Review for "Focal-TSMP: Deep learning for vegetation health prediction and agricultural drought assessment from a reginal climate simulation" By Eddin et al.

In this study, the authors proposed deep learning models to predict NDVI and brightness temperature (BT) with simulations of a coupled earth system model as inputs. As a downstream example, the predicted NDVI and BT are used to evaluate long term agricultural drought indices (VCI, TCI and VHI) and the importance of input explanatory variables are also explored with explainable artificial intelligence. The deep learning models, based on vision transformers and CNN, takes advantages of tangling complex relations as well as biases compared to traditional radiative transfer models (RT). The results indicate an overall MAE of 0.027 and 1.90 K with R2 scores of 0.88 and 0.92 in predicting NDVI and BT and the authors claims that this framework is an effective way to examine the overall predictive capability of the earth system model to predict agricultural drought events. It is novel to use DL models to predict drought related variables which are not directly simulated by earth system models. However, the manuscript is not well written and organized, and needs major revisions before being published at GMD. My major comments are as follows:

1. The introduction lacks rationales on conducting the study. I suggest integrating sections 2.1 and 2.2 into the introduction. In the introduction, I would like to emphasize that the authors need to summarize what previous studies did, not describe what they did one by one (first paragraph in section 2.1 and first paragraph in section 2.2, which are not very readable and not necessary), but summarize and provide rationales and research gaps that this study will fill.
2. Lines 190-192: what is the rationale on variable selection?
3. Section 4.1: the authors need to provide rationale and advantages about the proposed architecture at the beginning of this section compared to state of art architectures (information in lines 322 to 333 with adjustments should be provided at the beginning of section 4.1). I also suggest using jargons as less as possible in the methodology section.
4. I suggest that lines 403-464 should be in the methodology section as baseline approaches and evaluation metrics. I also suggest the italics section in section 5 to be separated section as 5.1, 5.2 and so on, which will be much clear.
5. It is not clear what the authors would like to emphasize in lines 438-444.
6. It is not clear in line 451. Based on the methodology (lines 285-289), both the input variables and outputs are in weekly time scale, why here claims average two days?
7. The results are not well organized and too many sections. I suggest organizing all the results into one section as subsections for each topic.
8. Lines 467-471: it is confusing about different climatology. Based on the dataset information in section 3.1, the model simulation is from 1989 to 2019. What does climatology from 1981 to 1988 come from? What is the main point for analyzing different climatology (climate change?)? Table 1 is also confusing for mixing climatology different periods and validation/test different years.
9. Based on Table 1, the metric differences for different DL models are trivial (mostly around percentage scale), although the DL models have very different building blocks. The proposed DL model is also not consistent better than others particularly for the test dataset. My doubt is whether the trivial differences are caused by stochasticity not due to DL model itself. The authors claimed fixed random seed for reproductivity, but what if the seed is not fixed and what does the results look like for running the model several times? Will the results be consistent with any conclusion got from the table?

10. I suggest combining section 8 and 9 as one section (conclusions and limitations). The current discussion section appears only limitations not discussion.

Minor comments

1. At the beginning of abstract, I suggest adding one or two sentences about research problem the authors would like to solve.
2. AVHRR in line 44 firstly appear without full name and also check for other short names.
3. Line 50: change synthesis to be 'synthesize'.
4. Line 195: What is DA represented? If the term only shows once in the manuscript, it is not necessary to use the short name.
5. Line 200: thema and lambda represents?
6. Line 253: Is it equation (5) is for NDVI for AVHRR instead of VIIRS as stated in line 253?
7. Figure 1: It is better to use TSMP instead of TerrSysMP since TSMP is used all over the manuscript.
8. Line 325: it should have a period between 'efficiently In'.
9. Line 368-369: it is not clear why the third gate focusses on the water area? How do you know that?
10. Add references on line 430.
11. Line 448-450: the sign N(0, I =0.02) is not clear. It is better to use normal distribution with zero mean and standard deviation.
12. Line 461: reported how long it takes for the focal model, what about other DL models?
13. Line 471-472: What is these non-ML baselines? It is not clear for this statement.
14. Line 516: change 'fro' to 'for'.
15. Figure 10: change 'sprint' to 'spring' in the left two plots in the first row.