Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2022-45-RC2, 2022

Comment on essd-2022-45

Simon Tett (Referee)

Referee comment on "1km Monthly Precipitation and Temperatures Dataset for China from 1952 to 2019 based on a Brand-New and High-Quality Baseline Climatology Surface" by Haibo Gong et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-45-RC2, 2022

Review of 1km Monthly Precipitation and Temperature Dataset for china from 1952 to 2019 based on a Brand-New and High-Quality Baseline Climatology Surface.

By Gong et al

This paper aims to produce a 1km resolution monthly dataset from 1952 to present for temperature and precipitation. It does this by first generating a 1km climatology for 1981 to 2010. This is done by interpolating *in situ* station data using various covariates including, when available, satellite estimates of precipitation and land surface temperature. The authors show that the data is an improvement on existing datasets. I think the paper and dataset are a useful contribution to knowledge of Chinese climate change and should eventually be published.

Major points

- Though the paper is not badly written, it would benefit from an edit to improve some of the English and clarify some points. In particular the discussion around model selection is difficult to follow.
- I am concerned about the lack of quality control on the data though the station data used in the dataset might have been quality controlled by CMA. Some discussion of this is needed in the final paper. The Impression that my Chinese collaborators give is the instrumental data prior to 1961 is not very reliable. Some discussion of this is needed

and could form part of the quality discussion on the data.

- I do not understand why satellite data was interpolated to 1km to then be used in the interpolation of the station data. It would be better to keep that on its original resolution (or have the same values for every 1km pixel within the satellite information foot print) rather than add false spatial information to it. Further, as this data is only present for some of the climatology period I worry about the quality of the analysis prior to that data being present.
- The section that describes the selection of the various interpolation models is unclear. My sense is that multiple models are tried and per calendar month the best one (smallest RMSE) is used. I am concerned that this will lead to over confidence. I appreciate the authors have kept 10% of stations, selected at random, back for testing. I *think** this data is used for testing the 11 different models and determining, for each month, which is the best model. In the absence of physical insight it would be better to have a single interpolation model for the whole period with data withhold from the model generation & selection, and then used to test the final model. This could also be used to test the error model used in the interpolation.
- The climatological analysis lacks a comprehensive uncertainty analysis. At the station level the authors are, I assume, using a white noise error. However, I think the interpolation will generate correlated error. The dataset would be an advance on existing approaches if it generated an ensemble of datasets that included the various error sources. That would allow users to determine how uncertain the results are, I suspect this will vary considerably depending on station density and other features of the data.
- The title is too long and over claims. It will not be "brand new" for long and the authors do not demonstrate its quality. I suggest excising such text from the title and a revised paper.
- Figures and associated text are rather small. I recommend that the authors create the figures at the size they expect them to be in the paper. Doing that will mean that text elements are of the appropriate size.
- As climatological precipitation has very large spatial gradients I think it would be better to compute RMSE statistics as fractions of estimated precipitation rather than in mm.
- I have similar concerns about the monthly-resolved model as I do about the climatology. That is the model selection and given the satellite data only includes a relatively small part of the entire dataset wonder why this is being used at all. Further there is a lack of uncertainty analysis.
- For the monthly resolved dataset I suspect that interpolating fraction of normal precipitation, rather than total precipitation, would be more accurate.
- The authors should consider making available an anomaly dataset (difference from normal) available whose effective resolution is likely low.
- The authors make the data available in a variety of formats (NC for the climatology and geotiff for the monthly resolved dataset). They also describe the scaling used in the NC file. I have not looked at the data in detail. Given my concerns about the work I will wait for a revised paper before doing this. I do not think they need to say in the revised paper what the scaling is – that should be in the NetCDF file. Unless they are reducing the precision it is not necessary to introduce a scaling factor either. I looked at the precipitation climatology dataset using xarray. I suggest that sensible names are used and a correct crs code (or name) So change "variable" to "pr" or "tas" as appropriate. And rather than z set it to month. I also recommend adding some metadata to the dataset pointing to the paper, the authors, and information on the period covered and possibly the statistical model used.