Reviewer 1:

General comments:

This manuscript describes an emulation of earth system model output, specific to the biogeophysical responses to forest cover change. Afforestation, reforestation, and reference scenarios are used to assess the model's representation of change in air temperature at and 2 metres above the land surface. The emulator appears to perform well against MPI-ESM, but substantial errors accumulate for extreme afforestation and deforestation scenarios in the other models and against observations.

The research topic is an important and relevant one, as the computational complexity of ESMs increases, as does the importance of assessing and planning land-based climate mitigation responses. The study is well-designed and clearly argued and the manuscript is well-structured and thorough. Overall I would recommend its publication with minor revisions.

We thank the reviewer for their encouraging comments and thorough review of the manuscript. We have tried to address their comments in the new version of the manuscript, as detailed in the answers below.

Specific comments:

What about anything other than surface and 2m air temperature? At least acknowledge its limitations as a metric for all "biophysical" climate impacts (e.g. that it conflates albedo- and latent heat-related surface temperature changes despite their quite different effects on the atmospheric energy balance).

Thank you for this point, and we agree that it merits mentioning. We have now added some lines on this under Section 6.2, L416. Particularly we highlight that:

“It should be noted that to provide complete representation of the biophysical effects of tree cover change, albedo and latent-heat fluxes would have to be considered as well. To this extent, the temperature responses provided by TIMBER arise from a combination of the effects of albedo and latent-heat fluxes responses to tree cover changes on the atmospheric energy balance.”

There is no mention of tree species/PFT and how the variation in tree types around the world influences biophysical properties in different biomes. Since tree PFT likely accounts for a major part of the variance of biophysical properties in the ESMs, I would have liked to know how this was translated to the emulator.

The figure captions should fully define all acronyms and variable names. It should not be necessary to read the text in detail to gain a basic understanding of what the figures show.

Thank you for this point, we now add more description to the ESM experiments as:

“The AFF (DEF) was implemented by removing the non required vegetation types (i.e. crops, grassland and shrubs for AFF and forest, grasslands and shrubs for DEF) and upscaling the remaining vegetation to fill up the grid cells. Bare land was conserved throughout this process in order to respect the biophysical limits of where vegetation can grow.”

We had initially also tried differentiating between PFTs of needle-leaf and broad-leaf trees (given the likely difference in biophysical responses). However, we quickly realised that between ESMs these were defined differently (e.g. MPI-ESM1-2-LR seemed to consider evergreen and deciduous as separated into tropical, boreal and temperate, whereas CESM2 mainly considered needleleaf and broadleaf). For sake of simplicity, we thus did not differentiate PFTs as much, however we do add a sentence on this in the Future Developments section.

We have also written out acronyms in the figure captions now.

Technical corrections:

L5: usage: suggest "accounting for them" to replace "their accountance"

Thank you, we have now implemented this suggestion.

L9 and L413: usage: suggest removing "agilely", which does not feel like a natural word, and its meaning is communicated later in the sentence already (with "by means of a computationally inexpensive manner"

Agilely has now been removed

L57: usage: suggest "hereafter" to replace "hereon"

We have implemented the suggested modification.

L59: typo: "biobphysical"

Thank you for spotting this, it is has now been corrected.

Reviewer 2:

Summary

The authors present a new approach for estimating the local surface temperature responses to changes in tree cover. They first fit and validate multiple GAMs to model outputs and observations (separately). They also generate additional GAMs to represent parametric uncertainty in these models. They asses the validity of using an existing observational relationship to determine 2-m surface temperature from ESM estimated min and max surface temperatures. They show reliable GAM results and state that the existing relationship is good enough for this study. These tools are then applied to two SSP scenarios of tree cover change to demonstrate the utility of this approach for estimating temperature change due to change in tree cover.

Overall review

I appreciate this effort and see this as a good contribution toward understanding the potential effects of land cover change on climate. The authors employ a clever mix of statistical techniques in a thorough framework that estimates surface temperature change due to tree cover change. My main concerns are with the paper organization and incomplete description of methods. Overall the methods are sound (as far as I can tell), although one aspect in particular could be improved, and comparison with an alternative method is warranted. I recommend addressing the following three main issues. See below for detailed comments.

We thank the reviewer for their thorough review of the manuscript and their constructive feedback. We have tried to address their comments in the new version of the manuscript, as detailed in the answers below.

1) The methods are incomplete. Some methods are out of place while others are missing or unclear.

We thank the reviewer for identifying room for improvements in the methods. We have now tried to be more detailed in the methods section, which hopefully strengthens the overall paper. Most importantly, we have:

- Provided more elaborate description on the ESM experiment in section 2.1 (see comments later on)
- Provided a Section 2.3 “Tree cover change scenarios in selected SSPs” that details the input data used by TIMBER to predict into different SSP scenarios
- Adjusted Section 3.2 so that the Blocked cross-validation and its use for model calibration and evaluation is clearer
- Adjusted Section 3.2.3 such that the blocked cross-validation performed to evaluate “no-analogue” tree cover change conditions is more detailed.
- Divided Section 3.4 into 2 subsections that clarify (1) how we finally arrive at 2-m air temperatures within TIMBER’s framework; and (2) how and why we quantify parametric uncertainties to finally identify significant temperature responses to tree cover change under different SSP scenarios, by means of the signal-to-noise ratio.
- Provided a more explicit introduction to Section 4.1 such that the differences between Figure 3 and 4 and their purposes for model calibration and model evaluation, respectively are clearer.

Some other modifications and additions were also made as later detailed.

2) I am not convinced that the Hooker relationship as used is adequate for the final estimation of 2m T. For the time being, it shows the potential of this overall approach, but it seems to have a rather large spread of error for operational use. The authors should consider taking up their suggestion and doing a model-specific fit of Tmin/max to 2m T, and comparing this to the current results.

We had initially thought about this, but realised it would get complicated since the observational 2-m data provided by the Global Historical Climatology Network monthly (GHCNm) would have to be regridded to the ESM grids. Furthermore, given that we look at relative changes in 2-m temperature we thought that any biases introduced would cancel out anyways (as long as the spread in biases are the same as checked for in Section 3.3.2). We have however added a line in the Conclusions and Outlook L394:

“It should be stressed that in its current set up TIMBER provides conceptual emulations of temperature responses to tree cover changes. In such its outputs should be caveated with the limitations of the current set up, specifically, that they are produced with limited amounts of training data, as well as that the 2m air temperature is diagnosed using observational relationships – as provided by Hooker et al. (2018) – directly applied to the ESM space”

3) The scenario results should be compared with other estimates of temperature response to land cover change. There are comparable examples of this using these three ESMs. While such examples use a simulation differencing method that encapsulates much more than the method in this study, One would expect some

similarities given the dominant role of tree cover change in surface temperature change. Maybe the comparison is between the TSmean change estimated by the GAMs for each scenario, rather than the 2m T, and the mean surface temperature change estimated by model differencing.

The LAMACLIMA experiments were designed specifically to enable the development of TIMBER. In such they are the only set of experiments that separate the local and non-local signals of climate responses to specific Land Cover Land Management (LCLM) forcings such as tree cover change. More so they represent a concerted effort with 3 participating ESMs that incorporate the LAMACLIMA experimental design in a self-consistent manner. To our knowledge other LCLM simulation exercises exist such as, LUMIP, LUCID, CORDEX-FPS LUCAS etc. However they combine multiple LCLM types (i.e. they do not only tree cover change but also irrigation, urban expansion etc) and they don't separate local and non-local signals. In such, we find it difficult to compare the final SSP results to these experiments due to the fundamental differences. We do however compare the overall emulator outputs to those obtained by deHertog et al. 2022 and Meier et al. 2018, which represent the closest possible reference points. We did some further looking into simulation differencing exercises but couldn't find any that directly isolated the local, biogeophysical responses of temperatures to tree cover change under the relevant SSPs (2-4.5 and 3-7.0). However, if there is a paper we have overlooked we would gladly like to incorporate it for comparison of TIMBER outputs within our study.

Specific suggestions and comments

Abstract

line 1:

"Society is set to experience…" is a bit overstated here. Evidence shows that the world is not even close to meeting the Paris Agreement. I suggest revising to something like "Land cover change has been proposed to play a significant role in achieving the temperature goals…"

We agree with this and will change accordingly to:

"Land cover change has been proposed to play significant role, alongside emission reductions, towards achieving the temperature goals"

line 6:

expand ESM

We have modified ESM to:

Earth System Model (ESM)

1. Introduction

lines 32-33:

you probably want to include the boysen 2020 paper in this citation also.

Thank you for this helpful suggestion, we have now included this reference

2. Training datasets

line 75-77:

Is there just one REF scenario? Then “The REF scenario spans 150 years…”

Yes there is only one REF scenario, we have modified the sentence as suggested

What do you mean by “full expansion?” Is 100% of the surface covered by forest or crop? Or do you just swap forest and crop in AFF and DEF?

We have revised the description of the ESM experiments on which the emulator is calibrated. More specifically we have added the exact land surface models used for each ESM and elaborated on the experimental design by adding L78-L81:

“The AFF (DEF) was implemented by removing the non required vegetation types (i.e. crops, grassland and shrubs for AFF and forest, grasslands and shrubs for DEF) and upscaling the remaining vegetation to fill up the grid cells. Bare land was conserved throughout this process in order to respect the biophysical limits of where vegetation can grow.”

lines 80-84:

Doesn't this require additional simulations to the ones listed in the previous paragraph? These additional simulations are different than the "full expansion,", correct?

You may want to reiterate here that the local biogeophysical responses are largely independent of the global land cover change scenario that drives the global biogeochemical response.

It appears that you don't do this local vs non-local decomposition.

The experimental design elaborated upon initiates full expansion following the checkerboard approach, which allows interpolation of both local and non-local signals as done in the peer-reviewed method of Winckler et al. (2017). In such, no additional experiments are required. We do specify that local responses:

"can be applied in any global tree cover change scenario,"

In L80-L84 and that non-local responses:

"depend on the global extent and patterns of land cover change."

We will however make this more explicit by modifying L83:

"Given that local responses are independent of land cover change scenario"

To:

"Given that local responses are independent of the global extent and patterns of land cover change, "

Maybe combine sections 2 and 3 into a Methods section? And relabel the included sections accordingly?

For sake of not having too many subsections we maintained sections 2 and 3 as separate (for GMD they do not allow subsubsubsections)

Add another section to the methods describing the creation and use (i.e. experimental design) of the tree change ssp scenarios. For example, are the changes used the changes over time from the start to the end of the scenario? Are these changes estimated from a particular model, as the scenario prescriptions do not include forest cover change?

We have added a section on "Tree cover change scenarios in selected SSPs" that clarifies the exact input data needed for TIMBER to predict temperature responses in other tree change ssp scenarios.

2. Training datasets

3. Statistical emulation of temperature response

Explain why you selected January and July for your analysis.

We will add a line to L139:

"The performance for example months of January and July, which are representative of the hottest and coldest months for the Northern Hemisphere and vice versa for the Southern Hemisphere, is demonstrated on"

Modeling surface temperature

lines 124-128:

Does this mean that a separate model is built for each orographical feature? This stratification needs to be clarified. How are the orographical features defined? Show some example features on a map? If these are separate models, then this should be clear in calibration, evaluation, and figure 1 also.

It technically refers to a "Varying Coefficient Model "(see Hastie and Tibshirani 1993. Varying Coefficient Model. Journal of the Royal Statistical Society. Series B (Methodological) Vol. 55, No. 4 (1993), pp. 757-796 (40 pages)). The input numeric vector of orography is element-wise multiplied with the smooth evaluated at the corresponding covariate values:

Y = beta_0+beta_1(orog_s)*X+....

Where X is the lon,lat,treefrac covariate matrix and beta_1() is a function of the coefficient space modulated by orog_s. For simplicity of the text we do not go too far into this, especially since it is rigorously elaborated upon in Wood (2014). However, we have now added the reference to Hastie and Tibshirani (1993)

We have now clarified that orography corresponds to "meters above sea level" and also provided and referred to maps of orography used in the appendix.

lines 159-162:

This isn't clear. I assume you take the final model and run it within each block, but I don't understand how you assign the treefrac changes and what the binning is for. Do you replace all grid cell changes within a particular bin with a specific number (as long as this number isn't the same as the actual grid cell fraction?). Then what do you calculate the RMSE against if there is no analogue? And are you just comparing the binned RMSE with the original RMSE - what does this tell you if you don't actually have a corresponding temperature change for the binned inputs? Do you compare these values for each bin separately? How does orographic stratification play into this?

As part of the broader effort to better describe the employed methodology, we have re-structured Section 3.2.3 and now hope that it gives a clearer picture of how the blocked cross validation procedure to evaluate the predictive ability of Γm for amounts of tree cover changes 'unseen' during the model calibration was conducted. To answer the reviewer's questions specifically, first a set of blocks is constructed by, within each of the seven continuous geographical regions visible on the top of Figure 1, separating grid cells according to the amount of tree cover change ∆2015treeFrac encountered between the REF and AFF or REF and DEF simulations, using bins of ∆2015treeFrac magnitudes: [0.01-0.15), [0.15-0.3), [0.3-0.5), [0.5-0.8) and [0.8-1.0], for both positive and negative signs of tree cover change. A blocked cross validation procedure is then conducted by using these blocks. Successively and for each block, Γm is trained on data for the whole land area except over that block. For each block, the RMSEs between the values predicted by Γm and the actual ESM or observational data are then calculated.

Diagnosing 2m surface temperature

line 166:

What do you mean by observational surface temperature? What is being measured?

I assume that the observational 2m temperature is literally 2m off the ground, regardless of canopy type and height, correct?

As referred to in Hooker et al. (2018), the Global Historical Climatology Network monthly (GHCNm) dataset is used (See doi:10.1175/JCLI-D-18-0094.1). This collects weather station measurements from all over the world and performs their own peer-reviewed interpolation to get a continuous dataset of air temperatures. The air temperatures are defined according to the International Organization for Standardization (ISO) as 2m above the surface. In such, it is indeed 2m off the ground, regardless of canopy type and height. We have now also added the reference to the GHCNm dataset.

lines 189-209:

This is a clever way to test the validity of using Tmin/max. Do you use the same gaussian kernel for both data sets? I am unfamiliar with this technique and the kernal selection seems a bit arbitrary (although you probably want it to look like a fit to each respective dataset, which indicates that the kernels would be different). In any case, you assume that each bivariate dataset for T is normal in order to use this technique. Is this the case? Did you do a regression for each dataset to check for normality and determine the appropriate kernels? Presumably you then use the each data set to train the respective covariance matrices with the respective training data error.

A separate kernel is fitted to each dataset. Indeed we assume that it is normally distributed, and we have added a Shapiro-Wilke test for normality to the appendix to substantiate this.

Emulating 2m air temperature under different land cover change scenarios

This section is misnamed, and it seems you intended to include the description of the scenario analysis

This section covers both how we draw temperature values from the GAM and subsequently derive 2m Temp to look at new SSP scenarios. For sake of chronological consistency with the results as well we would prefer to keep it as separate. We now clarify in this section that the procedure introduced is used to treat the signal-to-noise output of the final 2-m temperature responses under different land-cover scenarios so as to identify significant responses.

This section needs to be moved into section 3.2 and is about uncertainty in the GAMs model.

Section 3.4 covers both how the local response in mean 2m air temperature is derived by combining the methodologies described in Sections 3.2 and 3.3 and applying them to 2 SSP scenarios, as well as how the uncertainty in these final results that arise from the difficulty to select the right set of parameters during the GAM calibration is estimated. Assessing this uncertainty is done as a last step in order to identify the signal-to-noise ratio of the final result from the full emulator, after application on a tree cover change scenario that was not included in the model calibration. To make this clearer, we have renamed and restructured Section 3.4, and in particular divided it into two subsections: 3.4.1 Predicting mean temperature changes from land cover changes in the SSPs and 3.4.2 Estimating parametric uncertainty in the temperature change predictions. We hope that this overall makes clearer why this section is presented last in the Methods part.

lines 228-231:

Is this done globally for each relevant pixel? Or do you just take 200 samples total, from wherever (which doesn't make sense if you average them)?

This is done globally for each relevant pixel, we have clarified this now in L228.

4. Results

Blocked cross validation results

line 248:

Please explain this month-specific approach in the methods section.

Thank you for the suggestion, we have clarified the month-specific approach in the methods now.

lines 261-263:

Do you mean less than 2.5? Figure 3 shows a lot of area with RMSE well over 0.25 K.

You may want to quantify how much and which areas have RMSE < 0.25 K.

An RMSE over 0.5 could actually mean that this method is not very good at representing temperature changes due to tree cover change, as such changes are often estimated on the order of 1 degree or less.

Comparing with appendix C doesn't back up the statement about high RMSEs corresponding to extreme deforestation.

Also, based on the Obs in appendix C, extreme deforestation is realistic.

We now quantify area with RMSE < 0.25 K, as mentioned in the next comment we have also made the colour scale continuous so the RMSEs are more discernible. High RMSEs do happen for deforested areas within CESM2 and EC-EARTH, and otherwise occur less systematically in Obs. and MPI-ESM; we specify this now.

lines 265-278:

Figure 4 is just a statistical representation of figure 3, correct? Does figure 4 show the median and interquartile range? Figure 4 does show that the numbers are generally lower than they look in figure 3, in most cases. Maybe a different scale in figure 3 would be helpful.

Figure 3 provides the final model with the lambda and nbf calibrated according to blocked cross-validation. The performance shown is after it has trained on all available data. We will add a different scale to Figure 3 and agree that it may be misleading.

Figure 4 shows synthesis results of the final model's predictive ability into "no-analogue" background climate conditions. In such, blocked cross-validation is carried out with the final model again (with blocks composed of background climate and continuous geographical region), and we consider only the RMSEs obtained for its predictions into the test set blocks.

lines 280-294:

It is unclear how you did this.

We have now clarified this in the methods under the Evaluation section, renamed to "Blocked cross validation for model evaluation". More specifically, we now say that a second round of blocked cross validation is conducted on the final chosen models. Instead of climate and geographical regions however, the blocks are composed of the same seven geographical regions, and tree fraction changes (binned by magnitude and sign of change) in L160 on:

"Then, another set of blocks is constructed by splitting the same seven continuous geographical regions as in the previous section, but by dividing the grid cells constituting those according to the amount of tree cover change Δ2015treeFrac encountered between the REF and AFF or REF and DEF simulations, using bins of Δ2015treeFrac magnitudes: [0.01-0.15), [0.15-0.3), [0.3-0.5), [0.5-0.8) and [0.8-1.0], for both positive and negative signs of tree cover change. A similar procedure to that applied for the no-analogue background climate conditions is then conducted but using these newly constructed blocks: Successivel and for each block, Γm is trained on data for the whole land area except over that block, using the sets of parameters identified in Section 3.2.2. For each block, the RMSEs between the values predicted by Γm and the actual ESM or observational data are then calculated. They constitute an estimate of the predictive ability of Γm for tree cover change amounts unseen during raining and are presented in Section 4.1.3."

Illustration of TSmean outputs

lines 295-317:

Are these just the final models? Is the 95% interval just the distribution of the grid points within the latitudinal bands?

What about the parametric uncertainty in the model?

We only show the 95% interval for the distribution of the latitudinal bands as the main purpose of the figure is to show the GAM's representation of the expected responses across the complete spectrum of tree cover changes. Since the parametric uncertainty is mainly a tool to evaluate how significant these responses are (as now clarified and made more explicit in Section 3.4), we choose not to show it here so as to maintain the interpretability of the figure.

Surface 2m air temperature diagnosis

line 326 and figure 7:

95% is not an interquartile range

This was a typo, 95 has been removed now.

lines 327-333:

Are these results similar for the other ESMs?

Since other ESMs only outputted 6hourly TS values we had difficulty deriving TSmin and TSmax, so to remain within scope of this study we mainly focussed on CESM2 that provided 3hourly TS data.

Also, is this really good enough? ESMs can output more frequent temperature values, and so model values can be obtained closer to the measurement times. Maybe this improves the relationship? What are alternatives? Diurnal temperature modeling outside of the ESM? Is it possible to create a more appropriate basis using different data and Hooker's method? I see that later you suggest model-specific fitting.

We agree that this is sub-optimal, and provide further discussion on this under "Possible Improvements". We would like to stress that as in the title, the framework is conceptual and could benefit from continuous improvement. In this paper however, we focus on step by step setting up the conceptual framework, which could then be modified and improved in a continuous manner. We do however acknowledge these limitations and areas for improvement within our Conclusions and Outlook, and have further added a line caveating the outputs of TIMBER as being conceptual with room for improvement in L394:

"It should be stressed that in its current set up, TIMBER provides conceptual emulations of temperature responses to tree cover changes. In such its outputs should be caveated with the limitations of the current set up, specifically, that they are produced with limited amounts of training data, as well as that the 2m air temperature is diagnosed using observational relationships – as provided by Hooker et al. (2018) – directly applied to the ESM space.

lines 336-339:

This is an overly optimistic assessment. Clearly Tmax has higher biases, evidenced by double the plot scale.

Tmax generally has higher values (See Appendix B), hence compared to the magnitude the biases are still acceptable. The main point of consideration is that the model should not have higher RMSEs when predicting into no-analogue conditions as compared to its RMSEs after having seen the whole training data set. We will however add a point acknowledging that:

"TSmax show quite high biases, although in relation to their absolute values these are of the same relative magnitude as those of TSmin and TSmax. Moreover, RMSEs of \Gamma^{max}_m show similar magnitudes when predicting into no-analogue conditions as those obtained after having seen the whole training data set."

this in L339.

lines 340-348:

Are these results similar for the other ESMs?

Unfortunately, we could not calculate these values for other ESMs as the data was outputted on 6h level (vs 3h for CESM2). This made extraction of TSmin and TSmax difficult, we are however in discussion with the ESM modelling groups on the possibility to derive TSmin and TSmax based on outgoing Long Wave radiation, however this would require further calculations and analysis that is beyond the scope of this study.

5. Exploration of tree cover scenarios

These are also results and should be in the Results section.

This has been moved under results now

Why did you select CESM2 for this and the other examples?

CESM2 was the only ESM that had TSmin and TSmax output available and covered a larger range of tree cover change in the AFF and DEF experiments, hence we chose to use it as a best representation of TIMBER's first conceptual results.

Can you compare these results to other estimates? There may be some confounding factors but there are studies estimating changes in surface temperature due to land cover change. While it may not be useful as evaluation, it certainly would be interesting to see if there are corresponding patterns.

We agree that this would certainly be interesting, however as also mentioned there may be some confounding factors (i.e. inconsistent split in the local and non-local signals) which would make comparison difficult. Since this study's main aim is to build a conceptual framework with which to provide impact-relevant (i.e. 2m) temperature responses, we try to focus on the model evaluation and simply provide example end results as well as compare the outputs to the closest possible studies i.e. that of deHertog et al. 2022 and Meier et al. 2018. We acknowledge that comparison with other studies would be interesting, although out of scope, for this study, and provide further discussion on it in the Conclusions and Outlook section now.

lines 355-360:

This should be in methods.

This has now been moved to the methods section 3.4 now

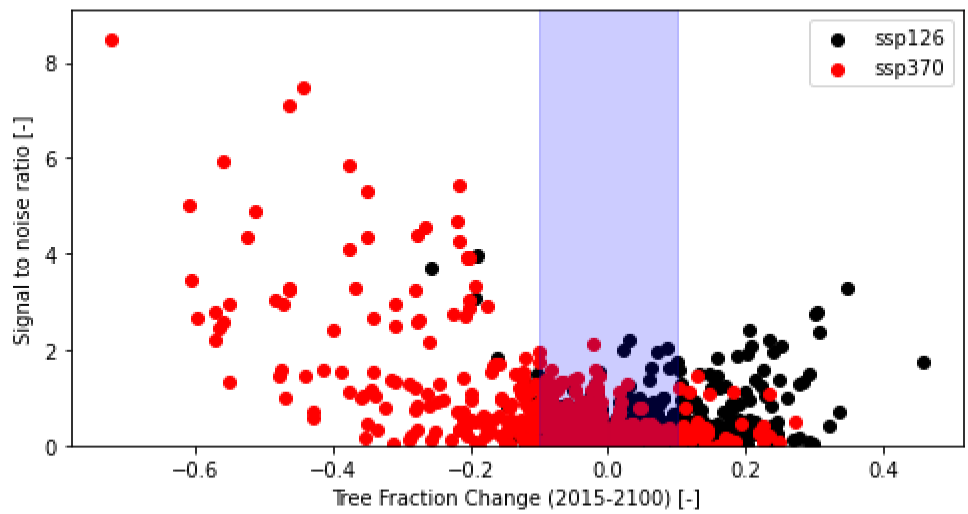
Also, figure 9 states 2015 tree frac change. What is this actual change? Is it the change over the century?

Yes this shows the final end-of-century changes in tree cover with respect to the year 2015 (2015-2100), we will clarify this in Section 3.4 now as well as have added a Section 2.3 “Tree cover change scenarios in selected SSPs” that provides more details on this.

lines 366-369:

It would be interesting to see scatter plots of signal to noise ratio vs change in tree frac. This may provide an estimate of a cover change threshold(s) that may result in significant temperature changes

As a preliminary step we made a plot that pools together all grid points (blue block indicates <|0.1| tree cover change interval):



It does show an increase in the signal-to-noise ratio at higher magnitudes of tree cover change, especially for deforestation (i.e. negative changes in tree cover). Some areas still however have low signals at higher magnitudes of tree cover change and this is dependent on the location e.g. in Figure 9, for both SSP 2-4.5 and 3-7.0 we see more than 10% afforestation in East Asia, however the signal-to-noise ratios remain below 0.5.

6. Conclusion

lines 417-424:

This is an important point. This method does provide an opportunity to look at model-specific physical responses as it isolates the tree cover change, which is often different across models. There are still challenges related to initial conditions and model-specific climates.

lines 425-430:

isn't there a winckler paper showing non-linearity in temperature response to change with changing initial forest fractions?

Winckler, J., C. H. Reick, and

J. Pongratz (2017), Why does the

locally induced temperature response

to land cover change differ across

scenarios?, Geophys. Res. Lett., 44,

doi:10.1002/2017GL072519.

Thank you for this reference, we have added it now to L430.