "climate". Actually this is what the authors are really after with their paper. They compare their computational results with surface observations (Vandenberghe et al., 2010) and those relate only to "near- surface" manifestations. (At least I assume this to be the case, the Vandenberghe et al. paper has not yet been published and I have not seen it.)

    - We agree with the referee that the overall or deep permafrost is (was) probably not in equilibrium with preindustrial (LGM) climate, or at least that it remains to be proven. As we are dealing with surface distribution of permafrost related to air temperature at the surface, we follow the referee suggestion to use "near-surface permafrost" instead of "permafrost" as a whole. We thus refer the term "permafrost" to "near-surface permafrost" in our introduction:

      "For both approaches, a strong hypothesis is to consider the climate as a steady-state and to assume that the near-surface permafrost (hereafter referred to as "permafrost") is in "pseudo-equilibrium" with it."

- *2) The second statement refers to the proposition that we can somehow provide a statistical tool on top of climate models (CM's) to predict past or current permafrost type distributions. Mean annual temperature is indeed a strong indicator for presence/absence of permafrost. Unfortunately, the nine PMIP2-models (as well as others) are known to perform rather poorly in Polar Regions as we can see in Chapter 8 of the Assessment Report 4, WG1 of the IPCC 2007 report, herein Figs. 8.2 (a, b), page 609 (English version). The calculated temperatures for Siberia today are in part off the true value by several degrees (it is worse in Greenland). If the CM have difficulties to simulate the present, how can we be certain that they do better during the LGM? The CM's have difficulties to handle mean annual air temperatures in Polar Regions and, in particular, on ice surfaces – see the deviations between CM's and reality in Antarctica and Greenland. Now, how does this affect the calculations for the LGM (lots of ice surfaces back then)?*

  The Levavasseur et al paper refers to this aspect on page 2254, where we read: …nine CM from PMIP2 cannot simulate a cold enough climate… warm bias that cannot be completely corrected by one of the two tested SDMs…

  The central weakness of this paper is in my opinion the use of the CM's, which are known to "misbehave" in Polar Regions. From the outset there should have been little hope to arrive at a "correct" result.

  The statement (page 2254) "…warm bias that cannot be completely corrected by one of the two tested SDMs..." is very problematic, because it seems to me that it is a rather dubious approach to correct an incorrect calculation by application of "statistical medicine". We do not even know much about the nature of this bias – is it a constant value or otherwise? I have therefore little hope that further refinement of the statistical methods as discussed at the end of the paper (page 2256) will help to improve the situation.

Anyway, the result is that both approaches (GAM-RV-downscaled and MLR down- scaled) end up with a rather poor performance with respect to %DP (the percentage of discontinuous permafrost (DP) in agreement with data) for current climate and end up with a dismal performance for LGM with respect to %DP (Tables 2 + 3). This indicates in my view a major flaw in the adopted approach. It is my clear impression that the only result this paper provides is a clear signal that CM's cannot properly reconstruct temperatures in Polar Regions for today and, in particular, for the LGM.

  - We do not agree with the reviewer comment:

    - At CTRL period, both statistical approaches reveal a significant contribution; particularly with multinomial logistic models, which clearly improves climate models results in terms of near-surface permafrost distribution (including discontinuous permafrost with means %DP around 55% which does not show "poor performance"). This improvement especially appears in Siberia with the used local-scale topography (see Fig. 4c). These results are already interesting.

    - Then, the poor performances of climate models affect only LGM results; but the example of the IPSL model (best LGM results, see Fig. 5b) proves that our downscaling methods contribute to improve climate models results. The better the model is, the larger the improvement.

    So, climate models are not solely responsible of our results and downscaling methods bring relevant improvements, especially in polar regions (as Siberia) with multinomial logistic models.

    We do not focus our study on polar regions but on the whole Eurasian continent. The climate models are still in relatively good agreement with Eurasian temperatures data from the Climate Research Unit.

    Finally, we cannot say that the differences with data (for continuous or discontinuous permafrost) only indicate that "CMs cannot properly reconstruct temperatures". Several factors are the by-product of these errors as: the RV relationship in association with the GAM method is responsible for about 24% of the difference with data (see Fig. 5a), the selected predictors used by logistic models, permafrost data uncertainties (unknown and probably significant), the supposed "equilibrium" between permafrost and climate, etc.

    "With the hypothesis that LGM and CTRL permafrost data have no uncertainties, that the simulated climates from climate models are at equilibrium with permafrost data and that the relationships between permafrost and chosen variables are stable with time, the nine climate models from PMIP2 cannot simulate a cold enough climate to represent the LGM period. Another study from Saito et al. (2010) confirms this result. Thus, the methods are limited by large-scale errors from climate models at the LGM time period. The better climate models are, the larger the improvement by the SDMs."

- *This situation puts any reviewer into a difficult position, because reviewers are not supposed to rewrite a paper. But in this case I think, a publication should only go ahead if it is clearly stated as key result that CM's need to be improved considerably before it can be hoped that such an analysis will be successful.*

  - We agree that our methods are limited by climate models simulations, especially for past climates. However, we conduct our study with the available climate data and the use of climate models does not discredit our analysis. We emphasised the fact that climate models need to be improved in our conclusions:

    "In order to obtain better contribution of the SDMs, climate models need to improve the representation of large-scale temperature on continents at LGM."

- *Abstract line 19-21 – The sentence "Nevertheless, this also proves that a simple relationship between permafrost and the SAT only is not always sufficient to represent local permafrost." could be deleted. What is stated here has been known for a long time. No new result.*

  - Has the reviewer any references on this? To our knowledge, showing that such a relationship (from Renssen and Vandenberghe 2003) does not contain enough information for local-scale studies is a new result. Precisely, a key point of our paper is to put in perspective this assumption.

- *Last three sentences in Abstract: see also my point under 2): The failure of the SDMs to "improve permafrost distribution" is probably not due to shortcomings of the chosen statistical methods ". . . which deserve further studies", but due to shortcomings of the CM's. I suggest a more specific formulation to the effect that we deal here with a CM problem and not so much with a statistical "mishap".*

  - We partly agree with the reviewer. The poor performances of the SDMs, especially for LGM period, are due to shortcomings of the climate models. However, we do not deal with a problem entirely due to climate models. Although they are partly responsible, the choice of other predictors (e.g. snow cover) could further enhance permafrost distribution. Finally determining the contribution of each factor (including climate models) to the "failure of SDMs" requires a sensitivity analysis beyond the scope of this paper. We clarified that at the end of the abstract*:*

    "LGM simulations from climate models lead to larger differences with permafrost data, than in the CTRL period. These differences reduce the contribution of downscaling and depend on several other factors deserving further studies."

- *Page 2235, line 6: The paper by Lawrence and Sclater (2005) adds very little to the question of permafrost reaction to climate change. The paper explicitly states in chapter 2. Model, that the applied CCSM3 – GCM includes a 3.43 m-deep soil model. Delisle (2007) has shown that such an approach is not appropriate, since permafrost is usually a lot deeper*

  - We followed the referee advice and deleted Lawrence and Sclater (2005) reference.

- *Page 2239, line 10 and follows: "the gridded temperature climatology by CRU" ... "permafrost is probably not in equilibrium with present climate"...*

  CRU considers a rather short time span (1961-1990). There is (my impression) general agreement of a cooling period in Siberia from 1930 – 1960 (about -2◦C) and warming of +2◦C in 1960 – 1990 (see e.g. Permafrost Temperature Dynamics Along the East Siberian and an Alaskan Transect (Extended Abstract) VLADIMIR E. ROMANOVSKY, T.E. OSTERKAMP, T.S. SAZONOVA, N.I. SHENDER and V.T. BALOBAEV, Tohoku Geophys. Journ. (Sci. Rep. Tohoku Univ., Ser. 5), Vol. 36, No. 2, pp. 224-229, 2001 or other sources) and subsequent cooling at least until 2003 (Fig. 1c in Comiso & Parkinson, Satellite-observed changes in the Arctic, Physics Today, Aug. 2004). So the near-surface permafrost today is probably in rough equilibrium with the average climate of Siberia in existence since 1930 apart from its reaction to the ±1◦C changes during one 60 year cycle.

  - We agree and reformulated the sentence according to the previous comment about "near-surface" permafrost in "pseudo-equilibrium" with climate:

    "Although the CRU climatology corresponds to the period of the permafrost observations, the overall permafrost system is not in equilibrium with present climate and more with preindustrial simulations from climate models. However, in the following, we will consider the climate as the steady-state and assume that near-surface permafrost is in rough equilibrium with it."

- *Page 2243, line 17-23: The lengthy explanation to thermal conditions at the base of glaciers should be deleted. The authors rightly observe in the follow-up sentence that their statistical methods do not add anything to solve this problem and cannot be used for above problem in any way.*

  - We followed the suggestion of the referee and deleted the corresponding paragraph.

- *I had initially difficulties with Tables 2 and 3. They are easier to understand by following the detailed explanations in the text. I would appreciate to add a sentence e.g. as follows: "For detailed explanation* *see text." In this context it is not helpful to use expressions such as "area of about..." Be specific: page 2244, line 24: the correct numbers are 1.1 x 106 km2 for model 4 and 0.8 x 106 km2 for model 7.*

  - We fully agree and revised tables 2 and 3. We added some text references into the captions. We also corrected the numbers at page 2244, line 24.

- *Table 3: Under "GAM-RV Downscaled" %DP is explained to represent "the percentage of discontinuous permafrost (DP) in agreement with data". For model 6 the percentage is 0, the other models are hardly better. The calculated area of DP by model 6 is 4.8 x 106 km2, the true value (DATA) is 4.5 x 106 km2. As I understand it the calculated DP is completely somewhere else as the DATA – DP (I could not work it out from the Figs. 7-8.). Hard to believe – could you explain in the text where the calculated DP ends up geographically in relation to the DATA-DP or what these numbers mean in reality?*

  If we look down to the %DP- line "MLR downscaled", the situation hardly improves. Apparently about 80-90 % of calculated DP is somewhere else as supposed.

  Once again I believe the value of this paper would be enhanced, if the authors would flatly state in their conclusions that this type of analysis should await the availability of vastly improved CM's.

  - We clarified the %DP and %CP calculation in the section 3.2.2.:

    "These percentages of common area between permafrost data and climate models are obtained by summing the surface of the grid-cells including continuous (discontinuous) permafrost for both. For example, 0%DP means that discontinuous permafrost from climate model and data are entirely non-overlapping."

Answer to referee #2

- *This paper did not convince me entirely. The basic aim of the paper, to assess how statistical downscaling methods can improve the representation of permafrost extent and type by large-scale climate models, seems rather hopeless to me: How should any physically meaningful method correct for large-scale errors of climate models so typical for the high latitudes? Large areas of Siberia are extremely flat, but small-scale topographical variability is certainly the main reason for permafrost extent and type variability. Under these conditions, it is not clear how any downscaling method should lead to better agreements with the data if this improved agreement is to be physically meaningful.*
    - Our downscaling methods are not entirely “statistical”, the selected predictors bring some physics to the relationships built between temperature and predictors for GAM and between permafrost and predictors for multinomial logistic models (ML-GAM and MLR). It is a statistical relationship based on some physically meaningful reasoning. We built a sort of physical models for the small scales based on simple statistical relationships. We tried to improve climate models results by bringing local-scale information (as topography) lacking in climate models and not correcting their large-scale errors. In that sense there are reasons to have some “hope”.

- *Applied to a small mountain region, one would expect the methods used here to yield much improved results, but on the large scale here; there is no reason to expect downscaling to yield anything useful. I still find the paper interesting, but I wonder whether it should be restricted to, say, the Himalaya or some other mountain region.*
    - We disagree with the referee. Applying our multinomial logistic models only to mountain regions as Himalaya leads to a calibration with only one permafrost category reducing the analogues needed for LGM projections. If we calibrate the multinomial logistic models on Himalaya with only discontinuous permafrost at CTRL period, we will not be able to predict continuous permafrost at the LGM period, especially on Himalaya. Similarly, to calibrate GAM only on mountain regions reduces the present temperature range for downscaling on the LGM climate.

      Moreover, another goal is to build the most global relationship between permafrost and predictors applying our methods on the largest possible region.

      “Applying logistic models on a large region as Eurasian continent allow us to build a global/generic relationship between permafrost and several factors.”

      “To calibrate on a large region as Eurasian continent also allows to build a global relationship, which could be tested on other region of interest.”

- *I found this paper quite hard to read, partly because of the unrestricted use of acronyms. The authors should try to find ways to change this. It seems to me that the English might be improved in some places (but I’m not a native speaker).*
    - We agree and tried to reduce the general acronyms without to burden the text.

- *Page 2235, line 10 – “the permafrost representation depends on the resolution of climate models which cannot reflect the local physical processes involved". The resolution is certainly not the only reason for the insufficient physics of climate models.*
    - We agree and reformulated the sentence as:

      “[…] but permafrost representation partly depends on the resolution of climate models which cannot reflect the local-scale physical processes involved.”

- *The large errors for the LGM permafrost extents (even in the initial fields) might be due to the fact that LGM precipitation rate were probably very low. This means little snow, and might therefore lead to a* *larger permafrost extent than what you would expect using a simple temperature-index method developed for the present.*

  - We agree with the referee that the larger errors for LGM permafrost extents from climate models have many reasons; precipitation is one among several causes of underestimated permafrost extents in climate models (even without downscaling methods). Quantification of these factors would need other sensibility studies, which are not the purpose of this article.

- *The basic hypothesis that permafrost depends solely on temperature is very strong. The authors acknowledge this by stating that future research should include snow cover, and that at least mountain permafrost is influenced by snow cover, but I think this could be stated more clearly.*

  - Done:

    "Future research should include snow cover and thickness and soil temperature, especially for mountain permafrost influenced by snow cover."

- *Page 2241, line 2: "The calibration is the fitting processes of the splines on present climate." What splines?*

  - This sentence was indeed not clear. We now define the spline function in the text as follow:

    "[…] Then, we define the nonlinear functions as cubic regression splines (piecewise third degree polynomials)."

- *Page 2242, line 17: The procedure used to construct the LGM topography is not clear to me. What does GRISLI do in this?*

  - We detailed the procedure to construct LGM topography. We need the ice-sheet model GRISLI only to compute the difference between present and LGM elevation. Using an ice-sheet model allows us to take into account the influence of different mechanisms as sea level variations or ice-sheets subsidence on topography. Then, we apply the difference between LGM and present orography on local-scale topography ETOPO2 to obtain a local-scale LGM topography.

    "We build the LGM topography from ETOPO2 adding in each grid- point a value corresponding to the difference between LGM and present orography. This difference is calculated with the elevation provided by present and LGM simulations of the ice-sheet model GRISLI (Peyaud et al., 2007) to account for the ice-sheet elevation and subsidence, and the sea-level changes."

- *Page 2242, line 21: The two different types of continentality are not clearly defined. They do not seem to used anyway, so either cut this or define the variables correctly if I missed something.*

  - We defined the two continentality indices according to Vrac et al. 2007:

    "The physical interpretation is the effect of coastal atmospheric circulation on temperature. DCO does not depend on time and is only affected by sea-level change (or land-sea distribution). The second continentality index is the "advective" continentality (ACO). ACO is somewhat similar to DCO albeit being modulated by the large-scale wind intensities and directions from climate models and represents an index of the continentalization of air masses. It is based on the hypothesis that an air parcel becomes progressively continental as it travels over land influencing temperature. Hence ACO depends on the changes of land-sea distribution and on wind fields coming from the climate models simulations."

- *Page 2248, line 24: "a study on the predictors choice ... could be an interesting prospect but is not the purpose of this article." Strange. To my mind, it should be one of the main purposes of this article.*

  - We disagree with the interpretation of the referee. Our aim is to determine if we could improve climate models results by changing spatial scale with downscaling methods. For temperature downscaling, our predictors choice is consistent with previous studies from Vrac et al. 2007 and Martin et al. 2010. To make a comparison we have to be consistent between both approaches and keep the same predictors. Testing several sets of predictors is a second step beyond the scope of this article.

- *Page 2272, Fig. 5: Figure 5 should be explained more clearly. What are the percentiles used in the diagram?*
    - Figure 5 has been revised without the box-and-whisker plots as recommended by referee 3.

Answer to referee #3

- *The discussion article attempts to quantify the performance of different empirical models to predict / hindcast permafrost distribution based on climate model outputs. I found several major issues in the discussion article: (1) The manuscript is hard to read because of often imprecise wording and confusing article structure.*

    - The article structure we have chosen follows the chronological reasoning of the study to facilitate the understanding, especially for non-statistical readers of TC:

      In section 3, we feel easier to present the first GAM-RV method and to discuss its results before to introduce the logistic models.

      In section 4, the use of logistic models results from GAM-RV observations: indeed, since it appeared that our results are clearly driven by temperature, it was interesting to use an alternative method in order to directly downscale permafrost. Then, we naturally confront the results from both approaches.

      We think that introducing both approaches together in one method section leads to confusions between SDMs, burdening the text. We feel more logical to describe and discuss the first GAM-RV method and to show the improvement (or not) of the alternative logistic approaches.

      However, we leave the choice of the most appropriated article structure to the editor of TC.

      About the imprecise wording we have followed step by step the comments of the three referees and hope that the current version is thus more precise.

- *(2) The quality of the input data is a major issue.*

  Regarding (2), the authors need to demonstrate that the permafrost maps used for model calibration and performance assessment are of good enough quality for the purposes of this study. I am not convinced that the authors will be able to demonstrate this in a revised paper.

    - Indeed, we are not able to estimate the quality of the input permafrost maps, because they are built from several sources. We clarified the corresponding paragraph:

      "To validate the statistical models on present climate we use geocryological observations reviewed and grouped into one circum-artic permafrost map by the International Permafrost Association (IPA) and the Frozen Ground Data Center (FGDC) (Brown et al., 1997). Most of compiled permafrost data are observations between 1960 and 1980 drawn on different maps with different scales by several authors (Heginbottom et al., 1993). In a similar way, LGM permafrost data correspond to a recent map of permafrost extent maximum in Europe and Asia around 21 ky BP, combining different geological observations from different maps as described in Vandenberghe et al. (2008) and Vandenberghe et al. (2011). The combined LGM maps are not always distinctive in describing the permafrost categories, which could have different definitions depending on the authors. Moreover, the age of LGM permafrost indicators is often not precisely defined. Consequently, it is difficult to judge the accuracy of the final maps and we have to keep in mind these restrictions in our interpretation."

- *(3) Some of the modelling decisions are inconsistent or hard to follow. Regarding (3), some important modelling decisions are incomprehensible to me. For example, why don't the authors use a multinomial logistic GAM? It appears arbitrary that the GAM is used to predict temperatures, and the MLR to predict probabilities. In the current paper it is impossible to attribute differences in model results to either the nonlinearity of the GAM or the choice of modelling temperatures versus modelling permafrost classes; this is certainly undesirable and unnecessary.*

    - We are thankful for this suggestion. We have, in the present version, restructured the manuscript to enclose Multinomial Logistic – GAM also in our study. The choice of ML-GAM as the main alternative method is more relevant because both methods are thus clearly based on GAM.

So, in section 4 ML-GAM is the main used multinomial logistic model. But we kept MLR in order to compare both linear and nonlinear approaches.

"To link a categorical variable, such as permafrost, with continuous variables, a common statistical technique is the use of logistic models representing the occurrence probability of an event (often binary, e.g., permafrost or no permafrost). This probability can take continuous values between 0 and 1. For instance, Calef et al. (2005) build a hierarchical logistic regression model (three binary logistic regression steps) to predict the potential equilibrium distribution of four major vegetation types. More classically, Fealy and Sweeney (2007) use the logistic regression as SDM to estimate the probabilities of wet and dry days occurrences. In the context of periglacial landforms, Brenning (2009) (rock glacier detection) or Luoto and Hjort (2005) (subartic geomorphological processes prediction) obtain good results with logistic GAM. Lewkowicz and Ednie (2004) use logistic regression to map mountain permafrost. So, logistic models can be based on linear or nonlinear combinations of the predictors depending on the context of the study. In the case of permafrost downscaling, at our knowledge, no evidence allows us to focus on linear or nonlinear relationships between permafrost and the predictors. To be consistent with section 3.2, we use a logistic GAM in its multinomial form (Multinomial Logistic GAM - ML-GAM) to model the occurrence probabilities of three permafrost indices (continuous, discontinuous and no permafrost) as illustrated in figure 2 (right half). Here, ML-GAM is used as a SDM to estimate the occurrence probabilities of the explained variable (Y, permafrost in our case) for each category or class j by a sum of nonlinear functions (f_k), conditionally on numerical or categorical predictors (X_k) (Hastie and Tibshirani, 1990): […]

where $P(Y_i = j)$ is the probability of the j^{th} permafrost category, f_k are defined as cubic splines, n is the number of predictors and i is the grid-cell. To use ML-GAM, we need to define a reference category (r). We obtain j−1 relationships and the occurrence probability of the reference category can be deduced with […] (considering m categories). To make a comparison with ML-GAM, we also apply in background a classical Multinomial Logistic Regression (MLR - Hilbe (2009); Hosmer and Lemeshow (2000)). The occurrence probability ($P(Y_i = j)$) of each category of the predictand are estimated by linear combinations of numerical or categorical predictors (X_k): […]

where β_k are the regression coefficients for the j^{th} permafrost category. The method is based on the use of a Generalized Linear Model (GLM - McCullagh and Nelder (1989)). GLM generalizes linear regression using a link function between predictand and predictors unifying various statistical regression models, including linear regression, Poisson regression and logistic regression. GAMs are simply a nonlinear extension of GLMs. ML-GAM and MLR are performed with the R package "VGAM" (Yee and Wild, 1996; Yee, 2010a,b)."

In section 4.1, on one hand, we confront GAM-RV vs. ML-GAM results. On the other hand, we compare ML-GAM vs. MLR results. For example:

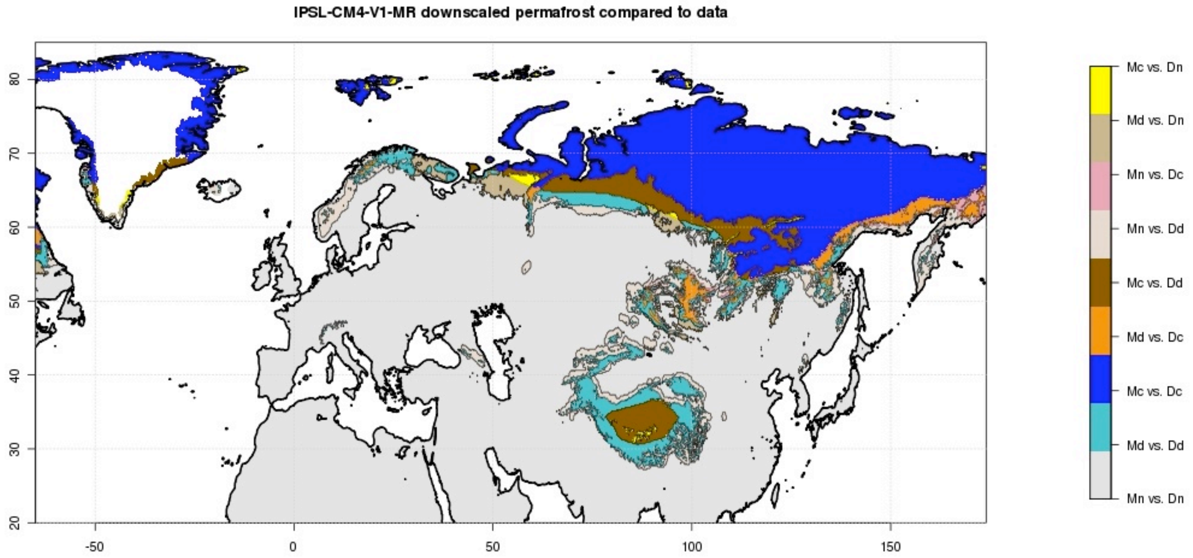
"In table 2, the ML-GAM downscaling improves the continuous and discontinuous permafrost areas for both climate models. In comparison with interpolated climate models, ML-GAM reduces the total permafrost difference with data from IPA/FGDC of $1.4\times10^6 km^2$ for ECHAM5 and $1.6\times10^6 km^2$ for IPSL-CM4."

"Note that very similar permafrost distributions appear using MLR (not shown) with permafrost areas (figure 5a) and kappa slightly weaker than ML-GAM results (see table 2). This observation is in agreement with Brenning (2009) which shows that GAMs can be slightly better than GLMs in the particular context of periglacial landforms prediction."

In section 5 ML-GAM do not change our conclusions because it faces difficulties to improve local-scale permafrost distribution too. We just precise that MLR obtains results slightly better.

"GAM-RV does not improve the statistical agreement, reflecting the weak potential of climate models to correctly represent permafrost limits for the LGM period. MLR obtains the best results, slightly improving the agreement with data for all climate models."

- *Another major issue is the inclusion of ETOPO2 elevation as a predictor variable in models that also include SAT from climate models. Since SAT is already included in the models, the only purpose of elevation can be to represent residual altitudinal effects that are neglected by the climate models' underlying coarse topography. Consequently, only the residual elevation (climate model elevation minus ETOPO2 elevation) should be used as a predictor variable for local bias correction. I disagree with the use of elevation per se. In addition, if I correctly understood the cross-validation approach proposed by the authors, I object to its use in its present form, and urge the authors to perform a cross-validation in the statistical sense of the word.*
    - In principle we fully agree with the referee, but in fact the residual elevation is not meaningful for the method used here. Indeed using the difference between orography from climate model and ETOPO2 leads to the same residual elevation for low altitude (as Europe with "no permafrost") and high altitude (as Himalaya with "discontinuous permafrost") regions. Nevertheless, we checked and tested our methods replacing the ETOPO2 predictor by the residual elevation: same residual elevations are so calibrated with different permafrost corrections leading to inconsistent modelling of permafrost. For example, the map below corresponds to the downscaling of ECHAM5 by ML-GAM using the residual elevation instead of the ETOPO2 topography as predictor for CTRL period. We clearly show a disagreement between downscaled climate model and permafrost data on Himalayas due to the contribution of the residual elevation. Moreover, lower permafrost surface, %DP/CP and statistical agreement support these results. For these reasons, we kept the elevation ETOPO2 as predictor variable.



- *Abstract - The focus of this paper is on quantification – The reader should therefore have an opportunity to find some "hard", dependable numerical results in the abstract: the overall measures of model performance (maybe the kappa coefficients) for each of the key models and both time periods.*
    - Done with the kappa indices for both time periods:

      "In average for the nine PMIP2 models, we measure a global agreement by kappa statistic of 0.80 with CTRL permafrost data, against 0.68 for the GAM method."

> "The prediction of the SDMs is not significantly better than large-scale fields with 0.46 (GAM) and 0.49 (ML-GAM) of global agreement with LGM permafrost data."

- *Page 2234, line 22: "We start with the hypothesis from Renssen and Vandenberghe (2003) that permafrost depends solely on surface air temperature" – (1) This doesn't seem to be stated as a hypothesis in the cited paper. (2) If it was a hypothesis, there is enough scientific evidence that it is wrong (e.g., numerous papers studying mountain permafrost distribution as related to MAAT, PISR, and vegetation). (3) The authors probably want to state this as an "assumption", not hypothesis (not only here but throughout the text).*
    - We agree with the comment (3) and reformulated the sentence accordingly:

      "For simplicity, we first assume that permafrost depends solely on air temperature at the surface (or temperature at 2 meters and hereafter referred to as "temperature") with the relationship from Renssen and Vandenberghe (2003) presented in section 2 with the used permafrost databases."

- *Page 2234, line 22: What is "surface air temperature"? MAAT, the entire time series, summer temperature, daily maxima?*
    - We agree that the "surface air temperature" variable is ambiguous. We modified it into "air temperature at the surface" which designates the air temperature at 2m as mentioned before. Then, to satisfy the temperature conditions from Renssen and Vandenberghe 2003, we compute the Mean Annual Air Temperature (MAAT) and the coldest month mean air temperature.

- *Page 2236, line 4-5 – "downscaling methods, bringing local information" – unclear if this refers to additional (non-SAT, maybe albedo or vegetation etc.) information, or simply to local ground-truth information that is used for local calibration? It seems to refer to the former (higher-resolution elevation data).*
    - We clarified the term "local" which always refers to "local-scale" or higher-resolution data.

- *Page 2236, line 15 – "smart": omit this word; it gives the sentence the flavour of expressing a preconceived opinion on the predictive performance (and the "intelligence" of the internal functioning) of the proposed methods.*
    - Done

- *Page 2237, line 8-10 – Meaning of this sentence remains unclear to me. The GAM is not limited to continuous predictor and response variables. Logistic GAMs can deal with binary response variables; any GAM can incorporate categorical predictor variables by representing these as dummy variables. GAMs are simply nonlinear extensions of GLMs, and the logistic GAM is a nonlinear extension of logistic regression. Valid arguments in favour or against the use of GAM vs. GLM would be related to the assumed or known existence of nonlinearities in empirical relationships of physical processes (or knowledge of the lack of such nonlinearities), or the greater tendency of more flexible models to overfit to the training data, if prediction is the main goal. A recent publication of Brenning (2009, in Remote Sensing of Environment) in the context of periglacial landforms suggests that, at least in this particular context, GAMs tended to be slightly better than GLMs, although they were not necessarily better than other linear methods. See also Luoto and Hjort (2005, in Geomorphology).*
    - As said before we now use ML-GAM and the suggested papers are now cited in section 4.

- *Page 2237, line 12-13 - This is quite different from permafrost presence / absence modelling. For example, Lewkowicz & Ednie (2004, in Permafrost and Periglacial Processes) used logistic regression to map mountain permafrost – this example would be more relevant to the present study than the cited examples and references.*

  - The purpose of the references cited in the manuscript is to introduce the different fields using logistic models in climatology. Thank you for this additional reference now cited and more relevant in our context.

- *Page 2237, line 17 – "hypothesis" -> "assumption"; I do not agree that steady-state climate and equilibrium conditions are or have to be assumed when using empirical models. This would more likely be the case when using (simplified) physically-based approaches, or when attributing physical interpretations to empirical findings.*

  - We agree with the referee but we modified the paragraph according to the referee 1.

- *Page 2238, line 1 – omit this sentence*

  - Done

- *Page 2238, line 4 – "freezing point of water" – permafrost is usually defined to be at or below 0 degrees Celsius for a certain time period, which is not exactly the same as the freezing point of water, which depends on pressure, salt content, etc. See e.g. p. 83 of the cited reference (French, 2007).*

  - Done, the paper now reads:

    "Permafrost is defined as ground permanently at or below 0°C for two or more consecutive years (French, 2007)."

- *Page 2238, line 5-8 - Insufficient detail on the origin, type and quality of "present" permafrost distribution data is provided. In addition to the original scale (or resolution) at which the map was prepared, it is essential to explain what method was used by its authors. Also, what are the uncertainties, possible or known systematic biases, and regional differences in the product's quality?*

  Page 2238, line 10-11 - Again, more details on quality and origin of the reconstructed permafrost extent are needed; any relevant information from the cited papers should be repeated here because it is essential for judging the utility of the data for the present article. Stating that both maps "have been drawn [sic!] in a similar way" is not sufficient and does not help support the implicit claim that these maps are of sufficient quality, which needs to be demonstrated.

  - We refer the reviewer to our answer to the second comment.

- *Page 2239, line 2 – GAMs are not limited to continuous response variables. Logistic GAMs for binary responses are readily available (for example, in the R packages gam and mgcv, probably the most widely used GAM implementations), and approaches for converting binary into multi-class classifiers are well documented in the statistical literature.*

  - We fully agree with the referee and reformulated the sentence. This is absolutely correct; GAMs are not limited to continuous response variables. We use a logistic GAM in a multinomial form available in the VGAM R package (cf. response to the third comment.)

    "To simulate a discrete variable such as permafrost, we first decide to downscale the temperatures from different climate models with the same approach by GAM as Vrac et al. (2007b) and Martin et al. (2010a)."

- *Page 2239, line 19 (and elsewhere) – "temperature" – should read "MAAT" (mean annual air temperature)*

  - Not exactly, the term "temperature" corresponds to the "air temperature at the surface" variable, as clarified before. Then, we compute MAAT and the coldest month mean temperature to satisfy the relationship from Renssen and Vandenberghe 2003. So we prefer to generalize both variables under the term "temperature conditions".

- *Page 2239, line 19-21 – Why using this particular relationship? What is it based on? Is there any reason to believe that it is more likely transferable to the LGM than the other decision rules that have been* *proposed? LGM permafrost predictions will critically depend on these thresholds. Given the large uncertainties inherent in the thresholds, it would appear reasonable to perform a sensitivity analysis or compare the results obtained with different decision rules.*

  - We agree with the referee but the application of this particular relationship is simple and it is the most used relationship in climate modelling to obtain permafrost from climate simulations. To perform sensitivity analysis about the uncertainties inherent in the thresholds is beyond the scope of the article. This relationship is one solution among others, the latter being not necessarily better. This was clarified here:

    "Several relationships exist in literature (e.g., Nechaev 1981, Huijzer 1997), the most employed in climate modelling are the following conditions from Renssen and Vandenberghe (2003): […]"

- *Page 2240, line 1-5 - This is not convincing. (1) Other variables (such as PISR, vegetation) are not empirically compared to permafrost distribution, and consequently the implicit claim that they are not important cannot be made. (2) The claim of an "obvious" empirical association between the permafrost map categories and MAAT should be supported by numerical evidence, i.e. suitable measures of association (AUROC, kappa and the like). On the other hand, it may be acceptable to simply state that MAAT is the only variable considered, that other influences are known to exist (provide references), but act mainly at the sub-pixel scale, and that the model's predictive performance will tell us in an objective way if the results are of an acceptable quality.*

  - We disagree with the referee. We never mentioned (even implicitly) that other variables as vegetation or PISR are not important for permafrost distribution. On the contrary, one of our key results is that local-scale permafrost distribution needs more information brought by other variables as vegetation, snow cover or PISR. Again, one of our aims is to demonstrate that considering only temperature (MAAT + coldest month mean temperature) is not sufficient to study local-scale permafrost.

- *Page 2240, line 10 - "is a reasonable approximation" – Such an affirmation is not acceptable in this part of the paper because the question whether and how reasonable this approximation is constitutes the main goal of this article.*

  - We fully agree with the referee comment and changed the sentence as follow:

    "Nevertheless to a first order, deriving permafrost from temperature will be the base assumption of this study for present-day conditions."

- *Page 2240, line 20 - "cubic splines" – in the general formulation of the GAM, the nonlinear transformation functions are not necessarily cubic splines. Use a more general expression here, and specify later the particular settings (cubic splines) used in this study.*

  - Done

    "The large-scale predictors will be described in section 3.2.1. More precisely, this kind of statistical model models the expectation of the explained variable Y (the predictand, temperature in our case) by a sum of regressions with nonlinear functions (fk), conditionally on the predictors Xk (Hastie and Tibshirani, 1990): […] Then, the nonlinear functions as cubic regression splines (piecewise by third degree polynomials)."

- *Page 2240, line 25, "gaussian" -> "Gaussian"*

  - Done

- *Page 2241, line 8-10 - Cross-validation is a well-defined and widely used statistical estimation technique that is based on resampling (partitioning) the observations, not the variables. It is incomprehensible to me* *what the described leave-one-variable-out cross-validation would do to help assess the technique's predictive performance. And again, to be clear, this is not a cross-validation.*

  - We disagree with the referee. As we understand it, the goal of the cross-validation in statistics is to calibrate a statistical model on a different sample of the prediction step. Contrary to the referee comment, our procedure resamples the observations, not the variables. For every predictor, we select eleven months from the climatology for the calibration step. Then we predict the remaining month of the downscaled temperature climatology, with all variables/predictors. In other words, here, one observation corresponds to a gridcell-month, i.e., a given CRU-gridcell in a given month. Months are not the predictor variables. So, we adapted a "cross-validation" as it is defined.

- *Page 2241, line 14 - Please see the citation() function in R – it provides recommendations for referencing R and its packages.*

  - Done

    "We perform this analyses within the statistical programming environment R (R Development Core Team, 2009) and its "mgcv" package (Wood, 2006)."

- *Section 3.2.1 in general: provide a rationale for the choices you make, instead of simply stating that only one physical predictor is used and that you "choose to work with nine" CMs, for example.*

  - We followed the recommendation of the referee. The predictors choice is based on previous study from Vrac et al. 2007 and Martin et al. 2010, which tested different sets of predictors to downscale temperature as stated here:

    "Previous studies from Vrac et al. (2007a) and Martin et al. (2010a) lead us to select four informative predictors for temperature downscaling, fully described in their studies."

    Then, some of PMIP2 models do not have the required LGM outputs to apply our downscaling; only nine models were available for our study for this period. This was clarified here:

    "The required LGM outputs for section 5 lead us to work with nine of them listed in table 1."

- *Page 2242, line 1-2 - SAT is mean annual SAT?*

  - As said before, we replaced the term "surface air temperature" by "air temperature at the surface" referred throughout the text to as "temperature". We used the temperature climatology from climate models (12 monthly means by grid-points). According to Renssen and Vandenberghe 2003, we computed the MAAT and the coldest month mean air temperature to derive permafrost.

- *Page 2242, line 2-4 - "This variable ... bilinearly interpolated at 10' resolution" – This tells me that the actual downscaling takes place here, by interpolating SAT. The GAM / MLR simply use this finer-resolution data to calibrate an empirical relationship, but they themselves do not perform down scaling.*

  - We disagree with the referee comment. We could make downscaling from the coarse resolution of climate models. Interpolation is just used to produce more spatial variability from the statistical models.

- *Page 2242, line 13-15 - Not present-day ETOPO2 elevation (or LGM elevation, respectively) should be considered here, but the difference between ETOPO2 and the elevation used by each CM should be used in order to reflect elevation-related temperature biases.*

  - We refer the reviewer to our answer to her/his fourth comment.

- *Page 2243, line 15-24 - This belongs into the Methods section. The decision to mask certain areas is part of the chosen method, not a result of the application of the chosen method.*
    - Done

      "Nevertheless, for permafrost representation we mask the ice-sheets (Greenland and Fennoscandia for LGM) as the presence of permafrost under an ice-sheet is not obvious and is currently debated. Moreover, since our estimate is based on surface temperature there is no reason why the permafrost under the ice-sheet shall be mainly driven by air temperature, above the ice-sheet."

- *Page 2244, first paragraph - I strongly recommend starting with a general quantitative performance summary before presenting the maps and analysing regional / altitudinal difficulties of the model.*
    - We did not follow the suggestion of the referee because the indices are computed from maps. We thus feel it is difficult to summarize quantitative performance of statistical models without describing maps beforehand.

- *Page 2244, second paragraph, to page 2247, line 2, and again page 2247, line 24 to page 2248, line 27: Introduce the performance measures and the MLR in the Methods section, not here. This will make the results section much easier to read.*
    - We refer the reviewer to our answer about the article structure.

- *Page 2244, line 16 - "To quantitatively assess the effect of our downscaling on permafrost representation, we measure the agreement between CMs and data..." – This sounds odd to me. First, what is "data", in this context? Second, if you want to assess the performance of the downscaling methods (=GAM, MLR), why do you say in the same sentence that you are going to assess the CMs, and not the downscaling methods?*
    - Thank you. We wanted to refer to "downscaled" climate models in order to assess the performance of the downscaling methods. This has been reformulated as follow:

      "To quantitatively assess the effect of the downscaling on CTRL permafrost representation, we measure the agreement between permafrost distributions from downscaled climate models and IPA/FGDC data with different numerical indices whose results are listed in table 2."

- *Page 2244, line 24: "CMs underestimate the permafrost area" – Well, CMs do not produce permafrost maps, but SAT maps, do how can they underestimate the permafrost area? Are the temperature thresholds of Renssen and Vandenberghe (2003) being used?*
    - We reformulated the sentence as:

      "Continuous permafrost derived from downscaled temperature is still underestimates and no or less discontinuous permafrost is predicted in right location (%DP ranges between 0 and 20%)."

- *Page 2244, line 26, and elsewhere - More generally used terminology is available to refer to %CP and %DP; e.g., concepts such as true positive rate (or sensitivity), or, using remote-sensing terminology user and producer accuracy. Using such general concepts would make the results much more readable.*
    - We did not follow the suggestions of the referee. The concepts of true positive rate or other are defined for binary variables and we don't know the remote-sensing terminology. This comment underlines again the problem of acronyms. We revised the text in order to delete some acronyms and to improve the readability.

- *Page 2245, line 2 - "...show the limits of the GAM method" – clearly, this is a discussion item, should not be presented as part of the results. I am also not convinced that this is an issue of the GAM, but maybe of the* *authors' choice of applying the GAM to temperature prediction and then applying fixed decision thresholds.*

  - We agree and rewrote the sentence as:

    "The responses of these two climate models show the limits of the GAM-RV method."

- *Page 2245, line 9 – I do not understand this interpretation.*

  - We deleted this unclear sentence.

- *Page 2245, line 10 – Standard deviation based on 9 observations, one of which (in the LGM situation) is pretty far off the mean value? I am avoiding the word "outlier" here, would rather interpret this as either supportive of a skewed distribution, or simply a small-sample effect. In any case these standard deviations are not very reliable statistically, they will be sensitive to the deviation from mean in any individual observation.*

  - We agree that the standard deviation is not very statistically reliable for a small-sample (9 models), nevertheless it gives a first (rough) indication about the inter-variation between climate models. We clarified that:

    "Indeed, in table 2 GAM-RV reduces the standard deviation for all area indices. Although standard deviation computed on small-sample is not very reliable statistically, it gives a first indication about the inter-variation between climate models."

- *Page 2246, line 9 – 13: motivation and calculation of kappa_max remains unclear to me.*

  - We clarified that:

    "Without downscaling, ECHAM5 obtains a κ of 0.64 and 0.68 for IPSL-CM4. These values are difficult to interpret because the kappa's scale (between 0 and 1) depends on number of categories and sample-size. To gauge the strength of agreement without an arbitrary scale, we use the kappa maximum (κmax)."

- *Page 2247, line 2 – "statistical significance": use this expression only in the context of statistical hypothesis testing.*

  - Yes, the exact term is "statistical meaning" modified in the text.

- *Page 2247, line 4 – "statistically relevant, in better agreement with data and not by chance" – The authors should distinguish between two questions: (1) is the GAM (or any other method) better than chance agreement achieved by, e.g., tossing a coin? (2) Is the GAM better than any of the other methods studied? The kappa coefficient is one possible approach for answering question (1) as it (is intended to) adjusts for chance agreement. In the case of pairwise comparisons between different methods (e.g., GAM versus MLR), it is harder to tell whether the estimated differences (in kappa coefficients, for example) can be attributed to random variations, i.e. chance. In my view, in this context most readers would interpret "not by chance" in the sense of statistically significant (sensu hypothesis testing), although even this would be problematic and not straightforward. In brief, the authors should be aware of the complex statistical implications of the apparently simple statement provided on l. 4, and should therefore refine the wording not only here but throughout the text in order to make sure that all interpretations are unambiguously supported by the results.*

  - We agree with the referee and revised the text, especially concerning the term "by chance". We want to measure the statistical relevance of the methods in comparison with chance agreement.

    "Consequently, the results obtained by GAM-RV are statistically relevant and in better agreement with permafrost data from IPA/FGDC."

- *Page 2248, line 19 - Why reference a GAM paper (Yee & Wild, 1996) and R package for logistic regression? Was the VGAM package used for MLR in this study? If yes, say so.*
    - Done

      "ML-GAM and MLR are performed with the R package "VGAM" (Yee and Wild, 1996; Yee, 2010a,b)."

- *Page 2249, line 27 - Please present quantitative results that support this statement.*
    - It is just an observation from the maps before to give quantitative indices.

      "Permafrost distribution obtained with ML-GAM shows better agreement with data than that obtained with GAM-RV (figures 3c and 4c)."

- *Page 2252, last line, to second line of page 2253 - Application of Fisher's test and Student's test to "variances from MLR and GAM-RV simulations" – unclear what "simulations" (or predictions) it refers to, and how these are tested. It has to be made clear what is tested for what reason, and if the test's distributional assumptions are honoured. The latter is unlikely here because the observations used for model fitting have to be assumed to be spatially autocorrelated.*

  Page 2253, line 2-3 - Even if the previously mentioned test results are statistically valid, the statistical non-significance does not show anything. Only positive (i.e. significant) test results should be interpreted, the inability to reject the null hypothesis shouldn't. However, even if, I do not see how this would allow an interpretation in terms of [reconstructed or hindcasted?] permafrost distribution being "strongly driven by the large-scale temperatures from CMs."
    - Thank you, we simplified our interpretation without Fisher's test or Student's test:

      "No significant decrease appears in terms of inter-variation between all climate models: the measured standard deviation (table 3) is higher than CTRL period and remains fairly stable around 3×10^6km^2, except for ML-GAM which halves the inter-variation between climate models."

- *Page 2253-2254 (itemized list): should be worked into a new Discussion section. In the Discussion, the present results and methods should also be discussed in the context of the broader literature.*
    - We did not follow the suggestion of the referee to create a new discussion section but we now present our results in the context of the broader literature and according to the short comment from Dr Saito.

      "Another study from Saito et al. (2010) confirms this result. Thus, the methods are limited by large-scale errors from climate models at the LGM time period. The better climate models are, the larger the improvement by the SDMs."

- *Page 2272, Fig. 5 – The box-and-whisker plots appear to be based on nine values each? I am not convinced that a box-and-whisker representation should be applied to such a small sample. I recommend omitting the box-and-whisker plots, and showing only the point symbols and for each model a line representing the median value.*
    - Done

- *Page 2273, Fig. 6 e) – The modelled relationship between discontinuous permafrost probability and annual mean temperature is clearly nonlinear and non-monotonic – how can this be accomplished by the MLR, which can only produce monotic relationships?*
    - Figure 6 shows the probabilities obtained by MLR not the linear relationship directly.

- *Wording:*

Some language editing is required (grammar, diction), the following list is not complete. Throughout the text: "inter-variability" -> "differences" [generic term], "inter-variation" [rarely used], "co-variation", "inter-relation" [may have a stronger interpretative or even causal connotation, which should probably be avoided in this context]; "variability" refers to the "ability" to vary, i.e. (in a statistical/stochastic context) to a property of the underlying random variable, not to empirical findings of very limited coverage.

  - We follow the referee's advice and replace the term "inter-variability" by "inter-variation" throughout the text.

- *Page 2234, line 19 – "proves" – avoid this strong word, I doubt that the comparison of several semi-physical models can "prove" the stated relationship between two physical phenomena, permafrost and SAT. Also, is it not trivially true that the permafrost – SAT relationship is not simple? Replace this statement with one that expresses a positive finding, e.g. that a particular relationship or process is particularly well represented by the chosen model(s)*

  - Done

- *Page 2234, line 24 - "Our SDMs do not significantly improve permafrost distribution" – of course they don't, they are models, not nature. Change wording: "...the prediction of... compared to..."*

  - Done

    "The smaller permafrost area predicted by GAM-RV is mainly explained…"

- *Page 2234, line 25 - "at this period" – omit, and write "LGM" earlier in this sentence.*

  - Done

- *Page 2235, line 11 – "local" – refers to sub-pixel / sub-grid-cell (I think so), or to locally varying as opposed to global (e.g., local trends across several grid cells, depending on particular geological or land-use conditions etc.); in l. 17, "local" seems to refer to "spatially varying"; on p. 2236 l. 14-15, "local" appears to mean "sub-grid-cell"*

  - We refer here to the "local-scale" (higher-resolution).

- *Page 2236, line 1 – "resolved" should presumably read "solved"*

  - Done

- *Page 2236, line 1 – "need a lot of computing time" -> "are computationally intensive"*

  - Done

- *Page 2237, line 7 – "simulated" should read "predicted" (same for most other occurrences of simulate / simulated elsewhere in the text, e.g., page 2248 line 6)*

  - We keep the term "simulated" for climate models and "predicted" for statistical models.

- *Page 2237, line 19 – "simulating" -> "hindcasting"*

  - The term "simulating" might be inappropriate strictly spoken, but the wide community of climate modelling uses the term "climate simulation" for past and future climates.

- *Page 2237, line 21 – "their" – whose? The paleo-environments of the CMs?*

  - Done

    "In paleoclimatology, discrepancies appear between large-scale climate models and data-proxies, the latter being intimately related to their close paleoenvironment."

- *Page 2237, line 23 – here and elsewhere in the text, the word "data" should be replaced with more precise, context-specific words. Here, for example, it may refer to direct or indirect observations of the climate variables to be predicted, so "observations" or "observational data" might be a better word choice; "data" could be anything.*

    - Done

- *Page 2238, line 5 – "geological" -> "cryospheric", or simply omit the word; not all permafrost researchers are geologists or consider their research or methods to be geological.*

    - The exact term is "geocryological" observations.

      "To validate the statistical models on present climate we use geocryological observations…"

- *Page 2239, line 2 – "simulate" -> "predict"*

    - Done

- *Page 2244, line 23 - "decrease of permafrost area" – use more precise wording; maybe the "smaller predicted permafrost area" or "negative bias in the predicted total permafrost area"; similar issues with "decrease" and "increase" elsewhere in the text*

    - Done

- *Throughout the text: "our" -> "the"*

    - Done

- *Page 2252, line 7-8 - "GAM slightly warms temperatures" – sloppy wording, the GAMs are not to blame for rising temperatures.*

    - We reformulated the sentence as:

      "IPSL-CM4 obtains slightly warmer temperatures with GAM"

Answer to short comment from Dr Saito

Dear Kazuyuki Saito, we read your paper with attention. Some of your conclusions are in total agreement with ours, especially about not enough colder LGM simulations from PMIP2 climate models. Consequently, we now cite your letter in our article.