Comment on gmd-2021-271

Anonymous Referee #3

Referee comment on "Predicting global terrestrial biomes with the LeNet convolutional neural network" by Hisashi Sato and Takeshi Ise, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-271-RC3, 2022

The authors present a concise study about the application of the LeNET convoluted neural network (CNN) to predict global terrestrial biome distribution from climate data and then apply the CNN to predict hypothetical future biomes from climate model output.

General comments

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)

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.

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

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

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.

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.

L113: "he 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.

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

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.

L195: "Table S9 compared the dependence of reconstruction accuracy on combinations of climate datasets for training and test195
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.

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 scikit-learn, which can easily be executed on a standard desktop PC.