Dear Referees,

On behalf of all the co-authors I thank you for the insightful and constructive comments directed to the manuscript "Improving the representation of high-latitude vegetation in Dynamic Global Vegetation Models". We have prepared point-by-point responses to each of the comments and believe that further implementation of these in the revision will improve the quality of our manuscript. For convenience and reference, we have numbered the Referee comments with "RC-x.x", where the first "x" corresponds to the referee number and the second "x" to the respective comment. Each of our responses is offered below the respective comment emphasized in blue italics.

Kind Regards,

Peter Horvath

Contents

1 Anonymous Referee #1

Received and published: 30 June 2020

General comments

The overall objective of this paper was to identify biases in a dynamic global vegetation model (DGVM) and, if possible, to find ways of reducing the biases. The analysis focused primarily on relatively undisturbed landscapes in Norway. The target model output was the within-gridcell plant functional type (PFT) distribution. One unique and valuable aspect of the manuscript was that the PFT distributions predicted by the DGVM were compared to multiple products, including field surveys, satellite products, and the output of species distribution models. Field surveys were much more similar to the satellite products and distribution models than to the DGVM. Improvement to the DGVM was realized by incorporation of a precipitation seasonality index, although it was clear that this improvement would not be the end of the story.

Given that PFT distribution is an important quantity that is still challenging for DGVMs to predict, I think that the manuscript covers a topic that will be interesting and useful to readers of Biogeosciences. I also appreciated how the DGVM was compared to multiple products and how the distribution model was leveraged. However, I think that the value of the manuscript could be increased by being more thorough with the methods (see below). Also, I think that more could be done to make the manuscript interesting to readers who use models other than CLM.

We are thankful to the Referee #1 for his/her positive response and constructive comments.

Specific comments

RC-1.1 - The title should be modified. It mentions “Dynamic Global Vegetation Models” in the plural, but only one model is discussed. I also think the title is too general. I would suggestion “high-latitude vegetation distributions” rather than simply “high-latitude vegetation”.

This is a good suggestion. We shall adjust the title to specify that high-latitude vegetation distributions are considered. With regard to the plural mention of DGVMs, we believe that even though we tested this particular exercise only on one DGVM (namely CLM4.5BGCDV), the procedures/methods of implementing variables from DM as new parameters in DGVM can be used in multiple DGVMs not just the tested one (thus the plural form).

RC-1.2 - Lines 83-84: This point is overstated. There are publications that have evaluated PFT distributions from dynamic vegetation models against field-based datasets, at least on regional and national scales.

In line with response to Referee #3 on this same point (see also comment RC-3.6), we will adjust the formulation of the sentence and add a reference.

RC-1.3 - Methods: I am puzzled by the limitation of the study to only 20 plots. Certainly these 20 plots span the range of mean annual temperature and precipitation, but other factors are also commonly perceived to be important. Indeed, the distribution model seems to take 100+ inputs. Some questions that come to mind is whether the plots span the range of observed precipitation seasonality (identified by this study as an important factor!), soil texture, and soil nutrients.

We agree that a higher number of plots would have been beneficial. Ideally, we would want 1000+ plots or perhaps a regional/global simulation. However, labor-demanding preparation of all data layers for each

plot was one of the critical factors for this study and we had to find a compromise between what was practically possible and what was considered robust in terms of the aim of the study. From a methodological perspective, our opinion is clearly that a representative sample of 20 plots is sufficient to demonstrate the differences between the three methods of representing the vegetation distribution.

The gradients of precipitation and temperature are known to be among the most influential for vegetation distribution (e.g., Ahti et al. 1968; Bakkestuen et al. 2008), thus we have chosen to include these particular two variables when selecting the 20 plots. However, we also agree with the Referee #1 in the argument that the 20 plots' representativity across the range of precipitation seasonality should be tested (since this is identified as an important factor). We will therefore include a comparable test and add a third diagram to the Supplementary Figure S3. Please also see the response RC-2.6 to Referee #2 with a similar request.

RC-1.4 - Line 157: Why not assign the observed soil texture to the 20 plots?

The observed data on the 20 plots unfortunately do not include information about soil texture. The plots were mapped using wall-to-wall vegetation mapping, where only data about the type of vegetation cover are available.

RC-1.5 - Section 2.4.3: I am concerned that the DGVM and the DM uses different driver data to represent the same phenomenon. For example, does one use SeNorge2 and the other reanalysis to represent precipitation? Does one use observed soil texture and the other "default" soil texture? If so, might differences in inputs account for differences in the DGVM and DM predictions?

Absolutely. Ideally, we would use the same climate input data for both DM and DGVM. However, there are technical obstacles: DM uses multi-year monthly averaged climate data as input, while DGVM requires 3-hourly meteorological data as the input. SeNorge2 dataset, which is used in DM, has only daily data available, therefore can only be used for DM but not for driving DGVM. For DGVM, we had to use available reanalysis or regional climate model data for present day climate (CORDEX data in this manuscript). To compare the differences between the driving data for DGVM and DM, we have listed mean annual temperature and precipitation for both datasets in the table S1 and figure S5 of the supplement. There are indeed some minor differences between the two sets of driving data, however it is beyond this study to quantify the effect of these differences. We will devote a paragraph to clarify the potential bias this may imply in the discussion.

Soil texture does not come in as an explanatory variable in the DM, whereas DGVM is using soil texture as an important parameter affecting various processes in soil, such as soil temperature, moisture and organic matter decomposition. We will add a comment on the differences between the input data in the paper and discuss its potential implications.

RC-1.6 - Line 183: Was the DM model previously tuned to these 20 plots? To Norway?

The DM was not tuned specifically to these 20 plots. The training data for DM included the whole set of 1081 plots (across Norway) at a different thematic resolution (detailed vegetation types instead of PFTs) and at a scale of one point per polygon. Although the 20 plots were included as a subset of the total 1081 plots, we believe the influence is minimal, since they have gone through a spatial and thematic conversion. Moreover, the DM was evaluated with a completely independent dataset.

RC-1.7 - Line 414: Might phenology also be an issue? Further, what is the light compensation point of the PFTs? Perhaps the authors can use the light compensation point to directly evaluate the relative shade tolerance of the different PFTs.

Please also see comments to Referee #2 (RC-2.14) and Referee #3 (RC-3.26) regarding this paragraph in the discussion. Phenology is likely to be an issue, as evergreen plants seem to have advantage in competing with deciduous plants in general in the high-latitude region in the model. It is therefore suggested that stress for evergreen plants in winter and spring may not be well represented in the model to limit the growth of boreal NET in some regions. However, we admit that this issue is not well documented through our results and therefore have decided to remove this paragraph from the discussion.

RC-1.8 - Discussion: Are there lessons for people who use other models? The more the authors can draw out such lessons, the broader the audience this paper would appeal to. The TEM model, which has a more detailed representation of boreal PFT diversity than CLM, immediately comes to mind as one example.

Thanks for the suggestions. The present-day vegetation distribution outputs from dynamical vegetation models could more often evaluated by use of multiple products complementing the RS, such as by including DM and AR as presented in this study. We also believe that the procedure of identifying new parameter values from DM, running a set of sensitivity tests and implementing the sensible new parameters into a DGVM is not limited to CLM4.5BGCDV (the DGVM tested here) but transferrable also to other DGVMs, such as the TEM model. We will make sure that this is stated more clearly and include more thorough discussions with regard to applicability to other models in the revised manuscript.

Technical corrections

RC-1.9 - The manuscript is very readable, but it should still be reviewed for grammar.

We will carefully search the manuscript for grammatical errors.

RC-1.10 - Page 3, Lines 43-45: There is a problem with word choice in this sentence. Vegetation distributions are not implemented in ESMs, but rather are predicted by ESMs. The ESM predictions can then be evaluated with satellite products (as done in the present analysis).

We will rewrite the sentence according to the referee's comment.

RC-1.11 - Section 2.4.1: It would be useful for the authors to briefly describe how the DGVM determines the amount of area to each PFT.

We will add a brief description on how the area of each PFT (i.e. percentage cover fraction %) is determined by DGVM in the revised manuscript. The percentage cover fraction of each PFT is equal to the average individual's fraction projective cover (FPC_{ind}) multiplied by the number of individuals (N_{ind}) and average individual's crown area ($CROWN_{ind}$). FPC_{ind} is a function of the maximum leaf carbon achieved in a year, while $CROWN_{ind}$ is related to dead stem carbon simulated by the model. N_{ind} is mainly determined by establishment and survival rate controlled by establishment and survival threshold conditions.

RC-1.12 - Data availability: Note that the GitHub link not up yet. I understand if the authors do not want to release the link prior to manuscript acceptance, but it is still important not to forget to release the link.

This is an important point. We shall keep in mind that the scripts are to be made available as soon as the manuscript is accepted.

2 Anonymous Referee #2

Received and published: 19 August 2020

General Comments

This study evaluates estimates of PFT distributions from a DGVM in comparison to those of remote sensing and empirical models, and against a field-based dataset, for 20 plots of high-latitude vegetation types across Norway. The topic investigated, approach taken, and results reported will be of interest to the modeling community. The paper could benefit from more or better explanation of the methods, especially the CLM simulations. For example, it is unclear whether or not this is intended to be any kind of 'temporally-explicit' analysis; this seems a sort of model estimation of some 'average' PFT distribution from the spin-up results that was compared to field plots and remote sensing data, both of which presumably represent a specific point in time (that is not specified in either case in the methods here).

Thank you for this to-the-point comment. We agree that more careful explanation of some aspects of the methods is necessary. We will adjust the manuscript with regard to the specific comments you provide here.

This study represents a temporally explicit analysis of the 'present-day' vegetation distribution. We agree that this needs to be emphasized more clearly. In line with further replies to RC-2.10, the temporal context will be specified for each of the three modelling methods as well as for the AR in the respective sub-chapter (2.4).

RC-2.1 - To properly interpret the results, the sensitivity tests need more explanation and clarification to justify and understand what was done here in this study (vs. previous work).

We will add more explanation and discuss on the sensitivity tests in the revised manuscript. Also, we shall review the formulations of what was done in this study vs previous work.

RC-2.2 - The "RS method" as one of the three methods compared here seems kind of out of place in this analysis since it is not a method for predicting future PFT distributions as with the DGVM and DM methods. What is the reasoning / purpose behind including RS in this comparison? Or could / should it be used in this study more as a 'reference' data set, like the AR data?

We understand the concern of Referee #2 on this point. We also agree that RS is often being used as a verification/reference dataset in land surface modelling. However, the emphasis of this work is on improving the DGVM for the 'present-day', based on the premise that the better DGVM are able to predict the present-day distribution of vegetation (based on the processes/parameters driving the DGVM), the more reliable predictions for the future will the model be able to produce. Moreover, RS is also of interest from the perspective that products derived from RS data may also be burdened with uncertainties, needing evaluation - just as DM and DGVM - against a ground-truth/reference data set, which in this case is AR (see also our response to RC-3.5). We will make this clearer in the revised version of the manuscript.

Specific Comments

RC-2.3 - 25-26. please consider this statement carefully; numerous authors could claim that this is untrue

Thank you for pointing this out. This comment accords with a comment of Referee #3 (RC-3.6) and we will modify this statement in the abstract of the revised manuscript as well as the introduction (lines 83-84) where the amended sentence will be supported by references (e.g., Druel et al 2017)

RC-2.4 - 34. can these three thresholds be named here, or at least hint at what they are (e.g. ". . . based on . . .)?

Yes, we agree that the thresholds should be mentioned here. Also, in line with your other comment (RC-2.15), we will adjust the text to clarify that only precipitation seasonality (bioclim_15) is influential.

RC-2.5 - 115-116. this is not quite clear and perhaps needs to be specified or qualified; i.e. don't many "countries" have national-scale inventory programs?

This will be re-worded. What is meant here is that wall-to-wall vegetation surveys on national scale are rarely made. AR (the reference dataset) is an example of an area-representative survey.

RC-2.6 - 126-131. Selecting only 20 plots seems limited, even if deemed acceptable for bioclimatic variation. There needs to be better explanation / justification for this choice, how "acceptable" was determined, and whether a kriging of temperature and precipitation really captures "bioclimatic" variation across the country.

We agree that a set of 20 plots is a rather limited number. Referee #1 raises the same issue (RC-1.3), and our response (and justification for the choice) is given in comments to Referee #1. We will amend the text to explain our choice better.

The representativeness was tested for and explained in supplements S3 and S4 (see also Fig.S3 and Table S4). By acceptable representativeness we mean that the selection of 20 plots does capture the variation across the whole range of temperature and precipitation (in the revised version we will add also "precipitation seasonality" - Fig.S3 – following comment RC-1.3) compared to the full set of 1081 AR plots. The representativeness of the 20 plots was also tested against the full dataset of 1081 AR plots with regard to PFT coverage, where a Chi-square test showed that the two datasets are much more similar than expected by chance.

We agree that the sentence on line 131 is not clearly formulated. Also, in line with the comment RC-2.1, we will make sure that it is clear what was done in this study vs. previous studies. Kriging was used in a previous study to interpolate the original SeNorge2 dataset from 1km down to 100m for the purpose of distribution modelling (a procedure which was done and described in Horvath et al. 2019). We agree that this information is not relevant for the representativeness comparison, and it is more important to include a specific description of how the representativeness test was done in this study (in addition to the existing description in supplement S3). We will reformulate this paragraph in the revised manuscript accordingly.

RC-2.7 - 150. curious decision to give a new acronym to CLM. why not just refer to it as "CLM"? and actually, you do, somewhat, as it seems to switch back-and-forth between "DGVM" and "CLM4.5" for the rest of the manuscript. I see the idea to associate the results from CLM as representative of the "DGVM" approach, but when describing or referring to the specifics of CLM then just call it "CLM" (or "CLM4.5")

We understand the confusion here. The terms will be further explained. CLM has an option to run will full vegetation dynamics (CLM4.5BGCDV), this option is further referred to as DGVM. The abbreviation of DGVM is used throughout the manuscript to refer to this particular setup of CLM. This will be clarified in the revision, and consistency in the use of terms will be carefully checked.

RC-2.8 - 154. it may be useful here to point out what these simple assumptions are, and how different (or not) they are from those for which the DM method is based on.

We will add more details of the assumptions used in DGVM in describing establishment, survival, mortality and light competitions. Compared to DM which uses statistical relationships (line 180) to predict the probability of VTs/PFTs from environmental variables, DGVM assume a simple environmental threshold for establishment, survival and mortality of a PFT to occur (see supplement S6) This will be motivated for and explained made clearly in the revised version of the manuscript.

RC-2.9 - 171. was soil C initialized somehow, or was it a separate (longer) spin-up? are these mostly undisturbed sites or was that taken into consideration for the vegetation spin-up at each site? was the CORDEX climate used for the spin-up? average or de-trended?

Thanks for pointing this out. In our experiments, soil C and N were firstly initialized using the restart file from an existing global present-day spin-up simulation with prescribed vegetation. Then, they are spun-up together with vegetation for 400 years. All the selected sites are mostly undisturbed. The 30-year CORDEX data were cycled during the spin-up. A 30-year period is consistent with WMO climatological normals based on the rational that 30 year is short enough to avoid large long-term trends while long enough to include the range of variability. Thus, the data is not de-trended or averaged. We noticed that vegetation distribution is insensitive to interannual variation or decadal variation of the climate forcing when it reaches equilibrium state in most of our study sites (see supplement S10). This will be now specified in more detail in this paragraph of the manuscript.

RC-2.10 - 174. what year / era does this RS map represent? Table 2. I don't think all of this detail is necessary in the main text.

A very good point, which should be clarified indeed. The RS product used in this study is created from satellite images covering the period of 1999-2006 (Johansen, 2009). We will make this clear in the manuscript.

We agree that Table 2 might be too detailed for the main text. We will move Table 2 into the supplement.

RC-2.11 - 278, 279 & 305 are confusing uses of sub-headings

We agree that further splitting the chapter 4 (Sensitivity experiments and model improvement) into methods and results might seem untraditional. We suppose that it has not been made clear that the paper falls into two parts: an analysis of data, and a sensitivity analysis which is based upon the results of the analysis. We will add a motivation sentence at the end of the introduction, clearly telling that the sensitivity experiments are a separate chapter, which builds upon the results of the analyses. In that case chapter 4 would remain, but the sub-headings would be removed and instead split into separate paragraphs. (see also reply to RC-3.23)

RC-2.12 - 287. swe_10 and tmin_5 make sense as described but can "precipitation seasonality" be explained? "bioclim_15" is not as obvious as the other two parameters

A very good point. We will include a description and a reference to how "precipitation seasonality" is calculated (O'Donnell & Ignizio, 2012). "Precipitation seasonality" is defined as the ratio of the standard deviation of the monthly total precipitation to the mean monthly total precipitation (also known as the coefficient of variation) and is expressed as a percentage.

RC-2.13 - 293-299. there just seems like so much of the justification and explanation of decisions and approaches for the sensitivity test are glossed over here. For example, why are these particular

parameters chosen, how was bioclim calculated, is the stepwise order important, what does it mean "three PFTs at the same time", how were the thresholds determined, etc etc. Perhaps a little more explanation than just "see Horvath et al 2019" (line 286) would be helpful.

We agree with the Referee #2. Since a lot of the sensitivity experiments is based on the results from the previous study by Horvath et al. 2019, referring to this article is necessary. However, we agree that explicitly describing the sensitivity experiments is important. We will add more detailed explanation on the reasoning behind the set-up of the sensitivity experiments, including the specific topics that Referee #2 is pointing to in this comment.

RC-2.14 - 414-415. this seems like a bit of a leap without a more direct connection to the results of this study.

We agree that the arguments in this paragraph are not supported by the results of this study. In line with the comments from Referee #3 (RC-3.26) and request from Referee #1 (RC-1.7) we decided to remove this argumentation from the revised version of the manuscript.

RC-2.15 - 468. but in line 312 it was stated that two of those three "had little effect"

Yes, this must be a remnant of a previous formulation. We will remove the two parameters that did not improve the DGVM performance from this sentence. We will also amend the abstract with regard to this (see also reply to a comment for RC-2.4 and RC-2.17)

RC-2.16 - 498-499. when are high-quality RS products ever not available anymore in this day-and-age?

We agree that this needs to be reformulated to explain the challenges clearly. It is not the "high-quality" of RS products in terms of resolution or coverage that we are concerned about, but rather in terms of being able to supply proxies of other properties (such as deriving parameter improvements, traits or in some cases vegetation distribution in high enough thematic resolution). In particular, at high latitudes low sun-angle results in large shadow effects. Furthermore, our results show that analyses of high spatial resolution RS images have limitations when it comes to thematic precision and resolution. We will reformulate this sentence.

RC-2.17 - 503. Just to be clear, it seems that these parameters were identified in a previous study, not this one, correct? And actually in this study only one of them (bioclim_15) was found to be useful, no? This same claim is made in the abstract, as well, and should be used with care.

Yes, we agree, and we will carefully re-formulate the sentences with this regard both in the conclusion and abstract. Please see also related comment RC-2.4 and RC-2.15.

Technical Corrections

RC-2.18 - - please review the grammar, wording and sentence structure throughout

All the technical and wording amendments suggested below will be implemented in the revised version of the manuscript and the text will be carefully searched for erroneous grammar.

42. please re-word and fix the grammar of this sentence one way or the other

55. remove "the" before DGVMs

60. latitudes

150. replace “further” with “hereafter”

170. “recalculated”

Table 2. “AR” is missing from the caption

292. change “NEB” to “NET”, I think

341. “spectre” should be “spectrum”?

412. “overprediction of Boreal NET”?

3 Anonymous Referee #3

Received and published: 25 August 2020

The manuscript "Improving the representation of high-latitude vegetation in Dynamic Global Vegetation Models" by Horvath et al analyses the performance of three different vegetation modeling approaches with regard to the spatial distribution and relative abundance of plant functional types (PFT) in Norway. The modeling approaches include a dynamic global vegetation model (DGVM), remote sensing (RM), and a statistical distribution model (DM), which relates occurrences of vegetation types to multiple environmental variables. The authors found that both RM and DM showed a better performance than the DGVM when compared to observational data from a range of field sites. They then tested if it was possible to use the DM to improve the predictions of the DGVM with regard to PFT composition and distribution. It was found that, through inclusion of three further bioclimatic constraints based on the analysis of the DM, the performance of the DGVM could be improved. The authors recommend DM as a complementary tool for the assessment and improvement of DGVMs.

RC-3.1 - The manuscript is well written and easy to understand in general. The research topic (assessing and improving DGVMs at high latitudes) is certainly relevant, and the chosen approach is original and seems useful to me. However, the description of the methods needs to be improved, with regard to the chosen statistical approaches, and also the motivation to carry out certain analyses. It often becomes clear only later in the manuscript why a certain method was applied. I therefore recommend minor revisions before a new version of the manuscript may be submitted.

We thank Referee#3 for a set of thorough comments. We will improve the sections of the manuscript in line with these comments.

Comments:

RC-3.2 - L 28 While the term ’DGVM’ is explained at the beginning of the abstract, the term ’distribution model (DM)’ is used in this sentence without previous explanation. Please explain shortly in the abstract what a DM is and how it differs from a DGVM, since some readers may not be familiar with the concept.

Good point. We will add a sentence about the difference between process based (DGVM) and correlative (DM) models.

RC-3.3 - L 58 Please define or explain in more detail what you mean by 'thematic resolution'. Furthermore, it should be mentioned that recently, specific high-latitude PFTs, such as mosses, for instance, have been added to a number of DGVMs, e.g. Jules (Chadburn et al, 2015, The Cryosphere), JSBACH (Porada et al 2016, The Cryosphere), or ORCHIDEE (Druel et al 2017, Geoscientific Model Development) and several more.

The term thematic resolution is meant to refer to number of classes (ex. PFTs) in a model. This will be explained in the revised version of the manuscript. Thank you for pointing to these references, we will consider including them as examples in this paragraph in the revised version of our manuscript.

RC-3.4 - L 60 Three examples are given for the difficulties of DGVMs to simulate extents of high-latitude PFTs correctly. However, I do not see how the underestimation of forest carbon storage by DGVMs relates to this, since this is rather a consequence, and not a reason for the incorrectly predicted extent. Please explain in more detail.

Good point. The sentence about carbon storage underestimation will be reformulated so that it will be clear that discrepancies in the DGVM have implications on different systems (e.g. carbon storage),

RC-3.5 - L 71 Please add a short statement to describe in which regard the RS products are not consistent.

The study by Myers-Smith et al. (2011) reports a mismatch in the spatial resolution between satellite observations and the spatial heterogeneity of vegetation patches in tundra ecosystems. This will be clarified in the introduction. Also, different satellite products produce varying results with regard to vegetation classification (Majasalmi, T. et al. 2018). We will shortly describe these inconsistencies in the manuscript (please, also see RC-2.2).

RC-3.6 - L 83 At least one study (Druel et al 2017, Geoscientific Model Development), uses site data to assess the DGVM's performance with regard to plant traits. Please be more specific in this regard, and explain what exactly is new in the validation method.

Yes, we will reformulate this sentence and make clear that our study focuses on evaluation of vegetation distributions between different models/methods. Also, we will mention the study by Druel et al. (2017) as an example of evaluation with field data.

RC-3.7 - L 121 I do not understand this sentence: If one plot is 0.9 km2 large, then 1081 plots are around 1000 km2, but 18x18 km are only 324 km2. Also, the plots are distributed throughout Norway, so the 18x18 km area has to mean something else. Is it the distance between the plots on a grid which covers Norway? Please explain.

Thank you for pointing this out! For us who have been working with these data for so long time, it is easy to forget that it is not obvious how they are structured! There is a regular grid covering the whole land area of Norway on which the plots (in total 1081 plots), each with a size of 0.9km2, is placed every 18 km (in latitude) by 18 km (in longitude). This will now be explained in more details in the revised paper.

RC-3.8 - L 129 To me it seems that low values of temperature and precipitation are underrepresented in the 20 selected plots compared to the full data set. This should be mentioned here briefly and then considered later in the Discussion section.

We agree that there is a slight underrepresentation in the frequency of plots with the lower values for temperature and precipitation. However, the most important factor was to include plots covering the range of the temperature and precipitation values experienced, which we have succeeded in (Fig S3). We will add a brief description in section 2.3 "Study plots" and in the Discussion.

RC-3.9 - L 156ff By using the default surface parameter values for CLM, the DGVM may miss some relevant information to correctly predict PFT distribution, compared to RS and DM. Furthermore, by using climate forcing from 1980-2010 and running the DGVM into a steady state with regard to this period, historical climatic effects, which may influence today's PFT distribution are not considered. These points should be mentioned in the Discussion section of the manuscript.

We understand the concern of the Referee #3 regarding this aspect. In line with replies to the RC-1.5 we will add more detailed discussion on the issues raised in this comment. As to the concern on the usage of the climate forcing data, we indeed overlooked the historical climate effects on vegetation distribution, which response usually lag several years or decades behind climate changes. However, this is considered to have minor impacts on the large biases observed in DGVM (e.g., too much boreal NET and too few shrubs), as historical climate effect (such as cooler temperature in the past) might actually favor more boreal shrub than boreal NET (please, also see our reasoning to comment RC-2.9). We will clarify this in the Discussion.

RC-3.10 - L 162 Why was the CORDEX data not also used for the DM method? This should be briefly mentioned here.

In a previous study (Horvath et al. 2019) the authors have created distribution models for vegetation types with a range of predictors (including SeNorgre2 data), where the statistically important predictors were selected in the forward selection procedure. At that point the SeNorge2 was the most reliable climate dataset available for the whole study area. It will be described further in the section 2.4.3. We will add a comment on the choice of climate data sets, including the choice of the CORDEX, respective the seNorge2 dataset.

RC-3.11 - L 175 Please explain 'supervised' and 'unsupervised' in more detail.

While with the supervised classification, training data is based on well labeled data from part of the study area, during the unsupervised classification algorithm is only supplied with the number of output classes. 'Supervised' and 'unsupervised' classification methods will be explained in more detail in the revised version of the manuscript.

RC-3.12 - L 182 the number of explanatory variables (116) is rather high. It should be shortly explained what these are, and why such a large number is necessary for the regression. Even if this information is provided in Horvath et al (2019), it should be summarized here.

A short description of the explanatory variables (grouped into categories) will be provided in this section of the revised manuscript. Also, a sentence about forward variable selection procedure will be added, to make clear that only a few of the 116 variables were actually included in each final DM.

RC-3.13 - L 183 It would be good to add a short summary of the evaluation method for the DM here, so the reader can assess the DM better.

A short summary of the evaluation procedure will be added. Evaluation of each model was carried out using an independent evaluation data set and by calculating the area under the receiver operator curve (AUC), a threshold-independent measure of model performance commonly used in Distribution modelling. AUC can be interpreted as the probability that the model predicts a higher suitability value for a random presence grid cell than for a random absence grid cell (Fielding & Bell, 1997).

RC-3.14 - L 186 I wonder if, by discarding all other VT except the most probable one, biases in the distribution of the VTs are introduced. Let us assume the logistic regression predicts a certain VT always with a slightly higher probability than a second one; according to the description, only the first VT would occur in the predicted map at all pixels, and all observations of the second one would be discarded, although this VT occurs quite frequently in reality. Please explain this in more detail.

This is an interesting and intriguing topic. As the Referee #3 rightfully points out, there is a possibility of slight biases in certain regions, for the reason outlined. However, as far as we are aware, this has not yet been closely investigated. We are preparing a manuscript covering this topic in more detail - The results so far suggest that the approach for compiling the wall-to-wall map from 31 DMs, which we also use here, is performing the best out of the tested approaches (Horvath et al., manuscript in prep.). Additionally, as the probability of presence for each VT is predicted separately for each grid-cell, the probability values for every VT varies independently of the probabilities for the other VTs, throughout the study area. Thus, we regard the chance that one VT consistently outperforms another VT over all the grid cells to be negligible.

RC-3.15 - L 200 I don't understand why an aggregated PFT profile is needed, I thought that the comparison of the 3 modeling approaches and the AR data is done for each of the 20 plots?

Indeed, the main comparison is between the 3 modelling approaches and AR on each of the 20 plots (this can be found in figure 2 and 3). But besides, it was also worth investigating the overall performance of the tree methods across the study area. In order to do that, we needed the aggregated PFT profiles!

RC-3.16 - L 208ff This sounds like one comparison was done with the aggregated profiles (one for each method, aggregated over all 20 plots), using the chi-square test. Then, for each of the 20 plots the profiles were compared regarding their dissimilarity. It is not clear to me, why two different statistical methods were used to compare the models (DM, RS, DGVM) to AR.

We will clarify this in the revised version of the manuscript. The point here is that we wanted to compare the three models (DM, RS, DGVM) to AR both with respect to the overall pattern (represented by the aggregated profiles) and with respect to their performance on each plot; the latter in order to identify the circumstances under which some of the models deviated strongly from the reference. Accordingly, the chi-square test was used to formally test if the models overall deviated from the reference, while the proportional dissimilarity index (which does not come with a statistical test) was calculated to address the purpose of identifying strongly deviating modelling results at plot scale.

RC-3.17 - L 222 I thought the dissimilarity index was used to assess the similarity between the 3 modeling approaches and the AR data. Why is it then necessary to do a pairwise Wilcoxon-Mann-Whitney test in addition? Please explain the reasons for the chosen statistical approach in a more detailed way.

Our statistical analyses serve several purposes of which one is to assess the goodness-of-fit of the modeling results to the reference (I.e., to assess their performance); another (which is addressed by the Wilcoxon-

Mann-Whitney tests) is to assess the degree to which the models produce pairwise similar differences. We will explain this in the paragraph.

RC-3.18 - L 230ff As mentioned above (L200), by aggregating the PFT profiles of the 20 plots, differences in profiles between plots are lost. Hence, it is not possible to evaluate the 3 models with respect to the correct prediction of differences in profiles between individual plots. Also, while the AR data (for each plot) can be interpreted as a random sample, it is not clear to me how the model approaches can be consistently included in this Chi-square test. Moreover, the number of elements (6 PFTs) is actually too small for a Chi-square test. The authors need to justify this better, or change their testing approach.

The mere purpose of analyzing the aggregated profiles is to assess the models' ability to produce overall predictions of PFTs that accord with the PFTs' overall frequency (as given by the reference). We do not see any reason why the chi-square test should not be useful for a contingency table of 6 classes.

RC-3.19 - L 249 If I understand Fig. 2 correctly, the lines which connect the dots denote the individual plots, which means that for one method (e.g. DGVM), the dissimilarity can be high (1.0), while for another method (e.g. RS) it can be much lower. The result that the goodness of the fit between a given method and AR data depends on the set of chosen plots may point to some underlying systematic deficiencies of each method and should be discussed later.

Exactly as you describe, the values of dissimilarity index portrayed as dots connected by lines in Fig.2 represent the similarity of each plot between a particular method and the reference dataset AR for that plot. While the individual dissimilarities may be high, we have good reasons to believe that the selection of 20 plots is sufficiently representative for the study area that the major patterns emerging from the analyses reflect real major patterns. Furthermore, you are right that systematic deficiencies in some of the methods are reflected in the single-plot patterns shown in Fig. 2. Some of these were discussed in the previous version of our manuscript and we will carefully search for more when we prepare our revision. These will then be taken into account in the discussion.

RC-3.20 - L 252 The statement in this sentence is not evident to me in Fig. 3, because this figure simply shows the profiles for each plot (which is a good way of illustrating the results, in my opinion). Wrong reference?

Absolutely. This typo will be corrected to Fig.2

RC-3.21 - L 254 Please see also my comment to L 222; I assume that the authors use the Wilcoxon test to assess if the median values of the dissimilarity indices for the 3 models are significantly different from each other. However, I think it is more relevant how the models differ to each other with regard to the AR data. This information is contained in the values of the dissimilarity index, and it should be reported more clearly here. The pairwise comparison of the 3 models seems to me of secondary importance to assess the goodness of the fit to AR data.

This is correct. The core result we report in this paragraph is the dissimilarity between the methods and the reference dataset. This is reported on lines 249-250 "While RS had the lowest median proportional dissimilarity with the AR reference (0.19, compared to 0.26 for DM and 0.41 for DGVM), …".

The pairwise comparison results of the Wilcoxon rank-sum tests are mentioned only after the core findings to support the similarity between RS and DM at most plots. We will ensure that this is clear in the revised paper.

RC-3.22 - L 262ff The visual comparison of the 3 models in Fig.3 and the associated description is more helpful to assess the modeling approaches than the statistical methods described before.

In the revised version of the manuscript, we will give more emphasis to the discussion of Fig. 3 in terms of model performance (do a carry out a joint assessment of the figure and the results of the statistical methods).

RC-3.23 - L 279ff This belongs into the Methods section. Explaining the sensitivity analysis earlier also makes it much easier to understand the goal of the overall approach.

Please see also our reply to RC-2.11. We will make clear in the introduction, that the sensitivity experiments are a separate chapter, which builds upon the results of the analyses. However, we will delete the subheadings 4.1 Methods and 4.2. Results to avoid confusion.

RC-3.24 - L 287 The term 'precipitation seasonality' should be better described, in particular since it is found later that it is important to improve DGVM parameterization.

Please see also our reply to RC-2.12. "Precipitation seasonality" is defined as the ratio of the standard deviation of the monthly total precipitation to the mean monthly total precipitation (also known as the coefficient of variation) and is expressed as a percentage. This will be added to the revised manuscript.

RC-3.25 - L 379ff The point about 'good' and 'poor' DMs is not clear to me. Why should poor DMs be used at all? Please explain, and also consider my comment above (L 186).

The terms 'good' and 'poor' refer to the predictive performance of the individual DMs (i.e. AUC - see also reply to comment RC-3.13). The study by Horvath et al. (2019) provides predictions of the distribution of a total of 31 vegetation types across the study area of Norway (with AUC values ranging from 0.671 to 0.989). Reasons for the low predictive performance of some DM may vary, but in this case is most likely caused by missing important predictors. The set of predictor variables used in the study (n=116) might seem excessive, but nevertheless the authors conclude that several important factors are not represented among these 116 (soil nutrients, NDVI, LiDAR etc.). The reason for this is that variables representing these factors were not available in the required formats/resolution/coverage at the time-point the study was carried out; a general problem in distribution modelling. By using the chosen set of predictor variables, statistical approach and settings, the authors obtained the best possible distribution models, even though with regard to the AUC values, some might be considered weak/poor. The direct answer to the comment is that the DM method requires estimates for the probabilities of occurrence for (almost) all vegetation types to create a seamless vegetation map, which in turn is required for making estimates for the PFT profiles as robust as possible. Thus, in this context, 'poor' models are better than no model. We will make this (important) point more clear in the revised version of the manuscript.

RC-3.26 - L 411 It may not be clear to readers why the lack of a shade-intolerant birch-PFT in the DGVM leads to the over-representation of NET in plots 17 and 18. The birch-PFT should rather have an advantage in mountainous regions compared to NET, which is currently lacking in DGVMs. Please clarify.

Please see also our reply to RC-1.7 and RC-2.14. We agree with the Referee #3 that this argument is not clear and without a clear support from our results. We will remove the argument from the revised manuscript.

RC-3.27 - L 450 Please check the literature for the recent progress in including high-latitude vegetation types into the PFT scheme of DGMVs, and add this to the discussion.

We will study the recent literature on this topic and add more recent references in the discussion. See also our reply to RC-1.8.

RC-3.28 - L 467 This sentence is hard to understand, please reformulate.

Yes, this will be reformulated.

RC-3.29 - L 475 It should be mentioned if increased seasonality promotes or impedes growth of NET.

Thanks for pointing this out. By applying the new threshold, the growth of NET is impeded if the value for precipitation seasonality is larger than 50 (Table 4, Supplement S6 and S11). This will be mentioned in the revised manuscript.

Supplement:

RC-3.30 - L 40 missing reference L 51 missing reference L 52 missing reference

Thanks for pointing this out. This is a remnant of splitting the document into manuscript and supplement. All the references are now fixed.

RC-3.31 - L 55 The PFTs for this study are not in bold font, but shaded grey, please make this consistent.

This will be fixed

RC-3.32 - L 56 The caption of Tab. S6 should be a bit more detailed: Is zbot the bottom height of the canopy (11.5 m above ground)? How is the coefficient of variation in precipitation seasonality computed?

We will adjust the caption to clarify all the mentioned abbreviations

RC-3.33 - L 90 The cover fractions in plots 801,2108,4268 are clearly not in a steady state. Please check if this significantly affects the results (e.g. by extrapolating the trends in cover), and repeat the DGVM runs, if necessary.

Thanks for pointing this out. We will extend the running time of our simulations for these sites to check when the vegetation distribution reaches a steady state, and we will investigate whether has an impact on our results.

RC-3.34 - L 122 missing reference

Comments on style:

All the following comments on style will be implemented in the revised version of the manuscript.

L 42 I think 'an' is not needed here.

L 55 'DGVMs' instead of 'the DGVMs'

L 60 ’at high latitudes’ instead of ’in the high latitude’

L 66 ’in’ not necessary

L 138 the second "of the" is not necessary

L 373 add ’the’ before ’reason’

L 401 ’differ’ instead of ’differs’

4 REFERENCES:

Ahti, T., Hämet-Ahti, L. & Jalas, J. 1968. Vegetation zones and their sections in northwestern Europe. – Annls bot. fenn. 5: 169-211.

Bakkestuen, V., Erikstad, L. & Halvorsen, R. 2008. Step-less models for regional environmental variation in Norway. – J. Biogeogr. 35: 1906-1922.

Fielding, A. H., & Bell, J. F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation, 24(1), 38–49.

Horvath, P., Halvorsen, R., Stordal, F., Tallaksen, L. M., Tang, H., and Bryn, A.: Distribution modelling of vegetation types based on area frame survey data, Applied Vegetation Science, 22, 547-560,

Johansen, B. E.: Satellittbasert vegetasjonskartlegging for Norge, Direktoratet for Naturforvaltning, Norsk Romsenter, 2009.

Majasalmi, T., Eisner, S., Astrup, R., Fridman, J., and Bright, R. M.: An enhanced forest classification scheme for modeling vegetation–climate interactions based on national forest inventory data, Biogeosciences, 15, 399–412,

O’Donnell, M.S., and Ignizio, D.A., 2012. Bioclimatic predictors for supporting ecological applications in the conterminous United States: U.S. Geological Survey Data Series 691, 10 p