reply:hess-2023-15-MS

May 2023

1 First reviewer

The paper "Data-driven estimates for geostatistical characterization of subsurface hydraulic properties" by Hesse et al. is a nice try to build a comprehensive data set of variogram parameters that could be used as reference values or as prior distributions for cases in which there are not enough data to construct an experimental variogram from the available observations. While the proposal seems interesting, I have many reservations about the paper's contents and how the results are obtained and analyzed.

The reviewer presents long and thorough discussion of our manuscript, which raises a number of points considered critical. At least some of these more critical are, in our opinion, caused by a misunderstanding of our aims and the methodology employed here. To avoid such a misunderstanding in the future, we will revise our manuscript and explain the points more accurately. Others are important and we will address them below as well as in the revised version of the manuscript.

First of all, pretending to find the "universal" variogram parameters distribution is a little bit pretentious. The analysis is made from a purely mathematical/computer science perspective without considering that variograms relate to the underlying geology. In this sense, one would not expect variograms computed in fractured media to be the same as those computed in sedimentary formations. Also, and very importantly, a parameter such as hydraulic conductivity is not expected to vary in space smoothly under any circumstances; for this reason, the use of Gaussian variograms or, in general, variograms with low roughness should never be advocated (unless they are accompanied by a nugget effect that would break the smoothness imposed by those variograms).

This comment presents several of the misunderstandings in our opinion. First, at no point do we use the term "universal" for the derivation of our variogram parameters. Quite the contrary! We make it clear numerous times throughout the manuscript that the parameters derived in our study strongly depend on both the model and the data being used to derive them. Second, of course the estimate would change to some degree if the data set would be split into, say, fractured and porous media. We did, for example, recently publish a manuscript where we investigate the impact of such a re-clustering of the data on a Bayesian estimation procedure [Kawa et al., 2022]. However, in this study, we determined that with the given sample size, only a modest gain in uncertainty reduction was achieved. We therefore, argued in the manuscript, among other things, for an increase of available data for subsurface characterization. Since the sample size used in this study was somewhat smaller than the data set used in the aforementioned study, we did not attempt to split the sample even further. We discuss this problem of accuracy vs. precision, which affects every statistical analysis in the discussion. Thirdly, we do explicitly not advocate for the use of the Gaussian model function. In fact, we do not advocate for the use of any model function in particular. The choice of variogram model functions in our study was only motivated by the fact that they are regularly employed. We do not, can not and don't want to make any comment on whether their use is justified or not. If the reviewer wants to argue in favor or against the inclusion of any model function based on the above criteria, we would happily include or exclude any model function. In fact, following the comment of two other reviewers, we did include and re-did the analysis with respect to one other model function, namely the Truncated Power Law variogram.

Next, the terminology used departs sufficiently from the one used in standard geostatistical literature (see, Journel and Huijbrets, 1978; Isaaks and Srivastava, 1989; Deutsch and Journel, 1991; or Goovaerts, 1997) to make it very confusing, and, at some points incorrect.

The terminology we follow in our manuscript follows the notation we used in Müller et al. [2022]. This notation in turn is a compromise between different notations used in the wider geostatistical community since this is the target audience for this Python package. We agree, that some notation decisions could be justified more and we will update the manuscript accordingly. Further details will follow in the responses below.

Let's dive into more detail about my reservations. The variogram is a vectorial function that depends on the separation distance vector h; only in the case of isotropic variograms in all directions does such a function becomes a scalar one. The definition in section 2.2 and the analysis thereof is made assuming that the variogram is a scalar function depending only on the modulus of h. This is a significant conceptual flaw, which is not cleaned with the anisotropy analysis made in section 3.3. The expression of the variogram function in line 135 is incorrect. You cannot separate the variance and the nugget effect. The variance of the random function model is (n+sigma2) (using the notation in that section), and the covariance is the variance minus de variogram ((n+sigma2)-gamma). The authors should recall that, generally, the variogram is modeled as a sum of variogram functions, each one contributing their share to the total variance (or variogram sill). They have chosen to model it as the sum of two variogram models, a nugget effect (with a contribution of n) and another parametric function (with a contribution of sigma2), which is perfectly OK; however, the authors should acknowledge that there could be fitting with more than two components. What is not OK is much of the later analysis in which this fact is disregarded: many of the findings are due to this fact, the fitting of a nugget plus another variogram, and some of the conclusions are said to be due to, say, the type of parametric model used, disregarding that a nugget had been used to fit the variogram.

As we said in the manuscript, the majority of data was already processed data, ie., in the form of empirical variogram clouds. In these cases, we therefore followed the discretion of the original authors who produced these variograms. We call them secondary data, and in the manuscript we do fully discuss the ramifications of their use. However, the reviewer makes an important point regarding the use of primary data, which needs to be better discussed in the manuscript.

Very important, sigma2 is not the variance of the process! It is only the contribution of one of the two structures fitted to the total variance.

The variogram computed as a function of the modulus of the separation vector is referred to as the omnidirectional variogram. It is generally used during the modeling process to assess the nugget effect and the sill of the variogram and to have an idea of the average range from the different directions, but it is not a variogram to be used in practice for anisotropic media. Variograms are vectorial functions, and as such, their value depends on the vector orientation reflecting the anisotropy in the correlation patterns of the variable under analysis. Such anisotropy is characterized by an ellipsoid (in 3D) or an ellipse (in 2D), which, in principle, may be arbitrarily oriented, although, in practice, it is assumed that one of the ellipsoid axes is oriented vertically and the other two in the XY plane, not necessarily parallel to the Cartesian axes. The lengths of the semiaxes of this ellipsoid are known as the anisotropy ranges and are crucial to characterizing the spatial continuity of the attribute. The authors disregard the vectorial nature of the variogram and introduce the "characteristic length" as one of the parameters of the parametric variogram components. The characteristic length, 'l' in the equations in the lines between 135 and 150, is not the range and may induce mistakes. In standard geostatistics, the range is defined as the distance at which the variogram value reaches the sill, or for variograms that are asymptotic to the sill, the distance at which the variogram reaches at least 95 % of the sill. This means the range will equal l for the spherical variogram but 3l for the Gaussian or exponential ones. Comparing the value of l for different variogram functions is like comparing two different quantities.

We wanted to keep the section about the theory and assumptions behind the used variogram formulation quite brief but we agree that certain aspects could be made more clear. First, we are mostly following the formulations of Rubin [2003] where a "non-dimensional" distance is used in the variogram formulation to reduce the spatial variogram to the omnidirectional one. For this, the lag distances along the main axes of correlation are rescaled with the respective anisotropy ratios and the axes are rotated to follow the principle axes. This was explained in detail in the paper about GSTools [Müller et al., 2022]. We will add some more explanation about this in the manuscript. Second, the criticism on the term "variance" is understandable and we will substitude the term "variance" with "correlated variability" as used by Rubin [2003]. There, this variability is also denoted with σ^2. The reviewer is right, that the total variance (and/or the sill of the variogram) s is then the sum of the sub-scale variability

(n) and the correlated variability (σ^2): $s = \sigma^2 + n$. GSTools includes the nugget in all models as a separate parameter since the pure nugget model would be the only one with a totally different parameter set. Cressie [1993] also includes the nugget in the model examples. Third, GSTools implements anisotropy by using rotation angles and anisotropy ratios that unify the variogram formulation with the omni-directional scalar variogram function by using the non-dimensional distance h as just described. We will add a paragraph clarifying this. We also do agree with the reviewer that comparing parameters between different model functions can be very elusive due to conceptual differences between these functions and the role these parameters play. This is why we constantly warn against the simple transfer of results derived for one model function to another and make it clear that every statement we make is contingent on the function being used.

I am very concerned with the many published theoretical papers in which the variogram is assumed to be Gaussian. This assumption is always a matter of convenience, especially in those papers more focused on analytical approaches, disregarding the fact that Gaussian variograms make only sense for variables that vary smoothly in space, such as the thickness of a layer, but never should be used to model a parameter characterized by a very large spread with significant short-scale variability like the hydraulic conductivity.

We are aware of the long standing debate regarding the Gaussian model function and the high smoothness it introduces into the modelling process. We explicitly mentions this topic and how it may be detrimental to the applicability of said model function. However, our study is not intended to make any statement regarding this issue. The only criterion for inclusion was whether the model was regularly used. This is certainly the case for the Gaussian model, however justified one may think this is. As such, we do not consider this inclusion of this model function to be an endorsement in any way.

It is not clear if the variograms analyzed are for K or logK. Line 157 mentions K, and nowhere is logK mentioned. The variogram of K may not of interest in geostatistical studies since it is common to use a lognormal random function to model the spatial variability of K. In addition, if K had been used, my criticism of using Gaussian variograms (or variograms with low roughness) is exacerbated.

This is an important observation, which was also criticized by another reviewer. Of course, all the analysis refers to the log-hydraulic conductivity. To better convey this important fact, we revised the manuscript accordingly.

Considering the significant uncertainties associated with experimental variograms, trying to distinguish whether a stable or a Matern variogram fits better than an exponential or a spherical one is meaningless. The analysis would have benefited from comparing just the spherical and exponential variograms (plus the nugget effect). Neither the Matern nor the stable models make sense when modeling hydraulic conductivity from a geological point of view. A much more meaningful fitting would have been a power variogram, which would relate to the fractal behavior claimed by Neuman.

Following the similar comments of the two other reviewers, one of them being Shlomo Neuman, we re-did the analysis by including the a fractal model function.

More specific comments

What are observable base rates? (line 22)

Observable base rates refers to the notion that informative prior distributions in Bayesian inference should be based on observable frequencies.

What is a variogram cloud? (line 39, line 96, and many more times in the following pages)

This term refers to the empirical variogram function. We added an explanation to the manuscript to make this more clear.

The paragraph from lines 52 to 55 is meaningless.

We do not agree. In fact, one other reviewer encourage us to make the comparison of established variogram models even more explicit.

After introducing the roughness coefficient in line 166, the authors analyze the roughness of the different parametric functions used in their variogram definitions, but they fail to recognize that as soon as a nugget effect is added, the roughness of the fitted variogram is zero. This explains why, later, some of the analyses seem to favor fitting a "Gaussian variogram" when in truth, they are fitting a variogram composed of a nugget effect plus a Gaussian component. The nugget removes the implicit roughness of the Gaussian variogram, justifying its application in practice.

This is a very important observation. What we saw in our study was that the Gaussian model can describe very well empirical variogram clouds. So from a pure fitting perspective, its use seems justified. We did not further comment of this but simple left the observation as is. The observation by the reviewer could be an explanation as to why the model seems to perform so well in this regard, despite the shortcoming often disucssed in the literature.

When introducing section 3, the authors discuss the length scale and vertical and horizontal anisotropy. For the first time, the authors recognize that anisotropy may be an issue, but it is poorly defined. The length scale could be the larger range, and the horizontal and vertical anisotropies could be the ratios of the ranges in the other two principal directions with respect to the largest range. But this is not what the authors analyze; they focus on the range of the omnidirectional variogram, plus the ranges in some arbitrary directions in the XY plane (not necessarily in the directions of the principal ranges) and the range in the vertical direction.

As said in the manuscript, the majority of data that analyzed anisotropy was secondary data, ie., the authors already did separate the data according to this criterion. Howver, the anisotropy for the horizontal directions is one result where primary data form a substantial amount of the overall data. We therefore would like to re-do this analysis. In general, what we saw from secondary data, the practice was to align the x- and y-direction with the physical directions indicated by the measurement campaign and not be some possible anisotropy observed in the data. This may not be the best way to perform this analysis, but we had to work with the data we have, not with the data we wished for.

In section 2.3, it would be nice to discuss the fitting procedure used to get the variogram parameters.

For all the analysis, we used the tools provided by GSTools. The estimation of variogram parameters is, for instance explained in Müller et al. [2022] and one the website of the project. What aspects of the fitting procedure would be relevant to mention explicitly in the manuscript?

In the fitting algorithm, the authors should use an algorithm capable of fitting a fully vectorial function and attempt to identify not just a single omnidirectional range but actually the whole anisotropy ellipsoid (or ellipse), that is, the three (or two) principal ranges and their orientations.

Again, the majority of the data was secondary data. In no case, where primary data was available did we simply fit an omnidirectional variogram. We did acknowledge the shortcomings of the used data in this regard and its ramifications on the topic of Bayesian inference, namely the fact that it increases the uncertainty in the prior distribution. However, the point raised here is important and needs to be better explained in the manuscript.

The statement in line 215 does not seem to be much justified. Why should all models fit data? There may be physical reasons why one kind of variogram is more appropriate than other. In some cases, as already mentioned, some variogram models are inconsistent with the sample data (such as the Gaussian variogram without nugget effect to fit the spatial variability of hydraulic conductivity).

Of course, not all models should be better at any given data set. However, this was only a statement in the aggregate.

Line 223. The authors realize that the nugget effect is responsible for the similar results yielded but the different parametric models.

We are not sure what this statement means.

The matrices of scatterplots must be explained. What does each point in each scatterplot represent? The diagonal scatterplots make no sense unless they represent something different from the rest of the scatterplots. If they show the same parameters, all points in the diagonal scatterplots should fall in the diagonal at 45 degrees.

We agree that the scatterplot is not sufficiently explained and needs to be revised to become a better contribution to the manuscript.

The end of the sentence in line 229 is very important. It is always so! The particularities of the site under analysis should drive the analysis. Trying to find the de "universal" parameter distribution is a mistake.

Again, at no point do we attempt to find a universal law underlying all aquifer and/or soil sites. This is a severe misunderstanding of our aims. Our main aim is to gather and subsequently use a large data set to derive informative prior distributions for Bayesian inference. In addition, we try to discern typical behavior of variogram functions, which may be of use for delineating some challenges of properties found in real-world data for geostatistical applications.

The second sentence in the paragraph starting at line 230 is hard to understand.

We revised this sentence in the manuscript.

Unclear what is meant in the sentence that starts in line 251 to the end of the paragraph.

We also revised this sentence in the manuscript.

End of page 12: A power variogram could explain this behavior.

Following this and the other reviewers' comments, we included a power-law variogram into the analysis.

In the paragraph in line 260, the relationship between the lengths should be at the 45-degree line when comparing spherical and Gaussian and at the y=3x for the comparison with the spherical variogram.

This is an interesting point, which we consider in the revised manuscript.

Section 3.3 makes absolutely no sense. Anisotropy is a fact and should be appropriately analyzed and accounted for. Analyzing lambdax and lambday when they do not represent the actual principal directions of the anisotropy ellipse is useless.

See our answers above.

The analysis of lambdaz would make sense if compared with the largest range in the plane but not when compared with an arbitrary range calculated in the x direction.

See our answers above.

In section 3.4, I hope that the normalization of the nugget was done with respect to the total variance (n+sigma2), not with regard to sigma2. There is no reason why n could not be larger than sigma2, therefore escaping from the [0,1] limits.

See our answers above.

The nugget not only accounts for measurement errors but also for short-scale variability. The discussion in this section is poor and, at times, flawed.

We mention several times in the manuscript that the nugget represents measurement error and sub-scale variability.

Line 354 is explained because the Gaussian model needs the nugget to fit most experimental variograms, whereas the exponential does not. Consequently, the results are not an artifact of the fitting procedure.

This is an interesting observation by the reviewer. But wouldn't it still mean that it is an artifact by the fitting procedure?

In the opening of section 3.5, the authors forget that the length parameter is not enough; there is an anisotropy ellipse.

See our answer above.

Why does Figure 15 not show a histogram as before?

Figure 15 does show a histogram like the other figures before.

The discussion about survivor bias is very interesting.

We agree.

The quality of the figures must be improved.

We agree. Several of the figures still show blurring and in some cases, the font needs to be equalized.

I think a reanalysis of the data accounting for the mistakes made in this evaluation, probably reducing the number of variogram functions and avoiding the construction of the prior probability functions for the different models, would be worth publishing.

References

Noel Cressie. *Statistics for Spatial Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey, September 1993. ISBN 9780471002550.

Nura Kawa, Karina Cucchi, Yoram Rubin, Sabine Attinger, and Falk Heße. Defining hydrogeological site similarity with hierarchical agglomerative clustering. *Groundwater*, 2022.

S. Müller, L. Schüler, A. Zech, and F. Heße. `GSTools` v1.3: a toolbox for geostatistical modelling in python. *Geoscientific Model Development*, 15(7):3161–3182, 2022. doi: 10.5194/gmd-15-3161-2022. URL `https://gmd.copernicus.org/articles/15/3161/2022/`.

Yoram Rubin. *Applied Stochastic Hydrogeology*. Oxford University Press, USA, 2003.