Rebuttal

Dear editor and referees,

thank you for the time and effort you put into reviewing our manuscript. In the following, we provide point-by-point replies (in blue) to the comments (in italic). While line numbers in the comments refer to the previous manuscript version, mentioned lines in the response relate to the revised manuscript. Attached to the response is a marked-up manuscript version with tracked changes.

With kind regards,

Alraune Zech
on behalf of the author-team.

Editor comment:

Dear Authors:
Your revised documents, together with your rebuttal, were evaluated by four reviewers, three of them had already participated in the discussion step of the journal. Three out of the present reviewers evaluated fairly well the scientific quality of your revised paper and the presentation of the relevant results. Instead, the scientific significance of your study received contrasting comments, and one reviewer was quite critical with respect to both the scientific significance and quality of your study.

After having read again all the documents pertaining to your submission, I recognize that you put a great effort into improving the original paper, by providing also satisfactory responses to all of the comments and concerns raised during the discussion step. However, I also concur with the present reviewers that even this revised version is still not in a shape to be published in HESS and should require some additional revisions.

By sharing the comment from one reviewer, I do believe that the present contribution is definitely not the only example of a hierarchical approach to transport in heterogeneous porous materials and soils, and it will not be an isolated case, I guess (to the benefit of Science). Anyhow, I do suggest the Authors should carefully consider revising their paper according to the concerns and points now raised by Ref's. #2 and #3. Moreover, due attention should be given to all of the comments received by Ref. #4.

Therefore, you are invited to upload new revised documents, together with point-by-point replies to all of the comments received by the reviewers. Should you disagree with some comments, please explain why clearly.

We followed the advice of the editor and revised the manuscript accordingly to the comments raised by the referees, particularly those of Ref. #3. While we adapted the manuscript to the criticism raised by Ref. #4, we see a significant discrepancy in the referee’s and our perception of the work, including its purpose and scientific methods. We provide detailed responses to all points raised by Ref. #4 within this rebuttal.

Referee #1 Evaluation: *For final publication, the manuscript should be accepted as is.*

Referee #2 (X. Sanchez-Vila) Evaluation: *For final publication, the manuscript should be accepted as is.*

Referee #3 Evaluation: *For final publication, the manuscript should be accepted subject to* ***minor revisions.***

This is an interesting paper trying to prove that stochastic groundwater modelling is possible. The authors discuss a modular approach to incorporate heterogeneity into groundwater flow modelling. Then they apply to one of the experiments in the well-known MADE site. I recommend publication after some "minor" corrections, in the sense that they will take little time to implement, but "major," in the sense that they lower the author's claims. I like the paper because it shows that stochastic modelling is possible. I do not like some of the claims and discussions. I do not like that the model is two-dimensional, either.

We want to thank the referee for his positive evaluation of our work. We appreciate the time and effort he put into reviewing our manuscript. The paper will benefit from revising it according to his constructive comments. We toned down the claim on originality with respect to hierarchical aquifer modelling and specified the purpose of the study alongside: we aim to provide an easy-applicable conceptualization for integrating heterogeneity quantitatively into models in line with the referee's statement "*that stochastic modelling is possible*". To underline that, we also made the numerical code for generating random binary inclusion structures public with reference provided in code availability section.

MAJOR CORRECTIONS

1) *There is nothing new in this modular approach to incorporating heterogeneity into aquifer modelling. Any claim of originality in this respect should be toned down, and references to similar approaches in groundwater modelling or reservoir engineering included. (A few references implementing this concept could be Damsleth et al., 1992, Huysmans and Dassargues, 2009, Neto et al., 1994, Proce et al., 2004, to cite a few dating back to the past century). Please, rewrite lines 64-65 with proper referencing.*

   We reformulated the abstract and other parts of the manuscript (particularly regarding the use of "novel") and specified the advantage of the modular approach we present.

   We reformulated (previous) lines 64-65 and integrated (provided and additional) references.

2) *Already in the abstract, the authors claim that their model is constructed with as minimal data as possible. This is clearly not so in the description of the application. The amount of data used is substantial and hardly available in most sites. Even if some of the data are not used as conditioning data, they are needed to infer the different parameters of the nested heterogeneous models.*

   In the course of the manuscript, we introduce several conceptual models for heterogeneity structure, with different levels of observation data requirements:

   1) deterministic (module A): piezometric surface map, pumping tests

   2) deterministic + binary (modules A+B): piezometric surface map, pumping tests, few flowmeter logs

3) deterministic + binary + random log-normal (modules A+B +C): piezometric surface map, pumping tests, multiple flowmeter logs (for geostatistical analysis)

While we agree with the referee on the aspect of available data for the third conceptual model (modules A+B +C), we think that the first two are based on a decent amount of field data which is often available at field site. Head observations and a few pumping tests are rather standard. While flowmeter logs might not always be available, there is often some geological information on contrasting layering of sand and clay. Nowadays, few depth profiles are easy to achieve from direct-push injection logging (DPIL)/hydraulic profiling (HPT) or cone-penetration tests (CPT).

However, we see that this was not outlined properly. For clarification, we specified that the level of data requirement is indeed high for the concepts including module C. Also, we agree that the word “minimal” is misleading in this context. It was thus eliminated in the text and sentences were rephrased. E.g. we reformulated the abstract: “The conductivity model is constructed step-wise following field evidence from observations; seeking a balance between model complexity and available field data.”

3) *In the introduction, some statements should be toned down, and a few historical references are missing.*

The corresponding part of the introduction was revised considerably.

1) *In line 40, it says that huge amounts of data are needed for kriging. Certainly, you need data to infer the variograms, but not as many as the one you need in the latter application, where you claim that hardly any data are used.*

The passages on Kriging were reformulated. We specified the aspect of required data alongside those of Gaussian models. See also the previous comment.

2) *In line 42, it says that stochastic methods, on the other hand, need a limited amount of data. Kriging is a stochastic method, which apparently needs a huge amount of data. In any case, the statement is not true, stochastic methods need data, large amounts, as it is shown later.*

We rephrased this inconsistent paragraphs profoundly.

3) *When listing the common methods, some historical references are missing, such as Gómez-Hernández and Gorelick, 1989, for Gaussian random fields; Journel and Gómez-Hernández, 1990, for indicator simulation; and Strebelle for multiple-point statistics. (By the way, Freeze, 1975 is not the best example of the use of Gaussian random fields, since he used uncorrelated values.)*

We rephrased and modified the references accordingly.

4) *Line 76, it seems that the number of data used is not minimal. Previous works have shown that properly accounting for hydraulic conductivity heterogeneity at the MADE site is sufficient to reproduce the mass transport behaviour at the MADE site (i.e., Salamon et al. (2006), or Li et al. (2011))*

The paragraph was rephrased, missing references were added.

4) *For the paper to really serve its purpose of enticing practitioners to use stochastic modelling, the model would have had to be three-dimensional. But it is not! How much does the dimensionality reduction influence the results? This must be discussed.*

   We extended the discussion on the impact of dimensionality reduction and clarified the difference between heterogeneity conceptualizations dominated by the binary structure and a log-normal distribution: When it comes to a predominantly log-normal heterogeneity structure, we agree that dimensionality makes a difference. When module C is the main component representing heterogeneity, models should actually be in 3D to not underestimate flow velocity and connectivity. For conductivity conceptualizations dominated by the binary structure (module B), the differences between model results for 2D and 3D are marginal. As the case for your application to the MADE site. This is the results of the binary layer structure, which does not increase connectivity in the third (y-)dimension when considering horizontal isotropy. In this sense, the 2D character of binary fields can even be more enticing for practitioners to use stochastic modelling at this reduced computational effort. However, we stressed, that when applying the proposed heterogeneity conceptualization for modelling flow in transport in other application, a 3D model setup should be considered first and a complexity reduction to 2D models should only be taken when warranted by the conductivity conceptualizations.

   We revised the paragraph in section 3.3 Numerical Model Setting, adapted the supporting information and added a paragraph in the discussion section 4.

References *(provided by the referee)*

Damsleth, E., Tjolsen, C. B., Omre, H., & Haldorsen, H. H. (1992). A two-stage stochastic model applied to a North Sea reservoir. Journal of Petroleum Technology, 44(04), 402-486.

Gómez-Hernández, J. J., & Gorelick, S. M. (1989). Effective groundwater model parameter values: Influence of spatial variability of hydraulic conductivity, leakance, and recharge. Water Resources Research, 25(3), 405-419.

Li, L., Zhou, H., & Gómez-Hernández, J. J. (2011). A comparative study of three-dimensional hydraulic conductivity upscaling at the macro-dispersion experiment (MADE) site, Columbus Air Force Base, Mississippi (USA). Journal of Hydrology, 404(3-4), 278-293.

Huysmans, M., & Dassargues, A. (2009). Application of multiple-point geostatistics on modelling groundwater flow and transport in a cross-bedded aquifer (Belgium). Hydrogeology Journal, 17(8), 1901.

Journel, A. G., & Gomez-Hernandez, J. J. (1993). Stochastic imaging of the Wilmington clastic sequence. SPE formation Evaluation, 8(01), 33-40.

Neton, M. J., Dorsch, J., Olson, C. D., & Young, S. C. (1994). Architecture and directional scales of heterogeneity in alluvial-fan aquifers. Journal of Sedimentary Research, 64(2b), 245-257.

Proce, C. J., Ritzi, R. W., Dominic, D. F., & Dai, Z. (2004). Modelling multi-scale heterogeneity and aquifer interconnectivity. Groundwater, 42(5), 658-670.

Salamon, P., Fernandez-Garcia, D., & Gómez-Hernández, J. J. (2007). Modelling tracer transport at the MADE site: the importance of heterogeneity. Water resources research, 43(8).

Strebelle, S. (2002). Conditional simulation of complex geological structures using multiple-point statistics. Mathematical geology, 34(1), 1-21.

Referee #4 (J. Herweijer) Evaluation: *For final publication, the manuscript should be **rejected.***

Summary

Below are the three main issues that I identified with respect to this paper:

1- The conceptualization is poor, it does not take into account available geological information. The modelling process does not adhere to well-published standard practices for this type of hierarchical modelling. Some key features of the data which underpin the conceptualization are not addressed, most critically, inconsistencies regarding the basic data (K values).
2- The analysis is conducted in a single 2D-vertical cross-section assuming symmetry in one of the horizontal directions. This is not in accordance to observed piezometric level and tracer test data showing flow and transport in both horizontal directions. The results of the 2D model cannot be used for quantitative analysis.
3- The paper incompletely references geological issues very relevant to conceptualization of heterogeneous aquifers. Especially when it comes to options for geological analysis and modelling, some statements in the paper are not well documented and potentially misleading.

Based on point 1 and point 2, I reject this manuscript.

The MADE data set is the result of significant efforts to collect a 3D dataset of hydraulic conductivity and tracer transport observations. Many papers have been published addressing various aspects of the data and the geological concepts, showing significant variability in 3D of hydraulic conductivity and tracer transport. Creating models in 3D that take into account these data is well within the realm of what is technically possible. No scientific rationale is presented to explain why the model has been restricted to only 2D.

Given the preceding discussion and the harsh critique of the referee, we are afraid that nothing we modify within the manuscript will lead to a different general evaluation of the referee. However, we tried our best to address all points raised and adapt the manuscript.

First, we like to address two key points, which we feel are the major source for misunderstandings and disagreement between the referee and us:

- **purpose of the manuscript:** we aim to provide an easy-applicable conceptualization for integrating heterogeneity quantitatively into flow and transport models which naturally differs from previously presented hierarchical approaches. We did **not aim** to provide a detailed heterogeneity conceptualization for MADE, e.g. **reconstruction of 3D structure**, which produces the actual 3D tracer plume of the transport experiment at that side. We follow a different path of heterogeneity conceptualization by generating **random conductivity** structures (in combination with Monte Carlo simulations) which capture the main feature of transport observed at the site which a decent amount of field data. By construction, these structures do not necessarily represent the actual aquifer structure at MADE.

- **Data availability in (hydrogeological) field sites** is well known to be much more reduced (given budget limitation) than in petroleum industry related research. In this line, we **purposely** did **not** use all (hydro-)geological data available for MADE. We rather tried to rely on a decent amount of data which is accessible through standard and/or (cost-efficient) novel monitoring methods at usual hydrogeological sites.

We see that these points might not have been clearly enough addressed in the manuscript. We thus, emphasized them in the revised manuscript version.

Detailed response:

1 – Conceptualization is incomplete, inadequate and not novel

We want to stress the point, that a model is always a simplification of the reality. We address the question: How simple can it be to still represent the relevant processes and to make reasonable predictions (given the aim of the study)? As addressed in the manuscript in l 256: *“We thereby aim to identify the "most simple" of our concepts which still provides a reasonable prediction of the complex observed mass distribution.”* We will refer to the aim of a “most simple” model conceptualization repeatedly.

We agree with the referee that (at an arbitrary site) a detailed site conceptualization is best when considering all available (geological) data. We are aware that several detailed conceptualizations of the heterogeneous structure for the MADE site exist including the work of the referee (Herweijer, 1997) or Dogan et al., (2011, 2014). However, a detailed model of the MADE site was not the purpose of our work. On the contrary, we aimed to present a simple approach for flow and transport modelling taking heterogeneity into account to reproduce non-uniform transport plumes with a decent level of field observation data. We chose the MADE site since it provides ample references data and studies to compare to.

This is explicitly stated in the manuscript several times: l.9, l. 90ff, l. 305ff (to just state a few).

The paper promises a 'novel' conceptualization. However, the paper falls short of an adequate conceptualization. As discussed in 1a below, very adequate conceptualizations have been provided for the MADE aquifer (e.g. Herweijer, 1997) and the MADE site in particular (e.g. Julian et al., 2001), and these should be referenced and compared with the proposed approach. Also, as discussed in 1d below, the 'nested' scale' approach chosen by the authors is not novel.

We rectified the formulation “novel”. We are aware that the conceptualization of heterogeneity into zones and binary structures is conceptually not new. Also its application to MADE is known to us. The novel aspect in our approach is the **quantitative** use of these components for a **predictive** model (without calibration) and giving guidance to use this conceptualization for transport models also at other sites, including software tools.

A comparison of your approach to the detailed aquifer conceptualizations, such as presented by Herweijer, 1997 is neither meaningful nor feasible:

- representation of aquifer heterogeneity differs conceptually: while Herweijer, 1997 aims to provide a reconstruction of the actual heterogeneity pattern, we do not, but make use of ensembles of random structures which do not directly represent the actual structures.
- quantitative transport model: our target is to quantitatively model the transport in a predictive manner (without calibration), while Herweijer, 1997 does not provide a quantitative transport model: *„It should be noted that the data are only used in a qualitative manner; in other words, no analysis method is attempted that aims at a direct quantitative duplication of field data“*., p. 98 of the thesis.

Similarly, the work of Julian et al., 2001 does not target at predictive modelling either but on structure reconstruction: *“Inverse analysis was conducted to estimate optimal K values”*, p. 543.

However, we provide direct reference to both publications.

1a) Creating K zones based on the change in piezometric surface was earlier discussed for the MADE site by Rehfeldt (1992) and Herweijer (1997). These K zones are related to the main geological feature at the site that has been reported about in several papers (e.g. Herweijer and Young, 1991; Young 1995; Herweijer, 1996, 1997; Julian 2001; Bowling et al., 2005). The single vertical K zone model presented in the paper is incorrect. Herweijer (1997) established that the MADE aquifer consists of a two layer system where a high K channel deposit incised in somewhat older lower K deposits (see figure below).

Bowling et al. (2005) shows via GPR data a similar two layer system with a unit of coarser sand and high energy depositional bed-forms overlaying a more stratified lower sand unit. The same two-layer system can also be seen in figure 2 of the paper under review, as the borehole flowmeter K data of wells F20 and F40 show a sharp increase of K above the 56-67 m depth level.

The two-layer system is also clearly reflected in the Tritium plume of the MADE2 experiments (see figure below from Boggs et al., 1993).

In order to obtain a much more realistic conceptual model, the authors should use the above references on the geological background of the high K contrast. The conceptual model should also include the vertical K contrast at ~ the 57   58 m depth level as indicated by figure 2 in their paper and shown by various references cited above. The paper should also present a cross-section of the modelled vertical tracer distribution to compare with the observed vertical tracer distribution.

The two zone model is a simplification of the complex geological structure at the MADE site which represents the large scale zonation indicated by head observations to conductivity conceptualizations, but of course not the actual complex pattern. We added the reference to Rehfeldt et al., 1992 explicitly for zonation based on changes in the piezometric surface.

Again, we refrain from using (and referring) to detailed investigations on geological structure for the conceptual conductivity setup to remain exemplarily for concept use at other (less intensively investigated) sites. We do not aim to construct a more realistic conceptual model, particularly not by calibrating to transport experiment results. We seek the “most simple” concept which still provides a reasonable prediction of the complex observed mass distribution. A comparison of modelled plume cross-sections does not make sense to us. At the level of only Module A it is a deterministic Gaussian plume (following the analytical solutions of ADE). For the conductivity conceptualizations including random components (Module B and/or C), plume simulation results from individual realizations do not reproduce the observed tracer distribution, as a result of the random conceptualization. We further refrain from including comparison to cross-section observations since we do not aim to reproduce the actual tracer plume, but focus on predicting the average longitudinal mass distribution.

1b) At the scale below the main zonation scale referred to in 1a above, the authors then insert binary K contrast 'inclusions' representing a medium scale of heterogeneity. This binary (Boolean) technique has been commonly used (e.g. Haldorsen and Lake, 1984, Desbarats 1987) to create a heterogeneous architecture at various scales. The dimensions of these ‘inclusions’ assumed by the authors, seem not to be based on any field evidence or analogue systems. Rehfeldt (1992) shows a section of a nearby quarry with potential dimensions of these inclusions. Herweijer and Young (1991) show how pumping test data reveal some insights regarding the hydraulic continuity of these *inclusions. Herweijer (1997) specifically mentions some scenarios for dimensions based on sedimentary analogues for the same aquifer. Bowling (2005) shows detailed data for the same site, where GPR sections show some of the sedimentary structures controlling heterogeneity on this scale.*

In order to improve the conceptual model, the authors should elaborate on the nature of these inclusions and use the dimensional information for the inclusions contained in the above references.

We agree that the use of binary techniques is not new. We provided reference to import statistical results, such as Rubin, 1995. We added further references in line with the suggestions by the referee.

Following our paradigm of constructing a simple structure based on a decent amount of field data, we consider the dimension of the inclusion to be unknown. Having no knowledge on horizontal correlation length or connectivity of longitudinal structures is a typical situation at sites. To cope with that we provide a strategy to come up with a reasonable range of values from those few measurements available and follow a parametric uncertainty approach: we suggest to include all of these length and determine their impact then. In this sense, the referee is right that "*The dimensions of these 'inclusions' [...] seem not to be based on any field evidence or analogue systems."* This is done on purpose and clearly outlined in the manuscript. Consequently, we do not aim to refine the conceptual model by adding information from more observational data. Again, at most field sites this amount of data (as available for MADE) is not given.
Our results further show that the precise inclusion length is not crucial for a reasonable prediction given the studies goal. The critical point is that preferential flow path are represented at all, in our case by the inclusion structure.

A clear statement on the role of field data for outlining the inclusion topology is given in the manuscript (l.191-195): *"The inclusion topology is a matter of choice and data availability. [...] More complex layering structures can be adapted if additional topological information is available. However, the specific topology often plays a subordinate role. When not having any information on spatial correlation of heterogeneity, it is beneficial to assume some instead of sticking to a homogeneous model."*

We further specified the aspect with regard to application at MADE in the manuscript (l. 303-305): *"We represent these structures making use of the binary inclusion structured described in section 2.2. We assume little to no information on horizontal structures and connectivity to mimic typical field situations – thereby deliberately ignoring the large amount of data at MADE. We make use of solely four flowmeter logs (Figure 2a).*"

1c) At the next lower scale level, the authors use randomized K values from the borehole flow meter data. Figure 4 shows several datasets for K distribution, and there is a significant discrepancy between the mean and the range of borehole flowmeter data and the K datasets. For the borehole flowmeter the log-normal mean is a factor 5 lower than the pumping test value that represents the bulk of these BHF data (the pumping test in the high K zone – the channel deposit). The mean of the K derived from grain-size is a factor 2 higher than the pumping test value, which could indicate that the K values derived from grainsize would be more reliable to represent the high end of K values. This type of differences between hydraulic conductivity data is not uncommon. Apart from data acquisition issues, the differences between differently measured K values can often be traced *back to scale effects, which are of utmost relevance to conceptualization, and should be addressed in the paper. The authors should also review for example Rehfeldt ea. (1989) and Young (1998) regarding some issues with the borehole flowmeter data specific to the MADE aquifer. Young (1995) and Herweijer (1997) show at the neighbouring MADE-1HA test site borehole flowmeter K-values for the same sediments. They publish values with a higher log-normal mean and maximum, which corroborate the grain-size K values for the MADE site as shown in figure 4 of the paper under review. This extreme end of the K distribution has at the MADE-1HA site a significant effect forming high-K pathways (Young, 1995; Herweijer, 1997)*

In order to support the conceptual model, the authors should review the meaning of the different hydraulic conductivity values as measured for the MADE site and how that impacts the conceptual model. They also should explain specifically why the borehole flowmeter data were selected, why the deviation between the mean of the borehole flowmeter data and the relevant pumping test is acceptable and, why the borehole flowmeter data are preferred to the grainsize K data (which seem to better represent this pumping test and the high end K values).

We fully agree with the referee on the discussion on differences between hydraulic conductivity data of various monitoring methods and their relation to acquisition and scale effects. That is one reason why we included Figure 4, although not using all of the mentioned data sets. We expended the discussion on Figure 4 accordingly (l. 222ff).

The referee is mistaken in the perception that we use borehole flow meter data for the sub-scale log-normal distribution. The only parameter we deduce from field data for module C, additional to those used for module A+B, is a log-conductivity variance. Here, we refer to the most recent DPIL data of Bohling et al., 2016 (l. 343).

Furthermore, we do not use the flowmeter data for determining the mean conductivities but both pumping tests (and other data) instead (module A). We only make use of 4 flowmeter logs to deduce structural information on layering and the binary character for module B. Also note, the similarities on the two-point statistics between the different methods (Figure 4), indicating that the observation methods well agree in the structural characteristics at MADE, although differing considerably in the mean.

We clarified the relation of our model to the different mean values reported for MADE (l. 304ff): “When fixing regional conductivities from pumping tests, model scale coincides with measurement scale. This way, our structures are independent from upscaling of method (and location) specific geometric means reported for MADE (Figure 4).”

1d) The hierarchical/nested scale approach using deterministic zonation with various levels of binary and continuous stochastic infill is not 'novel'. It has been widely used before, and the authors should reference some earlier work applying this approach (see e.g.: Damsleth et al, 1990; Herweijer, 1997 section 6.6, specific to the MADE aquifer – see also first figure of this review; Smith et al., 2001; Yupeng & Shenhe, 2013)

The authors should quote above references as examples of the hierarchical method they employ, and not refer to their approach as novel, neither in general nor specific to the MADE aquifer.

We toned down the claim on originality with respect to hierarchical aquifer modelling and clarified the purpose of the study alongside: providing an easy-applicable conceptualization for integrating heterogeneity quantitatively into models.

We reformulated the abstracts and several text passages accordingly. We further integrated additional and listed references. [Note that we could not find the reference of Yupeng & Shenhe, 2013.]

2 – Use of 2D model for a 3D plume

The paper presents a 2D cross-sectional model along the main axis of flow. The field data show a major component of flow transversal to the main flow. The tritium plume picture below from Boggs et al., 1993) shows that initially the tracer released at the 5 injection wells converges and subsequently diverges. An elongated finger of the plume shows further downstream, but probably already developed closer to the source area. This finger is probably related to some very high K pathways related to sedimentary structures that are highly anisotropic and directionally variable (Young, 1995; Herweijer, 1997).

The Bromide plume shows similar transversal movement (including a sharp sideways movement close to the source) and downstream patchiness (see e.g. Julian et al., 2001, fig 5 &13) flow lines at the boundary of two zones with very different K values and which occurs at an angle to the regional flow direction (Freeze and Cheery, 1979).

Given a clear horizontally anisotropic flow and transport pattern, a 2D analysis/model is very limited and unrepresentative. A 3D model should be used. The results of the 2D model cannot be used for quantitative analysis.

Again, we need to outline that we do NOT consider results of transport experiments to constrain the hydraulic conductivity distribution in order to keep our model predictive and free of calibration.

The referee nicely outlined how the results of the transport experiments revealed information on preferential flow and non-uniform flow in transverse horizontally direction. BUT this could not have been foreseen in the hydraulic observation data. This is confirmed by the setup of the monitoring network for the first transport experiment which relied on hydraulic data only.

We assume symmetry in the horizontal direction because hydraulic heads and hydraulic conductivity do not indicate anisotropy in horizontal direction. The observed piezometric levels (Figure 1, left) show a slight non-uniformaty in the horizontal flow pattern, but in general there is a clear main flow direction perpendicular to the head isolines. Thus, a complexity reduction from 3D to 2D in terms of flow is warranted.

The impact of a 2D instead of 3D model with regard to transport was studies. In short, we found that a 2D model is sufficient to resolve the binary structure we propose. We extended the discussion of that aspect further in the manuscript (sections 3.3 & 4) and in the supporting information: We specified details in on the 3D tests and clarified the difference between heterogeneity conceptualizations dominated by the binary structure and a log-normal distribution: When it comes to a predominantly log-normal heterogeneity structure dimensionality makes a difference. When module C is the main component representing heterogeneity, models should actually be in 3D to not underestimate flow velocity and connectivity. For conductivity conceptualizations dominated by the binary structure (module B), the differences between model results for 2D and 3D are marginal. As it is the case for your application to the MADE site. This is the results of the binary layer structure, which does not increase connectivity in the third (y-)dimension when considering horizontal isotropy. In this sense, the 2D character of binary fields can even be more enticing for practitioners to use stochastic modelling at this reduced computational effort. However, we stressed, that when applying the proposed heterogeneity conceptualization for modelling flow in transport in other application, a 3D model setup should be considered first and a complexity reduction to 2D models should only be taken when warranted by the conductivity conceptualizations.

Given the well fit of the predictive model results to observation data, we think that our model can in fact be used for a quantitative analysis. It stands in line with other transport models, explaining the complex flow patterns by an alternative conceptualizations and field data at MADE. At no point we wish to diminish the worth of any other model for the MADE site (including the work of the referee). We provide an alternative approach for predictive transport modelling at a significantly heterogeneous site with a simple conceptualization and decent observation effort.

3 – Potentially misleading remarks regarding use of geological data and geological models

The paper makes (line 53-57, 64) statements regarding the use of geological data (training images) and geological models referring papers that are 20+ years old (Koltermann and Gorelick, 1996; Herweijer, 1997).

In line 54, the authors dismiss training images as limited available and unrepresentative. This statement is incorrect: training images are quite widely available sourced from satellite images (Google Earth) and extensive literature on geology, sedimentology and paleogeography. This holds especially for relative recent shallow deposits which form the MADE aquifer, and which are often the subject of groundwater modelling efforts.

We reformulated these paragraph profoundly, also according to the comments of the referee. We want to stress that we do not dismiss training images as unrepresentative. We actually consider them as very useful tools for complex structure construction, when data (particularly vertical profiles) are available. We see that this was misleadingly formulated beforehand.

Herweijer (1997) provides a number of references to the sedimentology and paleogeography of the MADE site, which would provide a very good start with respect to representative training images. Ronayne et al. (2008) give a good example of a model based on training images to model a hydrogeological test site.

Again, we do not aim to construct a more realistic conceptual model on the costs of including observation data which is often not available at typical hydrogeological field sites.

The paper also states that geological models as used in the petroleum industry have not found their way into applied hydrogeology (line 64), a statement which is quite strong, and in my view incorrect. Alloisio, (2011), Dowling et al. (2013) and Peereboom (2018) are examples of applications for a variety of shallow to deep aquifers. Even if it is the case that this type of modelling has not ‘found its way’ into widespread use in hydrogeology, this is not a reason to simply to set it aside from a research point of view. The authors should explain why their approach is ‘better’ than the standard geological modelling methods used in the petroleum industry and the hydrogeological applications of these methods referenced earlier in this paragraph.

The authors should re-assess the literature on the above matter and correct their statements about the use of geological data and models.

We are happy when the referee could provide us references to papers in journals frequently read by hydrogeologists in science and practice. Given that all stated articles are conference proceeding or theses, this somehow confirms our statement. Anyway, the corresponding paragraph was reformulated.

In this regard, we aim to clarify that we do not wish to set the work done in petroleum industry aside. Our approach is not ‘better’ than the standard geological modelling methods used in the petroleum industry but has a different purpose (as mentioned repeatedly and stated in the manuscript). It is well known that financial limitations are much different in (purely) hydrogeological studies. Thus, we focus on aquifer heterogeneity construction for a level of available field data usual at hydrogeological sites.

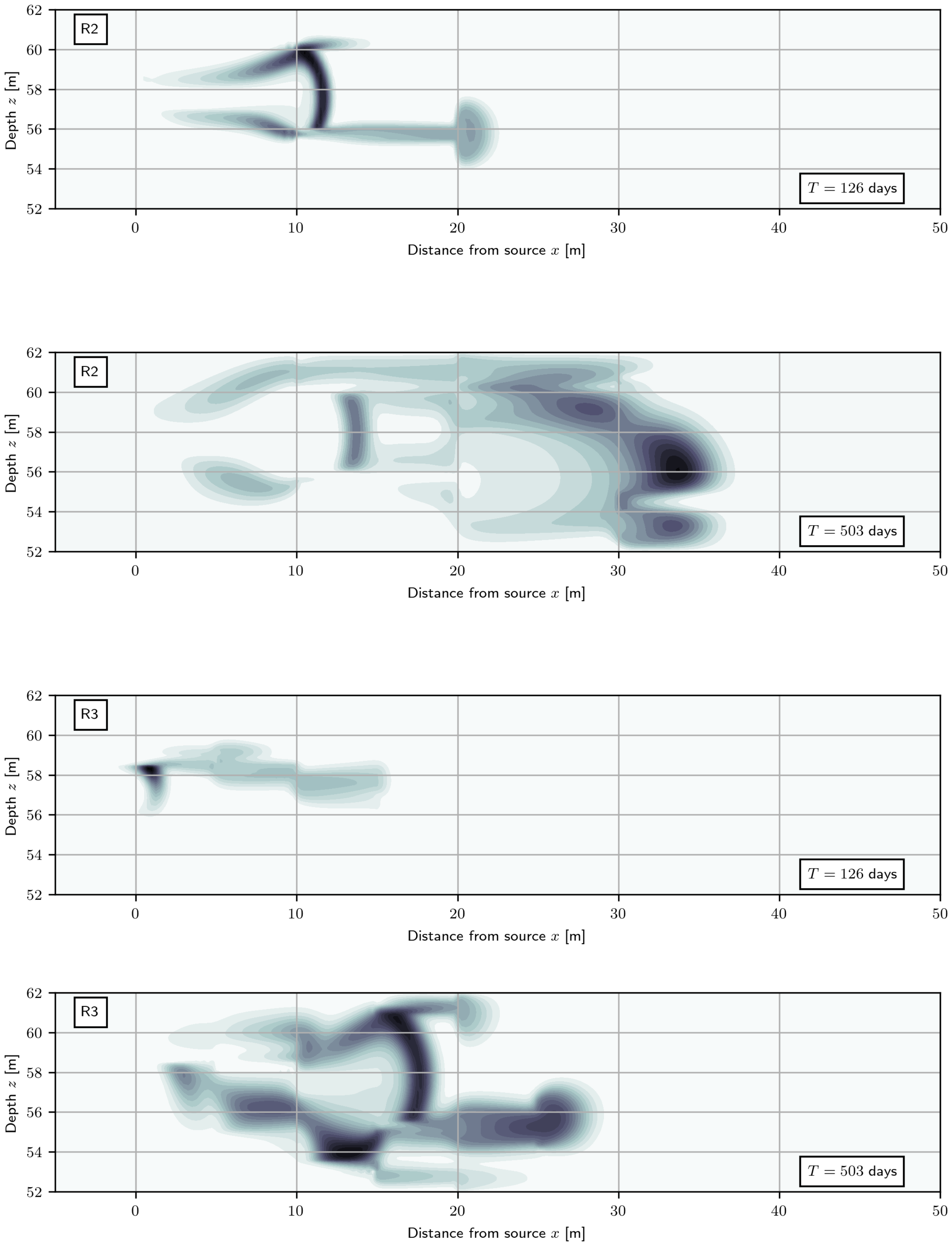
Some further issues in need of clarification:

Line 335 and last section of supplement: As earlier discussed, the MADE plumes are variable in 3D and should really be modelled in 3D. The sections explaining the 3D modelling effort to confirm the validity of the 2D model are confusing. It is unclear if the 2D model was simply copied in the 3 rd dimension, ie. is completely symmetric in the 3 rd dimension, or if some sort of heterogeneity in the 3 rd dimension was included. The supplement also states that the extension in the 3 rd dimensions has no impact because of the binary nature of the ‘inclusions’ vs. K values that change gradually. If transport being restricted to a 2D cross-section of a 3D model is a significant impact of the binary conceptualization, than the binary conceptualization should not be used, as gradual changes of K are the norm for non fractured/fissured aquifers such as the MADE aquifer It would also be informative if the authors would present a visualization (map or x-section slices) of the 3D model and the modelled plume. It should also be looked at if a small level of transversal dispersion (representing very small scale heterogeneity) would have a significant impact on the 3D version of the 2D model.

Data on the observed transport plumes (3d) at the 8 times of the MADE1 experiment are not available to us since they are generally not public. We therefore focused on the averaged longitudinal mass distribution (1D transects) which is available to us. We already added a comment on the data situation to section 3 (L. 245) in the previous revision.

We rewrote the paragraph concerning the model dimensionality in the manuscript and the section in the supporting information (see also comment above). We follow the referees advice, and specified the settings of the 3D inclusion model: specifically the y-direction was extended in a range of -15 to 15 m (with the source located at y=0). For an inclusion length of I_y = 10m, three blocks of different random inclusion structure is present in the transverse horizontal direction. We further added a figure of one realization of the 3D inclusion structure to the SM.

Since we follow a Monte Carlo approach with an ensemble of realizations, there is not a single simulated tracer plume, but a large variety which all look different depending on the binary inclusion structure realization. Our results are based on the ensemble average. To clarify that point we add here some simulated plume distributions of several realizations (of the 2D model). Note that a display of the 3D plume is hardly feasible and displaying the plume of the 3D model along the xz-cross section looks similar to those of the 2D realizations.



Mass distribution contours (cross sectional view, color levels of 5%) for two realizations R2 & R3 at two times T=126 and 503 days after injection.

We agree with the referee's statement on "gradual changes of K" in real aquifers. This is clearly not represented by a binary structure. We can just again repeat that we seek for a simple heterogeneity conceptualization which reproduces transport pattern sufficiently well given the defined study purpose. This does not imply that the conductivity structure actually looks realistic, which is a core characteristic of upscaling. If the purpose would have been the reconstruction of the aquifer heterogeneity in 3D and the reproduction of the 3D tracer plume, than we fully agree, that the model would need to be in 3D and that the heterogeneity would need a clearly more sophisticated conceptualization as we present. For such a study the referee is refereed to Dogan et al., 2014.

When modelling in 3D, clearly transverse horizontal dispersion takes place. Given that we resolve heterogeneity, we applied a small value for transverse horizontal dispersivity representing hydrodynamic dispersion effects solely (see also Zech et al., 2019). As expected it lead to a slight spreading of solutes in transverse horizontal direction. However, this effect is averaged out by post-processing the mass distribution to longitudinal mass transects. Again, we like to emphasize that we do not aim to provide a model which is able to give a proper plume reproduction, but an approach to reproduce main transport features.

Line 330-334 and section 'details on flow and transport' section in supplement: The injection rate modelled is Qin = 1.166e-5 m3/sec. The injection rate quoted by Boggs et. al. is 10.07 m3/48.5 hr which converts to 5.57e-5 m3/sec. If any adjustment has been made to the injection rate (perhaps to adapt for the 2D model setting?) this has to be clarified. The paper states: We use a flux related injection representing natural conditions. For technical details, the reader is referred to the Supporting Information. This is unclear and there seems to be no further specific discussion on 'flux related injection' in the supporting documentation.

The injection rate quoted by Boggs was 10.07 m3/48.5hr (=5.57e-5 m3/sec.) distributed over 5 wells (p.3285). Thus, when modelling the transport along the main flow transect, the actual injection rate in the source well is a fifth of 5.57e-5 m3/sec, thus Q_{in} = 1.166e-5 m3/sec.

We thought the reader familiar with typical injection modes (initial conditions) of transport models, which are typically in resident or flux proportional mode [Kreft and Zuber, 1978]: for a resident mode the initial mass is constant (e.g. along the well) while for a flux proportional mode it is distributed according to the local conductivity (e.g. along the well). When dealing with transport in heterogeneous media, it is particularly important to distinguished according to the given field conditions.

We specified both information in the SM, which were indeed missing so far.

Section 3.4: calibration and predictive capacity of model: The model matches data in a line downstream of the injection, but as discussed earlier in the review (point 2) does not represent any 3D movement of the plume close to the injection point and further downstream. The paper also does not show a vertical cross-section of the observed vs the modelled tracer distribution. Hence any calibration is quite limited and potentially has limited predictive value As discussed in the paper (line 270) there is significant uncertainty due to the mass balance issues with the bromide tracer (~50% in snapshots past 300 days). The calibration could be (should be?) tested using the MADE

2 tritium tracer test (Boggs et al., 1993), which seems to have resulted in a better mass balance (77% in final snapshot at 328 days).

We agree that our model does not provide information on mass distribution at point resolution level. Again, we did not aim to reproduce the actual plume distribution. The model matches the data in a line downstream of the injection, which is the perpendicularly averaged mass distribution along the main flow path. Thus, it should be considered as the average of observed values along a observation transect located at the distance x from the injection. When aiming to predict major plume properties such as leading mass and location of the bulk mass, this is highly useful information. In this line, we also see no point in comparing vertical cross-sections of modelled and observed data, which is anyway not possible for us since the concentration distribution data is not public available for all time snapshots of the MADE1 tracer experiment.

Again, we do NOT calibrate the model. We refrain from calibration to keep the model predictive mimicking field situation where no preliminary information on transport behaviour is available. We also refrain from repeating the study focusing on MADE-2 experimental results, which show in general the same pattern, although having a slightly better mass recovery. As the referee might be aware of, there are other issues with the MADE-2 experimental data, such as transient flow conditions due to significant seasonal water table fluctuations during the experiment. However, we are sure when running flow simulation adapted to the MADE-2 experimental settings for an ensemble of random conductivity with the binary inclusion structure, the observed longitudinal mass distribution would show the same characteristics as observed in the MADE-2 experiment, which are generally similar to those of the MADE-1 experiment.

Title: There is a grammar error in the title (the part which reads ‘in a Heterogeneous Aquifers’). It should be either ‘in a Heterogeneous Aquifer’ or ‘in Heterogeneous Aquifers’

We thank the reviewer for pointing out this flaw. We corrected it.

References:

Dogan, M., Van Dam, R. L., Bohling, G. C., Butler, J. J., and Hyndman, D. W.: Hydrostratigraphic analysis of the MADE site with full-resolution GPR and direct-push hydraulic profiling, Geophys. Res. Lett., 38, L06 405, https://doi.org/10.1029/2010GL046439, 2011.

Dogan, M., Van Dam, R. L., Liu, G., Meerschaert, M. M., Butler, J. J., Bohling, G. C., Benson, D. A., and Hyndman, D. W.: Predicting flow and transport in highly heterogeneous alluvial aquifers, Geophys. Res. Lett., 41, 7560–7565, https://doi.org/10.1002/2014GL061800,5152014.

Herweijer, J.C., 1997. Sedimentary heterogeneity and flow towards a well. Ph.D. dissertation (in English), Free University, Amsterdam, NL, https://www.hydrology.nl/images/docs/dutch/1997.01.07_Herweijer.pdf

Kreft, A., and Zuber, A. (1978). On the physical meaning of the dispersion equation and its solutions for different initial and boundary conditions. Chem. Eng. Sci. 33, 1471–1480.

Zech, A., Attinger, S., Bellin, A., Cvetkovic, V., Dietrich, P., Fiori, A., Teutsch, G., and Dagan, G.: A Critical Analysis of Transverse605Dispersivity Field Data, Groundwater, 57, 632–639, https://doi.org/10.1111/gwat.12838, 2019.

A Field Evidence Model: How to Predict Transport in ~~a~~ Heterogeneous Aquifers at Low Investigation Level?

Alraune Zech[1 2], Peter Dietrich[1 3], Sabine Attinger[1 4], and Georg Teutsch[1]

[1]Helmholtz-Centre for Environmental Research - UFZ, Leipzig, Germany
[2]Utrecht University, Department of Earth Science, Utrecht, The Netherlands
[3]Eberhard Karls University Tübingen, Germany
[4]University of Potsdam, Germany

Correspondence: Alraune Zech (a.zech@uu.nl)

Abstract. Aquifer heterogeneity in combination with data scarcity is a major challenge for reliable solute transport prediction. Velocity fluctuations cause non-regular plume shapes with potentially long tailing and/or fast travelling mass fractions. High monitoring cost and ~~presumably missing~~ a shortage of simple concepts have limited the incorporation of heterogeneity to many field transport models up to now.

We present ~~a novel~~ an easy-applicable hierarchical conceptualization strategy for ~~aquifer heterogeneity. The hierarchical~~ hydraulic conductivity to integrate aquifer heterogeneity into quantitative flow and transport modelling. The modular approach combines large-scale deterministic structures and ~~simple stochastic methods. Such a heterogeneous conductivity can easily be integrated into numerical models~~intermediate scale random structures. Depending on the modelling aim, the required structural complexity can be adapted. The same holds for the amount of ~~available field~~ monitoring data. The conductivity model is constructed step-wise following field evidence from observations; ~~though relying on as minimal dataas possible~~seeking a balance between model complexity and available field data. Starting point are deterministic blocks, derived from head profiles and pumping tests. Then, sub-scale heterogeneity in form of random binary inclusions are introduced to each block. Structural parameters can be determined e.g. from flowmeter measurements or hydraulic profiling.

As proof of concept, we implemented a predictive transport model for the heterogeneous MADE site. The proposed hierarchical aquifer structure reproduces the plume development of the MADE-1 transport experiment without calibration. Thus, classical ADE models are able to describe highly skewed tracer plumes by incorporating deterministic contrasts and effects of connectivity in a stochastic way even without using uni-modal heterogeneity models with high variances. The reliance of the conceptual model on few observations makes it appealing for a goal-oriented site specific transport analysis of less well investigated heterogeneous sites.

Copyright statement.

1 Introduction

Groundwater is extensively used worldwide as the major drinking water resource and consequently needs to be protected with respect to quantity and quality. Increasing pressure on the quality originates from the intensification of agriculture using agrochemicals (non-point sources), an increased urbanization with the resulting solid and liquid wastes and contaminant spills from industrial applications (point sources).

Essential for groundwater protection is the quantitative analysis of the fate and transport of various contaminants in the groundwater body. This can be either for a provisional risk assessment or for the clean-up of an already existing groundwater contamination. Numerical models are common tools to quantify the flow and transport, where partial differential equations are solved using initial and boundary conditions (Bear, 1972; Fetter, 2000).

For simplicity, we restrict ourselves to saturated flow and transport of a dissolved, non-reactive contaminant. The governing equation for its concentration $C(\boldsymbol{x},t)$ is the advection-dispersion equation (ADE) (Bear, 1972):

$$\frac{\partial C(\boldsymbol{x},t)}{\partial t} = -\boldsymbol{u}(\boldsymbol{x},t)\cdot\nabla C(\boldsymbol{x},t) + \nabla\left(\mathbf{D}\cdot\nabla C(\boldsymbol{x},t)\right) \tag{1}$$

given in space $\boldsymbol{x} = (x,y,z)$ and time t. $\mathbf{D}$ is the dispersion coefficient tensor and $\boldsymbol{u}(\boldsymbol{x},t)$ is the Darcy velocity vector. The latter is a function of the hydraulic gradient J and the heterogeneous hydraulic conductivity $K(\boldsymbol{x})$ through Darcy's Law. A proper description of the velocity field $\boldsymbol{u}(\boldsymbol{x},t)$, thus aquifer heterogeneity, is crucial for predicting the concentration distribution $C(\boldsymbol{x},t)$.

The adequate parametrization of the heterogeneous conductivity $K(\boldsymbol{x})$ poses a significant challenge in practical model ~~development due to the lack of data~~ setup due to data scarcity. Numerous deterministic and stochastic approaches have been developed to incorporate the effects of spatial heterogeneity of conductivity on flow and transport, particularly in the context of stochastic subsurface hydrology (Dagan, 1989; Gelhar, 1993; Koltermann and Gorelick, 1996). Representing conductivity by an effective uniform value is convenient for aquifers of low heterogeneity since it can be inferred from pumping tests with decent monitoring effort. But predicting transport in aquifers of significant variability fails when neglecting local effects of heterogeneity and preferential flow.

~~On one hand, fully deterministic approaches use either uniform (effective) conductivities in large domains or maps of heterogeneity, created by interpolation, e.g. Kriging (Kitanidis, 2008). The former approach requires only few data to the price of neglecting local effects of heterogeneity. The latter requires a huge amount of observation data which is hardly ever available in practical cases. Furthermore, conductivity fields from interpolation result in smooth structures lacking geological realism. On the other hand, stochastic methods allow to resolve heterogeneity based on a limited amount of data. Thus, they are able to capture the uncertainty in flow and transport predictions caused by heterogeneity. Common methods as (i) Gaussian random fields (Freeze, 1975; Dagan, 1989; Gelhar, 1993; Zinn and Harvey, 2003); (ii) indicator/hydrofacies models (Carle and Fogg, 1996; Fogg et al., 2000); or (iii) multi-point statistics/training images (Renard et al., 2011; Linde et al., 2015) allow to create spatially distributed conductivity fields of higher geological realism. Modelling flow and transport in ensembles of heterogeneous fields (Monte Carlo approach) do not only provide mean behavior but also uncertainty ranges.~~

Stochastic methods allow resolving heterogeneity and thus capture the induced uncertainty in flow and transport predictions. However, the amount of observation data required is usually high, depending on the method's complexity. Common methods are (i) Kriging (Kitanidis, 2008). (ii) Gaussian random fields (Dagan, 1989; Gómez-Hernández and Gorelick, 1989), (Zinn and Harvey, 2003), potentially combined with Kriging for conditioning to observations; (iii) indicator/hydrofacies models (Journel and Gómez-Hernández, 1993; Carle and Fogg, 1996; Fogg et al., 2000); or (iv) multi-point statistics/training images (Strebelle, 2002; Renard et al., 2011; Linde et al., 2015).

~~Log-normal random fields require a number of parameters like geometric mean, log-variance and spatial correlation lengths in horizontal and vertical directions. They result from geostatistical analysis of spatially distributed observations, e.g. from flowmeter, permeameter or injection logging, as DPIL (Dietrich et al., 2008). Despite increased efficiency in exploration methods, the cost and effort related to gather sufficient data hampers the application in practice. Alternatively, hydrofacies models use indicator geostatistics with transition probability to generate geological heterogeneity structures. Although conceptually different, the general amount of input data is similarly high. Training images are known for their geological realism, but depend strongly on the high resolution input data, e.g. reconstructed images from outcrop studies. Not only the availability of a training image limits their application, but particularly the question if it is representative for the larger aquifer domain where transport is modeled (Koltermann and Gorelick, 1996).~~

For many unconsolidated sediments, field observations showed that conductivity is approximately log-normal (Delhomme, 1979), (Gelhar, 1993; Rubin, 2003); characterized by the geometric mean K_G and the log-conductivity variance σ_Y^2. Variogram analysis provides structural parameters such as correlation length ℓ and anisotropy ratio e based on spatially distributed observations, e.g. from flowmeter, permeameter or injection logging. Despite increased efficiency in exploration methods, data is not even sufficient for variogram analysis in most practical cases thus hampering the practical application of Kriging and Gaussian random fields. Alternatively, hydrofacies models use indicator geostatistics with transition probability to generate geological heterogeneity structures. Although conceptually different, the required amount of input data is similarly high. Multi-point statistical methods provide heterogeneity structures of high geological realism, when training images are available. Although satellite data might provide areal training images, vertical structures rely on extensive literature on geology or outcrop studies. Both are hardly available at the scale representative for plume transport impeding the method's use at hydrogeological sites.

A recent debates series (Rajaram, 2016; Fiori et al., 2016; Fogg and Zhang, 2016; Cirpka and Valocchi, 2016; Sanchez-Vila and Fernàndez-Garcia, 2016) outlined the gap between the advanced research in stochastic subsurface hydrology and its application in the practice of groundwater flow and transport modeling. We see a significant reason in the lack of data for complex stochastic models. Thus, we advocate the use of hierarchical approaches, combining deterministic and stochastic hydraulic conductivity conceptualization. ~~In contrast to many application in the oil and gas industry (Bryant and Flint, 2009), they hardly found their way into applied hydrogeology (Herweijer, 1997).~~ Hierarchical approaches are regularly used in reservoir modelling (Damsleth et al., 1992; Smith et al., 2001; Bryant and Flint, 2009), particularly for consolidated sediments. Aside from qualitative approaches for multi-scale heterogeneity representation e.g. Neton et al. (1994); Herweijer (1997), or Koltermann and Gorelick (and references therein), only few quantitative approaches were proposed, such as: generating sequences of facies assemblages using indicator geostatistics and transition probability at various scales (Weissmann and Fogg, 1999; Proce et al., 2004), or combining training images for large-scale facies realizations with variogram-based geostatistical methods for random intrafacies permeability (Huysmans and Dassargues, 2009). Both approaches show a high level of model complexity and required (hydro-)geological input data.

Here, we present a ~~novel conceptualization strategy of aquifer heterogeneity in a hierarchical~~ parsimonious hierarchical aquifer heterogeneity conceptualization which is easy to apply in quantitative models for predicting flow and solute transport. The deterministic/stochastic framework combines descriptive zonation with statistical methods, following the lines of Gómez-Hernández and Gorelick (1989). Goal is to optimize the aquifer structure setup given the simulation target constrained by the available field data. Thereby, we aim to provide ~~a tool~~ tools making aquifer heterogeneity more accessible for practical applications~~. Our~~, including hands-on software. The approach is based on the fact that subsurface heterogeneity can be generally classified into (a) larger scale dominant features which primarily determine the general flow direction together with the average groundwater flow velocity; and (b) smaller scale features which are responsible for the dispersion, respectively the spatial spreading of a contaminant or solute.

We create a deliberate connection between the model parameterization requirements and the field characterization methods employed for measurement beyond a single method. Pumping tests, for example, are a recommended characterization method to determine the spatially averaged transmissivity respectively hydraulic conductivity, even in a heterogeneous aquifer environment (Herweijer, 1996; Zech et al., 2016). Together with the averaged gradient estimated from piezometric levels this yields good estimates of the mean groundwater flow velocities. ~~On the other hand, high~~ High resolution, small-scale borehole logs of hydraulic conductivity (e.g. from flowmeter or ~~DPIL) can~~ direct push methods) provide the data needed to estimate the variability of the hydraulic conductivity field and consequently the dispersion parameters needed. Here, we consider two stochastic methods representing spatial variability: Gaussian random field which requires sufficient data for variogram analysis and a simplistic binary structure which relies only on a few (e.g. 2-4) well-logs, but takes parametric uncertainty into account. The latter is developed as option for less investigated sites only requiring a decent amount of field data from standard monitoring methods for heterogeneous aquifer modelling.

We demonstrate the methodology using ~~field characterization~~ data from MADE, a heterogeneous, well investigated research ~~field~~ site (e.g. Boggs et al. (1990); Zheng et al. (2011); Gómez-Hernández et al. (2017)). Following our adaptive approach, we use ~~a minimum of field data on aquifer properties to construct a numerical transport model~~various amounts and types of hydraulic observation data for heterogeneity conceptualization to construct numerical transport models. Predictions are independently evaluated ~~using~~ against field tracer data from the MADE-1 experiment (Boggs et al., 1992). ~~In contrast to most other MADEtransport models, we~~ We do not reconstruct the actual conductivity structure at MADE, but predict tracer plume behavior following a Monte Carlo approach devoid of calibration. Model results ~~shows~~ show good agreement with observed plume data, also compared to other ~~complex~~ predictive transport models for MADE (~~i.e.~~e.g. Salamon et al. (2007), Fiori et al. (2013, 2017); Bianchi and Zheng (2016)). In this line, we provide an alternative approach for predictive transport modelling at a significantly heterogeneous site with a simple conceptualization and decent observation effort.

The course of the paper is the following: section 2 features the approach in light of different modeling aims. Section 3 is dedicated to the application of the methodology for the MADE aquifer. We close with a summary and conclusions in section 4.

2 Approach

Large scale hydraulic structures of hundreds or more meters determine the groundwater flow direction and magnitude in combination with groundwater catchment boundaries. Subsequently, they set the mean transport velocity. This is the key parameter to predict the location of the bulk mass of substances dissolved in the groundwater when input conditions are known.

Variations of hydraulic properties on intermediate scale, in the range of tens of meters, generate spatially variable flow fields. They also render transport velocities variable at these scales resulting in a larger spreading of plumes. This is particularly important for modeling tailing or leading mass fronts. Fluctuations on scales smaller than these intermediate scales have a blending effect, generally increasing local mixing and enhancing dispersion (Werth et al., 2006).

Following this conceptual view, we generate hydraulic conductivity fields composed of three components: Module (A), (B) and (C) which capture the effects at large, intermediate and small scale heterogeneity, respectively. Each component is selected according to the model aim and the data at hand to parametrize the hydraulic conductivity for this component.

The procedure is exemplified for the MADE site. This significantly heterogeneous site was intensively investigated with various measurement devices providing many different data sets, as pumping tests, flowmeter and DPIL measurements (Boggs et al., 1990; Bohling et al., 2016). Detailed information on MADE can be found in section 3 and the *Supporting Information*.

In the approach, we considers several steps:

1. Specifying the aim of the model: What do we want to predict?

2. Selecting processes and process components which need to be accounted for in the model: What does this imply for the conceptualization of hydraulic conductivity?

3. Selecting suitable measurement methods: Which method can deliver the data needed for parameterizing hydraulic conductivity with ~~minimal~~ affordable effort?

4. Conceptualizing hydraulic conductivity.

5. Calculating flow and transport.

Before specifying the hydraulic conductivity component Modules (A), (B) and (C), we illustrate our concept discussing two exemplary model aims.

2.1 Exemplary Model Aims

Model Aim "Mean Arrival"

1. **Aim**: Prediction of mean arrival of a contaminant from a point source.

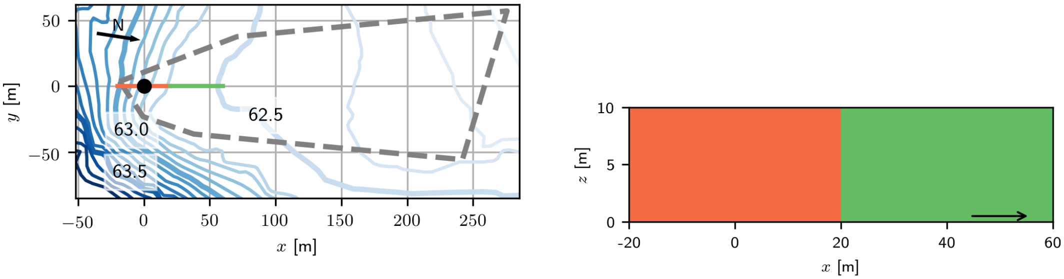


Figure 1. Left: Potentiometric surface map of head measurements according to Boggs et al. (1990). Orange-Green line indicates location of cross section displayed right: Concept (Module A) for large conductivity structure with deterministic zones of low (orange) and high (green) conductivity. Arrow indicates flow direction. Location of the interface between structures corresponds to change in hydraulic head pattern at ~~$20m$~~$x = 20m$.

2. **Processes:** Estimation of regional groundwater movement, direction and magnitude of flow making use of the groundwater flow equation and Darcy's law. Transport is modelled by advection. For sake of simplicity we do not consider reactivity.

3. **Field characterization:** Regionalized groundwater level measurements provide direction and magnitude of hydraulic gradient. It is critical to outline areas of different gradients (zones) indicating regional hydraulic conductivity trends and large scale heterogeneity. Pumping tests can provide independent values of effective transmissivity within each zone.

4. **Conceptualization of hydraulic conductivity:** Conductivity is considered homogeneous within each large scale zone. Effects of heterogeneity are captured in effective parameters representing average flow behavior, e.g. determined from pumping tests.

5. **Solving flow and transport:** Flow is solved either analytically, e.g. for one or two zones of different effective hydraulic conductivity, or numerically in case of a more complex spatial distribution of zones. Transport can be determined making use of analytical or numerically solutions of the ADE according to initial and boundary conditions.

2.1.1 Example MADE

The piezometric surface map of MADE (Boggs et al., 1992, Fig. 3) shows a significant non-uniform hydraulic head pattern. At 20 m downstream of the injection location, head isolines reduce abruptly. The reproduced head contours in Figure 1a allow to delineate ~~these~~ two major zones: an area of low conductivity upstream (left) and high conductivity downstream (right). Two large scale pumping tests confirm the contrast in mean conductivity of about two orders of magnitude (Boggs et al., 1992).

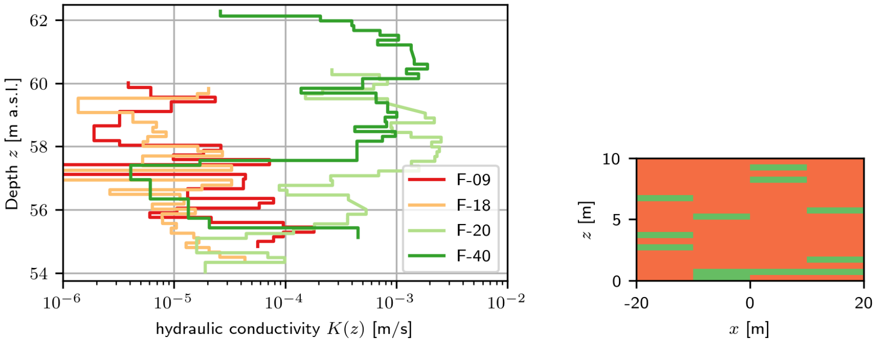


Figure 2. Left: Four flowmeter logs of hydraulic conductivity $K(z)$ versus depth z; the logs *F-09* and *F-18* are close to the tracer test injection location; *F-20* and *F-40* are several tenth of meters downstream (see Figure 3). Right: Concept of binary inclusion structure (Module B) with 15% high conductivity inclusions (green) embedded in the bulk of low conductivity (orange). Inclusion length are arbitrarily chosen as $I_h = 5\,\mathrm{m}$ and $I_v = 0.5 - 1\,\mathrm{m}$.

Consequently, flow should be modelled with distinct mean conductivity in two vertical zones (Figure 1b) when aiming to model mean arrival times for the MADE site.

Model Aim "Risk Assessment"

1. **Aim:** Prediction of early or late arrival of contaminants commonly used in risk assessments.
2. **Processes:** Flow and transport equations; it is particularly relevant to capture variability in transport velocity to estimate spreading behavior of plumes.
3. **Field characterization:** Detecting and delineating high and low conductivity subsurface structures with a characteristic horizontal length scale of several meters. Typical examples are channels formed in braided river systems. Typical investigation methods giving field evidence of such heterogeneity structures are small scale slug tests, borehole flowmeter logs or permeameter tests detecting strongly vertically varying conductivity.
4. **Conceptualization of hydraulic conductivity:** Spatially structured non-uniform conductivity.
5. **Solving flow and transport:** Small variations in conductivity allow to apply analytical solutions with effective measures, e.g. from first order theory (Dagan, 1989). Spatially resolved heterogeneity requires numerical solution of flow and transport with numerical tools (Monte Carlo approach).

2.1.2 Example MADE

Borehole flowmeter logs at MADE (Rehfeldt et al., 1989; Boggs et al., 1990) reveal horizontal layers with conductivity differences over 2 – 3 orders of magnitude. For instance, the flowmeter log *F-40* shown in Figure 2a has a bulk of high conductivity values with about 15% of values being two orders of magnitude smaller. Logs at other locations (*F-09* and *F-18*) show the inverse behavior: a bulk of low conductivity values with embeddings of high conductivity.

Such strong vertical variation indicate the presence of high conductivity channels acting as preferential flow path and low conductivity zones with stagnant flow which both impact strongly on plume spreading behavior. Consequently, when aiming to model early and late plume arrival these feature need to be accounted for in a flow and transport model for the MADE site.

2.2 Scale-dependent Conductivity Modules

Given the scale-dependency of hydraulic conductivity features and their distinct relevance for flow and transport predictions, we propose three components: Module (A), (B) and (C) which capture large, intermediate and small scale heterogeneity effects, respectively. Given a certain model aim, components are selected (or not) with regard to the available field data. We shortly discuss the Modules and motivate their use based on the data of the MADE site example for different aims.

Module A

The aquifer domain of interest is divided into deterministic zones of significantly different mean conductivity (i.e. more than one order of magnitude). The structure can comprise horizontal or vertical layering simply in blocks or complex zone geometries depending on information available. The use of Module A is warranted when observation data indicates significant areal conductivity contrasts.

The zones represent large scale geological structures exhibiting conductivity differences potentially over several orders of magnitude as a results of changes in deposition history or changes in the material's composition (Bear, 1972; Gelhar, 1993). Zones can be delineated using geologic maps, piezometric surface maps and geophysical methods providing information on aquifer structure, sedimentology and genesis. Pumping tests are suitable for identifying mean conductivities for each zone due to their large detection scale. Flow simulations on the deterministic zone structure should reproduce the observed head pattern.

The MADE site is an example where the concept of two zones of different mean hydraulic conductivity (Figure 1b) can reproduce conceptually the hydraulic head pattern. Details will be discussed in section 3.

Module B

When hydraulic conductivity shows heterogeneous features at the same length scale as the plume transport itself, they require proper resolution. A contaminant plume typically passes several of these intermediate scale features but not enough to ensure ergodic transport behavior. Thus, using effective parameters is not warranted. Since limited data availability precludes from a deterministic representation of these features, stochastic approaches suit best.

Binary stochastic models are ~~the simplest~~ a simple way to capture the effects of intermediate scale features (Haldorsen and Lake, 1984), (Dagan, 1986; Rubin, 1995). Figure 2b shows an example how to conceptualize a medium with two K values: inclusions (K_2) are embedded in the bulk conductivity (K_1), with p characterizing the percentage of K_2. Inclusions of high conductivity may represent preferential flow paths whereas inclusions of low conductivity can be obstacles like clay lenses.

The inclusion topology is a matter of choice and data availability. A simple design is a distribution of non-overlapping blocks with horizontal length I_h and thickness I_v. Figure 2b provides an impression with arbitrary choice of parameters. More complex layering structures can be adapted if additional topological information is available. However, the specific topology often plays a subordinate role. When not having any information on spatial correlation of heterogeneity, it is beneficial to assume some instead of sticking to a homogeneous model.

Characteristic length scales in vertical direction I_v are detectable with low effort from a few borehole logs (Figure 2a). Characteristic horizontal length as I_h are critical since they require spatially distributed observations. A parametric uncertainty approach can keep the effort low. A range of reasonable I_h values is estimated and applied in the random inclusion model. A sensitivity analysis reveals the impact of the parametric uncertainty of I_h on transport results. The estimates of I_h could results from auxiliary data such as vertical length scale in combination with anisotropy ratios. Another option is expert knowledge based on geological structures and similarities to outcrop studies. Methods such as diffusivity tests (Somogyvari et al., 2016) or novel approaches for pumping test interpretation (Zech et al., 2016) also offer options to gain estimates for I_h.

The binary structure as in Figure 2b is beneficial in its plain stochastic concept relying on few input data, simple implementation and low computational requirement. It can be combined with Module (A) by implementing it within every deterministic zone preserving the mean conductivities. As for MADE, the inclusions represent the contrasting vertical layers as observed in flowmeter logs (Figure 2a), from which the inclusion parameters can be deduced for every deterministic zone (section 3).

Module C

Variations in grain size and soil texture form small scale heterogeneities of characteristic length scales up to one meter. Their relevance for transport predictions depends on the degree of heterogeneity and ergodicity. A plume is considered ergodic when the behaviour within one realization is statistically representative, i.e. exchangeable with ensemble behaviour. Figuratively speaking, an ergodic plume has travelled long enough to sufficiently sample heterogeneity. This is usually assumed for transport distances of $10-100$ characteristic lengths (Dagan, 1989), ~~which increasing value~~ with higher values for increasing degree of heterogeneity. When ergodic, effective parameters can capture effects of heterogeneity. Otherwise, the use of a spatial random representation is warranted.

If required, small scale features can be conceptualized with a log-normal conductivity distribution $K(\boldsymbol{x}) \propto \mathcal{LN}(K_\mathrm{G}, \sigma_Y^2)$ with geometric mean K_G and log-variance σ_Y^2. Including a spatial correlation structure depends on the acquired complexity and the availability of two-point statistical data as correlation length and anisotropy. Figure 3b gives an example.

Geostatistical parameters can be inferred from spatially distributed observations (Figure 3a), e.g. permeameters, borehole flowmeter, or injection logging (Figure 4). This is related to high effort and costs. Novel techniques like DPIL (Dietrich et al., 2008; Bohling et al., 2016) can provide a large amount of data at acceptable costs and time, but they are only accessible for shallow sites. Alternatives can be approaches which derive geostatistical parameters directly from pumping tests (Zech and Attinger, 2016; Zech et al., 2016) or dipole tracer test (Zech et al., 2018). Note the discrepancy in geostatistical estimates among observation methods (Figure 4), which is not uncommon for heterogeneous sites. Differences are attributed to scale

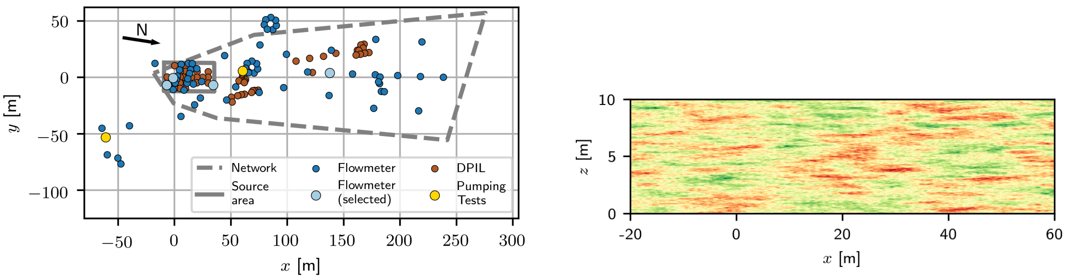


Figure 3. Left: Locations of measurements and tracer test observation network according to Boggs et al. (1990); Bohling et al. (2016). Right: Gaussian random field with exponential co-variance structure as conceptual module for small scale conductivity (Module C).

effects as a results of different method characteristics, such as support volume and resolution. Thus, caution has to be given to the appropriate use of observation data in conductivity conceptualization.

When combining with larger heterogeneity structures, small scale fluctuations are subordinate. In case of field evidence, Module (C) can be combined with Modules (A) and (B) by adding zero-mean fluctuations. According to Lu and Zhang (2002), the variances of heterogeneous sub-structures is additive. Thus, the log-normal variance relates to a 'variance gap' between the total variance, e.g. from a geostatistical analysis of the entire domain, and the binary model's variance (Module B). It can be interpreted as the system's variance which is not captured by intermediate and large scale heterogeneity. The length scales for a correlation structure should be significantly smaller than the inclusion lengths of Module (B). Including small-scale heterogeneity enhances the realism of conductivity structure – however, on the expanse of increasing investigation costs.

The MADE site is a rare example with geostatistics from multiple observation methods (Figures 3a and 4). Methods well suited for small scale heterogeneity show large variances from 4.5 up to 5.9. Given the high variance and the low mean conductivity, ergodic conditions cannot be assumed for transport within the range of a few hundred meters.

The large value in variance, as determined for MADE, can likely be the result of preferential flow and/or trends in mean conductivity. Thus, explicitly representing deterministic zones (Module A) and preferential flow paths (Module B) might render the representation of small scale features (Module C) redundant. Modeling hydraulic conductivity as log-normal fields solely based on Module (C) seems warranted when there is no indication for deterministic zones or preferential pathways.

Hierarchy of Scales

The hierarchy of scales poses an inherent problem for each groundwater model based on heterogeneous field data. Data interpretation often does not allow to clearly distinguish general trends from randomness.

The three modules provide a simple classification of transport relevant heterogeneity scales: (A) – beyond plume scale, i.e. above 100m; (B) – range of plume scale (about 10-100m); and (C) – sub-scale (<1m). ~~It will not be appropriate~~ This classification might not hold for every field and transport situation, but provides an orientation for developing site-specific heterogeneous conductivity structures.

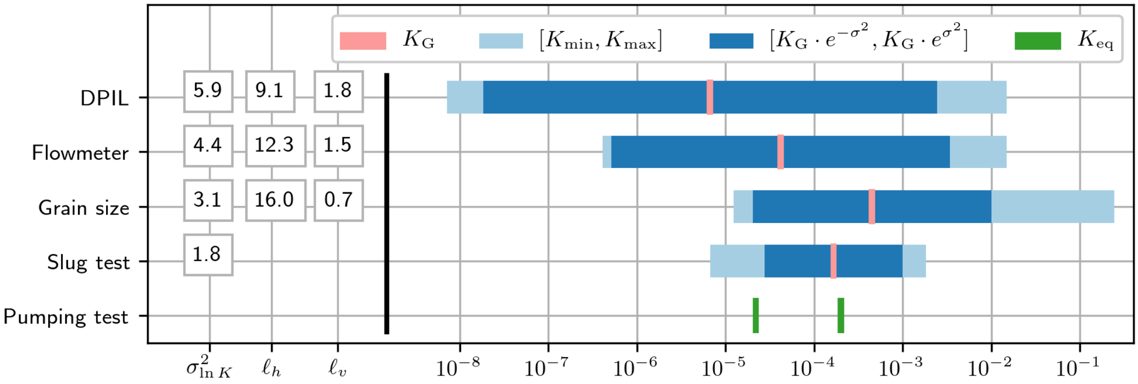


Figure 4. Geostatistical measures for MADE from DPIL (direct push injection logging) (Bohling et al., 2016), flowmeter, grain size analysis, slug tests (Rehfeldt et al., 1992) and effective mean values (K_{eff}) of two large scale pumping tests (Boggs et al., 1990): log-conductivity variance $\sigma^2_{\ln K}$, horizontal and vertical correlation length ℓ_h and ℓ_v, respectively. Visualization of range of observed values from minimal ($K_{\min}$) to maximal ($K_{\max}$), variance range and geometric mean K_{G}.

Which module to integrate at a specific site depends on multiple aspects: (i) Is there field data evidence for a heterogeneity structure of a certain length scale?; (ii) Is there sufficient data to parameterize a conceptual heterogeneity representation? And (iii) is it necessary to present the heterogeneity given the travel distance of the plume (ergodicity)? Having a positive answer to each of the question for a certain module warrants its consideration in the conductivity conceptual model

.

3 Predictive Transport Model for MADE

We validate our approach by performing flow and transport calculation for the MADE setting without parameter calibration. Although, many approaches to model the transport at the MADE site exist~~(Zheng et al., 2011)~~, including detailed aquifer conceptualizations (e.g. Herweijer (1997); Julian et al. (2001), for a detailed review see (Zheng et al., 2011)), only few of them have a predictive character, i.e. devoid of calibration to transport results (Fiori et al., 2013, 2017; Dogan et al., 2014; Bianchi and Zheng, 2016).

Based on the scale-dependent conductivity modules (section 2.2), we develop different conductivity structures according to the field evidence given the structural data at MADE. We thereby aim to identify the "most simple" of our concepts which still provides a reasonable prediction of the complex observed mass distribution. The computed tracer plumes are compared to the MADE-1 transport experimental results (Boggs et al., 1992; Adams and Gelhar, 1992). Since the observed spatial concentration distribution is not available, we make use of 1D longitudinal mass transects at specified times.

Following the approach steps outlined in section 2, we define our model aim broader then specified in section 2.1: The target is predicting the general plume behavior. This might serve different purposes as e.g remediation and includes the mean flow behavior. The fact that there is no break-through curve data available for MADE, inhibits to study the subject of arrival times. Particularly critical is first arrival as discussed in Adams and Gelhar (1992). Processes involved here are flow and transport governed by Darcy's Law and the Advection-Dispersion-Equation (Eq. 1).

3.1 MADE Field Data

The MADE site is located on the Columbus Air Force Base in Mississippi, U.S.A. The aquifer was characterized as shallow, unconfined, of about $10-11\,\mathrm{m}$ thickness (Boggs et al., 1992). It consists of alluvial terrace deposits composed of poorly sorted to well-sorted sandy gravel and gravely sand with significant amounts of silt and clay. The first extensive field campaign by Boggs et al. (1990) yielded a multitude of hydro-geological information, as e.g. piezometric surface maps and hydraulic conductivity observations from soil samples, flowmeter and pumping tests (Figure 4). Field campaigns in subsequent years supplemented observations and data interpretations. For an overview see e.g. Zheng et al. (2011); Bohling et al. (2016) or Table 1 in the *Supporting Information*. We apply a porosity of 0.31. Recharge is assumed uniform and very small (Boggs et al., 1990). Both quantities are kept constant due to the dominating effect of hydraulic conductivity given the significant variations and the uncertainty associated with observations (Figure 4).

The MADE-1 transport experiment was conducted in the years 1986–1988 (Boggs et al., 1990, 1992; Rehfeldt et al., 1992; Adams and Gelhar, 1992). A pulse of bromide was injected over a period of $48.5h$ applying a flow rate of 3.5 l/min. The forced input conditions enlarged the tracer body at the source. Transport then took place under ambient flow conditions.

Concentrations were observed within a spatially dense monitoring network at several times after injection. We focus on the reported longitudinal mass distribution of Adams and Gelhar (1992, Fig.7) at six times: 49, 126, 202, 279, 370, and 503 days after injection. Values are integrated measures over transverse planes and accumulated over slices of $10\,\mathrm{m}$ length, given at the centers of slices at $-5\,\mathrm{m}$, $5\,\mathrm{m}$, $15\,\mathrm{m}$, $\ldots$. The reported mass does not display mass recovery except at 126 days with recovery rates of $2.06, 0.99, 0.68, 0.62, 0.54$, and 0.43, for the six times, respectively. We do not normalize the reported mass to recovered mass, but stick to the actually observed values associating the mass loss to insufficient sampling in the downstream zone as discussed in details by Fiori (2014).

3.2 Hydraulic Conductivity Structures

Three hydraulic conductivity conceptualizations are designed in line with the specifications for MADE in section 2, which serve different model aims. Modules (A), (B) and (C) are combined successively to capture the scale hierarchy of heterogeneity at the MADE site. Figure 5 illustrated examples for each conceptualization.

3.2.1 Deterministic Zones (A)

~~As indicted by~~ Following the lines of Rehfeldt et al. (1992), we create conductivity zones based on the changes in the piezometric surface map (Figure 1~~, section 2.1.1), we~~). We chose two vertically arranged deterministic zones (Figure 5): a low in average conductivity zone Z_1 from upstream of the tracer input location to $x = 20\,\mathrm{m}$ downstream and zone Z_2 as high-

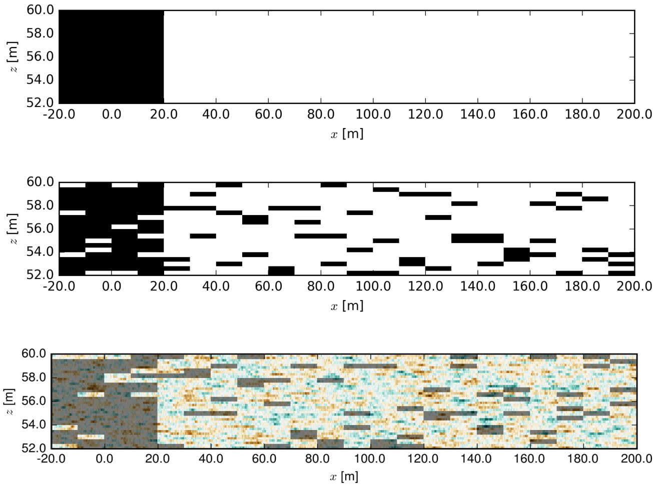


Figure 5. Realizations of hydraulic conductivity structures: (top) Deterministic zones (Module A), low K_1 in black, high K_2 in white. (center) Inclusions in deterministic zones (Modules A+B); amount of inclusions $p = 15\%$, inclusion lengths $I_h = 10\,\mathrm{m}$, $I_v = 0.5\,\mathrm{m}$. (bottom) Inclusions in deterministic zones and sub-scale heterogeneity (Modules A+B+C); correlation lengths $\lambda_h = 2.5\,\mathrm{m}$, $\lambda_v = 0.125\,\mathrm{m}$.

in-the-average conductivity area from 20 m downstream of the source ~~. At $x = 20\,\mathrm{m}$ is an abrupt change in the head isoline pattern~~(section 2.1.1).

We fix average conductivity values of $\bar{K}_{Z1} = 2e-6$ m/s and $\bar{K}_{Z2} = 2e-4$ m/s with a contrast of two orders of magnitude as stated by Boggs et al. (1992). The specific values are chosen according to the two large scale pumping test (Boggs et al., 1992) and the head level rise during injection which is particularly important for early plume development. Details are given in the *Supporting Information*. ~~This~~

When fixing mean conductivity from pumping tests, measurement scale coincides with model scale. This way, the mean conductivity in our structures is independent of the method specific averages reported for MADE (Figure 4). The deterministic conductivity conceptualization is suitable for properly modelling the regional groundwater in line with the model aim "Mean Arrival" as specified in section 2.1.

3.2.2 Inclusion Structure in Zones (A+B)

Flowmeter logs from MADE show a significant discontinuous heterogeneity in the layering (Figure 2). We represent these structures making use of the ~~simple~~ binary inclusion structured described in section 2.2. We assume little to no information on horizontal structures and connectivity to mimic typical field situations - thereby deliberately ignoring the large amount of field data to outline a detailed topology structure for MADE. We make use of solely four flowmeter logs (Figure 2a).

The binary conductivity distribution is constructed for the entire domain comprising both deterministic zones. The upstream zone Z_1 consists of a bulk of low conductivity K_1 with a percentage p of high conductivity K_2 inclusions; the downstream zone Z_2 vice versa (Figure 5).

We identify the specific values of K_1 and K_2 from the statistical relationship for binary structures (Rubin, 1995): $\ln \bar{K}_{Z1} = (1-p)\cdot \ln K_1 + p \cdot \ln K_2$ and $\ln \bar{K}_{Z2} = p \cdot \ln K_1 + (1-p)\cdot \ln K_2$ using the mean conductivities of the zones $\bar{K}_{Z1} = 2e-6$ m/s and $\bar{K}_{Z2} = 2e-4$. p is deduced from the flowmeter profiles (Figure 2a). Being from both zones Z_1 and Z_2, the profiles differ significantly in their average value. However, all show a tendencies of binary behavior with a significant spread over depth. The data is grouped into high and low values being at least two orders of magnitude apart. Then, p is the fraction of values in the minor group, which is $10-20\%$ for the MADE flowmeter profiles (Figure 2a) leading to $p = 15\%$ as default value.

The inclusions structure in both zones is designed according to the simplified block structure outlined in paragraph 2.2. The domain is divided into horizontal blocks of length I_h. Each block contains randomly located inclusions of thickness I_v. The flowmeter logs at MADE indicate a change in vertical layering every $0.25-1$ m (Figure 2a). Thus, we chose $I_v = 0.5$ m. In combination with a inclusion percentage of $p = 15\%$ and an aquifer thickness of 10 m this gives three inclusions per block.

The parameter I_h is the most difficult to extract from data, due to the limited amount of information on horizontal structures and connectivity. We specify I_h through a pragmatic, but stochastic meaningful approach by combining estimates with parametric uncertainty to rely on as little data as possible: A first guess results from auxiliary data analysis: An anisotropy ratio of $e = 0.1-0.025$ is given from the large scale pumping tests (Boggs et al., 1990)). Combining it with the inclusion thickness of $I_v = 0.5$ m gives a range of $I_h \in [5\,m, 20\,m]$. To cover parametric uncertainty we use three different values of I_h, namely 5 m, 10 m and 20 m instead of only one. The different inclusion ~~length~~ lengths produce distinct effects on connected pathways and thus on the mass distribution. ~~In the combined ensemble~~ A combined ensemble integrates the character of each inclusion ~~length is thus integrated.~~ lengths. Figure 5b shows an example structure for $I_h = 10$ m. Note that inclusion can touch, so some inclusions are thicker (e.g. $2I_v = 1$ m) and longer (e.g. $2I_h = 20$ m).

For the Monte Carlo Approach, we create ensembles of 600 individual random realizations, with 200 realizations of each inclusion length I_h, while all other parameters are fixed. Preliminary investigations showed that 200 realizations are sufficient to ensure ensembles convergence. Reported flow and transport results for the inclusion structure in zones (A+B) are ensemble means.

3.2.3 Sub-scale Heterogeneity in Zones (A+B+C)

We combine modules (A), (B), and (C) to an inclusion structure in deterministic zones with small-scale fluctuations (A+B+C), depicted in Figure 5, bottom. Structural aspects of modules (A) and (B) are the same as described before, including parametric uncertainty for the inclusion length $I_h \in \{5, 10, 20\}$ m. Module C is integrated as log-normal distributed conductivity fluctuations (section 2.2). The characterizing parameters for Module (C) depend on the statistics of the super-ordinate modules (A) and (B).

The log-normal fluctuations $\ln Y(\boldsymbol{x})$ are generated using *gstools* (Müller and Schüler, 2019) with zero mean, since the overall mean conductivity refers to $\bar{K}_{Z1}$ and $\bar{K}_{Z2}$ of the deterministic zones. The log-conductivity variance σ_Y^2 follows from the "variance gap", as difference between the variance of the inclusion structure and the overall variance. The binary inclusions for the chosen setting have a variance of $\sigma_Z^2 = 5.52$ resulting from $\sigma_Z^2 = p\cdot(1-p)\cdot(\ln K_1 - \ln K_2)^2$ (Rubin, 1995). With an overall variance of $\sigma_F^2 = 5.9$ as indicated by (Bohling et al., 2016) (Figure 4), we arrive at a fluctuation variance of $\sigma_Y^2 \approx 0.5$. We apply an exponential co-variance function with length scale parameters being a fraction of the inclusion length scales: $\lambda_h = 1/4 I_h$ and $\lambda_v = 1/4 I_v$. Testing several ratios, we saw that its impact on transport behavior is negligible. Ensembles consist of 600 realizations.

3.3 Numerical Model Settings

Flow and transport are calculated making use of the finite element solver OpenGeoSys (Kolditz et al., 2012) in the ogs5py python framework (Müller et al., 2020). The simulation domain is a 2D cross section within $x \in [-20, 200]$ m and $z \in [52, 62]$ m generously comprising the area of the MADE-1 tracer experiment (Boggs et al., 1992). We applied constant head boundary conditions at the left and right margin with a mean had gradient of $J = 0.003$. Tracer is injected at a well located at $x = 0$ with a central screen of 0.6 m depth. Injection takes place over a period of 48.5 h with an injection rate of $Q_{\text{in}} = 1.166e-5\,\text{m}^3$/s according to the initial conditions reported by Boggs et al. (1992). ~~It is~~ We use a flux related injection ~~being the realistic representation of~~ representing natural conditions. For technical details, the reader is referred to the *Supporting Information*.

We checked the impact of dimensionality~~(2D instead of 3D) and~~. A detailed discussion is provided in the *Supporting Information*. We found almost no differences between 2D and 3D simulation setups ~~. This is in contrast to known results for log-normal distributed fields, but can be explained by the conceptualization of the heterogeneous binary structure~~ where the binary structure (Module B) dominates. Extending the binary structure in the horizontal direction perpendicular to main flow does not provide additional degrees of freedom for the flow. Thus, extending the model hardly impacts the flow and thus transport pattern~~. A detailed discussion is provided in the *Supporting Information*~~, while significantly increasing computational effort. However, dimensionality effects hold for conductivity conceptualization with prevailing log-normal distribution, i.e. dominated by Module C. The option of complexity reduction by using 2D instead of 3D models is warranted for this application by the fact that conductivity conceptualizations is dominated by the binary structure (module B).

Simulation results are processed like the MADE-1 experimental data. Longitudinal mass distributions are vertical averages and accumulated horizontally over 10 m slices. Note that the simulated distributions show a full mass recovery. Besides spatial mass distributions for the six times where experimental data is available, we present the break through curves (BTCs) as temporal mass evolution at critical distances, although no BTCs data is reported for the MADE-1 experiment.

3.4 Simulation Results

Figure 6 shows the simulated longitudinal mass distributions $M(x)/M_0$ of the specified conductivity conceptualizations (section 3.2) at $T = 126$ days after injection. They are compared to the MADE-1 experiment data, which had a mass recovery of 99% at that time.

The mass distribution for the deterministic structure (concept A, yellow) shows a sharp peak close to the injection location and no mass downstream. The conductivity structures with inclusions in deterministic zones (A+B, blue) and with sub-scale

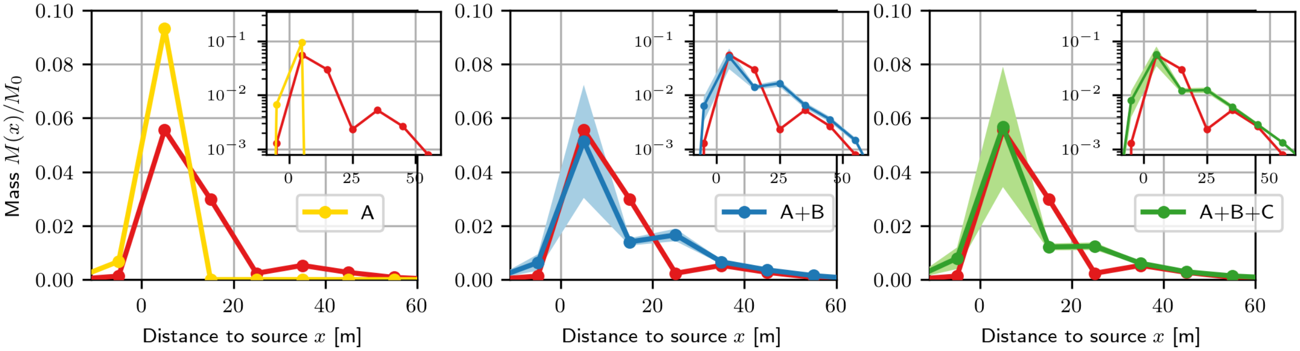


Figure 6. Longitudinal mass distribution at $T = 126$ days for conductivity concepts: (A) deterministic zones, (A+B) inclusions in zones, (A+B+C) inclusion in zones with sub-scale heterogeneity (Figure 5). Shaded areas (light blue and green) indicate parametric uncertainty bands. Mass distribution observed at MADE experiment in red. Linear scale and log-scale in subplot.

heterogeneity (A+B+C, green) result in skewed mass distributions with a peak close to the injection area and a small amount of mass ahead of the bulk. Shaded areas indicate parametric uncertainty due to the variable inclusion length I_h. The shade area margins refer to ±3 ensemble standard deviations, which is similar to the 99% confidence intervals, considering a Gaussian distribution of variations.

A direct comparison of the mass distributions $M(x)/M_0$ for the structures are depicted in Figure 7 for six temporal snapshots, including $T = 1000$d, where no experimental data is available. The general form of the mass distributions is persistent in time for all conductivity structures.

Figure 8 shows simulated breakthrough curves (BTCs) for the deterministic block and inclusion conductivity structure at three distances to the injection location. The results for concept (A+B+C) are very close to those of concept (A+B), thus not displayed. Apparent differences to the longitudinal mass distributions as in Figure 7 are due to the spatial data aggregation. The BTC for Module A has the expected Gaussian shape with a late breakthrough at $x = 5$ m given the very low conductivity in the injection area. The stochastic models have an earlier breakthrough and strong tailing at all distances.

BTCs are not available for the MADE-1 transport experiment. However, we added the aggregated mass values at the three locations for the six reported times in a subplot to indicate a trend of temporal mass development. Note that mass values of the btcs and those at MADE are at different scales due to data aggregation and mass recovery.

3.5 Discussion

All conductivity structures were able to reproduce the skewed hydraulic head distribution as observed at MADE (Figure 1a). The corresponding mean flow velocity determines the travel time. As a results, all models properly reproduced the spatial position of the mass peak (Figure 6).

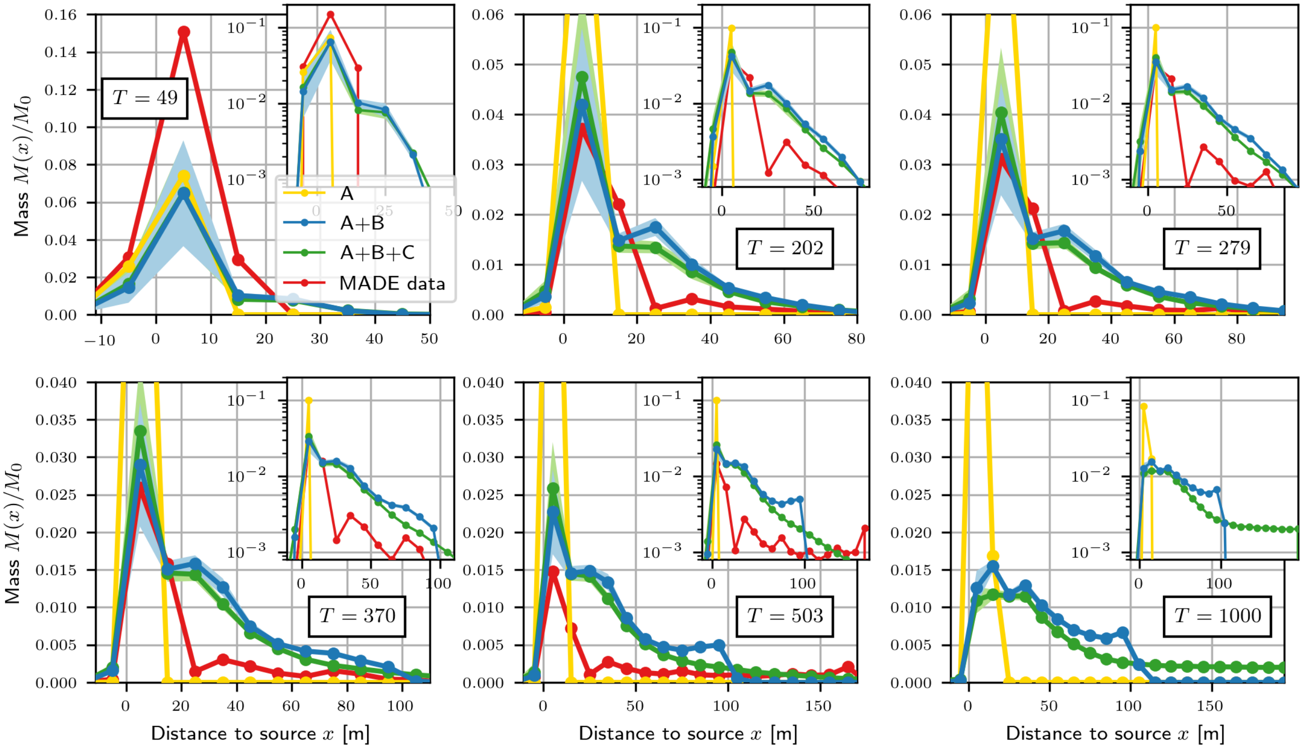


Figure 7. Mass distributions at times $T = 49$, 202, 279, 370, 503, and 1000 days (panels): red = MADE-1 experiment; yellow = concept (A); blue = concept (A+B); green = concept (A+B+C). Shaded areas (light blue and green) indicate parametric uncertainty bands; semi-log scale in subplot.

The deterministic block structure (A) failed to reproduce the skewed mass distribution observed at MADE. The leading front mass traveling through fast flow channels could not be predicted (Figure 7) solely using average K values in zones. In line 425 with model aim "Mean Arrival" (section 2.1), the simple structure allows to estimate the regional groundwater movement and to predict the location of the bulk mass. However, in case of aiming at "Risk Assessment", the arrival times of mass would be significantly underestimated, as clearly be observable comparing BTCs (Figure 8).

Tracer transport in a binary conductivity structure with inclusions (concept A+B) reproduces the observed mass, both for the peak near the injection site and the leading front. The simulated longitudinal mass distribution shows a second peak downstream 430 (Figure 7), which increases with time. The position is related to the interface between the low and high conductivity zones at 20 m distance to the source. Such a second peak is absent in the observed MADE-plume, however it might be associated with the mass loss for the later times. The skewed mass distribution is related to significantly smaller first arrival times as can be seen for the BTCs in Figure 8 compared to the deterministic structure. The BTCs are clearly non-Gaussian with heavy tailing. It shows the same temporal as the MADE experiment data.

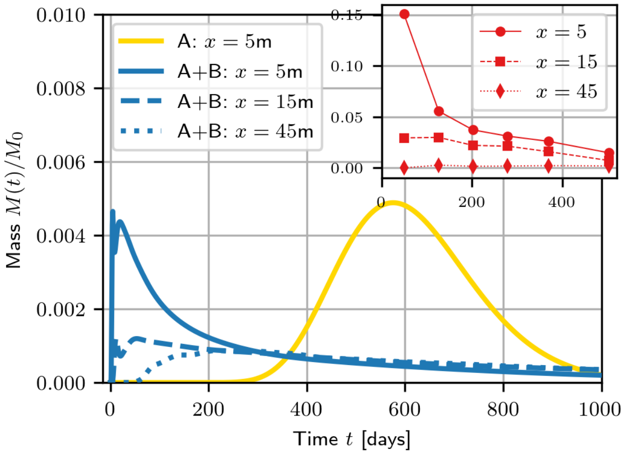


Figure 8. Breakthrough curves: Total mass $M(t)/M_0$ versus time at selected control plane locations for inclusion structure (A+B), (blue) at $x = 5$ m (solid), $x = 15$ m (dashed), $x = 45$ m (dotted); and for deterministic structure (A) at $x = 5$ m (solid yellow line). Reported mass values for MADE at the three locations (red markers) given in subplot. Regard the difference in scale due to the spatial averaging of experimental data.

The horizontal inclusion length I_h for structure (A+B) was not fixed, but was varied over the range of $I_h \in \{5, 10, 20\}$ m. The uncertainty bands in Figure 6b indicate that I_h mostly influences the height of the mass peak close to the source. I_h characterized the connectivity of the source area Z_1 to the high conductivity zone Z_2. Thus, it determines the distance of the bulk mass being trapped in the low conductivity area. The larger I_h the higher is the amount of mass transported downstream. The shape of the leading front is less impacted by I_h giving that its value does not influence the effect of the inclusions as preferential flow per se.

The predicted plume shape for the conductivity structure with inclusions and subscale heterogeneity (A+B+C) is almost similar to the one without sub-scale heterogneity (A+B). ~~consequently~~Consequently, the inclusion structure is the one which determines the shape of the distribution, whereas the impact of sub-scale heterogeneity is minor. Given the model aim of plume prediction, the additional effort for determining characterising geostatistical parameters for the sub-scale heterogeneity is not warranted.

The binary conductivity conceptualization (A+B) was derived for MADE with ~~minimal data from field investigations, thus with a high~~ few observations from standard methods, as can be expected to be present at many field sites. The price for the limited amount of data is parametric uncertainty. A sensitivity study revealed that the mass distribution resulting from the binary conductivity structure is very robust against the choice of parameters. The inclusion length I_h and the choice of the K contrast between the zones show the highest impact. The latter was expected as the mean conductivity determines the average flow velocity and by that the peak location and the general distribution shape. The impact of I_h is represented in the uncertainty bands (Figures 6b, 7). Other parameters as amount of inclusion p and sub-scale heterogeneity parameters as the variance have minor effects. For details, the reader is referred to the *Supporting Information*. In this regard, the binary structure is very stable towards parametric uncertainty.

4 Summary and Conclusions

~~When aquifer heterogeneity is at a similar scale as solute transport, predictive transport models need to incorporate spatially distributed hydraulic conductivity.~~ We introduce a modular concept of heterogeneous hydraulic conductivity for predictive modeling of field scale subsurface flow and transport. Central idea is to combine deterministic structures with simple stochastic approaches to rely on ~~a minimal amount of~~ few measurements and to forgo calibration. The scale hierarchy of hydraulic conductivity induces three structure modules which represent: (A) deterministic large scale features like facies; (B) intermediate scale heterogeneity like preferential pathways or low conductivity inclusions; (C) small-scale random fluctuations. Field evidence of heterogeneity features and module's input parameters are provided by observation methods with the appropriate detection scale. The specific form of the scale-dependent features depends on the site characteristics and field data. ~~Generally, we~~ We propose a deterministic model for large-scale features, a simple binary statistical model for intermediate and a ~~geostatistical~~ log-normal random model for small-scale features. However, the integration of alternative conductivity structures is possible. Thereby, the concept is easily adaptable to any field site making aquifer heterogeneity accessible for practical applications.

An illustrative example is given for the heterogeneous MADE site. Three modular conductivity structures are constructed, based on two observations: (i) the existence of distinct zones of mean flow velocity, and (ii) high conductivity contrasts in depth profiles suggesting local inclusions acting as fast flow channels. The structures are used in a predictive flow and transport model which is free of calibration. The comparison of results to the MADE-1 field tracer experiment showed that all conceptualizations can be of value depending on the modelling aim. However, predicting the mass plume behaviour required to take heterogeneity into account.

The combination of deterministic and simple binary stochastic showed the best ~~result~~ results given the trade-off between transport prediction and need for measurements. Realizations of hydraulic conductivity composed of binary inclusions in two blocks with different average conductivity. Details on the topology are thereby secondary, since binary structures show robustness towards the choice of specific parameters.

~~This rather~~ The simple binary structure was able to capture the overall characteristics of the MADE tracer plume with reasonable accuracy requiring only a small amount of observations. Among the few predictive transport models for the MADE site, the presented approach shows a higher level of simulation effort due to the Monte Carlo simulations. However, the lower level of data requirements makes it attractive for application at less investigated sites. Note that when applying the proposed heterogeneity conceptualization in other modelling application, a 3D model setup should be considered first, particular when heterogeneity is conceptualized by a log-normal distribution (modules C). A complexity reduction to 2D models is warranted when the heterogeneous conductivity conceptualizations does not impact the flow pattern in transverse horizontal direction, such as the binary structure. The generality of the binary concept makes it easily transferable to other sites; particularly when focusing on a few, but scale-related measurements.

A hierarchical conductivity structure allows to balance between complexity and available data. Large scale structures determine the mean flow behavior, which is most critical for flow predictions. They can be integrated to a model with reasonable low effort. Structural complexity increases with decreasing heterogeneity scale where small-scale features have the highest demand on observation data. However, even with limited information on the conductivity structure, simple stochastic modules can be used to incorporate the effect of heterogeneity. Considering small scale feature, the conductivity structure can be extended by including modules when additional measurements are available.

Distinguishing the effects of the scale-specific features on flow and transport also allows to identify the need for further field investigations and potential strategies. The adaptive construction based on scale-specific modules allows to create a conductivity structure model as complex as necessary but as simple as possible.

The use of simple binary models is very powerful when dealing with strongly heterogeneous aquifers. They require less observation data compared to uni-modal heterogeneity models, as log-normal conductivity with high variances. Binary models also allow to incorporate effects of dual-domain transport models without the drawback of having non-measurable input parameters which require model calibration. Our work shows that highly skewed solute plumes can be reproduced with classical ADE models by incorporating deterministic contrasts and effects of connectivity ~~stochastically~~statistically. ~~specific transport analysis of less well investigated heterogeneous sites.~~ In summary, we conclude:

- Modular concepts of conductivity structure allow to separate the multiple scales of heterogeneity. Scale related investigation methods provide field evidence and ~~characterizing~~ conductivity model parameters. A hierarchical approach for conductivity can thus ~~minimize the~~ reduce observation effort by focusing on the model aim.

- Site specific heterogeneous hydraulic conductivity can be easily constructed with simple methods taking the (limited) amount of data into account. For aquifers with high conductivity contrast, we recommend combining large-scale deterministic structures and simple binary ~~stochastics~~ stochastic models.

- The application example at MADE showed that complex field structures can be represented appropriately for transport predictions with an economic use of investigation data.

This work aims to contribute to bridging the gap between the advanced research in stochastic hydrogeology and its limited use by practitioners, being a subject of recent debate (e.g. Rajaram (2016)). We advocate the use of heterogeneity in transport models for successfully predicting solute behavior, particularly in complex aquifers. This can be done with few data and simple tools: adaptive structures allowing to combine deterministic, ~~simple stochastic~~ random binary and geostatistical models depending on the available data and the site-specific modelling aim.

Code and data availability. Study related python scripts are public available at *https://github.com/GeoStat-Examples/Binary_Inclusions* (Zech and Müller, 2020), including scrips for (i) generating and (ii) visualizing binary inclusion structures as well as (iii) scripts for transport simulations in the random inclusion structure adapted to the MADE-1 site settings. The python API *ogs5py* Müller (2019) and the geostatistics packages *gstools* Müller and Schüler (2019) used in this study are both available on *https://github.com/GeoStat-Framework*. Data on the MADE aquifer can be accessed via the stated literature sources. Data generated for this study is available upon request to the corresponding author.

Author contributions. All authors contributed to developing the approach and writing the paper. Simulations and figure preparation was performed by AZ.

Competing interests. No competing interests.

Acknowledgements. The authors like to thank Jeff Bohling for background information on Flowmeter and DPIL data. We thank the editor and reviewers for their helpful comments.

References

Adams, E. E. and Gelhar, L. W.: Field study of dipersion in a heterogeneous aquifer: 2. Spatial moments analysis, Water Resour. Res., 28, 3293–3307, https://doi.org/10.1029/92WR01757, 1992.

Bear, J.: Dynamics of Fluids in Porous Media, Elsevier, New York, 1972.

Bianchi, M. and Zheng, C.: A lithofacies approach for modeling non-Fickian solute transport in a heterogeneous alluvial aquifer, Water Resour. Res., 52, 552–565, https://doi.org/10.1002/2015WR018186, 2016.

Boggs, J. M., Young, S., Benton, D., and Chung, Y.: Hydrogeologic Characterization of the MADE Site, Tech. Rep. EN-6915, EPRI, Palo Alto, CA, 1990.

Boggs, J. M., Young, S. C., Beard, L. M., Gelhar, L. W., Rehfeldt, K. R., and Adams, E. E.: Field study of dispersion in a heterogeneous aquifer: 1. Overview and site description, Water Resour. Res., 28, 3281–3291, https://doi.org/10.1029/92WR01756, 1992.

Bohling, G. C., Liu, G., Dietrich, P., and Butler, J. J.: Reassessing the MADE direct-push hydraulic conductivity data using a revised calibration procedure, Water Resour. Res., 52, 8970–8985, https://doi.org/10.1002/2016WR019008, 2016.

Bryant, I. D. and Flint, S. S.: Quantitative Clastic Reservoir Geological Modelling: Problems and Perspectives, pp. 1–20, John Wiley & Sons, Ltd, https://doi.org/10.1002/9781444303957.ch1, 2009.

Carle, S. F. and Fogg, G. E.: Transition probability-based indicator geostatistics, Mathematical Geology, 28, 453–476, https://doi.org/10.1007/BF02083656, 1996.

Cirpka, O. A. and Valocchi, A. J.: Debates – Stochastic subsurface hydrology from theory to practice: Does stochastic subsurface hydrology help solving practical problems of contaminant hydrogeology?, Water Resour. Res., 52, 9218–9227, https://doi.org/10.1002/2016WR019087, 2016.

Dagan, G.: Statistical theory of groundwater flow and transport: Pore to laboratory, laboratory to formation, and formation to regional scale, Water Resour. Res., 22, 120S–134S, 1986.

Dagan, G.: Flow and Transport on Porous Formations, Springer, New York, 1989.

Damsleth, E., Tjolsen, C. B., Omre, H., and Haldorsen, H. H.: A Two-Stage Stochastic Model Applied to a North Sea Reservoir, Journal of Petroleum Technology, 44, 402–486, https://doi.org/10.2118/20605-PA, 1992.

Delhomme, J. P.: Spatial variability and uncertainty in groundwater flow parameters: A geostatistical approach, Water Resour. Res., 15, 269–280, https://doi.org/10.1029/WR015i002p00269, 1979.

Dietrich, P., Butler, J. J., and Faiss, K.: A rapid method for hydraulic profiling in unconsolidated formations, Ground Water, 46, 323–328, https://doi.org/10.1111/j.1745-6584.2007.00377.x, 2008.

Dogan, M., Van Dam, R. L., Liu, G., Meerschaert, M. M., Butler, J. J., Bohling, G. C., Benson, D. A., and Hyndman, D. W.: Predicting flow and transport in highly heterogeneous alluvial aquifers, Geophys. Res. Lett., 41, 7560–7565, https://doi.org/10.1002/2014GL061800, 2014.

Fetter, C. W.: Applied Hydrogeology, Prentice Hall, 2000.

Fiori, A.: Channeling, channel density and mass recovery in aquifer transport, with application to the MADE experiment, Water Resour. Res., 50, 9148–9161, https://doi.org/10.1002/2014WR015950, 2014.

Fiori, A., Dagan, G., Jankovic, I., and Zarlenga, A.: The plume spreading in the MADE transport experiment: Could it be predicted by stochastic models?, Water Resour. Res., 49, 2497–2507, https://doi.org/10.1002/wrcr.20128, 2013.

Fiori, A., Cvetkovic, V., Dagan, G., Attinger, S., Bellin, A., Dietrich, P., Zech, A., and Teutsch, G.: Debates – Stochastic subsurface hydrology from theory to practice: The relevance of stochastic subsurface hydrology to practical problems of contaminant transport and remediation. What is characterization and stochastic theory good for?, Water Resour. Res., 52, 9228–9234, https://doi.org/10.1002/2015WR017525, 2016.

Fiori, A., Zarlenga, A., Jankovic, I., and Dagan, G.: Solute transport in aquifers: The comeback of the advection dispersion equation and the First Order Approximation, Adv. Water Resour., 110, 349–359, https://doi.org/10.1016/j.advwatres.2017.10.025, 2017.

Fogg, G. E. and Zhang, Y.: Debates – Stochastic subsurface hydrology from theory to practice: A geologic perspective, Water Resour. Res., 52, 9235–9245, https://doi.org/10.1002/2016WR019699, 2016.

Fogg, G. E., Carle, S. F., and Green, C.: Connected-network paradigm for the alluvial aquifer system, Geological Society of America Special Papers, 348, 25–42, https://doi.org/10.1130/0-8137-2348-5.25, 2000.

Freeze, R. A.: A stochastic-conceptual analysis of the one-dimensional groundwater flow in nonuniform homogeneous media, Water Resour. Res., 11, 725–742, https://doi.org/10.1029/WR011i005p00725, 1975.

Gelhar, L.: Stochastic Subsurface Hydrology, Prentice Hall, Englewood Cliffs, N. Y., 1993.

Gómez-Hernández, J., Butler, J. J., Fiori, A., Bolster, D., Cvetkovic, V., Dagan, G., and Hyndman, D.: Introduction to special section on Modeling highly heterogeneous aquifers: Lessons learned in the last 30 years from the MADE experiments and others, Water Resour. Res., 53, 2581–2584, https://doi.org/10.1002/2017WR020774, 2017.

Gómez-Hernández, J. J. and Gorelick, S. M.: Effective groundwater model parameter values: Influence of spatial variability of hydraulic conductivity, leakance and recharge, Water Resour. Res., 25, 405–419, 1989.

Haldorsen, H. and Lake, L.: A New Approach to Shale Management in Field-Scale Models, Society of Petroleum Engineers Journal, 24, 447–457, https://doi.org/10.2118/10976-PA, 1984.

Herweijer, J. C.: Constraining uncertainty of groundwater flow and transport models using pumping tests, in: Calibration and Reliability in Groundwater Modelling, vol. 237, pp. 473–482, IAHS Publ. no. 237, Golden, Colorado, 1996.

Herweijer, J. C.: Sedimentary Heterogeneity and Flow Towards a Well, Ph.D. thesis, Vrije Universiteit Amsterdam, Amsterdam, 1997.

Huysmans, M. and Dassargues, A.: Application of multiple-point geostatistics on modelling groundwater flow and transport in a cross-bedded aquifer (Belgium), Hydrogeol J, 17, 1901–1911, https://doi.org/10.1007/s10040-009-0495-2, 2009.

Journel, A. G. and Gómez-Hernández, J. J.: Stochastic Imaging of the Wilmington Clastic Sequence, SPE Formation Evaluation, 8, 33–40, https://doi.org/10.2118/19857-PA, 1993.

Julian, H. E., Boggs, J. M., Zheng, C., and Feehley, C. E.: Numerical Simulation of a Natural Gradient Tracer Experiment for the Natural Attenuation Study: Flow and Physical Transport, Ground Water, 39, 534–545, https://doi.org/10.1111/j.1745-6584.2001.tb02342.x, 2001.

Kitanidis, P.: Introduction to Geostatistics: Applications in Hydrogeology, Cambridge University Press, Cambridge ; New York, 2008.

Kolditz, O., Bauer, S., Bilke, L., Böttcher, N., Delfs, J.-O., Fischer, T., Görke, U. J., Kalbacher, T., Kosakowski, G., McDermott, C. I., Park, C. H., Radu, F., Rink, K., Shao, H., Shao, H. B., Sun, F., Sun, Y. Y., Singh, A. K., Taron, J., Walther, M., Wang, W., Watanabe, N., Wu, Y., Xie, M., Xu, W., and Zehner, B.: OpenGeoSys: An open-source initiative for numerical simulation of thermo-hydro-mechanical/chemical (THM/C) processes in porous media, Environ. Earth Sci., 67, 589–599, https://doi.org/10.1007/s12665-012-1546-x, 2012.

Koltermann, C. E. and Gorelick, S. M.: Heterogeneity in Sedimentary Deposits: A Review of Structure-Imitating, Process-Imitating, and Descriptive Approaches, Water Resour. Res., 32, 2617–2658, https://doi.org/10.1029/96WR00025, 1996.

Linde, N., Renard, P., Mukerji, T., and Caers, J.: Geological realism in hydrogeological and geophysical inverse modeling: A review, Adv. Water Resour., 86, 86–101, https://doi.org/10.1016/j.advwatres.2015.09.019, 2015.

Lu, Z. and Zhang, D.: On stochastic modeling of flow in multimodal heterogeneous formations, Water Resour. Res., 38, 1190, https://doi.org/10.1029/2001WR001026, 2002.

Müller, S.: ogs5py v1.0.5, https://doi.org/10.5281/zenodo.3546035, 2019.

Müller, S. and Schüler, L.: GSTools: Reverberating Red, https://doi.org/10.5281/zenodo.3532946, 2019.

Müller, S., Zech, A., and Heße, F.: `ogs5py`: A Python-API for the OpenGeoSys 5 scientific modeling package, Ground Water, under revision, 2020.

Neton, M. J., Dorsch, J., Olson, C. D., and Young, S. C.: Architecture and directional scales of heterogeneity in alluvial-fan aquifers, Journal of Sedimentary Research, 64, 245–257, https://doi.org/10.1306/D4267FA0-2B26-11D7-8648000102C1865D, 1994.

Proce, C. J., Ritzi, R. W., Dominic, D. F., and Dai, Z.: Modeling Multiscale Heterogeneity and Aquifer Interconnectivity, Groundwater, 42, 658–670, https://doi.org/10.1111/j.1745-6584.2004.tb02720.x, 2004.

Rajaram, H.: Debates – Stochastic subsurface hydrology from theory to practice: Introduction, Water Resour. Res., 52, 9215–9217, https://doi.org/10.1002/2016WR020066, 2016.

Rehfeldt, K. R., Hufschmied, P., Gelhar, L. W., and Schaefer, M.: Measuring hydraulic conductivity with the borehole flowmeter, Tech. Rep. EN-6511, EPRI, Palo Alto, CA, 1989.

Rehfeldt, K. R., Boggs, J. M., and Gelhar, L. W.: Field study of dispersion in a heterogeneous aquifer: 3. Geostatistical analysis of hydraulic conductivity, Water Resour. Res., 28, 3309–3324, https://doi.org/10.1029/92WR01758, 1992.

Renard, P., Straubhaar, J., Caers, J., and Mariethoz, G.: Conditioning Facies Simulations with Connectivity Data, Math Geosci., 43, 879–903, https://doi.org/10.1007/s11004-011-9363-4, 2011.

Rubin, Y.: Flow and Transport in Bimodal Heterogeneous Formations, Water Resour. Res., 31, 2461–2468, https://doi.org/10.1029/95WR01953, 1995.

Rubin, Y.: Applied Stochastic Hydrogeology, Oxford Univ. Press, New York, 2003.

Salamon, P., Fernàndez-Garcia, D., and Gòmez-Hernández, J. J.: Modeling tracer transport at the MADE site: The importance of heterogeneity, Water Resour. Res., 43, W08 404, https://doi.org/10.1029/2006WR005522, 2007.

Sanchez-Vila, X. and Fernàndez-Garcia, D.: Debates – Stochastic subsurface hydrology from theory to practice: Why stochastic modeling has not yet permeated into practitioners?, Water Resour. Res., 52, 9246–9258, https://doi.org/10.1002/2016WR019302, 2016.

Smith, R., Bard, W., Corredor, J., Herweijer, J., McGuire, S., Antunez, A., Block, T., and Lazarde, N.: Geostatistical Modeling and Simulation of a Compartmentalized Deltaic Sequence, Ceuta Tomoporo Field, Lake Maracaibo, Venezuela, in: Proceedings on SPE Latin American and Caribbean Petroleum Engineering Conference, Society of Petroleum Engineers, https://doi.org/10.2118/69572-MS, 2001.

Somogyvari, M., Bayer, P., and Brauchler, R.: Travel-time-based thermal tracer tomography, Hydrol. Earth Syst. Sci., 20, 1885–1901, https://doi.org/10.5194/hess-20-1885-2016, 2016.

Strebelle, S.: Conditional Simulation of Complex Geological Structures Using Multiple-Point Statistics, Mathematical Geology, 34, 1–21, https://doi.org/10.1023/A:1014009426274, 2002.

Weissmann, G. S. and Fogg, G. E.: Multi-scale alluvial fan heterogeneity modeled with transition probability geostatistics in a sequence stratigraphic framework, J. Hydrol., 226, 48–65, https://doi.org/10.1016/S0022-1694(99)00160-2, 1999.

Werth, C. J., Cirpka, O. A., and Grathwohl, P.: Enhanced mixing and reaction through flow focusing in heterogeneous porous media, Water Resour. Res., 42, W12 414, https://doi.org/10.1029/2005WR004511, 2006.

Zech, A. and Attinger, S.: Technical note: Analytical drawdown solution for steady-state pumping tests in two-dimensional isotropic heterogeneous aquifers, Hydrol. Earth Syst. Sci., 20, 1655 – 1667, https://doi.org/10.5194/hess-20-1655-2016, 2016.

Zech, A. and Müller, S.: GeoStat-Examples/Binary_Inclusions, https://doi.org/10.5281/zenodo.4134627, 2020.

Zech, A., Müller, S., Mai, J., Heße, F., and Attinger, S.: Extending Theis' Solution: Using Transient Pumping Tests to Estimate Parameters of Aquifer Heterogeneity, Water Resour. Res., 52, 6156–6170, https://doi.org/10.1002/2015WR018509, 2016.

Zech, A., D'Angelo, C., Attinger, S., and Fiori, A.: Revisitation of the dipole tracer test for heterogeneous porous formations, Adv. Water Resour., 115, 198–206, https://doi.org/10.1016/j.advwatres.2018.03.006, 2018.

Zheng, C., Bianchi, M., and Gorelick, S. M.: Lessons Learned from 25 Years of Research at the MADE Site, Ground Water, 49, 649–662, https://doi.org/10.1111/j.1745-6584.2010.00753.x, 2011.

Zinn, B. and Harvey, C. F.: When good statistical models of aquifer heterogeneity go bad: A comparison of flow, dispersion, and mass transfer in connected and multivariate Gaussian hydraulic conductivity fields, Water Resour. Res., 39, 1051, 2003.