Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2021-15-RC2, 2021

Comment on essd-2021-15

Anonymous Referee #2

Referee comment on "The three-dimensional groundwater salinity distribution and fresh groundwater volumes in the Mekong Delta, Vietnam, inferred from geostatistical analyses" by Jan L. Gunnink et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-15-RC2, 2021

Dear Authors,

I'm a geostatistician with no expertise on groundwater dynamics, then my review is on the geostatistical methods used.

The paper describes the work done to produce datasets on the fresh groundwater volumes in the Mekong Delta. The article is well structured and the presentation is clear and concise. The Figures and Tables complement in an excellent way the text in the manuscript. As far as I can judge, there are no major flaws in the statistical analysis, though some more work on the text is required. The conclusions are well supported by the results. The accuracy and precision of the results are reasonable and the Authors provide uncertainty estimates of their predictions.

The presented interpolation method is not original, the Authors state in the Introduction that "To our knowledge, this is the first time that the fresh groundwater volume on a large, delta scale is estimated by means of the geostatistical interpolation technique indicator kriging", which is something I cannot judge. The final dataset is publicly available and this is a great merit of the authors. It would be great if the authors will regularly update this dataset. They should state more on their future plans in their conclusions.

The study is valuable. My advice to the editor is to consider it for publication after a revision addressing the comments that follow.

Comments:

- I have been able to download from Zenodo and extract data from the netcdf files with tools such as "ncview" and "ncdump". The two files are: hydrogeology_Mekong_Vietnam.nc, TDS_Mekong_Vietnam.nc. The second file contains the two variables: Etype_estimate_TDS, Probability_TDS_smaller_1g_L. The data are made available on grids with horizontal spacing of 1 km and vertical distance of 5 m.

- Sec. 2.1. I like Fig.2, very informative. One weak point, It is difficult to link the information in the netcdf files with your description in Sec. 2.1. For instance, from Fig.2 it looks like you are distributing much more data that the variables that are actually present in the files. Please, make explicit the connection between the data presented in Fig.2 and the variables in the files on Zenodo.

- Sec. 2.2. This is a rather general description of the methods, where you made the reader aware of the names of the kriging schemes you have decided to use. I agree that ordinary kriging is not well suited for skewed variables. Lines 140-141 are simply wrong, the equation reported has problems with the parentheses and it does not make too much sense to me, because you have not explicitly stated the meaning of "z" with respect to "Z". What does it mean that z is conditional to the number of observations "n"? I can guess its meaning, though I think that you'd better describe this equation or you remove it from the text. (in the first place, is the equation needed?)

- Sec. 2.4. In your study, there is a small number of high quality data, which are presented in Sec 2.4.1, and a much more numerous set of what you call "soft" data. The soft data are unevenly distributed in clusters here and there along the Delta. You implicitly recognize that this unbalance between high and low quality data can introduce bias in your estimates, then the soft data are spatially aggregated. How different is the quality of the estimates from the soft data to those of the high-quality data? I understand that estimates from the soft data rely on some assumptions you make, but how good are these assumptions. Can the mismatch between different data sources introduce biases in your results? Please, elaborate more on these points.

- The unbalance between the small set of high-quality data and the large sef of soft data is so massive that the semivariograms are probably determined by your soft data only. This may explain the evident "bull's eye" effects, which are present in Fig.11. In some regions, where more "soft" observations are present, the fields are reconstructed in a more realistic way. In data sparse regions, red or blue "bubbles" can be seen here and there (see the attached image). Can you comment more on the impact on the inhomogeneities in the spatial data distribution on your results? In the attached Figure, one can clearly see sharp transitions between nearby gridpoints, which I believe is a consequence of indicator kriging. I think the Author should explicitly adress this issue and include some comments about it in the text. For instance, should experience users expect spatial continuity of the final fields?

- Sec. 3.2.1. Please move lines 295-299 (discussion on Gaussian transformation) into Sec.2.2.

- Sec. 3.2.1. Lines 283-284 "So, the geostatistical analysis and interpolation of TDS is performed within the volume of each individual aquifer/aquitard by using data that is located within that unit." This is an extremely important information, you should state that your spatial interpolation is performed this way in Sec. 2.2. For instance, by looking at the fields of "Etype_estimate_TDS" in your netcdf files, one can see sharp transitions between different spatial trends. This could probably be explained by this choice you made.

- Sec. 3.2.2. Please comment on the characteristic horizontal length scales inferred from the semivariogram. Are they comparable to the average distance between your high-quality observations? Are they closer to your spatial aggregation of the soft observations? Remember that the real (effective) resolution of the predicted fields are determined by your semivariograms, not by the grid spacing.

-Sec. 3. Please make clear to the reader which are the variables that are stored in the netcdf files, by making explicit reference to them when you mention them in the text.

Please also note the supplement to this comment:
https://essd.copernicus.org/preprints/essd-2021-15/essd-2021-15-RC2-supplement.zip