Dear topical editor,

We thank the referees for their constructive comments, and we appreciate your help and patience. We have revised the manuscript according to the referee's comments and suggestions. Below is our one-to-one response to your concerns. Throughout this letter, given words are written in blue and numbered consecutively.

Kind regards,
Hisashi SATO on behalf of all authors

Referee #1

(1) The authors have addressed all my comments from the discussion paper. I believe there is one new sentence that needs to be corrected (line 276 following), other than that the manuscript can be accepted as is:
For quantifying methodological uncertainty might also result from comparison of performances between correlative and process-based models in 'unsuitable' outside the environmental space of the training data (Yates et al., 2009).

Response:
We greatly appreciate your positive evaluation for our manuscript! For your mentioned sentence, we changed as followings (L277).

-- *For quantifying methodological uncertainty might also result from comparing performances between correlative and process-based models in 'unsuitable' outside the environmental space of the training data (Yates et al., 2009).*

Referee #2

(2) The authors have provide responses to most of my comments and I believe that the manuscript is improved. I am still somewhat reserved about the validation part of the model (see below) but acknowledge that this is the first application of CNN to biome classification which makes this novel and therefore a valid contribution to scientific discourse.

Major comments:

Regarding the HLZ model validation: I am grateful for the additional model comparison. I think that this is crucial. I also believe that a comparison between HLZ and CNN should be part of the main body of the paper (i.e. a figure or a table) rather than just providing supplementary data for the HLZ model. Similarly one could ask what were to happen if the HLZ would be run more finegrained. I am assuming that changing the bin sizes for the HLZ would potentially improve things until smaller bin sizes would not longer have an effect (e.g. convergence to a best possible model). I just want to make sure that the comparison between the CNN and the HLZ is as fair as possible.

Response:

We greatly appreciate your positive evaluation of our manuscript!
As you suggested, we conducted additional experiments of the HLZ scheme using finer-grained bin sizes. Its results are presented as table 2 in the main-body of this manuscript.

For describing the new experiment, we inserted following phrases in the L175.
-- *The bin sizes of the HLZ scheme are six for the annual mean bio-temperature class and eight for the annual precipitation class. As these coarse-grained bin-sizes would potentially depress the accuracy of the PNV simulation, we also developed look-up tables of 12 × 16, 24 × 32, and 48 × 64 bin-sizes to ensure that the comparison between our model and the HLZ scheme is as fair as possible.*

With this new experiment, the corresponding paragraph (L235-242 of the previous manuscript) was replaced by the following paragraph (L240-243).
-- *Accuracies of PNV reconstructions using the HLZ look-up tables for each climate data-set increase with the resolution of bin sizes for climate classifications (Table 2). It reaches quasi-equilibrium at 24 bio-temperature classes × 32 precipitation classes, which delivers nearly identical results with the CNN model. This result demonstrates that our VCE method extracts the best possible distribution of the most plausible PNV in a two-dimensional space of climatic variables.*

A new table, which explain the experiment, was inserted.

Table 2

CNN model and HLZ models accuracies for biome distribution simulations. CNN model corresponds to the top row model of table S2 (a RGB colour tile). Four HLZ models have different bin sizes for climate classifications (bio-temperature class × precipitation class). Each model was trained with the CRU dataset and adapted to the all-climate datasets (i.e., agreements of the CRU dataset correspond to the training accuracy, while other climate data correspond to the test accuracy). In addition, for each climate dataset, the annual mean bio-temperature and annual precipitation from 1971 to 1980 were log-transformed before use.

	CNN model	*HLZ models*			
		6×8	*12×16*	*24×32*	*48×64*
CRU	*58.3 %*	*50.0 %*	*54.9 %*	*58.1 %*	*60.4%*
NCEP/NCAR	*45.6 %*	*43.2 %*	*45.8 %*	*44.9 %*	*44.0%*
Miroc-ESM	*48.6 %*	*44.7 %*	*46.8 %*	*48.2 %*	*46.9%*
HadGEM2-ES	*41.3 %*	*37.2 %*	*40.1 %*	*40.8 %*	*39.5%*

With these modifications, we deleted table S11 and Fig. S5 from the previous manuscript, as they are no longer needed very much. We renumbered figure S6-S9 on the previous manuscript as figure S5-S8.

(3) Regarding my original point on the ability of the CNN (response 26): I am still not sure to what extent using a CNN is really the appropriate tool with respect to complexity.

In my mind the complexity of approach would be:

HZL > Logistic Regression > ANN > CNN

I therefore have no problem acknowledging that the CNN is capable of extracting all information from the images. As the authors say the CNN was developed to classify handwritten numbers and the benefit of the CNN is that the CNN is capable of using image information to extract complex features from images (in the number example this would for example be shadings/ edges/ vertical and horizontally oriented lines/ curves). In the case of the biome classification: there are exactly two features that are prescribed. T and P information as two-channel pseudocolor. The exactly same information could be presented to a logistic regression or an ANN as numeric

information (without having to first convert numeric information into images, which is - while straightforward in nature - laborious.

An ANN implemented in any ML package (sklearn, tensorflow, ...) or logistic regression in sklearn or any other package and an ANN in Keras or tensorflow could be implemented fairly easily. Given the limited amount of data and variables, I don't think that training such an ANN would take long. I have trained similarly sized ANN on current generation PCs.

Given that simple models allow for better explainability of results, I am really wondering whether CNN is the right tool. Therefore, it would have been nice to see how the CNN compares with for example logistic regression, given that logistic regression is also data driven, would make use of all the data (but remains a linear model). ANN on the other size is capable of extracting non-linear information. So an interesting question would be to see whether adding non-linearity adds benefit to the classification. Beyond that I am doubtful that the CNN would improve classification over the ANN, since the ANN should theoretically be capable of extracting all the information from T an P channels.

I know that this is not the manuscript the authors have written.

In that light, I don't think that the authors manuscript has any methodological problems and could be published subject to minor revisions.

Response:

If seasonal patterns are not considered for simulating biome maps (such as HLZ), we agree with you. All ANNs (Artificial Neural Networks) that can treat non-linearity would deliver the same result.

However, using our method, the seasonal pattern of multiple climatic variables can be employed for machine learning without any indexical expression. This is the most significant advantage of our method because indexical expression reduces the amount of information and adds a source of arbitrary. Among image classification ANNs, LeNet CNN has the lowest complexity, so we believe LeNet CNN is the right tool for this issue. To clarify the above advantage of our method, the introduction section was modified as follows.

Following paragraph in the L70-72 of previous manuscript was moved to L62 in the new manuscript.

--*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.

The sentences in L54-58 in the previous manuscript were moved to the L68 in the new manuscript, with some adjustments of conjunctions as follows.
-- *To account for seasonal variability, previous correlative climate-vegetation models needed to pre-define representative variables. For example, Levavasseur et al. (2013) divided each climatic variable into four "seasonal" predictors by averaging data corresponding 3-month periods (i.e., DJF for winter, MAM for spring, JJA for summer, and SON for fall). By contrast, the method we employed can automatically extract non-linear seasonal patterns for climatic variables that are relevant in biome classification.*

The following new sentence was inserted in L73.
-- *In other words, it enables CNNs to learn the seasonal pattern of multiple climatic variables without any indexical expression, which would reduce the amount of information and add a source of arbitrary.*

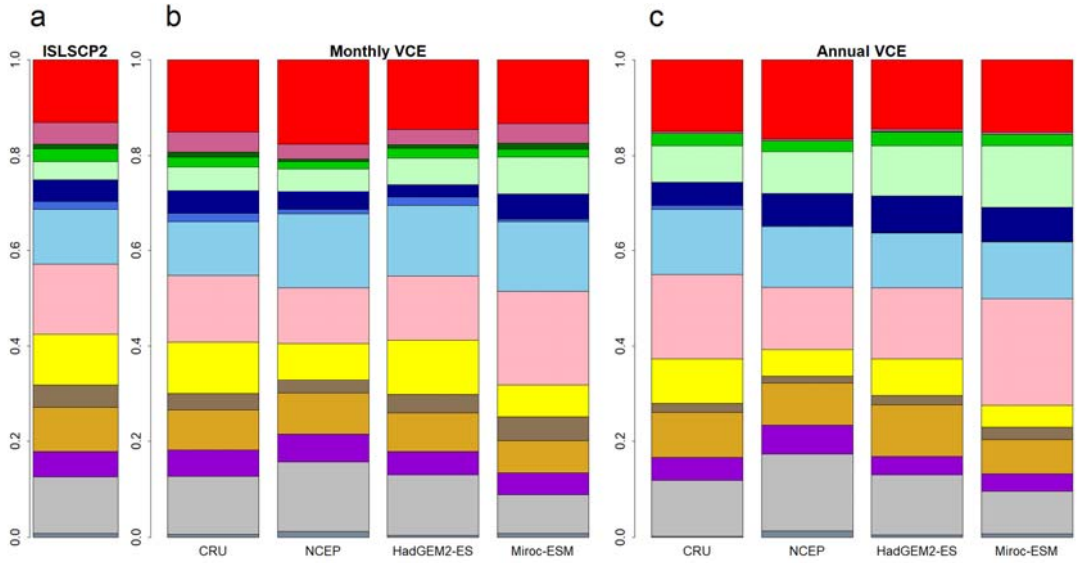
(4) Response 32: I am referring here to the visualization of results. The authors chose maps and summary statistics (accuracy) as their primary tool to convey information and I feel that for example looking at which biomes are more likely subject t allocation disagreement may be another way of better understanding why the model produces the results it produces, which for data driven methods is important but less straight forward.

Response:

We inserted the following figure to directly compare the biome compositions between maps. This figure is referred at the first paragraph of section 3.1. With this new figure 2, we renumbered figures 2-4 on the previous manuscript as figures 3-5.

Figure 2

Global biome compositions of the observation-based map (a) and simulated maps from CNN models trained by monthly mean climate (b) and annual mean climate (c) of CRU climate data spanning of 1971 to 1980. These CNN models were adapted to the four climatic datasets (CRU, NCEP, Had2GEM-ES, and MIROC-ESM) spanning the same period of the training data.



(5) L74: "We follow Ise and Oba (2019) and Ise and Oba (2020), a vital option for training CNN with a small number of input variables." > I suggest neutral language: "We follow the method of Ise and Oba (2019) and Ise and Oba (2020) for training ... "

Response:

We agree. The mentioned phrase was changed in accordance to your suggestion (L67).

(6) On a side note to comment 30: Including elevation may add additional information for the model to make use of (our ANN model in: Gerken, T., Ruddell, B.L., Yu, R. et al. Robust observations of land-to-atmosphere feedbacks using the information flows of FLUXNET. npj Clim Atmos Sci 2, 37 (2019). https://doi.org/10.1038/s41612-019-0094-4) did include for example elevation.

Response:

We cited the Gerken et al. (2019) as an example for how including elevation improve the model performance (L313).

-- For example, for geographically extrapolating flux data observed at flux tower sites, Gerken et al. (2019) trained artificial neural networks (ANN) using the elevation of each tower site.