Response to comments of reviewer #2:

Major Comments:

1.I agree with the other reviewer, that the snow issue should be better investigated. The results shown here are against many other publications stating the effect of snow insulation on soil temperature.

We refer the referee to our answer to question 2 of referee #1 where we show that the results are not in fact against previous publications, but fully in line with them when the effects of albedo and thin snow covers are considered.

2.One suggestion I can make is to calculate a "snow season" instead of using DJF values since snow can be persistent over spring. You can simply do it by air temperatures [days Tair <0] or if available, model show depth data [days snow depth >a threshold value like 1-5 cm]. Then compare the air vs surface temperature offsets during this snow season and see the results.

Unfortunately daily modeling output data under the Permafrost Carbon Research Network is not stored.

However we agree it is a good idea to use snow depth thresholds, using 4cm snow depth to explain the mechanisms of ISBA and JULES. See figures 9 and 10 of the revised paper and the detailed reply to question 2 of referee #1.

3.There needs to be subsection describing each model used in this study. It doesn't has to be long but at least give some important details about which processes they utilize and what major differences (grid-size/soil discretization/physical-biogeochemical processes/ snow schemes etc.) they posses compared to other models here.

We add the following paragraph in section 3.1.

"The LSMs in this study considered the following processes: dynamic vegetation, carbon cycling (Rawlins et al., 2015), snow, near-surface hydrological budget, soil thermal dynamics (Peng et al., 2015) and the treatment of freezing soil. Sophistication in the treatment of these processes varies amongst the models with each having specific parameterizations, In this study we investigate some key schemes and parameters that are important for permafrost simulation: 1) Unfrozen water / phase change. All models calculate soil thermal properties as a function of soil moisture and consider the phase change of water/ice, but CoLM and LPJ-GUESS do not consider transformation to ice of water solute mixtures below 0 ℃, which is a key feature in soil freezing and thawing. 2) Surface organic layer insulation. Only CLM4.5 and ISBA consider the insulating effect of moss. 3) Soil texture parameterization. The specified fraction of clay and sand in soil differs. LPJ-GUESS specifies the same soil texture for the TP as for the Arctic. 4) Organic soil fraction treatment. The organic content of soil differs among the models. LPJ-GUESS sets the same value for TP as for the more organically rich permafrost of the Arctic. 5) Snow processes. ISBA, LPJ-GUESS and UVic set static snow layers. UVic uses an implicit snow scheme while LPJ-GUESS uses the Bulk-layer scheme, which are both simpler than the dynamic multi-layer snow scheme of some other land models."

We also improve the Table 5 accordingly.

Revised Table 5. Year-round relative model characteristics on TP.

Model	Snow cover[1]	Albedo[2]	Soil water[3]	Unfrozen water effect during phase change[4]	Surface organic layer insulation	Snow scheme[5]
CLM4.5	Medium	Medium	Medium	Yes	Yes	Dynamic & ML
CoLM	Medium	Medium	Medium	No	No	Dynamic & ML
ISBA	Low	Low	Medium	Yes	Yes	Static & ML
JULES	Low	Low	Medium	Yes	No	Dynamic & ML
LPJ-GUESS	Medium	Low	High	No	No	Static & BL
UVic	None	Low	High	No	No	Static & L

[1] Low snow cover is confined to high elevations, medium tends to be on western TP
[2] LPJ-GUESS has constant albedo everywhere and UVic albedo varies slightly due to vegetation, year-round albedo variability for other models depends mainly on snow cover in winter and soil moisture, vegetation, etc in summer
[3] soil water content includes both liquid and ice fractions
[4] all models calculate soil thermal properties depending on soil moisture and also phase change of water, but CoLM and LPJ-GUESS ignore solute dependent freezing processes
[5] ML: Multi-layer, BL: Bulk-layer, I: Implicit; according to *Slater et al*. [2001]

Also some reference papers for each model should be included.

We add references for models in Table 1.

Table 1. The six land surface models, analyzed over the Tibetan plateau (TP).

Model	Native Resolution	Number of soil layers	Depth of soil column (m)	Spatial domain	Atmospheric Forcing Data
CLM4.5 Swenson and Lawrence, 2012 Oleson et al., 2013	1 °×1.25 °	30	38.1	Whole TP	CRUNCEP4[1]
CoLM Dai et al., 2003	1 °×1 °	10	2.86	Whole TP	Princeton[2]
ISBA Boone and Etchevers, 2001	0.5 °×0.5 °	14	10	Permafrost region follow IPA map	WATCH [3]
JULES Best et al., 2011	0.5 °×0.5 °	30	2.95	Whole TP	WATCH [3]
LPJ-GUESS Gerten et al., 2004 Wania et al., 2009	0.5 °×0.5 °	25	3	Permafrost region follow IPA map	CRU TS 3.1[4]
UVic Meissner et al., 2003	1.8 °×3.6 °	14	198.1	Whole TP	CRUNCEP4[1]

[1]Viovy and Ciais (http://dods.extra.cea.fr/)
[2]Sheffield et al. (2006) (http://hydrology.princeton.edu/data.pgf.php)
[3]Weedon et al. (2011) (http://www.waterandclimatechange.eu/about/watch-forcing-data-20th-century)
[4]Harris et al. (2013), University of East Anglia Climate Research Unit (2013)

4.Same for the 3 site locations. There is already good information on Table 4 about the sites but still it will be good to include a small subsection describing the similarities/differences among these sites. Especially at Fig.4, the 0.04m observation of D105 and 2.63m observation of D110 are missing. You can better explain the reasons in a subsection. Especially at Fig.4, the 0.04m observation of D105 and 2.63m observation of D110 are missing. You can better explain the reasons in a subsection.

We add sentences in P1781 L15, "... model results. The three stations are located along the Qinghai-Tibet Highway. D66 station is in the front edge of alluvial fan, with almost no vegetation. The soil is mainly composed of gravels, sands and pebbles. D110 is in the southern bank of ZhaJiaZangBu River. The ground is a wetland covered with short-stature emergent vegetation. The upper layer soil is composed of coarse and fine sand. The lower soil layer is mainly composed of fine sand. D105 is in the northern side of the Tanggula Mountain range. The ground surface is relatively flat, covered with plateau meadow. The soil is composed of both coarse and fine sand. The vertical profile of observed soil temperature of D66 extends from 0.04 m to 2.63 m, of D110 from 0.04 m to 1.8 m, and of D105 from 0 to 3 m. However the data continuity of the top layer temperature in D105 is not good. To examine modeled ground temperatures, we present the top ... ..."

and why you choose to compare these sites.

We have only the 3 observation sites and we have addressed this in P1781 L12-L13 "..., field studies on TP are quite limited, and we have only short duration (1996-2000) ground temperature profiles obtained from the GEWEX Asian Monsoon Experiment..."

Also, you can explain that you have used cutout of global simulations instead of running the models with the observed forcing for these sites and its consequent implications to the results.

We have addressed this in the 3.4 section, see P1781 L16-L17 "...temperatures (modeled temperatures were weighted bilinear interpolated onto the station locations) in Fig.4 and Table 4..."

5.I understand that the model results are gathered from RCN database and are restricted to the procedure of that project. However, monthly soil temperatures are not always enough for TSL style permafrost calculations. You can either request daily results from the modeling groups or at least mention this fact as one important reason for the performance of TSL method.

Unfortunately daily modeling output data under the Permafrost Carbon Research Network is not stored. Actually we have mentioned this point when addressing TSL method in the text. See P1775 L13-L18.

6.Your calculations are limited by model soil depth (3m) and you have mentioned that shortly in your text. However you can make more analysis with the models that have deeper soil layers. And maybe transfer the soil depth paragraph from conclusions to discussions. I leave this issue to the authors' choice.

This question is essentially the same as referee #1 question 3 to which we refer for our detailed answer. We address the point in the paper with a new Table 6 and discussion in a new section 6.2.

7.Observational map has its own uncertainty originating from the MAGT and statistical extrapolations. This should be mentioned more precisely in the text especially in your discussions.

Actually there is a paragraph in section 3.2 to specially address the uncertainty of Wang06 map. See P1780 L14-L18.

It is also good to address the uncertainty among different permafrost maps. We do this by describing the specific numbers about TP permafrost area reported by previous studies. We add a sentence in P1779 L24 as "...... different studies (Ran et al., 2012). Thus there is a large spread of observation-based TP permafrost area estimates from 110 $\times 10^4$ km^2 (Wang et al., 2006) to 150 $\times 10^4$ km^2 (Shi and Mi, 1988; Li and Cheng, 1996)."

To lower the impact of mismatches to Wang06 map, you might consider discussing inter-model range of TP permafrost area more. Fig.2b, for example, gives too much impact on mismatching Wang06 map.

The Wang06 map was re-gridded onto each model's resolution. To compare between the models would require each model to be re-gridded to each other's model resolution before comparison can be made. This would be confusing to describe. We also think that there is a fundamental difference in comparing models with an observation-based map rather than simply between each other.

8.To improve the scientific value of your model inter-comparison results you have to tackle each of the following issues: 1.forcing data, 2.model spatial resolution, 3.model time step, 4.model spin-up, 5.model soil layer dicretization,6.model soil depth, and finally 7.model processes. I assume it is most valuable to confine the differences to model processes and for that, one needs to make sure the others are the same or at least they have negligible differences. From your experiment, I see that only point 3 (time step) is the same. And you have mentioned point 1 and point 6 in your text. Although you have shown points 2 (spatial resolution) and 5 (soil layer discretization) in your Table 1, you did not mention them in the discussions.

So you should clear the issues regarding to points 2,4, and 5.

For point 2, we reduce the impact of spatial resolution on the results by re-gridding Wang06 map onto each model's spatial resolution. Although the spatial resolution varies among these models, the models are evaluated objectively and separately by comparing with Wang06 map of the their own spatial resolution, when comparing permafrost area and calculating kappa coefficients.

For point 5, we interpolate and extrapolate the ground temperature onto the common layers: 0, 0.05, 0.1, 0.2, 0.5, 1, 2, 3m. So there is no difference among models in calculating permafrost area when referring to soil layer discretization.

For point 4, model spin-ups could be of course different, but in all models were run for long enough (around 1 000 years) to ensure that the deep carbon is in equilibrium. Accordingly, we add one sentence in P1778 L25 to address this point, as "Model spin-ups are also different, but as they are long enough to ensure that the deep carbon is in equilibrium, about 1000 years, the spin up impacts should be small...."

9.Would it be possible (or useful) to include the correlation coefficients next to kappa metric?

We think kappa coefficient is already a good and common applied method for mapping similarity. So it is not necessary to use other statistics here.

I can suggest you to prepare a soil temperature plot showing annual mean, minimum, and maximum values of each soil layer temperature at the sites and maybe also the selected region or common region. With the soil temperature envelopes plotted in this style, we can see the mismatches of each model more clearly than the time series plots in Fig.4 and Fig.5.

Our plots of annual cycle in Fig. 4 and 5 show not only the maximum and minimum yearly temperatures (as requested) but also usefully show the time delays of models. And we compare the annual mean, and seasonal cycle amplitude of soil temperatures in detail in Table 4.

10.Why is LPJ-GUESS always the coldest? Your explanation in Sect 5 is very hypothetical. Unless you have the actual soil conductivity values or soil water content to compare, these are just candidates for the mismatch. This might as well be related to other soil processes like **type of soil heat transfer, coupling of soil water and vegetation cover** and several other soil parameters.

We agree other mechanisms you list are possible, and mention them as possibilities in the text. Unfortunately, the soil moisture output of LPJ-GUESS includes both liquid water and solid ice, not separately, and there is no soil conductivity output.

That must be one simple process that is uniquely different than other models.

As we discussed, the underestimation of soil temperature in LPJ-GUESS can be due to many factors, and the most possible are inappropriate prescriptions of soil thermal properties, poor representation of soil hydrology, and mis-match of vegetation types **for the Tibetan Plateau**. See P1786 L18 - P1787 L7. Fig. 8, Fig. 4a and c, and the soil texture setting of LPJ-GUESS (Table 5) which all support our explanation.

One obvious problem for this model's results is that why is it colder even though it has higher snow depth.

We think that the cold ground temperature of LPJ-GUESS can also be explained by its snow depth, snow density and snow scheme. Accordingly, we added one paragraph to address this point in section 5.. "LPJ-GUESS shows a similarly thick snow depth in the western part of Tibetan Plateau as CLM4.5 and CoLM (Fig. 6), but does not show as large surface temperature offset as those two models (Fig. 7). That is because LPJ-GUESS has a fixed snow density (362 kg/m3) which is higher than used in other models, and a relatively simple Bulk-layer snow scheme, with one static snow layer, unlike the dynamic multi-layer snow scheme of CLM4.5 and CoLM (Table 5)."

11.UVic is the warmest among models. You say UVic has no snow cover, then what is shown in Fig.6?

Please notice the color bar in Fig. 6 is not linear. The color in UVic represents some values smaller than 0.1. Actually the snow depth of UVic is less than 0.001 cm. But UVic does have snowfall and a simple snow scheme. Thus we treat the snow depth of UVic as 0. We improve Fig.6 to avoid such misinterpretation.

This is one other reason to explain models in a different section. You attribute the overestimated soil temperatures of UVic to snow sublimation. Then I don't understand why the soil is warmer. The longwave radiation should be used for this sublimation you mention, not to warm the soil. And since there is less snow cover in UVic (Fig.6), we should expect cooler ground temperatures, which is not visible in Fig 7.

Yes. The sublimation will cool the ground. But since there is no snow in UVic, the rather low albedo in winter (0.25) can absorb much more solar radiation, which will provide energy for sublimation and warm the ground. Although sublimation can cool the ground, the warmer ground indicate the low albedo effect is stronger in UVic.

Please see our detailed reply to question 2 of Referee #1.

12.I also don't understand the explanation of JULES and ISBA models being cooler at the surface even though they have much deeper snow depths.

Fig.6 shows that in most places the snow depth of ISBA and JULES is less than 10 cm. We argue this cooling is due to the “thin snow cooling effect”.

Please see our detailed reply to question 2 of Referee #1 and figures 9 and 10. .

13.What is the point of using MIROC-ESM results in this intercomparison? I don't see an immediate relevance comparing a fully coupled model to offline simulations of different models. Please justify your choice or remove that model.

We delete Miroc-ESM in Fig.5. And to avoid misinterpretation, we deleted the sentences (P1778 L26 - P1779 L6) “We also analyzed (but do not show here) ... ... with the stand-alone LSMs”.

14.What is the message to model developers for a better TP estimate? What needs to be improved according to your results?

We address this in the conclusions: "Although most models can capture the threshold value of MAGT and SFI, their ground temperatures still show various biases, both in the mean annual value and the seasonal variation. Therefore, most models produce worse permafrost maps with the TSL method. The TSL method is a more demanding, and to date, elusive target" and "If the observation sites for soil temperature are representative, then LPJ-GUESS and UVic have substantial biases in their soil temperature simulations, mainly attributable to inappropriate description of the surface (vegetation, snow cover) and soil properties (soil texture, hydrology). Other models (ISBA, JULES) show biases in the simulation of winter soil temperature"

Minor Comments:

P1771 L17 and L20: produce “better” permafrost maps of the TP you mean?

Here we mean with MAGT and SFI 1) the derived permafrost area is nearer to observation-based estimate and 2) the permafrost distribution is also better, with higher kappa coefficient. We have addressed this through Sect.4.

P1772 L9: lose the comma

Done.

L9: “plays” -> “play important roles”

Done.

L11: lose the comma

Done.

P1773 L20: majority of your models must be tuned for several different sites around the world. What do you mean “different from where they were tuned”? Maybe you can mention that they are mostly used to estimate Arctic permafrost and not the TP. That can clarify the aims of this work. But these are global models and they are not tuned only to NH areas...

Agree. We rephrase the sentence as "We note that this approach provides information on the modeling ability of current models on the warmer and physically unique TP permafrost in a NH simulation, hence providing a test of reliability for simulations of present and future global permafrost over TP."

P1774 L6: model’s -> models’

Done.

P1775 L9: remain -> remains

Done.

L10: model studies -> model-based studies

Done.

L14: most of these models can provide daily temps or even sub-daily temps. You should at least mention the restriction of the model results that are available from RCN.

Done. We change the sentence in L13-L14 with “... on TP. Data at higher than monthly temporal resolution are not stored by the models in the PCN archive. Therefore TSL diagnosis......”

P1776 L3: you can provide one supplementary plot/table to show that 38 m vs 3 m does not affect the MAGT method results for CLM.

Done. Please see our detailed reply to Question 3 of Referee #1 and Table 6.

P1779 L1: What is the reason to use MIROC-ESM here? As you say it is not comparable to offline-forced models. I don't see the input of mentioning that to this manuscript.

Sorry, this is misleading. We delete all mention of MIROC-ESM.

P1781 L3: lose the comma

Done.

L4: if you are talking about Cohen's paper, then you should put the reference out of parenthesis

Done.

L18: sites -> sites'

Done.

P1783 L24: I don't understand what you are talking about, when you choose K>0.4, then all models except CLM passes for the MAAT method. And for the criteria K>0.2, UVic also passes for MAAT and F methods. Please clarify which methods you are talking about here.

This is a misunderstanding. We should keep in mind that the permafrost derived with TSL, MAGT and SFI are modeling results, but those of MAAT and F not. Thus here we only focus on K associated with TSL, MAGT and SFI to evaluate the modeling ability.

To avoid this misinterpretation, we added a sentence in P1783 L23, "......K (Sect. 3.3), and we limit discussion to the K associated with TSL, MAGT and SFI, which are calculated with simulated soil temperatures. If we take the (arbitrary) ......"

P1784 L3: Please mention which figure or table you are referring to. In which figure do we see the seasonal cycle amplitude of ISBA is better matched than others? In Fig 4d, ISBA results are not so similar to the observed in terms of amplitude. In Fig 4a and 4c, almost all models (except LPJ-GUESS) have good matching amplitudes. And in Fig 4b is the only plot where we can see a better match of ISBA. If this is the case, you only plot where we can see a better match of ISBA. If this is the case , you should revise this sentence. Yes in Table 4, we can see ISBA is the only one that satisfies the <2C condition for all sites/ depths but considering there is only 2 sites for the lower depth (2.63), it is hard to generalize

This is a misunderstanding. 1) all the numbers of Table 4 are calculated from Fig. 4. So we can just focus on Table 4 here. 2) We didn't mean to emphasize ISBA is better than others on seasonal cycle amplitude simulation. "Bias is $\leq \pm 2.0\,^{\circ}$C" is a very loose threshold. Only one model (ISBA) seems to meet this criterion. The sentence is changed : ".. Only one model (ISBA) is consistent with the limited observations".

P1785 L14: CoLM model does not show lower mean annual temperatures than CLM or JULES according to Table 4

Yes you are right, we neglected to mention that we were referring to the selected region. We rephrase this part:

" We investigate both the air and ground temperature (Fig. 5) of the selected region (the region shown in Fig. 1), which is the coldest part of TP and should be permafrost. CoLM simulates no permafrost in the selected region despite CoLM having lower mean annual ground temperatures for the 3 m layer than many other models (ISBA, CLM4.5 and JULES) (Fig. 5). "

L18: classed -> classified

Done.

L18: permafrost -> non-permafrost?

No. "**precluding** it being classed as permafrost with the TSL method" means it is non-permafrost.

For the last paragraph of section 4.4, you should mention that you are talking about the selected region rather than the observational sites.

Done. We added a sentence in L13, "...variability (Table 3). We investigate both the air and ground temperature (Fig. 5) of the selected region (the region shown in Fig. 1), which is the coldest part of TP and should be permafrost. CoLM simulates no permafrost in the selected region despite CoLM having lower mean annual ground temperatures for the 3 m layer than many other models (ISBA, CLM4.5 and JULES) (Fig. 5). However, ..."

L21: revise the first sentence of sect 5. Too long to deliver the message clearly.

Done. It reads now, "As discussed in Sect. 4, the most noticeable ground temperature discrepancies among the 6 models are the underestimation of soil temperature by LPJ-GUESS and the overestimation of soil temperature by UVic, which lead to the largest biases in simulated permafrost area."

P1788 L10-12: sentence is too long to make sense. Separate the last part starting with "observation-based Wang06..."

Done.

P1789 L2: give references to show the need for model improvements and model depths extensions.

The reference for model depths is mentioned below, see P1789 L3, "Nicolsky et al. (2007) recommend a soil column of at least 80 m for models applied to arctic and boreal regions."

Section 4 contains both results and discussions. Put the title correctly or make a better separation between pure results and discussion points.

Done. We change the title and add a new discussion section 6.

You mention model soil depth could be a reason but you don't discuss that in your discussion sections.

Done. This question is essentially the same as referee #1 question 3 to which we refer for our detailed answer. We address the point in the paper with a new Table 6 and discussion in a new section 6.2.

Fig 1: your legend is not clear. What is the "selected region"? It is only described later in the section 3.4. You should describe it also in the figure caption.

Done. It now reads:

"Figure 1. Permafrost maps derived from different diagnostic methods and models compared with Wang06 map. Permafrost inside the common modeling region is used for all-models inter-comparison, while permafrost outside allows further evaluation over the whole TP for CLM4.5, CoLM, JULES and UVic. The observation-based map of permafrost (Wang et al., 2006) is re-gridded to match model resolution. The selected area in the western TP (33 °- 36 °N, 82.5 °- 85.5 °E) is used to examine across-model differences in Figure 5. Insets show location map of TP and how the common region is related to the TP."

It doesn't make sense to put Wang06 map in between methods. You should make a separation between methods and observational map.

The intent is to separate the direct and indirect methods. This also placed observations in the middle of the plot where it is most easy to compare across all the model simulations

Can you also put description to the smaller two maps under the panel (Tibet and common region).

We added a description to the caption: "We also show a location map of TP and how the common region is related to the TP."

Site locations are not very visible. Try to choose another marker and make them bolder

Done.

Fig 2: Can you explain how you calculated the error bars from resolution differences?

The error bar is calculated as half of the averaged grid cell area of the model, so is model resolution dependent. We add this clarification into the legend for Fig. 2.

Fig 4: Mention the reason of using only upper soil temp for D110 and only the subsoil temp of D105 sites in the caption

Done. It reads now:

"Monthly soil temperature variations at 3 stations from models and observations. (a) and (c) soil temperature of top layer. (b) and (d) soil temperature of deeper layer, 1996-2000. "Mean" denotes annual average temperature. We use the topmost available soil temperatures (0.04 m at D66 and D110, no good data for D105) and lowest available ones (2.63m at D66, 3m of D105), while D110 has only temperatures at 2 m depth."

Fig 6: You should mention the source and description of observations in the figure caption. Explain OBS_0.5, OBS_clm4.5, OBS_uvic in the caption.

Done. It reads now:

“Figure 6. Winter snow depth for the common region, averaged over 1980–2000. Note the nonlinear color scale. We use the Long Time Series Snow Dataset of China (Che et al., 2008)(http://westdc.westgis.ac.cn) as observed snow depth. The observed snow depth plot is further interpolated onto the models’ resolutions as “OBS_”. The OBS_05 is in 0.5°resolution for CoLM, ISBA, JULES and LPJ-GUESS. The OBS_CLM4.5 and OBS_UVic are in the resolutions of CLM4.5 and UVic separately.”

Table 5. Start all words with capital letter.

Done. It reads now “Table 5. Description of Model Characteristics Relevant to Soil Temperatures.”

What does “snow cover: none” mean for UVic? No snow representation? This has to be mentioned because it affects everything for soil thermal dynamics...

UVic has both snow fall and a snow scheme, but zero snow cover depth. On TP the snow is removed by sublimation, as we mentioned. Please see our detailed reply to your first comment. And we improve Fig.6 accordingly.

What do you mean by “unfrozen water effect during phase change”? Does that mean no freezing/ thawing occurs in CoLM, LPJ-GUESS, and UVic?

We clarified this in section 3.1" That is all models calculate soil thermal properties as a function of soil moisture and consider the phase change of water/ice, but CoLM and LPJ-GUESS do not consider transformation to ice of water solute mixtures below 0 ℃."