Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2021-267-RC1, 2021

Comment on essd-2021-267

Anonymous Referee #1

Referee comment on "GPRChinaTemp1km: a high-resolution monthly air temperature dataset for China (1951–2020) based on machine learning" by Qian He et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-267-RC1, 2021

The manuscript describes ML tools for spatial interpolation of air temperature data in China to a 1k resolution from meteorological stations. This topic is essential. However, many significant issues must be addressed. Here are some comments that hopefully can help to improve the manuscript qualtiy:

- It is not clear what is the spatial resolution of the meteorological station? Since the covered period and the temporal resolution, missing data of the monitoring stations may differ, it may be useful to provide a table or figure to summarize all this information.
- The methods used in this study are inappropriate, and the experiments lack sufficient detail. In ML, the dataset should be divided into train, validation, and test subsets. The validation can be adopted to evaluate the model while tuning model hyperparameters. Hyperparameters are crucial to obtain the best performance model, which is missing in this manuscript.
- Missing details regarding how to process the data., i.e., it is unclear how to deal with the missing data, how to normalize the data, etc.
- The model is compared based on one dataset, and without a statistical test, I would like to say there is a high chance that the GPR outperforms others by accident.
- Why the error RMSE, MAE and R2 shows a cycle pattern? Any reason for that?

Minor comments and questions?

1) Line 7 need to clarify why to use the "subset features" option of Geostatistic Analysis tools. Is it used to split features or datasets?

2) The explanation of SVM is not clear and needs to be further improved.

3) In Line 189, the sentence is not understandable.