Review of 'EMO-5: A high-resolution multi-variable gridded meteorological data set for Europe' by Vera Thiemig et al.

The authors intend to introduce and describe a gridded high resolution dataset over Europe produced from in-situ sub-daily meteorological observations supplemented by information from reanalyses and various other products. The dataset is used by CEMS in support of operational real-time emergency management and in principle could have many other applications. The dataset description paper is therefore, in principal, a useful contribution to the literature.

Aspects around data collection, quality assurance and gridding / interpolation are undoubtedly valuable. However, in almost all cases they fall consistently short of state-of-the-art practices undertaken elsewhere and, therefore, it is hard to justify claims that this is a uniquely valuable dataset. There is considerable room for future improvement in all aspects of the analysis.

That said, there is also value in version pointing at this point and thus publishing the current approach as an initial basis upon which further improvements can subsequently be made. To do so will require substantive revisions to the present manuscript to:

I. Better document key aspects
II. Apply caveats and draw comparisons to state-of-the-art approaches as relevant
III. Describe far more holistically how to access the data, its available formats, file structures, version control etc.
IV. Incorporate a substantive discussion of future requirements and present limitations

The journal is undoubtedly the appropriate place to publish this work. But the authors would need to redraft comprehensively addressing all the enumerated major points that follow for me to be able to recommend eventual publication. Given the volume of major points which will require substantive redrafting I will not make additional minor comments at this juncture.

Major points

1. The introduction is too short. Greater effort is required on scene setting including a discussion of the role and limitations of both reanalysis and satellite observations to provide the reader with necessary context to then properly interpret the paper.

   Indeed, there is in general a lack of sufficient acknowledgement of prior and ongoing valuable work in this area across the submitted manuscript as a whole to provide the reader with necessary context to properly and fairly evaluate the value of your work. For example ECA&D and the associated E-OBS gridded product served via the C3S CDS should be noted much more prominently and far earlier and calls into substantial question your assertion on lines 39-41. I would be hard pressed to defend the statement you make there given the availability of the well documented and highly utilised E-OBS product. Similarly, the efforts under C3S to collate station data as documented in e.g. Noone et al., 2021 warrants more attention than is currently given. Failing to acknowledge these – Copernicus funded – efforts just gives an impression of dysfunctionality across Copernicus program activities as well as –

for those with the knowledge – giving the distinct impression of over-selling what you have produced and doing so in glorious isolation. This is disappointing as a knowledgeable reviewer and equally does not bode well for realising programmatic synergies across Copernicus services in the new Copernicus budgetary cycle activities.

The paragraph starting at line 55 should come earlier and be expanded. Then in numerous places comparisons to existing products / approaches should be made. For example, the QC method in 3.1 should make reference to QC procedures of E-OBS, GHCND and HadISD at a minimum.

2. The reason why the analysis is limited to 1970 needs to be made far more explicit. There are many data that extend back further than 1970 and the choice of 1970 is arbitrary with no clear rationale given to justify the choice. From an application perspective there would be considerable value to extending the analysis to earlier periods to the extent possible.

   Given that the product is reputedly available from 1970 onwards it is problematic that Figures 1-3 only show availability post-1990. This gives the impression that you may be wishing to hide / obscure availability of data in the first 20 years of the series. The figures should show 1970 onwards or the product should be cut at 1990 and published as such. Later at the end of section 3 you suggest this is the case. Please clarify whether the product extends 1970 to date or 1990 to date and ensure consistently stating so throughout the paper.

3. The methodology in almost all aspects needs to be strengthened and much more explicit. The method must be outlined to the extent that the analysis could be to first order independently reproduced based upon the description. This is necessary and a central tenet of the scientific method itself. Please revise the method to be considerably more comprehensive such that based upon the method description an independent researcher may reasonably be able to recreate your approach.

4. The validation activity would be stronger if it could be shown in more detail for all variables and would, undoubtedly, serve to increase user uptake. However, the only real validation in the strict scientific sense is actually the leave-one-out analysis which is not even contained within the validation section at all. What is presently classed as being validation is actually characterisation of the product. Validation requires a value that is of known quality to compare against and none of the three sub-sections in the section purportedly to do with validation actually undertake such a comparison. That section should be retitled to Characterisation accordingly. Then the leave-one-out validation should be considerably expanded, shown for numerous parameters, and aggregated across well-sampled and sparsely-sampled regions separately.

5. In section 2 the input data is outlined to a very perfunctory level – greater detail here would help including a breakdown by source, data policy etc and links to sources where available publicly by elevating Table A1 to the main text and

augmenting with additional information. Also, the maps in Figures 1 to 3 conflate station and gridded input in ways that are really inaccessible. Maps of just primary station source by time and / or variable would be more accessible and avoid proverbially mixing apples and oranges together.

Worse, Section 2 is also completely silent on sharing of these station data onwards to e.g. the World Data Centre for Meteorology at NOAA NCEI or the C3S in-situ database effort or ECA&D and thus E-OBS. This limits the potential utility of this data collection activity if you are not actively trying to share data where you can with the activities being undertaken to improve access globally to meteorological holdings, many of which are funded by Copernicus. Where is the joined up thinking? How does this serve Copernicus if you are not actively sharing across Copernicus Services?

6. The quality control is a little limited, in particular in not including any form of neighbour buddy checking as is state of the art in e.g. GHCND or HadISD. Either that or, between the text and the table, the presence of buddy checks is obscured. You should at a minimum be explicit that the applied checks consist of a minimum set of record-based, logic and spike-based checks and do not include a number of other checks including repeat strings, distributional checks, frequent value checks and neighbour-based checks as is the case in state-of-the-art QC procedures such as GHCND or HadISD. You should aim to incorporate a broader array of such checks in future.

   Furthermore, I would expect the discussion of QC in section 3.1 to include some summary of the frequency with which different values are flagged and some consideration of heterogeneity of flagging of values across sources and regionally. Does the frequency of QC test failure look reasonable? Does it raise any flags for particular sources etc. etc. are all questions I would expect this paper to address if it is to build user confidence in applicability of the resulting product and yet it is presently silent on these issues.

7. The interpolation scheme discussion in 3.2 requires substantive additional detail. No discussion is forthcoming around why the different schemes have different skill per parameter. This must, intrinsically, be to do with the spatio-temporal correlation structures of the different parameters and how effectively the three schemes can handle these. In particular the parameters that vary more smoothly in space and time clearly are better suited to IDW whereas the much smaller scale correlation structure precipitation is better suited to SPHEREMAP. The lack of geophysical interpretation of the findings does not build the necessary user confidence that you understand what underlies your methods.

   The analysis should also remark upon the similarity in skill measures for most variables and diagnostics as this implies that interpolation choice isn’t a first order effect presumably? Furthermore, is there any gradation in skill between well sampled and sparsely sampled regions? One would presume so but your present aggregation just lumps all cases together. It would be considerably more informative to split this analysis and repeat for well-sampled and sparsely-sampled regions. I would expect no impact of choice in well-sampled regions where interpolation choice makes little difference but much larger impacts in sparsely sampled regions. Can you repeat that analysis but splitting out by regions of greater and lesser station density? What does that tell you?

8. Given that ERA-Interim/Land values may be biased in terms of variance and mean state isn't the time-varying use of this product to infill in data sparse regions potentially problematic for long-term product homogeneity? Also, ERA5-Land has replaced ERA-Interim -land having higher resolution and improved quality. Should the algorithm not use this product instead? In particular because ERA-Interim land is no longer produced in CDAS mode introducing potentially a major discontinuity into your product upon cessation of ERA-Interim land production?

9. The assumptions around 06-18 and 18-06 for Tn and Tx will miss many true Tx and Tn values particularly in the winter half season when daily maxima or minima often occur outside these windows. Particularly so in the higher latitudes of the domain. They also risk introducing a discontinuity with the mean temperatures calculated over the full 24 hours. At a minimum caveats need to be added but, ideally, you would calculate Tx, Tn and Tm consistently for full 24 hour periods to enable cases where Tx and / or Tn fall outside of nominal day / night.

10. More details are required upon the land-sea mask used and how it is aggregated to the 5km grid. Are stations omitted based upon whether land or ocean at the native e.g. 5 minute resolution of the mask but then the gridded product produced for all 5km gridcells that contain any land? The method is not reproducible or understandable absent of such details.

11. Section 3.3.4 is a little basic compared to state-of-the-art spline fitting techniques as used e.g. in HadEX or CRUTS. Why has such a relatively speaking simple approach been used as is described in 3.3.4 and how does it compare to spline fitting or other techniques?

12. The mean temperature discussion in 3.3.5 would benefit from discussing various pieces of literature about how to create mean temperatures and the difference between the mean of a 24 hour period, the mean of max / min and various other alternative methods. There have been several papers on the different ways to calculate means and the random and systematic effects they can have. For the purpose of your dataset the key issue is whether the different ways you have done so may impart systematic or random effects. This can be ascertained from a representative (climatologically) set of well-sampled stations being deliberately degraded to calculate the mean from the 24(+) instantaneous values and then various approximations. This can then at least bound your uncertainties.

13. I would expect considerably more in the data availability section. It is grossly insufficient to just point to a website. You must describe things such as the data format(s) (ideally there should be at least two to cater to a range of users), any

software tools available, version control, version retention policies, whether files are available as spatial or temporal aggregations, whether the files contain uncertainty information, whether they are univariate or multivariate. Otherwise no user is going to go to that site on the off chance. You are not really helping yourselves to advertise the availability if all you do is point to a URL. Users need to know what they are being offered and how well it is ultimately managed.

14. seNorge2 and seNorge2018 are as I understand it consecutive versions of a single dataset with 2018 replacing 2. You should use just the 2018 version as the other version has been deprecated. Having also reviewed that paper and in particular considering their discussion it would furthermore not be appropriate to treat that product as truth given the substantial caveats made around interpolation in data sparse areas of complex topography. Therefore the seNorge2018 validation must be noted to be conditional on the verity of that product which cannot be assured a priori. You cannot treat either seNorge version as 'truth' which significantly inhibits their value as a means of validation. There are many other national gridded products and comparison to a range of such products e.g. the UK analysis from the Met Office may be instructive and reduce the dependency upon a single product which may itself contain substantive systematic biases.

15. I cannot make head or tail of Figure 11 from the caption and the individual images. The caption needs expanding and the panels given better titles. I cannot understand why the panels in each case vary in the manner they do and it is non-intuitive to me and therefore will be so to your readers. A figure extending over several pages is also barely legible. You would probably be better to concentrate upon 1 or 2 events that could be given more space in the main text and place the remaining cases in supplemental materials.

16. Lacking in entirety is any meaningful discussion section which may highlight caveats, future work avenues, future priorities, synergies with other Copernicus and global activities (such as WMO) that may yield improvements etc. etc. I would expect a substantive discussion section which covered such points.