This study compares simulation output of the DGVM LPJ-GUESS using forcing data with different spatial resolution (approx.. 25 km² vs. approx.. 2500 km²). In particular, the authors emphasize on a comparison between two focal regions (one region with a high relief energy vs. another relatively flat region) as well as a pan-European simulation. The authors find that particularly in mountain regions (such as the Alps) the higher spatial resolution of the input data results in relatively large (up to almost 50%) differences in key output variables such as NEP, standing carbon mass, and LAI. Moreover, they emphasize on effects associated with coastal regions, where the coarse spatial resolution results in an overestimation of land-area and consequently related output variables, yet almost an order of magnitude lower as the effect reported for mountain regions. Based on this, the authors conclude that the biases introduced by coarse resolution should be taken into consideration when interpreting DGVM output since they rightfully claim this not to be a phenomenon specifically related to LPJ-GUESS.

As such, the study brings up an important aspect of dynamic vegetation modelling and consequently matches the scope of GMD very well. While I generally recommend publication of the study, the manuscript yet has to undergo substantial improvements regarding the overall structure and in particular the presentation of methods and results. In particular, I sometimes found the level of mathematical details overwhelming, whereas some textual parts of the manuscript lack sufficient detail to allow for reproduction of the approach. A general recommendation – in terms of readability – would therefore be to move mathematical deductions to the supplementary and elaborate textual descriptions. On a related note, I strongly recommend to transform the partly heavy tables into visual output (as done for Table 5 and Fig. 3) and present the tables in the supplementary. Finally, I wonder whether the effect of spatial resolution in coastal regions cannot be resolved more efficiently (see my specific comment on section 5.2 below).

We thank the reviewer for the overall positive assessment of our manuscript, and the detailed comments made, which have led to a substantial improvement of our draft. Please, find our replies in green text below. For quotations of the text we use *Italic* font, while the newly introduced amendments are in ***bold Italic***.

In the following, I provide more specific suggestions on how to improve the manuscript. Once these issues have been resolved, the manuscript in my opinion is acceptable for publication. Please note, that since the line numbers are not continuous (only every 5th line is indicated) I mostly based my comments on section numbers and not line numbers.

Section 1:

The introduction is relatively short and would benefit from elaborating in depth, why higher resolution climate input is required to more accurately simulate ecosystems. For instance, examples on topographic effects on temperature and precipitation can be mentioned, as well as their consequences for simulating impacts of extreme events such as late-spring frosts and droughts.

Also, some relevant studies which have previously used high-resolution climate-data input for

DGVMs deserve a mention. For instance, (Meyer et al., 2024) used a 250 m x 250 m spatially resolved thin-plate spline interpolation for single-point simulations as well as a downscaled 5 km x 5 km set of forcing data for spatial simulations to better resolve the impact of late-spring frost which represents a phenomenon that requires high-resolution forcing data to account for small-scale variations in micro-climate as discussed in Meyer et al. (2024). Additionally, the work by (Levin, 1992) and (Müller and Lucht, 2007) deserve a brief mention in the introduction and a discussion when interpreting the results. Müler and Lucht (2007) do not simulate at an as high spatial resolution as you do here, but they discuss the impacts of spatial resolution on simulation output, which is the main point of your paper.

Levin, S.A., 1992. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. Ecology 73, 1943–1967. https://doi.org/10.2307/1941447

Meyer, B.F., Buras, A., Gregor, K., Layritz, L.S., Principe, A., Kreyling, J., Rammig, A., Zang, C.S., 2024. Frost matters: incorporating late-spring frost into a dynamic vegetation model regulates regional productivity dynamics in European beech forests. Biogeosciences 21, 1355–1370. https://doi.org/10.5194/bg-21-1355-2024

Müller, C., Lucht, W., 2007. Robustness of terrestrial carbon and water cycle simulations against variations in spatial resolution. Journal of Geophysical Research: Atmospheres 112. https://doi.org/10.1029/2006JD007875

Based on such an elaboration you may want to consider to present specific questions/hypotheses that your work addresses, e.g. that higher resolved climatic input allows for more precisely mapping spatial heterogeneity of key model output variables in mountainous/coastal regions. Thereby, readers would already get a better glimpse of the topics the paper actually touches.

Both the abstract and the introduction were elaborated. In particular, an overview of latest downscaling methods, an explanation of the physics behind CHELSA algorithm and a short summary of our findings were included. The abstract was augmented by adding the following text:

Distinctive features of this algorithm include orographic nature of formation of precipitation, a negative derivative of temperatures with respect to elevation, and also, detailed consideration of shadowing and exposure of the terrain to the Sun in computations of solar radiation. We design a custom experiment protocol and use it to perform LPJ-GUESS simulations on both resolutions. *Comparative analysis reveals significant systematic discrepancies between the two resolutions.* ***In mountainous areas, all of the considered output variables show statistically significant differences. In particular, carbon pools are smaller on the high resolution, with the total carbon pool being 37–39% smaller. Furthermore, we quantify the extent to which the under-representation of orographic climate variation affects regional predictions across the European Union. This is expressed as a difference in the total value, which ranges from -3.8% for the net ecosystem productivity to 2.9% for the litter and soil C pools. These values are found to be comparable to differences caused by miss-representation of water bodies and shorelines on the low resolution.***

We thank Reviewer 1 for bringing the additional studies to our attention. In the revised manuscript, we refer to Meyer et al., 2024 in the introduction, which provides a neat additional example for an LPJ-GUESS application that requires high-resolution data:

For instance, using the dynamic global vegetation model LPJ-GUESS, Lagergren et al. (2024) explored how climate change and CO2 impacts of different vegetation types in Fennoscandia would affect habitats of rare and threatened species and also how reindeer grazing (an important source of income for the local population) would be affected. ***Another study based on LPJ-GUESS simulated the negative impacts of late-spring frosts on forest productivity, yielding a decline of NPP in frost years of around 50% compared to non-frost years (Meyer et al., 2024).***

Müller and Lucht demonstrated little impact on model results when running the DGVM LPJ between 10 and 0.5 degrees, indicating that the latter resolution is still too coarse to account for relevant effects of spatial heterogeneity. We include these points in beginning of the revised discussion (Sect. 6) as follows:

Earlier work by Müller and Lucht (2007) showed little impact on model results when running the LPJ DGVM between 10◦ and 0.5◦, at 0.5◦ intervals, suggesting that a resolution of 0.5◦ is still too coarse to account for relevant effects of spatial heterogeneity. Our study suggests that the impacts of resolution on the modeled output, linked to the influence of orography on the input climate, become noticeable at higher resolutions.

Section 2.1:

It is not clear whether this section describes a data source or an algorithm to process data (reading on, I understood it's the latter). Please refine the section to make this clear. Recall, that CHELSA typically refers to a ready-to-use downscaled climate grid and most readers will likely initially interpret it as a data-set (as did I).

In line 45 there is an odd (3) behind the spatial resolution. I assume this is a LaTex typo.

We have changed the title of the section to "CHELSA downscaling algorithm", so that it is clearer what exactly it describes. We note that the text of the Section unambiguously discusses the algorithm only. Specifically, the section starts with the following sentence:

CHELSA (Karger et al., 2017, 2021, 2023) is a family of semi-mechanistic algorithms designed to perform spatial downscaling of near-surface climate data.

At the same time, the section never mentions CHELSA data, which appears later in the text in Section 3.

As to (3), it is a common convention for denoting periodical decimals after the coma. For instance, 1/3 = 0.3(3).

Section 2.1.1:

The adiabatic lapse rate depends on the moisture content, with more humid air featuring a lower lapse rate compared to dry air (roughly 0.65K/100m vs. 1K/100m). From the description, it seems you did not take this into consideration but simply used elevation and pressure to derive lapse rates. I wonder how much error is introduced by this approach and I propose to at least mention the applied lapse rate (dry vs. moist) and discuss the potential implications of this or ideally - if feasible - resolve it. But I understand that this might be too labor intensive, so possibly a thorough description and discussion is sufficient at this point. In any case, since this effect is larger in mountainous regions, i.e. where you reported the largest effect of topography, it deserves a critical discussion and suggestions for solutions in future work.

The applied lapse rate is not a constant "dry" or "moist" value. Instead, it is "empirically" calculated by CHELSA for each gridcell from the 3D information of the CMIP6 model. Specifically, the algorithm uses the difference in temperature values between atmospheric pressure levels at 850 and 950hPa to derive a daily average lapse rate, which is then applied to the surface-interpolated temperature data, as described in Karger et al. (2023). For the details of CHELSA V2.1 and its parametrization, we refer our readers to the original study in Karger et al. (2023).

Section 2.1.2:

The downscaling of precipitation is not reproducible. For instance, I wonder whether CMIP6 wind data is used to derive the wind effect index or whether this is a purely topographic measure. I guess the former, since otherwise luv and lee - which depend on wind direction - cannot be identified. So, this certainly needs to be better elaborated. Ideally, you add equations as for the previous section from which the actual data processing and input variables can be reproduced and refine the textual description of the processing.

We understand where confusion arises. To tackle it, we have elaborated the beginning of Section 2.1 together with Section 2.1.2 to make them clearer.

We note that CHELSA algorithm we employ in this study is fully reproducible. We provide links to the original CHELSA articles where it was featured, as well as a link to the actual software implementation we use. To highlight these references, we added the following text in the beginning of Section 2.1:

For this study, we choose CHELSA V2.1 presented in Karger et al. (2023) and its original software implementation (Karger, 2022), that *scales ISIMIP3b temperature, precipitation, and downwelling shortwave radiation from an input resolution of 0.5◦ down to 0.0083(3)◦.*

In Section 2.1.2 we write that the precipitation algorithm is fully described in Karger et al. (2023 and 2021). We also note, that since we use an algorithm that has been thoroughly described in separate scientific articles, there is no need to repeat exactly the same description in our article. Instead, we provide a brief explanation of how it works, and what physics it captures. A reader, interested in more details, can follow the links provided in our manuscript. Nevertheless, in order to give more insight into how downscaling of precipitation works, we include formulas for computation of index H together with a textual description (see Section 2.1.2 in the supplement to this reply).

Please note, that it is not recommended to use the same variable nomenclature for different variables. In section 2.1.1 ‘H’ refers to elevation, here ‘H’ refers to the wind effect index. Please revise.

This was fixed.

Section 2.1.3:

I do not fully get whether slope aspect and inclination are considered in the downscaling of rsds. Since this can make quite a difference in mountainous regions - which is a focal aspect of the paper - it should to the least be discussed and ideally implemented. But from the description on the 'adjustment according to the surrounding topography' it is not clear whether slope and aspect are included, too. It rather reads as taking into consideration shadow effects but not slope aspect and inclination.

This Section was enlarged, and now includes a detailed explanation of how rsds downscaling works. See Section 2.1.3 in the appendix to this reply. In short, the downscaling procedure takes into account shadowing and obstruction of light, the position of the Sun, the slope and the aspect of the terrain, and cloud cover resulting from orographic precipitation formation. An interested reader can follow the link to the original CHELSA article Karger et al. (2023) in order to learn fine details of the algorithm.

General question: what spatial resolution does the underlying soil information have? Was this adjusted to match the spatial resolution of the forcing data? If not, this might explain some weird patterns observable in Fig. 4 (see my specific comment below).

Please include the relevant response already here.

The soil data was derived from the Digitized Soil Map of the World (Zobler, 1986; FAO, 1991), following Sitch et al. (2003). The underlying resolution is 0.5°x0.5°, like the climate used to feed the low resolution simulations. For the high-resolution simulations we used the same soil information at low-resolution to avoid introducing a confounding factor in the experiments. The same applies to the nitrogen deposition data used to force the European experiment (Tian et al. 2018). We expanded the text to clarify these points:

(L164) The low-resolution simulations were forced with ISIMIP3b climate, while the high-resolution simulations were forced with the downscaled dataset. Both simulations use the same soil properties dataset, derived from the Digitized Soil Map of the World (Zobler, 1986; FAO, 1991), as in Sitch et al. (2003). In order to prevent introducing possible confounding factors, the soil information was not downscaled, and we kept nitrogen deposition at a constant pre-industrial rate of 2 kgN ha-1 year-1.

And for the European experiment:

(L242) The input to the model is as in the ensemble experiment, except now we use historical ISIMIP nitrogen deposition data (Tian et al. 2018). Both simulations were fed with the original 0.5°x0.5° data.

Please, see also our response to the comment regarding Fig. 4.

Section 2.2:

This section lacks a clear rationale/message. The level of detail to which bootstrapping is explained is comparably high (and I wonder whether bootstrapping – which is a commonly applied procedure really needs that level of detail in the main text) but the purpose for running a bootstrapped hypothesis test is not clear. What is the main aim of bootstrapping and which data are used? Is this to show agreement or disagreement between the data from different spatial resolutions? This does not become clear the way it currently is presented.

And I wonder whether a wilcoxon rank-sum test (also known as Mann-Whitney U-test) would not perform equally robust since it has been designed for non-normally distributed data with low sample size.

In the beginning of Section 2.2, we added a few sentences explaining how we use the testing procedure later in our study:

In Sect. 4, we try to find systematic differences between high and low resolutions by comparing the corresponding regional averages of LPJ-GUESS output variables. We do this by testing if the mean values of the samples of the output variables are equal on both resolutions. Since on the 2 resolutions LPJ-GUESS produces outputs with different distribution variance, we are interested in the mean values only instead of the whole distributions.

There are a couple of reasons for including a detailed description of the bootstrap test used in this study. First, bootstrap tests exist in many variants, and it is hard to find a single reference that would be easily readable by non-statisticians. Second, the test is one of the key components in our study protocol. In an analogous study, the downscaling techniques can be changed, but the hypothesis testing procedure may be changed only under very specific circumstances, e.g. if the number of simulations is much higher.

Our task is to test whether the mean values of 2 samples are equal while knowing nothing about the distributions behind the samples. In our case, the distributions of high- and low-resolution samples are always different. For this reason, we need a test of the class of two-sample heterogeneous location tests. Mann-Whitney U-test is designed to test if 2 samples come from the same distribution or that 1 of them is stochastically greater than the other. This test simply cannot answer our question.

Section 2.3.1:

In contrast to the previous sections, this section stands out due to its clarity in describing LPJ-GUESS. I recommend to adopt the style of writing and presentation of methodological details from this section to the previous sections.

We introduced major changes in the manuscript in order to improve clarity.

Section 3:

I wonder why this section deserves its own main header (3). Why not simply adding this to section 2 and term section 2 ‘material and methods’?

Section 2 describes existing methods that we adopted for our study without significant changes. Sections 3, 4 and 5 are our own work. Section 3 in particular describes the preparation of data for our experiment. This is not material that we had before we started the study.

Section 3.1:

I don't understand why you used a different downscaling approach for wind and relative humidity. Wind-speed is spatially quite heterogeneous so a detailed discussion on possibly introduced artifacts is certainly required if using a bilinear interpolation of wind-speed. Ideally, the authors would make suggestions on how to improve the downscaling of wind and relative humidity.

We did not use a different approach for these two variables. CHELSA algorithm uses B-spline interpolation for wind. We also use an interpolation, but in our case it is bi-linear. This is because CHELSA articles never mention the exact parameters for the B-spline. It is not so important because both techniques are from the same class- polynomial interpolation, and there is definitely no loss of heterogeneity since B-splines do not capture those effects in the first place. As for humidity, it is not a part of the CHELSA algorithm V2.1 that we use, so our downscaling method for humidity is not different from it. We added the following to Section 3.1:

The CHELSA original algorithm depends on a B-spline interpolation for wind, while we adopt here bilinear interpolation. Both techniques derive from the same class-polynomial interpolation, and bi-linear interpolation is expected to capture better terrain heterogeneity. Relative humidity is not included in the original CHELSA approach.

Sections 4 and 5:

I understand, that the authors decided to present the methodological approach for each of their two experiments before presenting the experiment outcome. Yet, I wonder whether these methodological aspects should not go into section 2 (to which section 3 is added, see my comment above) and then emphasize on the main findings in section 3 – the results. I personally would find this way of presentation more intuitive than the current version.

Sections 3, 4 and 5 present our original contribution. Experienced researchers in the field might wish to skip to this section and not to read Sections 1 and 2. Furthermore, we do not see the benefit of combining the Sections on methods, data preparation and the experiments into one section.

Section 4.1 – line 168: ‘The latter condition was intended to prevent significant global differences in climate between the two areas’ - This statement does not make sense. The Pannonian basin features a very distinct climate than the Alpine Arc. Yet, I wonder whether this similarity is really required or even possible for your analyses.

It was quite challenging to find a control region that is flat enough and at the same time comparable in size with Alpine region. We have modified the text in the beginning of Section 4.1 to make the choice of our control region more obvious to the reader:

The control region, located between the Dinaric Alps and the Carpathian Mountains, was chosen to contain comparatively little mountainous terrain (Table 2), while being in close proximity to the Alpine region and of approximately the same size. ***The climate between the Pannonian basin and the European alps naturally differs but is still influenced by similar, large-scale circulation patterns that affect the European continent and the choice of the control region intended to prevent significant global differences.***

Very last statement on page 9: Only now it becomes clear why you applied a bootstrapping. As above, I recommend to restructure the methods section to link all of this related information more clearly, possibly in a specific section termed statistical evaluation or alike. And again, I wonder whether Wilcoxon rank-sum test might not also do the job. But this is more a philosophic question.

Regarding the Section on bootstraping, see our replies to the comment “Section 2.2”. As to restructuring, we refer to our answer to notes titled “Section 3” and “Sections 4 and 5”.

Line 190: why not running the whole experiment with these data from the very beginning? Please clarify why two different experiments are needed.

CHELSA is known to produce results that are close to the reality, but nonetheless it reveals a little bias. The latter is at least theoretically possible. We needed to prove that the difference in the Alpine region is due to the better representation of real climate, and not to the presence of bias. The results of the second experiment (Table 6 in the original manuscript) eliminate the influence of the potential bias on the differences, but their outcomes cannot be considered as realistic as those of the previous experiment. This is a simple and widely used control technique in statistical analysis, so we did not introduce additional explanations in the manuscript.

Section 4.2:

The tables presented in this section are difficult to digest and I wonder why tables 6 and 7 are not accompanied by figures as is table 5 with fig. 3. The authors may want to visualize tables 6 and 7 to then move the tables to the supplementary information and focus on the visual interpretation, which still can contain information on test-statistics if significance stars are added.

We chose to present the results of this section as tables as it allowed us to show all the necessary information in one place next to the text describing it. In comparison, the same results would occupy 8 separate images, which we would have to put in the appendix. This would make reading more tedious. But in case a reader needs visualizations, we provide the data in the supplementary materials, that can be used to either reproduce the table or to make the corresponding plots.

Table 5: While the table is quite informative, I personally find it to better fit into the supplementary

information. Instead, I would add significance stars to Fig. 3 to make clear which variables showed a significant effect of the downscaling. In the text, I would also emphasize on the actual fractions observed, i.e. down to approx. -50% for the mountain region and only down to -10% for the Pannonian basin. This provides readers with a better relative impression on how much precision is gained for a given parameter when using finer-grained forcing data.

We placed Tables 5, 6 and 7 next to the text discussing values shown in those tables. This way, a reader would easily switch between the flow of ideas (text) and the source of data (table) and can easily make comparisons, e.g. between variables. Moving tables to the supplement would impede readability. A mere addition of stars to the image would be misleading- the plot depicts delta/mu_hr, while the statistical tests were for delta values, not delta/mu_hr.

Section 5.1:

Line 242: please indicate clearly which domain you're referring to. If you would move section 5.1 to the methods you probably don't have to make this link because you can generally describe your domain and then elaborate on the experiments.

As mentioned above, we find that the current structure of the manuscript benefits the readability and explanatory organization of our paper. However, to be clearer about the domain in question, we added the word "European" (as opposed to Alpine, study or control), and added a reference to the table where the coordinate box is specified. The text now reads:

*In order to assess the impact of systematic biases in low-resolution LPJ-GUESS outputs on a European-regional level, we ran two simulations, at high and low resolutions, in the **European** domain specified in Sect. 3.1 **(Table 1)**.*

Figure 4: I wonder why the authors have chosen to not show fractions of the mean value as in section 4/Fig 3.

We chose to represent absolute change values on the map, rather than relative change values, to give the reader an impression of the magnitude of the figures involved. The tables contain also relative change values to give an idea of how large an effect the downscaled climate has on regional estimations. We feel that giving both values, absolute and relative, is more informative than sticking to only relative change values.

Moreover, it seems there are some weird pixels, e.g. in Norway or Finland, where a clear fingerprint of the LR data can be seen in between high delta values. I recommend the authors inspect these grid-cells to check for potential artifacts. Could this be related to the resolution of the underlying soil information in case this was not spatially downscaled? Did you downscale soil information?

These features are visible because the map in the figure represents the difference between the high resolution and the low resolution simulations, i.e., there *is* a low-resolution signal in the map, which is more visible in regions around the Alps or the Norwegian mountains. However, as pointed out by the reviewer, some of it might be related to the low-resolution input that we still use in the high-resolution simulations, namely soil properties and nitrogen deposition data.

This comment prompted us to review the input data used in the high resolution simulation, and we realized we had made the mistake of downscaling the nitrogen deposition data for the high resolution simulation (this only concerns the European simulation, as the nitrogen deposition is kept constant in the stylized ensemble experiments). As pointed up above, downscaling the nitrogen deposition data introduced a confounding factor. We have therefore repeated the high resolution simulation, this time using the same low-resolution nitrogen deposition data as in the low-resolution simulation. To clarify these points, the text was modified as follows:

(L164) The low-resolution simulations were forced with ISIMIP3b climate, while the high-resolution simulations were forced with the downscaled dataset. ***Both simulations use the same soil properties dataset, derived from the Digitized Soil Map of the World (Zobler, 1986; FAO, 1991), as in Sitch et al. (2003). In order to prevent introducing possible confounding factors, the soil information was not downscaled, and we kept nitrogen deposition at a constant pre-industrial rate of 2 kgN ha^{-1} year^{-1}.***

(L242) ***The input to the model is as in the ensemble experiment, except now we use historical ISIMIP nitrogen deposition data (Tian et al. 2018). Both simulations were fed with the original 0.5°x0.5° data.***

Section 5.2:

Line 260: The climate effect alone is only 2.1%, i.e. much less compared to the topographic effect of mountains.

The "climate" effect referred to in this section is the effect derived from topographical downscaling in the previous section, whereas the "geographical" effect refers to that derived from the poor representation of the shorelines. We see how this choice of nomenclature can be confusing, **so we propose to change the word “geographical bias” with “shoreline representation bias”.**

The climate effect when considering the full European domain is smaller than the value derived for the Alpine region because in the former simulation there are large areas with low elevation variability that keep the overall bias lower in relative terms. We continue this discussion and describe the pertinent changes to the text in the next answer.

Since the geographic effect seems to be dominant (3.4 % vs. 2.1 %), I wonder whether this bias cannot be accounted for by adjusting the values for coastal grid-cells according to actual land-mass. So, in your example of Fig. 5 the output of the northeastern LR-grid-cell could be weighed by a factor of 1-25/64 (25 out of 64 grid cells are water pixels) to better represent the actual land-mass contribution in coastal regions. This might be a more efficient way of treating spatial effects in coastal regions. So, for coastal regions there might be a relatively quick fix to improve simulation accuracy, since the remaining 2.1 % of climate effects probably are within the ballpark of general uncertainty of DGVMs. This aspect deserves more attention in the discussion, i.e. the current section 6 (which I would intuitively see as section 4). For mountain regions I however fully agree, that a spatial downscaling is required to improve accuracy given the comparably stronger effects.

We fully agree with the reviewer's observation that the shoreline bias could be mitigated by simply rescaling the low-resolution model output in those gridcells by the fraction of land area, given as an extra input to the model. However, some gridcells may have both water and high elevation variability, in which case downscaling the climate would be more appropriate. A criterion of whether to downscale a specific gridcell based only on elevation variability, independently of the shorelines, plus a rescaling of the model output on low-resolution shoreline cells by the fraction of land-surface area, as suggested, would completely address this problem.

We also agree that the climate-induced bias in the wider European region is comparatively small. Studies have shown that the spread of climate models used to force DGVMs leads to substantial uncertainty in carbon budget estimations (see citations in the modified text below). The impact in mountainous regions is much higher, and must be accounted for when the region of interest presents high orographical variability.

We have addressed these points by expanding the discussion as follows:

(L270) Earlier work by Müller and Lucht (2007) showed little impact on model results when running the LPJ DGVM between 10◦ and 0.5◦, at 0.5◦ intervals, suggesting that a resolution of 0.5◦ is still too coarse to account for relevant effects of spatial heterogeneity. Our study suggests that the impacts of resolution on the modeled output, linked to the influence of orography on the input climate, become noticeable at higher resolutions. The relative importance of these effects depends strongly on the focus region. Europe-wide simulations show an impact of resolution on aggregated ecosystem pools and fluxes of ∼ 3%, likely smaller than the uncertainty derived from the spread in climate forcings by different GCMs (see, e.g., Schaphoff et al., 2006; Morales et al., 2007; Schurgers et al., 2018). By contrast, these differences increase up to ∼ 46% in an Alpine region. Additional bias may result from poor representation of shorelines and small inland water bodies, but this effect could be mitigated by scaling the model output by the land-cover fraction in the affected gridcells. In areas of low variability in surface elevation, the difference between LPJ-GUESS outputs at different resolutions is much smaller and may be safely ignored in calculations involving regional averages of ecosystem variables. For this type of studies, one could optimize the resource requirements of the simulations by using a coarser resolution in areas with low elevation variability.

Additionally, the summary was modified as follows:

(L323) *We studied systematic differences between high-resolution LPJ-GUESS simulations, forced with the new dataset, and low-resolution simulations. We found that low-resolution simulations are systematically biased. Two main sources of bias were identified: (a) bias associated to the non-linear response of the model to orographical climate variability, and (b) bias associated to the poor representation of coastlines and inland water bodies on a coarse grid.* ***While the latter may be mitigated by rescaling the output by the land cover fraction in the affected gridcells, reducing the climate-response bias requires a finer grid resolution.*** *These sources of bias are independent of the downscaling algorithm, and apply to other DGVMs, insofar as their response to climate forcings is non-linear. Climate-response bias can be very large in mountainous areas; low-*

resolution simulations overestimated average predictions between ~ 4% and ~ 45% in an alpine region, as opposed to a mean bias of ~ 1.4% in a nearly-flat control region. Biases as large as in the alpine region were shown to be vanishingly unlikely in the control region. ***On a European scale, climate-response bias led to an overestimation of regional averages of ~ 3%. This suggests that this type of bias is very sensitive to overall changes in elevation, and should be accounted for when the focus region presents high orographical variability.***

Line 267: I do not fully understand why LAI and FPC cannot be quantified in a similar manner. Please elaborate.

We thank the reviewer for pointing this out. This was a mistake on our part. Indeed, LAI and FPC can be separated into climate-input and shoreline-bias contributions. We have added the details of the calculation as an appendix to the manuscript, and attached it to this document as well (please see below)

Section 6:

I personally believe, that the topographic effect is more important than the coastlines based on your results shown above. In the Alps you showed fractions up to 50% deviation from the mean, whereas the effects of coastlines at most were 10.3 % which could partly be resolved by accounting for actual land-mass within the LR grid-cell (see my comment section 5.2 above). This aspect deserves more attention (see also my comment above).

Please refer to our comment above.

Line 276: I don't get the implication of this sentence. Why should it not affect other models? And below you even state that other models should be affected, too. Please clarify.

With this sentence we wanted to highlight that internal processes in LPJ-GUESS are not sensitive to gridcell size, and LPJ-GUESS gridcells are completely independent of each other. This might not be the case for other models. If there is, for example, lateral flow of matter between gridcells, the model processes themselves will be sensitive to the resolution of the grid, and hence the climate effects discussed in this paper will be entangled with those of the lateral information flow. In other words, all models whose processes are non-linear with respect to the climate forcings will be affected through the different, downscaled input as discussed in the manuscript, but those with lateral information flow will be additionally affected through the gridsize dependence on lateral transport processes.

We suggest the following rephrasing to make this point clearer in the manuscript:

We note, however, that gridcells in LPJ-GUESS are independent from each other (there is no lateral information flow) and completely unaware of gridcell size. Hence, resolution only affects LPJ-GUESS simulations through the resolution of the input data, which is not necessarily the case for other models. ***By contrast, other models may include processes, such as lateral matter transport, which are sensitive to the coarseness of the grid. This introduces an additional dependence of the***

output on resolution, on top of the effects related to higher resolution climate forcings discussed in this study.

Instead of ‘growth season’ I would refer to ‘growing season’

We thank the reviewer for the suggestion. We have implemented this change in our revised manuscript.

Line 283: Spatial PFT realization is likely affected, too, beyond productivity and vegetation cover in general. Please include this aspect into your discussion.

In this study we focused on evaluating the likely magnitude of the impact of resolution on aggregated diagnostics. The spatial PFT distribution was consistent between the two simulations, but a full evaluation of species distribution, including a comparison with observations and with results of previous versions of the model, will be the object of future work.

Line 293: See my comment above. It should be possible to weigh the output achieved for coastlines according to the actual land fraction of a coarse grid cell. This does not resolve topographic effects but for coastlines it should do the job. Please discuss.

We refer the reviewer to the related comment above.

Line 300-316: I wonder whether this level of mathematical detail is required for a hypothetical framework which is designed for a future study. It does not really harm to have it, but it distracts from the actual point of the current manuscript and the discussion of its findings. I therefore suggest to simplify this paragraph and omit the theoretical/mathematical framework.

We agree with this point of view, and we have significantly simplified the end of the section by removing the proposed testing protocol and mathematical notation, while leaving only short textual description of the proposed studies. The revised section now reads:

Systematic biases in model outputs may arise as a consequence of differences in forcings other than resolution. For instance, high-resolution simulations might be sensitive to the algorithm used to downscale the forcings. In the context of climate change mitigation, correlations between different climate variables might influence relevant modeled variables (Zscheischler et al., 2019). To give an example of mechanisms responsible for these correlations, we notice that at points where light is obstructed, the temperature is lower than at neighboring points with no obstruction. Analogously, a spot with a significant amount of precipitation would be colder and darker than the same spot without precipitation. Such correlations are not built into univariate methods like CHELSA but can be captured by dynamical or multivariate downscaling methods. These methods are, however, generally more complex, and might require intensive use of computational resources. Therefore, it might be of interest to find systematic differences between simulations forced by the different methods. This could be done with the help of the methodology

presented in Sect. 2.2 and 4. A similar setup could also be employed to investigate systematic differences originating from alternative model configurations. For example, one could assess whether the modeled impacts of two different forest managing strategies on regional carbon sinks are significantly different from each other.

References:

Karger, Dirk Nikolaus, Stefan Lange, Chantal Hari, et al. “CHELSA-W5E5: Daily 1 Km Meteorological Forcing Data for Climate Impact Studies.” Earth System Science Data 15, no. 6 (2023): 2445–64. https://doi.org/10.5194/essd-15-2445-2023.

Morales, Pablo, Thomas Hickler, David P. Rowell, Benjamin Smith, and Martin T. Sykes. “Changes in European Ecosystem Productivity and Carbon Balance Driven by Regional Climate Model Output.” Global Change Biology 13, no. 1 (2007): 108–22. https://doi.org/10.1111/j.1365-2486.2006.01289.x.

Müller, Christoph, and Wolfgang Lucht. “Robustness of Terrestrial Carbon and Water Cycle Simulations against Variations in Spatial Resolution.” Journal of Geophysical Research: Atmospheres 112, no. D6 (2007). https://doi.org/10.1029/2006JD007875.

Schaphoff, Sibyll, Wolfgang Lucht, Dieter Gerten, Stephen Sitch, Wolfgang Cramer, and I. Colin Prentice. “Terrestrial Biosphere Carbon Storage under Alternative Climate Projections.” Climatic Change 74, no. 1 (2006): 97–122. https://doi.org/10.1007/s10584-005-9002-5.

Schurgers, Guy, Anders Ahlström, Almut Arneth, Thomas A. M. Pugh, and Benjamin Smith. “Climate Sensitivity Controls Uncertainty in Future Terrestrial Carbon Sink.” Geophysical Research Letters 45, no. 9 (2018): 4329–36. https://doi.org/10.1029/2018GL077528.

Sitch, S., B. Smith, I. C. Prentice, et al. “Evaluation of Ecosystem Dynamics, Plant Geography and Terrestrial Carbon Cycling in the LPJ Dynamic Global Vegetation Model.” *Global Change Biology* 9, no. 2 (2003): 161–85. https://doi.org/10.1046/j.1365-2486.2003.00569.x.

Smith, B., D. Wårlind, A. Arneth, et al. “Implications of Incorporating N Cycling and N Limitations on Primary Production in an Individual-Based Dynamic Vegetation Model.” Biogeosciences 11, no. 7 (2014): 2027–54. https://doi.org/10.5194/bg-11-2027-2014.

Tian, Hanqin, Jia Yang, Chaoqun Lu, et al. “The Global N2O Model Intercomparison Project.”

Bulletin of the American Meteorological Society. *Bulletin of the American Meteorological Society* 99, no. 6 (2018): 1231–51. https://doi.org/10.1175/BAMS-D-17-0212.1.

Zobler, L. “A World Soil File Grobal Climate Modeling.” *NASA Technical Memorandum* 32, no. 87802 (1986).

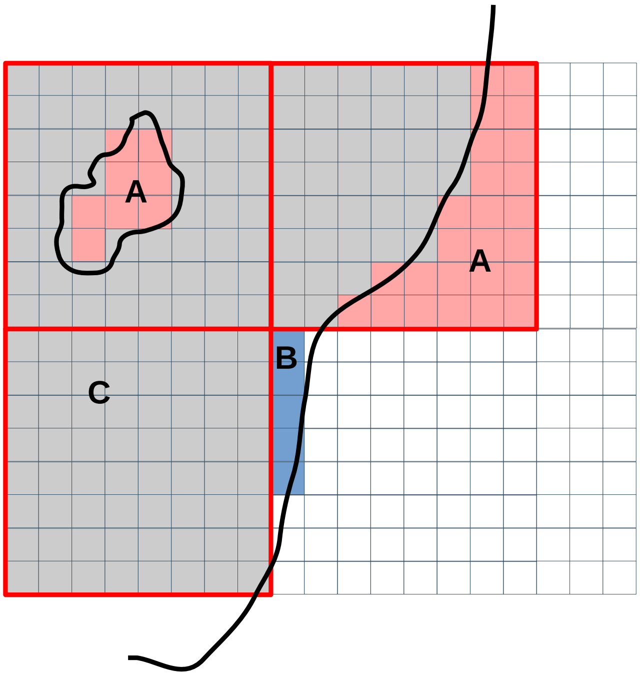

Figure 1: High- and low- resolution gridcells overlayed on the high resolution grid. The low-resolution gridcells are outlined in red. The thick black line represents the shoreline. *Gray:* areas present in the low-res simulation but not in the hi-res simulation. *Blue:* Areas present only in the hi-res simulation. *Red:* Areas present only in the low-res simulation.

1 Bias decomposition

Let X be a modeled variable, S_X the aggregated value of X over the simulated domain, and μ_X the domain-average. In order to calculate the climate-response and shoreline-representation components of the bias, we consider the following quantities, defined in the high resolution grid:

1. X_{ij}^{HR}: Value of the high-resolution output at grid point (i, j).
2. X_{ij}^{LR}: Value of the low-resolution output at grid point (i, j). We note that this value will be the same for all (i, j) within the same low-resolution gridcell (see Fig. 1.
3. A_{ij}: Surface area of the gridcell at gridpoint (i, j)
4. $M_{ij}^{\mathrm{LR,HR}}$: *Overlap mask*. It takes the value 1 at land points where low-resolution values and high-resolution values overlap (gray cells in Fig. 1), and 0 everywhere else.
5. $M_{ij}^{\overline{\mathrm{LR}},\mathrm{HR}}$: Only *high-resolution mask*. It takes the value 1 at land points present in the high-resolution simulation, but not present in the low resolution one (blue cells in Fig. 1) and 0 everywhere else.

6. $M_{ij}^{\mathrm{LR},\overline{\mathrm{HR}}}$: Only *high-resolution mask*. It takes the value 1 at land points present in the low-resolution simulation, but not present in the high resolution one (red cells in Fig. 1) and 0 everywhere else.

1.1 Regionally aggregated quantities

For regionally aggregated variables, such as the carbon fluxes and pools, the bias between high- and low- resolution outputs is:

$$\begin{aligned}\delta &= S_X^{\mathrm{LR}} - S_X^{\mathrm{HR}} \\ &= \sum_{i,j} X_{ij}^{\mathrm{LR}} A_{ij} (M_{ij}^{\mathrm{LR,HR}} + M_{ij}^{\mathrm{LR},\overline{\mathrm{HR}}}) \\ &- \sum_{i,j} X_{ij}^{\mathrm{HR}} A_{ij} (M_{ij}^{\mathrm{LR,HR}} + M_{ij}^{\overline{\mathrm{LR}},\mathrm{HR}}),\end{aligned} \tag{1}$$

where the indices (i, j) cover the whole domain. In this equation, the first sum represents the regional sum of the low resolution values, and the second term is the regional sum of the high-resolution values. Rearranging terms yields:

$$\begin{aligned}\delta = &\underbrace{\sum_{i,j} (X_{ij}^{\mathrm{LR}} - X_{ij}^{\mathrm{HR}}) A_{ij} M_{ij}^{\mathrm{LR,HR}}}_{\delta_{\mathrm{cli}}} \\ &\underbrace{+ \sum_{i,j} A_{ij} (X_{ij}^{\mathrm{LR}} M_{ij}^{\mathrm{LR},\overline{\mathrm{HR}}} - X_{ij}^{\mathrm{HR}} M_{ij}^{\overline{\mathrm{LR}},\mathrm{HR}})}_{\delta_{\mathrm{sho}}} .\end{aligned} \tag{2}$$

The first term of the above equation, labeled as δ_{cli}, involves values of X at overlapping gridcells exclusively (shown as gray cells in Fig. 1). Hence this term can be attributed to the difference in climate forcings between the two simulations. The second term, labeled δ_{sho} involves values of X at non-overlapping gridcells between the high- and low- resolution simulations. These gridcells are the red and blue gridcells from Fig. 1, and are associated with poor shoreline representation at low resolution.

1.2 Regionally averaged quantities

The variables FPC and LAI are averaged across the domain, rather than aggregated. The bias in this case is calculated as:

$$\begin{aligned}\delta &= \mu_X^{\mathrm{LR}} - \mu_X^{\mathrm{HR}} \\ &= \frac{\sum_{i,j} X_{ij}^{\mathrm{LR}} A_{ij} (M_{ij}^{\mathrm{LR,HR}} + M_{ij}^{\mathrm{LR},\overline{\mathrm{HR}}})}{\sum_{i,j} A_{ij} (M_{ij}^{\mathrm{LR,HR}} + M_{ij}^{\mathrm{LR},\overline{\mathrm{HR}}})} \\ &- \frac{\sum_{i,j} X_{ij}^{\mathrm{HR}} A_{ij} (M_{ij}^{\mathrm{LR,HR}} + M_{ij}^{\overline{\mathrm{LR}},\mathrm{HR}})}{\sum_{i,j} A_{ij} (M_{ij}^{\mathrm{LR,HR}} + M_{ij}^{\overline{\mathrm{LR}},\mathrm{HR}})},\end{aligned} \tag{3}$$

where the first term is the low-resolution regional average, and the second term is the high-resolution regional average. Rearranging terms yields

$$\delta = \delta_{\mathrm{cli}} + \delta_{\mathrm{sho}}, \tag{4}$$

where

$$\begin{aligned}\delta_{\mathrm{cli}} = & \frac{\sum_{i,j} X_{ij}^{\mathrm{LR}} A_{ij} M_{ij}^{\mathrm{LR},\mathrm{HR}}}{\sum_{i,j} A_{ij}(M_{ij}^{\mathrm{LR},\mathrm{HR}} + M_{ij}^{\mathrm{LR},\overline{\mathrm{HR}}})} \\ & - \frac{\sum_{i,j} X_{ij}^{\mathrm{HR}} A_{ij} M_{ij}^{\mathrm{LR},\mathrm{HR}}}{\sum_{i,j} A_{ij}(M_{ij}^{\mathrm{LR},\mathrm{HR}} + M_{ij}^{\overline{\mathrm{LR}},\mathrm{HR}})},\end{aligned} \tag{5}$$

and

$$\begin{aligned}\delta_{\mathrm{sho}} = & \frac{\sum_{i,j} X_{ij}^{\mathrm{LR}} A_{ij} M_{ij}^{\mathrm{LR},\overline{\mathrm{HR}}}}{\sum_{i,j} A_{ij}(M_{ij}^{\mathrm{LR},\mathrm{HR}} + M_{ij}^{\mathrm{LR},\overline{\mathrm{HR}}})} \\ & - \frac{\sum_{i,j} X_{ij}^{\mathrm{HR}} A_{ij} M_{ij}^{\overline{\mathrm{LR}},\mathrm{HR}}}{\sum_{i,j} A_{ij}(M_{ij}^{\mathrm{LR},\mathrm{HR}} + M_{ij}^{\overline{\mathrm{LR}},\mathrm{HR}})}.\end{aligned} \tag{6}$$

1 2.1.2 Precipitation

CHELSA considers only orographic precipitation (Karger et al., 2023), which is done by computing the wind effect index H for each high-resolution cell. This index reflects how much moisture gets pushed up towards the top of a mountain as well as rain shadow in its leeward direction, and it is computed using u-wind and v-wind components from CMIP6 data. Those components were interpolated to the high-resolution grid with a B-spline, and then were projected onto a world Mercator projection.

$$H = H_{W,L} \to d_{LH_i} < 0 \times H_{W,L} \to d_{LH_i} \geq 0, \tag{1}$$

$$H_W = \frac{\sum_{i=1}^{n} \frac{1}{d_{WHi}} \tan^{-1}\left(\frac{d_{WZi}}{d_{WHi}^{0.5}}\right)}{\sum_{i=1}^{n} \frac{1}{d_{LHi}}} + \frac{\sum_{i=1}^{n} \frac{1}{d_{LHi}} \tan^{-1}\left(\frac{d_{LZi}}{d_{LHi}^{0.5}}\right)}{\sum_{i=1}^{n} \frac{1}{d_{LHi}}} \tag{2}$$

$$H_L = \frac{\sum_{i=1}^{n} \frac{1}{\ln(d_{WHi})} \tan^{-1}\left(\frac{d_{LZi}}{d_{WHi}^{0.5}}\right)}{\sum_{i=1}^{n} \frac{1}{\ln(d_{LHi})}} \tag{3}$$

, where d_{WHi} and d_{LHi} denote the horizontal distances in windward and leeward direction, while d_{WZi} and d_{LZi} are the corresponding vertical distances. The summations in (2) and (3) are performed within a circle with the radius of 75 kilometers.

The H index is then corrected according to the atmospheric boundary layer height to account for the contribution of the surface pressure level to the wind effect. Lastly, the low-resolution precipitation p_{lr} is multiplied by the corresponding H indices and normalized to obtain high-resolution precipitations p_{hr}, so that within each low-resolution grid cell the sum of the values p_{hr} remains equal to p_{lr} (see section Methods in Karger et al. (2021)).

2 2.1.3 Surface downwelling shortwave radiation (RSDS)

The total shortwave radiation, measured in W/m^2 is represented as in (Karger et al., 2023), Sect. 2.2.2:

$$S_{\mathrm{n}} = S_{\mathrm{s}} + S_{\mathrm{h}}. \tag{4}$$

Here, S_{s} is direct solar radiation reaching the surface, computed according to the position of the Sun with respect to the high-resolution grid cell. Diffuse solar radiation S_{h}, which is the energy re-emitted by the atmosphere, takes into account the percentage of the sky observable from a grid cell.

Computation of S_s component starts with astronomical equations. For the sun elevation angle θ, sun azimuth φ, latitude λ, the solar declination angle δ, the Julian day number J, hour h, and the hour angle in degrees $\bar{\omega}$, we have the following:

$$\begin{aligned}
&\sin\theta = \cos\lambda\cos\delta\cos\bar{\omega} + \sin\lambda\sin\delta \\
&\cos\varphi = \frac{\cos\delta\cos\bar{\omega} - \sin\theta\cos\lambda}{\sin\lambda\cos\theta} \\
&\delta = 23.45 \cdot \sin\left(\frac{360°[284+J]}{365}\right) \\
&\bar{\omega} = 15°(12-h).
\end{aligned} \tag{5}$$

Using these identities, $\cos\gamma$ is computed as

$$\cos\gamma = \cos\beta \cdot \sin\theta + \sin\beta \cdot \cos\theta \cdot \cos(\varphi - \alpha), \tag{6}$$

where γ is the angle between the Sun beam and the normal to the terrain, while α and β are surface slope and aspect. Then, S_s is computed using constants $G_{sc} = 1367\ kW \cdot m^2$, $\tau = 0.8$, and air optical thickness m defined in formula (13) of Karger et al. (2023):

$$S_s(h) = \varsigma(h) \cdot G_{sc} \cdot \tau^m \cdot \cos\gamma. \tag{7}$$

Diffuse solar radiation is calculated as

$$S_h = (0.271 - 0.294\tau^m) G_{sc} \Psi_s, \tag{8}$$

where Ψ_s is the sky view factor computed as

$$\Psi_s = \frac{1}{N} \sum_{i=1}^{N} [\cos\beta\cos\varphi_i + \sin\beta\cos(\Phi_i - \alpha) \cdot (90 - \varphi_i - \sin\varphi_i\cos\varphi_i)] \tag{9}$$

for $N = 8$ azimuth directions Φ_i and the corresponding horizon angles φ_i.

$$rsds = \bar{S}_n(1 - 0.75 \cdot clt^{3.4}), \tag{10}$$

where $\bar{S}_n$ is an average of S_n over 24 hours, and clt is the cloud cover computed according to formulas (19)–(22) of Karger et al. (2023).

To summarize this procedure, we note that the S_{s} and S_{h} components are obtained by taking into account shadowing and obstruction of light, the position of the Sun, the slope and the aspect of the terrain, and cloud cover resulting from orographic precipitation formation.