Comment on essd-2021-418
Anonymous Referee #2

General comments

This manuscript used five deep-learning (DL) methods to downscale 31 ESMs-simulated surface air temperature (SAT) from a coarse spatial resolution (~2 degree) to a higher spatial resolution 0.5 degree. However, the work is not innovative and no important and robust findings were obtained. The authors mechanically downscaled surface temperatures using 5 different DL methods, which have been widely applied in this field. The results do not convince me since this method highly depends on the training data, including sample numbers, spatial and temporal scales, etc., as presented in table 2. The only observed training data CRU TS also have large uncertainties, since it is derived from unevenly distributed stations, which may deliver wrong signals to the DL methods, especially at complex terrain areas. Furthermore, the SAT is highly dependent on local climate conditions, terrain factors, as well as large-scale atmospheric and local circulation. No such physical-based ancillary data were used in this study, which limited the further applications of produced 0.5 degree data, especially in the mountain regions. Therefore, I cannot recommend publishing this manuscript in ESSD at the current stage.

Specific comments

- In Abstract, the authors claim “The SR algorithms are designed to enhance image quality and outperform much better than the traditional methods.” The authors did not
use any traditional methods, so this conclusion is not evidentially supported. Furthermore, what do the “traditional methods” mean?

- The authors concluded a ‘tertiary class echelon’ condition based on MAE. Can this conclusion be supported by RMSE or R, as both are used to assess the errors in this manuscript?
- Some abbreviations are not defined, what is ECM in the Introduction?
- The authors claim the “traditional SR methods”. What are traditional SR methods? What is the difference between traditional and non-traditional methods?
- At 2.1.1 section, 80% CRU TS (1901-1992) was assigned as training data and the rest 20% (1993-2014) was assigned as validating datasets. This is not appropriate. Because the short-term validate data may be highly influenced by climate variability.
- All the ESMs outputs were interpolated at 2 degree may bring new errors. For this reason, the poor performance of downscaled 0.5 degree data may originate from this step.
- For the applications of 5 DL methods. How many parameters? How to tune these parameters? More information should be given for the methods.