

Review of Berkeley Earth Surface Temperature - High-Resolution (BEST-HR): a 0.25° Global Gridded Temperature Data Set for Climate Monitoring by Robert A. Rohde, Bruce Calvert, and Zeke Hausfather

The authors present a substantial methodological update to the Berkeley Earth product including, amongst others, improved input data and resolution. The paper subject is clearly in scope for the journal. The paper is overall well structured with no obvious omissions or issues that would preclude its eventual publication. I have a small number of recommendations for the authors that may improve the paper.

Major comments

1. I assume that the group will switch to using the latest version of HadSST4 that accounts for the mid-20th century data issues potentially better (as documented in that group's most recent paper) as I doubt the Met Office will continue producing both variants. To avoid an immediate disconnect between the paper description here and the dataset it would be advisable to switch to the use of this new version of the product and update the description in 2.2 and amend the results and their discussion accordingly. Maybe you already do looking at lines 734-735 but if so you need to state this much more explicitly and earlier. The description in 2.2 should be much more explicit either way.
2. The use of HadISST and its assumptions viz Antarctic sea-ice in the pre-satellite era probably warrants considerably more attention and caveats than is currently given in the paragraph starting line 235 and then the results. My understanding is that the pre-satellite era extent estimates are unduly permissive and the implications of this for at the least local if not hemispheric estimates likely warrants some care and attention and quantification in the revision beyond the current very brief discussion in lines 630-635.
3. To my reading of section 3 there is no physical consistency check built in that T_{avg} must equal $T_x + T_n / 2$ and that your estimate of $T_n < T_x$ etc. Maybe its not possible to specify these in the method, in which case how far the fields vary from these basic premises would be worth some consideration. The text in lines 555-557 suggest that there is no formal constraint applied?
4. How independent are ensemble members in the super ensemble described in lines 645-652? Are your ten member ensembles recreated each month? Is this some random seed process? And if so how does this affect / complicate issues around reproducibility for end-use applications? Does it de facto force you to archive the full dataset and ensemble each month to enable reproducibility and is this (even remotely) feasible? It might be advisable for ensembles to be frozen over the historical epoch to avoid some undesirable issues with the construction of a new set of ensembles triggered by specific circumstances (a new set of data etc) rather than reconstructed from scratch each and every month? Are the grids also similarly reconstructed and does like the NOAA product this introduce month-to-month flutter in the deep past?

Minor comments

1. The paragraph starting line 27 would I think make better sense if integrated into the abstract rather than being placed where it is. These are key details that I would expect to see in the abstract and their current placement feels odd and breaks the paper flow.
2. The updated Thorne et al paper under revision and shortly to be resubmitted now uses a different blend of datasets than that described starting line 40 as documented in their public response to reviewers and this list should be updated accordingly.
3. The discussion on line 70-76 is a little misleading. As described in Hersbach et al for reanalyses the surface temperature and humidity are used to nudge the soil physical characteristics so they are assimilated but not directly. I would suggest re-reading the Hersbach et al or discussing this with the ERA reanalysis team to ensure this text faithfully reflects how these variables are used in the reanalysis systems.
4. On line 92-97 I would suggest adding the high resolution land only replay ERA5-land and regional products such as CERRA and UERRA to the description for completeness.
5. In lines 218-220 it may be worth noting that these summary of the day products are principally derived from ISD and the replacement GHCNh product and that SSOd is perhaps a more appropriate self-descriptive name of what these are – a daily summary as permitted from underlying ‘synoptic’ instantaneous reports.
6. Line 269 station i? Not iii?
7. How sensitive is the analysis described in 6.1 to the choice of CESM2 over another climate model?
8. In lines 739-742 I would be a little less definitive that DCENT is necessarily better. Its definitely different but I think the 19th century differences in particular its less clear to me and I believe the community as a whole whether it is definitively better than existing SST products.
9. Line 763 I think should be GlobTemp and not GlobalTemp
10. The remark on line 790-791 would be worth comparing to existing datasets of Tx, Tn and DTR more holistically than is currently done. See e.g the review in <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2015JD024584>