The manuscript uses five deep learning models to train CMIP6 air temperature to CRU TS gridded air temperature. The statistical evaluation shows that CMIP6 air temperature is improved. However, I have some concerns about the methodology. Machine learning models have been widely applied in many fields and can easily beat traditional methods and raw data inputs concerning statistical accuracy. I am not surprised that DL-based CMIP6 temperature is closer to CRU TS than raw CMIP6 temperature because raw models are not designed to approximate CRU TS. But the DL model design in this study has some problems. This makes me worry that the results are right for wrong reasons. Probably using simpler methods such as random forest or even linear regression can achieve the same improvement. Besides, the writing of the manuscript is a big problem. The manuscript needs extensive revision to be publishable.

- The language of the manuscript needs substantial improvement. Reading the manuscript is a challenge for me. The manuscript needs thorough language revision to meet the standard of a publishable paper. I tried to list some examples of grammar errors, format problems, or awkward expressions, but failed because there are too many problems.
- Temperature data from all models are interpolated to the 2° resolution using the bilinear interpolation, which is incorrect from my perspective. First, since air temperature shows a good relationship with elevation, temperature downscaling is often achieved using the temperature lapse rate method. You cannot simply use the bilinear method like for other variables such as precipitation. This method can cause notable bias in complex terrain. Second, the interpolation of model data will cause information loss. In other word, the interpolated air temperature is worse than the raw data due to my first question. Downscaling the temperature from 2° to 0.5° can show
some improvement, but how to consider the contribution of information loss to this improvement is difficult. Therefore, the authors should use a more appropriate temperature downscaling method.

- CRU TS uses limited stations to estimate air temperature. Most regions of the world only have sparse stations, and thus the quality of CRU TS is different. Observation-based datasets can show large differences particularly in regions with complex topography and few stations. Sometimes, model outputs could be more reliable than interpolation-based datasets such as CRU TS. Training the model in those regions using CRU TS as the true value does not sound reliable to me.

- The study divides the world into five parts to train the model (Figure 2). However, the division is problematic. EUR-AF contains Europe, Africa, and part of Asia. The three sub-regions are very different concerning their climate, area, and station density. Europe has dense stations and thus CRU TS has a good quality. In contrast, Africa has sparse stations and thus CRU TS has a low quality. If you train them as a whole, the quality in Europe could be degraded because most training samples are from Africa and part of Asia which are not reliable. Moreover, each part covers a broad domain (especially EUR-AF) with quite different climate conditions and temperature schemes, while training a model in such a large domain cannot consider those complexities. That's why I say you may get right results for wrong reasons.

- The introduction part is overlong and needs reorganization. Many contents introduce the background such as climate change and data which are widely known and do not have close relation to this study. In contrast, the deep learning techniques have been widely applied in Earth science fields recently and cover many variables including air temperature. Literature review of those studies is important but insufficient.

- The introduction to the five DL models in 2.2.1 is too abstract. The authors should focus more on the implementation of DL models such as model structures, parameters, training and testing strategies. In short, the method part should ensure that readers are able to reproduce the work. Some contents in “3 Results and discussion” can be moved to the method part.

- It probably better to adopt a traditional method as the benchmark in this study. The work shows results from five DL models, but readers can hardly know whether the improvement is good enough without comparing to a widely used method.

Some minor comments:

- The manuscript states that compared to traditional methods, the new DL methods can do downscaling, bias correction, and merging as a whole, while traditional methods need to address them separately. There are two problems here. First, some traditional methods such as geographically weighted regression do the same thing. Some traditional machine learning methods such as RF and ANN can also do this job. Second, black-box models makes data production much easier for researchers. However, this can sometimes block understanding of the world. Therefore, I recommend the authors revise some relevant descriptions in the manuscript.

- There is no need to separate SPAEF and other metrics. The concept of SPAEF and similar metrics are widely used in many studies. Please merge 2.2.2.1 and 2.2.2.2.

- How the 0.5-degree merged data is produced? I did not see a detailed description about how DL models realize this.
Please unify the usage of air temperature and surface temperature in this manuscript. They can be confusing to some readers. It is better to always use air temperature throughout the manuscript.

The authors claim the resolution of most ESMs are close to 2-degree. The word “most” is inappropriate to some extent considering some products in Table 1 have higher resolution. It is better to say that “most ESMs have a low resolution” than “most ESMs have a resolution close to 2 degree”