Review of the paper *"EMO-5: A high-resolution multi-variable gridded meteorological data set for Europe"* by Vera Thiemig , Goncalo N. Gomes, Jon O. Skøien, Markus Ziese, Armin Rauthe-Schöch, Elke Rustemeier, Kira Rehfeldt, Jakub P. Walawender, Christine Kolbe, Damien Pichon, Christoph Schweim and Peter Salamon

The paper describes EMO-5, a collection of gridded datasets for seven daily aggregated variables. The dataset covers Europe. The time period ranges from 1990 to 2019 and the grid spacing is 5 km. I understood that originally the aim of the developers was to provide the atmospheric forcings to a hydrological rainfall-runoff model, then EMO-5 was also distributed as a stand-alone product and made available to the general public. For this reason, the authors consider precipitation as the most important of the seven EMO-5 variables.

The input data is a blending of point observations, grid points from both observational gridded datasets and reanalyses. This implementation choice is one of the most interesting aspects of the work and -at the same time- one of its critical points.

Three different statistical interpolation schemes have been considered for EMO-5 production and SPHEREMAP was chosen as the best one. At the same time, the comparison of the three schemes allows the authors to present an evaluation of the products based on leave-one out cross-validation. The authors compare EMO-5 against seNorge over the Norwegian mainland. Then, 15 extreme events have been presented as case studies and EMO-5 has been evaluated qualitatively over them.

I believe that the topic of the paper is of extreme interest. Data fusion experiments, such as the one presented here, meet the demand of many users of weather data, which is to bridge that last mile that allows them to use the best input in a given application. In fact, the best input is often a combination of data sources. The weak point of the paper is that in its present form many of the claims the authors make in their conclusions are not sufficiently reflected in the results presented.

In conclusion, I think the manuscript can be published once the following comments are resolved.

Major Comments:

- *What is a "best" estimate for you?* In the paper, the concept of best estimate (or best guess) is often mentioned. The authors should explicitly state, in the Introduction, in what sense a predicted value is "the best" for them. This "declaration of intent" can be useful for some users to decide if your dataset is suited for their applications. From the text, It seems that in data dense regions the best estimate is an average of the observations, while in data sparse regions the accuracy and precision of the representation will be the same as for ERA-Interim. Note that other definitions are possible, for instance it might be that the best prediction is the one that allowed the authors to get the most realistic output out of a hydrological model.

  As a further remark that shows the importance of understanding what is a best estimate for you, I may discuss the reasons that led you to discard ordinary Kriging as the optimal interpolation method. Ordinary Kriging implements a spatially consistent low-pass filter over the entire domain. The properties of the filter vary in time, because they are defined by the time-varying semi-variograms. However, for a given day, ordinary Kriging provides results that are much more comparable between different regions across Europe because the “smoothing” makes them more comparable. On the other hand, the smoothing filters out the local scales that one can represent in data-dense regions, which is where the 4000 observations used for evaluation most likely are located. The spatial coherence provided by Kriging is not among the properties of the optimal “best estimate” that is interesting for the authors, then ordinary Kriging is discarded.

- *Validation approach.* The validation focuses almost exclusively on precipitation. The other variables are not considered, except for Table 4. A further weakness of the current validation approach (Table 4) is that the whole domain is considered simultaneously and only aggregated results are presented. However, as written in the previous comment, the predicted values do have very different meanings across the domain. A distinction between dense- and sparse- data regions is required to have an idea of the actual quality of the fields. The validation against seNorge is difficult to interpret (and representative of precipitation only): in some regions precipitation is smaller than seNorge, in others it is higher. Then, the conclusion is that EMO-5 is in good agreement with seNorge2, which seems not to be the case for some climatological indices you present (i.e. CDD, CWD in Fig. 10). Finally, the evaluation over the case studies is only qualitative: for instance, does the spatial distribution of the observations have an influence on the way extremes are represented?

  I may suggest a different approach that could solve most of the problems I see with the current approach. I have the impression that the aim of the validation is to show that EMO-5 is fit-for-purpose for hydrological simulations. If this is the aim, I think this should be stated explicitly. The authors should present their results on precipitation separately for observational dense and sparse regions. Furthermore, the authors should include a section on “indirect validation” of the meteorological fields, where the output of a hydrological model that is using EMO-5 fields as forcings is evaluated. In this way, all the seven variables will be evaluated simultaneously on their potential in providing useful results for the application they were designed to serve.

- *Interpolation uncertainty.* I have never mentioned interpolation uncertainty until now. I think that the authors do not provide enough evidence that the uncertainty fields they provide are actually characterizing the actual uncertainty in a reliable way. The only results they show is Figure 4, linked to a short discussion in Sec. 5.1. A proper validation of the uncertainty field would require the presentation of much more results (see e.g. Wilks (2019), Chapter 9). In my opinion, the status of the EMO-5 uncertainty products are “still under development” and they can be the topic of future works. Furthermore, if I understand well, the uncertainty is provided for all of the 7 variables but in the paper it is presented only for precipitation. In its present form, I suggest the paper should not mention the interpolation uncertainty as a product available to the users.

Minor comments:

- *Verification scores used in Sec. 3.2.* The verification scores considered in the cross-validation exercises are not well suited for the validation of precipitation, especially over very large areas. I would have used something like the Nash–Sutcliffe model efficiency coefficient (NSE) or the Equitable threat score (ETS) for some significant thresholds (e.g. 1 mm/day, 10 mm/day,....).

- *Interpolation methods.* It is not clear why SPHEREMAP is used for all variables. Since IDW is better for some variables, then why not use it for those variables?

- *Quality control.* The division between “suspect” and “good” data is not used at all, since EMO-5 uses both data. Then, why are the QCs flagging some data as suspicious?