The manuscript aims to produce a long term dataset of monthly 2m temperature over China at high spatial resolution on a 1x1 km grid. While the objective is appealing due to the challenges related to the complex topography and the irregular data availability in the target region, the applied methods show up with significant issues. The major issues are listed subsequently:

[1] The introduction discusses advantages and disadvantages of different information sources for the targeted dataset. While strong arguments for point-wise observational data are presented, long term reanalysis data products are not considered despite they provide consistent and spatio-temporally coherent information on the atmospheric state. It is unclear why such data is not considered to provide predictor variables.

[2] The method of data splitting leads to strong autocorrelation between the training and test dataset. Due to the spatial proximity of stations in both dataset, a fundamental requirement is hurt, that is the independency (or at least a minimization of dependency) between the training and test dataset. This is especially true for the stations located in the flat eastern parts of China with a dense observational network. Thus, the statistical model are prone to learn nearest neighbor-relations rather than learning real abstractions from the features, see, e.g. Kleinert et al., 2021 for a more detailed discussion on the requirement of splitting the test and training data temporally when stations are located close to each other.

[3] Only static features are used as predictors which implies that a model must trained for each month (!) of the period under consideration. Thus, dynamic information on the atmospheric state can exclusively deduced from the optimization procedure on the predictand. It is strongly recommended to introduce dynamical data as a predictor variable instead. Besides, the chosen predictors have periods with neglectable correlation with respect to the target quantity and important features such as the ambient topography (is the meteorological station located in a valley) is absent (see, e.g., Sha et al., 2020).

[4] The evaluation does not serve the objectives of the study. The stations in the test dataset are dominated by stations over flat terrain with a dense observational network. Thus, potential deficiencies in capturing the variations due to underlying complex topography are hidden. Indeed, Figure 5 indicates that residuals are considerably larger over the mountainous region.

[5] Several issues in the follow-up study are present such as (a) a focus on large-scale
temperature patterns instead of fine-scale patterns in Section 4.2. to reason the high spatial resolution of the dataset, (b) the interpretation of patterns in the Xinjiang region which look like artefacts (bulls-eye pattern in winter months) and (c) the missing notification on the better performance of the reference method ANUSPLIN for July-months in the 70s, 80s and 90s.

[6] The comparison to the competing datasets ERA5 and FLADS is misleading due to the much coarser spatial resolution of these two datasets. A fair comparison would consult datasets with similar spatial resolution such as the dataset described in Peng et al., 2019.

Further minor issues are:
* Splitting into three distinct datasets is unnecessary. Rather merge it to one dataset with one DOI.
* Refer to statistical and dynamical downscaling techniques in the introduction.
* Provide references to the problems related to remote sensing data (see l.66).
* Describe the remapping of the STRM DEM data onto the 1x1 km grid (should be an averaging method).
* The used software tool MATLAB should be only mentioned once rather than being repeated three times. More details on the respective ML-technique would be appreciated.
* l.149: Should be ‘ensemble machine learning’
* l.219: "Unnecessary reference to Equation 4 which directly follows the sentence.
* l.343f. This is sentence is barely comprehensible.

Mentioned literature references:


Please also note the supplement to this comment: