The authors performed a study to determine potential biome shifts under climate change. They make various projections, using 3 RCP scenarios, 5 GVMs and 31 biome maps. The novelty of this study is that they use random forest classification models to translate GVMs' LAI results to determine biome distributions.

Overall, the paper is well written. My main concern is regarding the focus of this study in relation to its novelty. Why do you focus so much on the future predictions, while the novelty of this study focuses on how to use GVM output in a more meaningful or coherent way. I would much rather see that the study focusses on that, and describes the differences between biome maps (e.g. why do maps differ so much, and maybe if the future scenarios remain, what causes the larger biome shifts at the borders of biomes – is that really just temperature change?). There are already quite some future biome map studies, and while a more coherent one would be nice, this study doesn't specifically focus on why the maps are different, so this study simply adds to the uncertainty in future biome mapping. I am not saying the study is not relevant, I am merely pointing out that, as I see it, the strong suit of this study is away from RCP scenarios and zooming in on how to make the best of GVM outputs and use them more directly, more consistently. The discussion now reads like a comparison of different papers on what biome changes where, which is not the main focus of this study and therefor perhaps not necessary (and not what you want to present as your main take away for this modelling effort).

We thank the reviewer for this important comment. We agree that the balance between methodological innovation and results on potential biome shifts is critical, and we also agree that understanding why biome maps differ is interesting and relevant. Our starting point was the observation that different GVMs use different sets of PFTs and model-specific biome classification schemes, which makes an objective comparison of modelled biomes for current and future conditions difficult. While comparisons of continuous outputs such as productivity or carbon stocks are straightforward, biomes always involve a post-processing and, somewhat arbitrary, classification step. This classification also leads to a lowest common denominator effect, i.e., the DGVM with the coarsest PFT and biome definition scheme constrains the overall biome classification. To enable an objective comparison of potential biome changes across models and scenarios, which does not suffer from the artefacts of more classical approaches, we first needed a consistent classification framework, i.e. the random forest approach. We see this method as a necessary step towards robust model comparison rather than as the main focus of the study. Robust biome comparisons require unified classification schemes. Our random forest framework provides such a unified, observation-informed classification of GVM outputs into existing biome schemes, enabling, to our knowledge, the first fully comparable multi-model assessment of future biome dynamics. Our approach therefore does not simply "add to the uncertainty", but demonstrates and quantifies the uncertainty related to biome classification and highlights that the choice of classification strongly affects projected biome shifts under future climate change. Such knowlege is essential for the development of conservation or climate change mitigation strategies because mis-classification and lead to inappropriate management decisions.

Furthermore, we think that it would be a lost opportunity to only focus on these methodological advances. Our results are related to issues like Amazon forest or boreal forest die-back and tipping elements in the Earth climate system, and our model results differ from the current mainstream story that forest systems will tip. Therefore, we think that we need this double focus, the first step to derive consistent biome changes from several models, and then the discussion of the relevant results.

In the revised manuscript, we will carefully consider the points raised by the reviewer and (i) strengthen the introduction to emphasize that the methodological innovation serves the primary goal of delivering robust, comparable projections of biome shifts across multiple GVMs and RCP scenarios, (ii) keep the focus on future projections and model inter-comparisons as the core results while clarifying how the method enables these analyses, and (iii) expand the analysis of spatial patterns of biome shifts (particularly at biome boundaries) to better address what drives these transitions beyond simple temperature effects. We will streamline the discussion to focus more on our multi-model results rather than on extensive comparisons with previous work.

Abstract

There are a few questions that come to mind.

We appreciate these clarifications and will revise the abstract accordingly.

First, why do we need biome projections if we have 31 observation-based maps already and we have 5 GVMs that project future vegetation. What is the benefit of the biome maps created in this study?

The 31 observation-based maps represent current conditions based on different data and methods, but they do not provide future projections. Future biome distributions can only be inferred from models, but GVM-specific classification schemes differ and are not directly comparable. To do so in a consistent way, we first need to classify GVM results into biomes, evaluate how well they reproduce observation-based biome patterns, and then apply the same classification to future GVM outputs. Our contribution is a multi-model GVM ensemble of future biome projections constructed with a unified classification framework. This enables direct comparison of GVMs within each biome scheme and quantification of how projected biome shifts depend on both model choice and reference biome map.

Second, the 4-56% projected change is huge, and as I read it, it is a certainty check of the outcomes: depending on the GVM and observation based biome map the random forest is build on, you get these large variations in projected biome change. The RCP's are simply scenarios you put in, so this is a simulation and not a model check. Thus, this sentence needs to be rephrased, as I probably do not understand what message you are actually trying to convey here.

The model evaluation step in our study is the comparison between the 31 observation-based biome maps and the biome maps derived from classifying current GVM results, where our approach shows high agreement (average κ across all models 0.77, maximum>0.9, see Fig. 1). The 4–56% range refers only to projections and expresses the spread in the fraction of grid cells undergoing biome transitions across GVMs, RCPs and the 31 biome maps used for training of random forest models. These numbers are not used for model evaluation, as suitable future data for evaluation are not available.

We will rephrase the corresponding sentence in the abstract to make clear that the 4–56% range reflects the full future ensemble spread across GVMs, RCP scenarios and biome maps, and should not be read as a probabilistic uncertainty range for future biome shifts. The fact that this range is large is itself a key result. It illustrates large uncertainties in future biome projections, here arising from different climate scenarios, GVM structures and biome classifications.

Third, what are your novel results? There are other studies that have projected biome changes under climate change, so what is new? I do read that Equatorial rainforest remains stable while other studies find forest dieback (yet I also know a study that projects increasing area for this biome). But what else? What is your method or result different from published literature?

The most important novel results are: (a) the first directly comparable multi-GVM assessment of biome shift likelihood under RCP scenarios using a unified classification framework, (b) the identification of regions with consistently high biome shift susceptibility across models, including under RCP2.6, and (c) the robust pattern across all GVMs that equatorial rainforests remain relatively stable in terms of biome type, whereas the largest and most consistent biome transitions occur in mid- to high latitudes, (d) that also most of the boreal forest remains stable, i.e., no evidence for two highly debated tipping elements in the Earth climate system. We will highlight these points more clearly in the abstract and discussion and place them in the context of previous, mostly single-model studies. For the tipping element issue, we will also cite more modelling studies that only focused on part of the tropical forest or the boreal forest.

Introduction

It doesn't really become clear why this approach is necessary. Sure, many different models with different outcomes currently exist, but the introduction reads like a summation of what goes wrong instead of a delineation of what these maps are actually used for and how they are created. The observation based biome maps is still an unclear aspect to me, how do you have a global (?right – not mentioned until now) observational map of biomes? Especially since it exists in 31 different versions?

The 31 observation-based biome maps we use were compiled by Fischer et al. (2021); our study builds on this dataset rather than aiming to re-derive or re-evaluate these maps in detail. We use these maps to inform the classification of GVM results into biomes. The fact that at least 31 global biome maps exist (and likely more) with different conceptual bases and biome types illustrates the need for an objective framework to analyze and compare biome projections. This is what we develop and apply here. We did not create those 31 maps, and it is not our aim to compare or assess those observation-based biome maps.

In the revised manuscript, we will better explain that global observation-based biome maps are derived from different combinations of floristic information, remote-sensing-based land-cover products, bioclimatic zoning, and expert mapping, and why this has led to multiple parallel global products with differing biome concepts. We will also provide more examples of how biome maps, both observation-based and modelled, are used (e.g. in assessments of ecosystem services, biodiversity patterns, land-use impacts and biome stability) to better motivate why robust projections are needed.

Additionally, the uncertainties mentioned in the introduction around GVMs seep through in the future predictions. Not only the aftermath (going from model results on biomes) is what causes uncertainty, but also various aspects within the GVMs. How do you propose that is handled by your random forest? I don't understand how a more direct link between GVM output and biomes helps reduce most of the uncertainty. Relatively speaking, how much of the uncertainty in GVM results comes from the final aggregation of data into biomes?

We agree that uncertainties from GVM structure and forcing propagate into future projections, and we see this as additional motivation for an objective framework that allows us to compare biome shifts across a range of GVMs and biome schemes. Our approach does not reduce the underlying GVM or climate-forcing uncertainty; rather, it standardizes the classification step so that these uncertainties become more transparent and comparable. The results can help guide model development, for example by highlighting biomes and regions where agreement with observation-based maps is particularly low.

In the revised version, we will explicitly distinguish three main uncertainty sources: (i) GVM process representation and parameterization, (ii) climate forcing and scenario uncertainty, and (iii) the biome-classification step. Our random forest approach primarily addresses (iii) by eliminating subjective, model-specific biome classification rules and applying a consistent, observation-informed scheme across models. While we cannot robustly quantify the exact fraction of total uncertainty attributable to each component, our ensemble results show that both GVM structure and the choice of the biome map used to inform the random forest models substantially influence projected biome shifts.

Methods

Is there a way to indicate how LAI and PFT are linked?

Yes. In the revised methods section, we will add a brief subsection explaining that each GVM simulates LAI separately for each PFT and that LAI of a given PFT emerges from its simulated performance under the prevailing environmental conditions. We will describe that we first weight PFT-specific LAI by cover fraction and then rescale to represent only natural vegetation, and that the resulting vector of PFT-specific LAI values is used as predictor input for the random forest classification.

The biome classification was done for each combination of GVM result and biome map. Why not make one model, to enhance comparison? You mention that the aftermath of GVM results into biomes creates much uncertainty, but now you do the same but with a random forest instead of the method GVM people decided to use. What is the difference?

Separate random forest models per biome map are necessary because the 31 maps define incompatible biome types (different numbers and definitions of classes). Constructing a single merged biome scheme would itself require a series of subjective decisions and would likely introduce additional uncertainty. Our goal is instead to enable systematic GVM comparisons within each established biome scheme and, in a second step, to assess the sensitivity of projected shifts across schemes. The key advance over model-specific schemes is that we use identical predictors and algorithms across all GVMs and train them objectively against the observation-based maps, instead of relying on model-specific thresholds. For more detailed analyses, we then select one widely used map (Olson et al.) as a common reference. We will clarify this logic in the revised manuscript.

I wonder how the random forest models are dependent on the prevalence of the different biomes. Are biomes that are more frequent on a global map not also predicted more often, simply because the change that a 'large' biome occurs is bigger? Random forests can only predict 'small' biomes in the case that the predictors there are extremely specific.

This is an important methodological point. For the Olson map we already report class-specific performance metrics (biome-wise κ values and confusion matrices) and discuss how global biome prevalence affects classification performance. In the revised version, we will more clearly state this in the methods/results and systematically assess whether the number of grid cells covered by a biome type is related to the κ value across biomes.

Results

The results presented in Figure S2 are the basis of the rest of the results, right? I am not sure it is best placed in the SI in that case. Also, the fact that some GVMs have such large disagreements, I find it tricky to understand what this means and why all GVMs are all still used for the rest of the analyses.

We agree that Figure S2 is important, as it shows the spatial agreement between observation-based biomes and biomes derived from the random forest models for all 31 maps. We prefer to keep it in the Supplement for two reasons: (i) the associated κ values summarizing this agreement are already presented in Fig. 1 in the main text, which highlights differences among the 31 biome maps, and (ii) the subsequent results focus on projected biome shifts (Fig. S3, aggregated in Fig. 2), which are central to the objectives of our study. Given our primary aim of quantifying uncertainties in projected biome shifts, we see Fig. 2 as the key figure.

We retain all GVMs, including those with lower agreement, because we explicitly aim to document and analyze ensemble spread and uncertainty rather than focusing only on the best-performing model.

Additionally, you mention that biome shifts are mostly present at borders of biomes. That is quite logical, but what I would like to see (perhaps in the discussion you mention it already) is that the tropical regions don't change.

We agree that biome shifts at boundaries are expected, and we will emphasize more clearly in the text and figure captions that the grey areas in Figs. 2 and 5 illustrate the relative stability of tropical and boreal regions in terms of biome type. As the reviewer indicates above, large-scale biome tipping would also imply biome changes in the core areas of a biome.

Great impact of climate change in e.g. the Amazon has been reported, but this is not reflected. How do GVMs deal with novel climate combinations? Is there an upper limit in climate suitability for the warmest biomes or lower limit for the driest biomes? Or are all warmer or drier conditions simply outputted as the biome with the warmest or driest climate? This links to the statement in lines 322-325 in the discussion.

The GVMs indeed include implicit or explicit upper limits for climate suitability. For example, in many models NPP decreases at high temperatures due to the temperature response of photosynthesis and respiration, which constrains the thermal niche of vegetation. Similarly, limited water availability reduces growth and can trigger drought-induced mortality. However, representation of drought impacts and mortality remains a challenge in GVM development, and potential heat- or drought-related decreases in productivity are often partially compensated by CO_2 fertilization, which can contribute to tropical forest stability in the models. How novel climates influence biomes in our analysis is determined by the implementation of ecophysiological processes in the GVMs; biome types themselves are not a model output but are derived in a post-processing step via our classification. It is generally assumed that, within limits, process-based GVMs can be meaningfully extrapolated to novel climates because they are built on mechanistic formulations rather than on empirical and static climate envelopes. Note, that these models are also often applied to climates different from today, such as the Last Glacial Maximum or the Pliocene to Miocene warm period. They are designed to be applicalble to novel climates, albeit, of course, with larger uncertainties than when applying them under today's climate, for which we also have most data to test the models. We will add these points in the revised version and explicitly link them to the Discussion passage mentioned by the reviewer.

In line 219, you mention susceptibility category and link this to consensus between models. I don't understand that. Biome change and model agreement is something else, and the link you refer to is not made specific.

Here, "susceptibility" refers to the number (or fraction) of model–biome-map combinations (GVMs × 31 biome maps) that project a biome change in each grid cell, i.e., an ensemble change frequency. Higher susceptibility thus means that more combinations project a change, and lower susceptibility means that more combinations project stability. This is distinct from data–model agreement for current conditions shown eg in Fig 1. We will clarify this definition explicitly in the manuscript and avoid ambiguous wording.

I am wondering about the comparison with the Olsen map you make in section 3.3. I know that somewhere in the methods you describe that nothing changes in biome structure between early 2000 and 2020, but it is strange that you look into the future until 75 years from now (2099) and ignore the first 20 years. Coming back to section 3.3, you describe that the olson map is reproduced with high agreement. But that map was made before 2001, so should the GVM result from the current climate really match 'such an old map' that well?

We did not intend to imply that nothing changed in biome structure between the early 2000s and 2020. In Sect. 2.2 we state: "We ignored potential biome shifts after the publication of biome maps and compared all maps to the model results for current conditions (2006–2020)." As noted there, the biome maps in Fischer et al. (2021) were published between 1964 and 2020, and a map published in, for example, 2001 is based on data from earlier years. Thus, none of the maps is perfectly matched to 2006–2020 conditions. To maintain consistency, we chose a single reference period (2006–2020) for all GVMs and used this for training and evaluation against all 31 maps, rather than using different temporal subsets for different maps. We agree that this introduces a temporal mismatch and we will clarify this more explicitly in the revised manuscript and discuss it as a limitation. Our interpretation of the high agreement with the Olson map is therefore that GVMs reproduce the broad biogeographic patterns encoded in that map, not that they replicate the exact present-day vegetation.

I do not understand the difference between figure 2 and 5.

The analysis proceeds in two steps. First, we build random forest models for every combination of the five GVMs and the 31 observation-based biome maps and apply them to the three RCP scenarios. Figure 2 summarizes these full ensemble results by showing, for each grid cell and RCP, how many of these combinations project a biome change. In a second step, we focus on one specific biome map, the widely used Olson et al. map, for a more detailed analysis of transitions between biome types. Figure 5 shows only these Olson-based results. The difference between the figures is therefore that Fig. 2 integrates over all 31 biome maps, whereas Fig. 5 shows detailed results for a single scheme. We will clarify this more clearly in the text and in the figure captions.