Review comments for HESS manuscript 2020-444 by Donald A. Keefer
Bødker Madsen et al., **<u>version hess-2020-444-ATC1.pdf</u>**.

To the authors, well done. You've done a good job with this paper and made some very significant changes. I do still have a few concerns. Nothing insurmountable, but definitely issues that need to be addressed in some manner.

I also want to recommend a publication. It's a chapter by Mohan Srivastava in a book about geostatistical simulation. It's fantastic and has some perspectives about simulations that are really insightful. The whole book is very good, but some of the methods are a bit dated now. Srivastava's chapter is still as relevant as when he wrote it, and can be extended to MPS, which he and Guardiano had developed two years earlier.

Srivastava, R. M. (1994). An overview of stochastic methods for reservoir characterization. In J. M. Yarus & R. L. Chambers (Eds.), Stochastic Modeling and Geostatistics: Principles, Methods, and Case Studies (pp. 3–16). The American Association of Petroleum Geologists.

Manuscript Evaluation **<u>version hess-2020-444-ATC1.pdf</u>**
<u>Review Criteria:</u>
Scientific Significance. I rate this paper as 1.5, Good-Excellent.
Scientific Quality. I rate this paper as 2.5, Fair-Good.
Presentation Quality. I rate this paper as 2, Good.

I suggest that this manuscript would be an excellent choice for publication in HESS, if the authors address my comments to the satisfaction of the Editor. I do not need to see it again.

<u>Aspects for Consideration:</u>

1. The paper does address relevant questions within the scope of HESS. Characterization of spatial distribution of geologic deposits and redox status.
2. The paper presents novel ideas and methods.
3. The conclusions are substantial: this method offers a way to effectively model the sediment distribution and associated redox conditions.
4. The methods and assumptions are generally valid and clearly outlined.
5. The results are generally sufficient to support interpretations and conclusions. However, some of the interpretations are not sufficiently established by this study. See comments below.
6. The description of experiments and calculations is sufficiently complete and precise to allow reproduction by fellow scientists.
7. The authors give proper credit to related work. They indicate their own contribution fairly well. There are a few places where they offer conclusions that I do not feel are substantiated by either this experimental design or by the results. I comment further on this below.
8. The title is a good reflection of the study.
9. The abstract is a good reflection of the study.
10. The overall presentation is very good and clear.
11. The language is generally fluent, but occasionally not sufficiently precise. I have comments to this below.
12. The formulae, symbols, abbreviations and units are correctly defined and used.
13. I believe that section 7.4 could be reduced or eliminated, as it is misleading and seems at odds with the bulk of the literature and even earlier passages in the paper. I address this more in my comments, below.
14. The number and quality of references are sufficient.
15. There is no supplementary material.

Major comments: **version hess-2020-444-ATC1.pdf**

The authors state unequivocally that geostatistical simulation can quantify uncertainty in the geological model. This capability is overstated. Geostatistics and geostatistical simulation rely on different statistical models that are, or should be, based on data and conceptual models of likely rock property distributions. While MPS can be more effective at simulating more-realistic sediment texture patterns than other geostatistical simulation techniques, these all are only approximations. And the statistical models they are based are only constrained estimates of the properties being modeled – to the degree that the selected models reflect aspects of the real system; and, that they sufficiently sample, with high reliability, the properties being simulated. I suggest the authors adopt this perspective, and add the word, 'estimate' or 'estimating', when talking about quantifying uncertainty. 'Quantitative estimate of uncertainty' is a good example.

The authors are inconsistent in their representation of the impacts of bias, subjectivity, and independence. They often suggest that subjectivity induces biases that are always undesirable. At other times, they advocate clearly for choices or methods that induce a specific bias, often based on the subjective justification that the choice better represents a preferred conceptual understanding. This inconsistency, and more importantly the erroneous representation, are significant in that they lead to confusion about the authors' advocacy regarding these concepts, and even to how the authors understand them. On lines 84-87 they state "In fact, subjective biases are accepted as one of the weak points of cognitive geological modeling…" In this situation the authors cite Bond 2015 and Wycisk et al. 2007 for this statement. It appears, however, that the authors have misrepresented both of these papers. Bond does not make this statement in her paper, and frequently recognizes the importance of geologic knowledge for successful modeling. Wycisk et al. never use the words subjective or bias, and state that knowledge of the system is required to make any modeling successful. Biases are endemic to any form of modeling – cognitive or quantitative (numerical, probabilistic, geostatistical). Generally, you want to add a specific bias as a constraint to interpretation (e.g., texture distributions in different training images, or the proportion of specific sediment textures in a proximal proglacial sedimentary setting). The main down side of expert insight in cognitive modeling is the difficulty in elucidating the various biases that the experts use. However, expert biases can be evaluated for likely correctness, and these 'constrained' biases can be very helpful in ruling out improbable sediment distributions; both Bond (2015) and Wycisk et al. (2007) recognize this. The implication of the authors' phrasing suggests that quantitative modeling is always better and less biased than cognitive modeling. I suggest that the literature demonstrates that both can be very useful if done correctly and properly qualified – and both can be similarly erroneous and misleading if not done correctly or properly qualified. I recommend language like, 'Subjective biases are seen by some as one of the weak points of cognitive geological modeling (Bond, 2015; Wycisk et al., 2009). It can be difficult to identify and quantify uncertainties due to biases in subsurface predictions from cognitive modeling, and so these biases cannot be fully accounted for in subsequent analysis or process modeling efforts (i.e., hydrologic modeling).'

On line 1000 the authors also seem to misunderstand importance and need for independence, in a statistical sense, and how bias is intentionally inserted in MPS. The goal of a TI is to introduce a bias, subjectively generated based on a combination of geological knowledge of the area and depositional settings, and observational data of the location. The reason MPS is being recommended, rather than other categorical geostatistical methods, is that these other methods don't introduce the right biases. In fact, they impose biases that make the models more incorrect than is generally desired. MPS is a way to introduce biases of more spatial continuity and large- to medium-scale heterogeneity into simulations, in order to generate model realizations that are more biased towards patterns of heterogeneity that are observed in outcrop and modern depositional analogues. Assumptions (or requirements) of statistical independence are part of the theoretical underpinnings of the tests. However, these methods are being used as approximations of more complex systems, and these assumptions can be (and usually must be) relaxed without making the MPS simulations worthwhile.

Figure 2: Input data…Soil Map. Should be Surficial Geology Map

Figure 8 revision is a great contribution. However, the Y axis label on 8a, 8b, c are wrong. Should be a frequency or count. I'm a bit confused about how this was generated. See other comments in this section.

Figure 12, the X axes are incorrectly labeled. Should be 'percent' or maybe 'percent occurrence' or change x-axis values to be 0-1 for probability.
Figure 12b includes results of two realizations. This isn't sufficient estimator of the prior. You need to label it with some other label.

In general the figure revisions are much better. Good job on these. The patterns are much more readable and easier to follow with the discussions. A few more fence diagrams might be helpful in showing more of the internal architecture.

There seems to be a bit of misunderstanding about the relationships between the data used for modeling redox conditions and about what one of the data streams means. On lines 235-236 the authors state, "The benefit of using the two types of data is that they provide independent measurements of redox conditions." This is incorrect, these data are not independent. The redox colors of the sediments reflect long-term redox conditions which are consistent with some period of pore-water chemistry. Importantly, the authors do not mention that the redox sediment colors are indicators of long-term historic redox conditions and not necessarily reflective of the redox dynamics within the recent, human altered subsurface. This may be why they feel the two data streams are independent. This study is not structured to evaluate this independence, and prudence suggests it is better to assume some level of dependence. The benefits of using the redox colors are that there are more of those data values than there are of the water chemistry, and the authors may assume that the relatively smooth variations in lateral redox sediment colors allow more reliable interpolation of the colors to unsampled locations than can be assumed for the recent shallow water chemistry. It would be helpful if the authors could add information or interpretations on how the redox colors correlate to water chemistry. The results might show a good agreement, or not such a good agreement. Importantly, the water chemistry has a temporal dynamic that is not captured by the colors.

Also regarding the redox discussion, on lines 406-411, the authors make statements about inferring groundwater flow conditions from the redox colors of the sediments in the lowland (buried valley). However, the authors haven't presented any data that lets them make inferences about flow in the groundwater system. They are obviously able to cite other work in the area that did study gw flow, but no conclusions about flow can come from this work. Instead, the authors could say their work is, or is not, compatible with hydrologic observations and conceptual models documented by others (e.g., Kim et al., 2019). As a case in point, the authors suggest no horizontal flow because of clay-dominating conditions. However, a plausible alternate interpretation could be that soil development has provided preferential flow paths through the upper 3 meters of material across the landscape, that this network is exploited for groundwater flow, and the extended flowpaths from infiltrating up-gradient positions might provide sufficient residence time of nitrate-laden water for significantly more reduction. I am not advocating for either hypothesis, just noting that there are no data to prove or refute either of these ideas.

There is more clarity of explanation needed for the discussions about how soft data are used and processed. Importantly the authors made some very helpful changes from an earlier version. This is in lines 533-575 of the ATC1 version. I particularly like the clarification of the potential impacts of non-stationarity and how this was handled. I like the use of the three zones. That's all very good. However, I'm particularly concerned and confused about the processes surrounding figure 8. It looks like the histograms were generated from assumptions of resistivity/lithology and the resistivity grid. The authors need to clarify the uncertainties or potential for error in the assumptions made and how these relationships were applied. There doesn't seem to be any calibration or validation around these resistivity/lithology relationships. This seems very subjective. Not necessarily wrong, but it seems to involve a several significant assumptions that warrant some recognition.

Table 4: I don't understand what the upper row signifies. I feel like I understand the text general discussion that supposedly addresses the table, but I don't understand what the values mean. The first row needs a better explanation.

On line 660 the authors cite Høyer et al. (2017) in their use of 'unconditional simulation'. In fact the simulations are conditioned by the Tis, but not on any data. However, Høyer et al. don't refer to this unconditional simulation as representing the prior distribution. They used this approach to evaluate how effectively the TI constrained the realizations. While the authors are using TI-conditioned simulation for this purpose as well, it is not correct to suggest that two realizations are able to provide sufficient representation of the whole prior distribution. This is important on Figure 12b, as you suggest it is the prior, but again, only two realizations (i.e., samples).

On line 791 the authors state, '…and hard data increase the information content (lowers entropy)…'. I believe this is incorrect. This should be written as, 'and hard data **decreases** the information content (lowers entropy)…' Remember, the equation is for Shannon's information entropy, not thermodynamic entropy. (cf. https://stats.stackexchange.com/questions/101351/entropy-and-information-content). Alternatively, the authors could think of entropy in terms of bits – a la Shannon. It takes more bits to describe more complexity, and fewer bits to describe less complexity. More bits = more information content. Fewer bits = less information content.

On lines 808-809 the authors state, 'This imply that simulation artifacts are not reoccurring in simulations…' I would strike this sentence. It's not phrased properly and is incorrect; analysis of the mode, as a statistic, doesn't prove or disprove either the frequency of artifacts or a bias. If the authors want to keep a reference to these figures, the text should be fixed so it doesn't erroneously say, 'no overall bias is found'. Maybe something like, 'The algorithmic biases that create these artifacts are generating a small number of artifacts that are not significantly affecting the posterior histograms.' However, this point was already made when the authors looked at the histograms earlier in the paper, so this sentence is not necessary or constructive. The statement, 'This is clearly an unwanted side-effect of the soft data conversion…' seems like an inference based on assumptions rather than an observation based on the algorithm. If that is the case, this statement is unsubstantiated and should be deleted. Consider a hypothetical example: any additions to the distribution would have no effect on the mode as long as they occurred less frequently than the modal value. These additions could be due to a systematic error (i.e., bias).

On lines 815-832, I think this passage is mistaken. Importantly, the redox scale only has three categories. The posterior distribution showed almost no occurrence of the middle, reducing, category. Don't forget the post-glacial seds were assigned a highly reducing value. And even in other textures, there was a non-negligible proportion of reduced occurrences. These all probably contribute to generate a higher entropy at land surface. The prior and TI for redox were strongly biased towards reduced conditions at a certain point in the subsurface. This is reflected in the low entropy. I don't see this as counter intuitive.

On line 959, the authors state, 'Although the small size of the TI may pose a problem for reproducing the intended variability…' They have been a bit inconsistent with how they refer to the Tis in terms of their level of detail, the intended goals with respect to the level of detail the authors are trying to represent, and with the recommendations on how much detail to put into a TI. These decisions do involve a bit of professional judgement, more clarity is needed about the goals for the TIs, what the authors tried to do with the TIs, and how they are accommodating the limitations that the choices caused.

On line 1016 the authors state, '…allows the quantification of uncertainties in the input data…' No it does not do this. It is not correct to use statistical models that were created by a data set to evaluate the uncertainties of that same data set. There isn't that much real data to begin with, so there really isn't enough data to prove any distributional assumptions or models. The manuscript details how the uncertainty in the geologic sediment and redox condition distributions have been quantitatively estimated.

Technical Concerns:
I present here a list of grammatical and minor technical concerns, listed by line number from the **version hess-2020-444-ATC1.pdf**.

Line Number: Comment
8: 'nitrogen (N) losses'…you are only dealing with one loss pathway, leaching. should use 'nitrogen (N) leaching' rather than 'losses'.
13-15: 'and 2) information of lithology and redox conditions…' this list is presented incorrectly. it should be: '2) lithologies from borehole observations, 3) redox conditions from colors reported in borehole observations, and 4) chemistry analyses from water samples.'
23: '...and conditional data'…should be: '…and conditioning data…' See Straubhaar, Renard, Mariethoz, 2016.
26: maybe 'consistent' rather than 'coherent'?
30: 'which may be fundamental to better understanding of the retention of the subsurface…" might read better as: which may lead to a better understanding of nitrogen (N) fate in the subsurface…"
34: maybe 'loss' instead of 'escape'…leaching is not ideal here because your citations also include runoff losses
50: questions: Do you mean nitrate concentrations, instead of conditions? If 'conditions' is correct, I'm not sure what that means. Can you clarify. Assuming concentrations is correct, shouldn't it also be 'leaching concentrations'?
56: 'contaminants: 1)' should be something like, "contaminants. Modeling approaches have included: 1)…"
66: I would suggest a better wording than, '...space thus requires…" might be, "…space would benefit from more-detailed…" Mostly, this seems to imply that your 25x25x2 is sufficient to capture all detailed heterogeneities in redox conditions, which isn't accurate (it's probably also not what you mean…it just seems implied in the current wording).
80: 'is attributed to uncertainties such as' should be 'contains uncertainties from sources that include'
84-87: "In fact, subjective biases are accepted as one of the weak points of cognitive geological modeling…" The authors cite Bond 2015 and Wycisk et al. 2007 for this statement. The authors have misrepresented both papers. Bond does not make this statemente and recognizes the importance of geologic knowledge. Wycisk et al. never use the words subjective or bias and state that knowledge of the system is required to make any modeling successful. See comments above about the authors' use of subjectivity, bias, and uncertainty concepts.
105: the geostatistical simulation doesn't "quantify uncertainty in…data". It does allow you to quantitatively estimate uncertainty in the unsampled locations, but not the sampled ones.
106: possible realizations of the subsurface…not a quantitative measurement of uncertainty?
235-236: "The benefit of using the two types of data is that they provide independent measurements of redox conditions." No that is not correct. The colors clearly reflect long-term redox conditions which are consistent with the pore water chemistry. So they aren't independent. See comments above.
324-325: '…based on geological event chronology.' should it be '…based on sediment heterogeneity and geological event chronology.' ??

406-411: You haven’t presented data that lets you make any inferences about flow in the groundwater system. See comments above.

458: ‘5.4 Conditional data’ should be ‘5.4 Conditioning data’ There are conditional simulations and conditional realizations that are conditioned to the data. This means the data become condition**ing** data.

533-575. I like the revisions to this portion of the manuscript. See comments above.

586+ I’m a little concerned about the term, ’soft probability’. I realize Caers and Mariethoz may use it, but qualifying the word ‘probabilities' with ’soft’ seems inaccurate or at least unclear. ’Soft data probabilities’ is more intuitive and seems more correct. I notice you use 'soft data probability’ on page 10. Please check on the usage of this term and consider either using ‘probabilities’ or ’soft data probabilities’. I leave it to your discretion.

587: 'resistivity grid in Figure 7’ probably makes more sense as, ‘resistivity grid from Figure 7’

614-616: While the mode shows layers in the buried valley, which might be bad, these same voxels show a lot of entropy, which implies a wide distribution over these cells. Which is good. You might make this point in the text. I think it’s helpful analysis discussion for readers.

623: Table 4. I don’t understand the upper row. I understand from the text, what general topic it’s addressing, and I think it is probably helpful. However, I don’t understand what the values mean. The first row needs a better explanation.

633: ‘In direct sampling, the nodes in the simulation grid…’ You never discuss this term (direct sampling) nor contrast it with alternative MPS simulation modes. Maybe back up when you introduce MPS, put in a couple sentences about direct sampling and snesim and that your version uses direct sampling. Just for context here.

635: can you add a citation or citations that you used to generate this list of parameters?

660: Høyer et al. (2017) do not refer to this unconditional simulation as the prior distribution. See comments above

667: ‘only occur near the surface in accordance with the TIs, but more infrequent than portrayed.’ …with some corrections… ‘only occur near the surface in accordance with the TI, much more infrequently than portrayed.'

686: ‘…incorporation of conditional data…’ should be, ‘…incorporation of conditioning data…'

709: ‘conditioned data’ should be 'conditioning data’

718: ‘…the soft data is particularly dominating the realizations…’ could be ’…the soft data is the dominant constraint on the realizations…'

726: ‘…This coherency explains…’ could be ‘…This consistency explains…'

740: question: why don't you more completely filter these oxic artifacts? even in post-simulation processing?

778-779: ‘…which together represent the full posterior model.’ I think it would be more correct to say something like, ‘…which together are used to represent the posterior model.’

787: ‘…indicating that other categories are almost equally probable…’ probably should be something like, ‘...indicating that these voxels have a near uniform distribution among several different categories.'

791: ‘…and hard data increase the information content (lowers entropy)…’ should be, ‘and hard data decreases the information content (lowers entropy)…’ See comments above.

808-809: I would strike this sentence. See comments above.

815-832: I think you’re mistaken here. See comments above.

836: ‘examples of mapping redox conditions with multiple-point geostatistical simulation…’ I don’t think the use of mapping is appropriate here. It’s the wrong audience for this use. A better wording would be, ‘examples of simulating redox conditions with sediment-texture distributions using multiple-point geostatistical methods…’

848: ’The spatial variability of the TIs is well represented in the prior realizations…” Not really. You only generated two realizations, which is insufficient to estimate the prior distribution, and the prior distribution that you list is not a good estimate of the redox TI - the green zone (reducing) is not well simulated at all.

848: ‘conditional data’…’conditioning data’

854, 857: ‘conditional data’ …’conditioning data'

847-865: Nice job on this section. It’s a good contribution.

868: ’The random path have a tendency to underestimate soft data…’ ? You later suggest this is why the
877: ‘randomly converted a fraction of the soft data…’ can you specify the fraction? You are defining a novel method and this would probably be helpful to people.
939-961: 7.3 ’The inclusion of geological mapping experts in the creation of ThIs introduces modeling subjectivity.’ See comments above.
959: Although the small size of the TI may pose a problem for reproducing the intended variability…” earlier, you stated that you wanted a generalized model that didn’t fully capture all of the possible geometries. can’t then make poor fit a criteria.
1000 “…would ensure independence of information.” See comments above.
1003: conditional >> conditioning data
1007: computationally feasible… 80 seconds per realization is trivial. Could have done many more than two realizations to estimate the prior.
1016 7.7 …allows the quantification of uncertainties in the input data…” No it does not. See comments above.