Dear Referee #3

Thank you for taking the time to evaluate our manuscript. Your comments will lead to a thorough revision of the paper. Following is our one-to-one response to your concerns. Throughout this letter, your comments are written in blue color and are numbered.

(1) Scope: The manuscript applies and tests the LeNET CNN, which is a published and widely applied CNN to predicting biomes, which is a new application to for this particular CNN. I am not sure to what extent such an experiment falls under model development and therefore the scope of GMD. It is my understanding that validation/ model evaluation manuscripts are also permissible in GMD, but I find the model validation part somewhat lacking (see next point)

Response:

The authors' instruction of the Geoscientific Model Development (https://www.geoscientific-model-development.net/) defines scopes of manuscript types for considering peer-reviewed publication. Our manuscript satisfies the following items, and hence we are sure this manuscript falls within the scopes of Geoscientific Model Development.

-- Geoscientific model descriptions, from statistical models to box models to GCMs
-- Model experiment descriptions, including experimental details and project protocols

Besides, we conducted an additional experiment to address to your concern. Please refer to our response in the next section.

(2) Validation/ comparison to other methods

The authors test the CNN predictions against true biomes, which produces an approximate 50% success rate. The authors also address the limitations of the model, such as the fact that biome change are transient and that real world biomes are much more fractured compared to modeled biomes due to human managements and climate conditions suitable to several biomes/ or plant functional types. Apart from this 1-1 comparison, there is no additional validation against other methods. The authors outline in the introduction several methods for predicting biomes (either empirically or based on pyhsiological limits of vegetation), but never address how their method compares to methods of less, similar or higher complexity. For example, does their method outperform the HLZ scheme or what else is being gained by throwing machine learning at this problem? I want to be clear here: I am not saying that this is not a

valid and useful approach, but I don't think that the authors provide sufficient discussion to establish this.

Response:

We conducted an additional experiment for comparing the accuracy of PNV map reconstruction between the HLZ scheme and our method. Other PNV mapping schemes introduced in our manuscript cannot be compared directly with our method because they require additional data such as soil physics and topography. The scheme of Woodward & Williams (1987) is the only exception, but it gives multiple vegetation types for a given climatic condition, and hence it cannot be compared with our method. Following sentences, new tables S11, and new figure S5, will be added to the revised manuscript.

Following sentences will be inserted in the L159:

--Finally, we conducted an additional experiment for comparing the accuracy of PNV map reconstruction between the HLZ scheme and our method using common training data set. We developed a look-up table of the most common PNV for each combination of annual precipitation class and annual mean bio-temperature class, consistent with the HLZ scheme. Note that the HLZ scheme employs a hexagon table, but we employed a cross-tabulation table for simplicity. CRU annual climate and ISLSCP2 PNV map were used for generating the table. Then the table was applied to all climatic datasets we employed in this study, drawing reconstructed PNV maps for comparison.

Following sentences will be inserted in the L:

-- Following sentences will be inserted in the L214:

The accuracy of PNV reconstruction using the HLZ look-up table for each climate data set is 50.0% for CRU, 43.2% for NCEP/NCAR, 44.71% for HadGEM2-ES, and 37.2% for MIROC-ESM. These values are lower than any of our models trained with annual precipitation and annual mean bio-temperature（all models of Table S2 and model 5 in Table S3). This comparison shows that our method delivers a more accurate reconstruction of the PNV map even if seasonality was not taken into consideration.

Following table and caption will be added as new Table S11.

-- Most common PNV type and its probability (in parenthesis) for each combination of eight annual precipitation classes and six bio-temperatures classes. For cells with less than five grids are indicated as "NA." Definitions of PNV type numbers are as follows.

1, Tropical Evergreen Forest/Woodland

2, Tropical Deciduous Forest/Woodland

3, Temperate Broadleaf Evergreen Forest/ Woodland

4, Temperate Needleleaf Evergreen Forest/Woodland

5, Temperate Deciduous Forest/Woodland

6, Boreal Evergreen Forest/Woodland

7, Boreal Deciduous Forest/Woodland

8, Evergreen/Deciduous Mixed Forest

9, Savanna

10, Grassland/Steppe

11, Dense Shrubland

12, Open Shrubland

13, Tundra

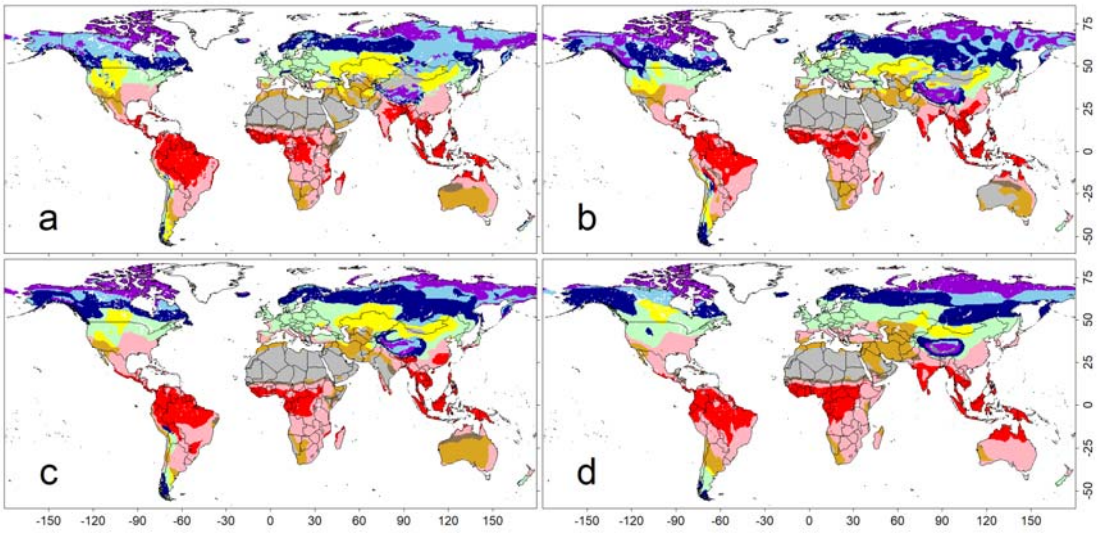
14, Desert

15, Polar Desert/Rock/Ice

		Precipitation class (mm/yr)							
		62.5~	125~	250~	500~	1000~	2000~	4000~	8000~
Bio-temperature class (Celsius)	0.75~	13 (66.3 %)	13 (67.9 %)	13 (62.7 %)	13 (47.7 %)	15 (85.7 %)	NA	NA	NA
	1.5~	13 (36.5 %)	8 (49.4 %)	13 (47.5 %)	8 (43.4 %)	13 (55.2 %)	NA	NA	NA
	3.0~	14 (65.2 %)	8 (35.1 %)	8 (43.0 %)	6 (46.1 %)	6 (57.6 %)	6 (58.8 %)	NA	NA
	6.0~	14 (60.9 %)	10 (58.0 %)	10 (56.4 %)	5 (27.1 %)	5 (43.1 %)	6 (42.5 %)	NA	NA
	12.0~	14 (86.7 %)	12 (52.2 %)	12 (42.8 %)	9 (41.0 %)	9 (23.8 %)	1 (72.9 %)	1 (59.1%)	NA
	24.0~	14 (93.5 %)	14 (48.5 %)	11 (25.9 %)	9 (47.8 %)	1 (41.1 %)	1 (93.3 %)	1 (96.2 %)	NA

Following figure and its caption will be added as new Figure S5.

-- Biome map generated by the look-up table of the most common PNV for each combination of annual precipitation class and annual mean bio-temperature class, consistent with the HLZ scheme (Table S11). Historical CRU annual climate and ISLSCP2 PNV map were used for generating the table. Then the table was applied to historical data of the (a) CRU, (b) NCEP/NCAR reanalysis, (c) Had2GEM-ES data, and (d) MIROC-ESM.



(3) Implementation of CNN

Based on the supplementary information, the CNN (LeNET) is run with default parameters and input parameters are air temperature and precipitation visualized as RGB images, which each image encoding a log transformed and normalized value for the two variables as color. The authors also conduct several experiments (see supplementary tables), but overall all of these have almost equal performance. I am wondering in this context, why this is the case. Is this something that has to do with the CNN that could be overcome by changes to model training/ model architecture changes or has this to do with the fact that the CNN is already extracting all the information that is extractable from the training data.

I feel that this may be the case, considering that CNNs are conventionally used to classify images/photos that are very complex (such as is this a dog or a cat), while the images fed into the CNN are very simple monocolor images. Once again this is an open question that could be addressed in additional discussion.

Specific comments

Response:

LeNet is the first CNN, and it was originally developed for classifying handwritten digits (i.e., ten categories). Still, LeNet seems to have sufficient ability to extract most of all the information contained in our training data irrespective of how it is visualized. For adding this point of view, we will insert the following sentences into the last paragraph of the discussion (L274.)

-- We compared performances of models trained by four different types of VCE

representation of annual precipitation and average annual bio-temperature, and all models have an almost equal performance (Table S2). This result might indicate that LeNet perfectly extracts at least two variables irrespective of how visualized.

(4) Introduction: I am missing some information about what motivates this model application and why predicting future biomes using AI may be useful.

Response:

For clarifying our motivations of this study, we will replace the last paragraph of the introduction (L63-68) with following sentences. In addition, the sentence in paragraph in the L 45-46 will be removed, as it's a duplicated description of our research purpose.

-- Using a CNN approach, we demonstrate an accurate and practical method to construct empirical models for operational global biome mapping. To the best of our knowledge, this is the first application of CNN to reconstruct a global biome map. We only employed a small number of climatic variables for input to examine how CNN improves the reconstruction accuracy compared to the classical HLZ scheme. We follow Ise and Oba (2019) and Ise and Oba (2020), a vital option for training CNN with a small number of input variables. This method represents climatic conditions using graphical images and employs them as training data for CNN models. After evaluating the accuracy of the biome map reconstructed by this method, we applied the trained CNN to climatic scenarios toward the end of the 21st century to demonstrate a possible model's application to predict the shift in the global biome map under changing climate.

(5) L72: ISLSCP2: Given that ISLSCP2 is potential land cover for the training, it would be good to discuss any potential issues with this dataset. Is this an unbiased representation of the true potential land cover.

Response:

We inserted the following sentence that explains the nature of this data set (L74):
-- The ISLSCP2 dataset represents the world's vegetation cover that would most likely exist now in equilibrium with present-day climate and natural disturbance in the absence of human activities.

We also prepared additional explanations for the ISLSCP2 dataset, but we think it is too much for this manuscript. If you think it's better to add these sentences on the manuscript, please let us know via editor, then we will do so!
-- This PNV data was delivered using the global 1km land cover classification data set

of Loveland et al. (2000). The most dominant "remnant" land cover type for each grid box was assigned as the PNV type. For grid boxes dominated by land use, simulation output of a process-based vegetation model (Haxeltine and Prentice, 1996) was employed to fill the PNV.

(6) L103: "mean of positive air temperature" > I am a bit confused about the positive. how are negative air temperatures treated? I would also encourage to replace positive with 'above freezing' for clarity.

Response:

Negative air temperatures were treated as zero for calculating the bio-temperature. According to your suggestion, we replaced the "positive" with "above freezing" in the definition of the bio-temperature. Besides, we found a mistake in the definition of the bio-temperature: It was calculated based on monthly air temperature, not daily air temperature, as was explained in our previous manuscript. Therefore, we will change your mentioned phrases as follows.

Previous sentence (L101-102):

-- Here, bio-temperature was defined as the mean of positive daily air temperatures.

New sentence:

-- Here, bio-temperature was defined as the mean of above freezing monthly air temperatures.

Previous phrase (L24):

-- mean of positive air temperature

New phrase:

-- mean of above-freezing temperature

(7) L113: "the model with monthly mean air temperature and monthly precipitation had the highest test accuracy"
> given that biomes are most often visualized along air temperature and precipitation axes, this does not seem to be surprising. Humidity and SW radiation may somewhat covary with T and P. I am wondering given that the CNN allows for 3 channels, whether there is some other variable (either climate or altitude) that may be useful to add.

Response:

Right. Adding variables (other than humidity and short wave radiation) may improve the vegetation map reconstruction. Especially, adding a topographical variable would be pretty helpful at the sub-grid scale. I inserted the following sentence at the end of the L269:
-- Another possible extension is simply adding one more variable that tightly controls PNV at sub-grid scales (such as altitude, slopeness, or slope aspect) into the VCE because one of the three RGB channels is empty in our model.

(8) Section 2.3: Training of the monthly CNN. The authors should elaborate here on the procedure for using monthly data.

Response:
For guiding readers, we added the following phrase at the end of section 2.3 (L129).
-- The annual and monthly climate training procedures are identical except for its VCEs.

(9) L190-194: I am not fully following this reasoning which seems to completly discout allocation disagreement. What the authors say may be true, but I don't think this is proven based on the information provided in the manuscript. One problem with this may be the map representation of results, which makes in depth comparisons and deep dive into potential reasons for model misses difficult.

Response:
Honestly, we cannot understand the point here. We feel simple map comparisons are not enough, so we calculated the quantity disagreement and allocation disagreement, which (we believe) make in-depth comparisons and deep dive into potential reasons for model failures.

(10) L195: "Table S9 compared the dependence of reconstruction accuracy on combinations of climate datasets for training and test climate datasets"
> I am a bit confused by this, given that the authors reasoned that using the same dataset for train and test could lead to overfitting and then argue here that using the same dataset for train and fit leads to higher accuracies which show robustness of the approach.

Response:
We agree. We will change the following phrases, which interpret the result of this experiment.

Previous phrases L197-199:
-- These results suggest that uncertainty in historical climate reconstruction is a larger source of failure in reconstructing biome distribution than the dependency of training on a particular climate dataset.

New phrase:
-- These results suggest that uncertainty in historical climate reconstruction and over-fitting are more significant sources of failure in reconstructing biome distribution than the dependency of training on a particular climate dataset.

(11) L282: "Since this method is simply an application of image classification AI, it demands much less technical skill and computer resources compared to other modern techniques such as those evaluated by Levavasseur et al. (2012), Levavasseur et al. (2013), and Hengl et al. (2018), for example."
> I am not sure that this is a fair comparison. One could similarly run a versy simple logistic regression or ANN from a standard package such as scikitlearn, which can easily be executed on a standard desktop PC.

Response:
We agree. We will replace your mentioned sentence (L280-283) as follows.
-- Since this method is simply an application of image classification AI, it does not demand much technical skill and computer resources.

Correction of Erratum:
We realized that the caption of figure 4 was a duplication of that of figure 3. Therefore, taking the opportunity of this revision, let us replace it with the following correct one.
-- Predicted biome maps under climatic scenarios from 2091 to 2100. Monthly means of four sets of forecasted climatic conditions derived from combinations of two climate models (i.e., Had2GEM-ES and MIROC-ESM) and two RCP scenarios (i.e., RCP2.6 and RCP8.5). These means were applied to the CNN model that was trained by the current biome distribution map, as well as the present climatic condition derived from the CRU dataset. Color definitions are available in Figure 1.

Best,
Hisashi SATO (on the behalf of all co-authors)