Reviewer comments (RC1) in black and answers in blue

Positive Aspects

- The graph-based approach simplifies complex geological models and reduces the computational costs.
- Distance map provides information about the potential pathways of contaminant transport.
- A new similarity measure used to compare the distance map to the cumulative mass distribution.

Thanks for your positive assessment of this work.

General Comments

- The term "groundwater" is often associated with specific subsurface conditions and flow regimes. While the principles of flow and transport in porous media can be applied to groundwater systems, the broader context of the study seems to be more general. It's important to use more accurate and inclusive terminology to avoid potential misunderstandings, a suggestion could be to use *porous media*.

  Our work is motivated by groundwater applications, as stated in the introduction. It is also illustrated with a groundwater synthetic case. While such work could have more general applications to flow and transport in porous media, it has not been tested. Nonetheless it could be tested and applied in other setting. This will be mentioned in the discussion at the revision of the manuscript.

- Including fault scenarios might seem unnecessary if the method doesn't perform well for cases without faults, as Appendix A shows.
    - Justify the Fault Scenarios: If the fault scenarios are crucial for real-world applications, provide stronger justification. Perhaps there are specific geological settings where faults significantly impact flow and transport.
    - Under this specific scenario, explore the limitations of the graph-based approach to justify the range of the metric that is considered acceptable.
    - Appendix A needs to include details of parametrization for the MODFLOW simulation.

  Exploring conceptual uncertainty such as different fault scenarios is a key motivation of our work, as highlighted in the introduction. As explained in Appendix A, the absence of very high conductivity paths (or very low conductivity barriers), which the graph approximates quite well can explain the mitigated performance of a graph-based approach in a multi-Gaussian setting.

So, the use of the method is particularly interesting to tests scenarios displaying different types of hydraulic conductivity contrasts or pathways. This explanation will be added to the discussion at the revision of the manuscript. In Appendix A, we already state that the same parameters are used but we could precise ‘same flow and transport boundary conditions’.

- The method still relies on a 3D simulation (MODFLOW) to generate the "ground truth" against which the graph-based method is compared. This limits the method's independence and its potential for significant computational savings. While the graph-based method can provide a quick and potentially accurate approximation, perhaps consider validation with simplified Analytical Solutions, Sensitivity Analysis or Machine Learning techniques. This would provide a more rigorous comparison without relying on numerical simulations.

  The generation of a synthetic “ground truth” (here using MODFLOW) is indeed necessary to test our approach. In a real case study, is seems reasonable to have monitoring wells at the outlet of the model boundaries, such as to interpolate cumulated mass at the outlet. It would not require the use of flow and transport simulations.

- Similarity measure: A similarity coefficient of 0.3 might seem low, especially considering that a perfect match would be 1.0. While a higher similarity coefficient would be ideal, a value of 0.3 can still be considered reasonable but needs to be explicitly acknowledged, especially given the complexity of the problem. The authors should provide a detailed discussion of the factors influencing the similarity coefficient and explain why this value is acceptable in the context of their study. Additionally, the authors could explore ways to improve the accuracy of the graph-based method, such as refining the graph construction by experimenting with different graph configurations to capture the underlying geological features better.

  Lines 263 to 266, we explain how we set this threshold of 0.3. Indeed, the proposed similarity metric is very sensitive to slight changes: a small shift both decreases the Wasserstein component of the similarity and decreases the Jaccard index. The advantage and drawback of each component are given in section 2.4 and justify the proposed metric. These additional explanations and a reminder of the dynamic of each component of the metric will be added to the discussion at the revision of the manuscript.

- A comprehensive evaluation of the graph-based method requires a clear understanding of the underlying physics-based model, including its setup and initial conditions. The authors should provide a detailed description of the MODFLOW simulations, including:
  - Model Domain: The spatial extent and discretization of the model domain.
  - Hydrogeological Properties: The values assigned to hydraulic conductivity, porosity, and other relevant parameters.
  - Boundary Conditions: The types of boundary conditions applied to the model boundaries.

- Initial Conditions: The initial distribution of hydraulic head and contaminant concentration.

All these characteristics are already described in section 2.1. Figure 2 illustrates sections of hydraulic conductivity for one of the fault scenarios. All scenarios are provided along with the code to generate them as explained in the Code and data availability section (line 430).

- Comparing a single MODFLOW scenario to multiple graph-based scenarios can be misleading, as it doesn't directly assess the accuracy of each individual graph-based scenario. A more appropriate approach would be to compare each corresponding pair of scenarios.

As stated in lines 260-261, “The similarity index described in Sect. 2.4 has been applied to analyse the results of the 80 scenarios. A representative sample of the results can be found in Fig. 5, and the distribution of similarities is shown in Fig. 6.” It means that for each of the 80 scenarios, we compute a similarity index (illustrated in Fig 5) between the MODFLOW cumulative mass map and the GRAPHFLOW distance map. The histogram of the 80 resulting similarity indices is displayed in Fig. 6.

Regarding scenario selection, as explained in section 2.5.1, for a given contaminant source position all pairs of scenarios are tested against each other in the scenario selection. The success rate and number of scenarios are averaged across the 80 possible scenarios. We should precise that the average is over all pairs of scenarios and all contaminant sources.

This will be added at the revision of the manuscript.

- The paper should be understandable to a broad audience without requiring extensive external references. Consider providing a brief explanation of the algorithms used:
  - Dijkstra's Algorithm
  - Other Algorithms (Jaccard dissimilarity, Wasserstein distance,Otsu thresholding)

We already provide explanations of the Jaccard dissimilarity and Wasserstein distance in section 2.4 as we combine them to propose a new metric. For Dijkstra’s algorithm and Otsu Thresholding, simple sentences (around lines 117 and 154 respectively) already describe the purpose of these methods, which is sufficient for the reader.

Specific Comments

- **Abstract:**

[2] The phrase "large-scale structural features" could be more specific. Explicitly mention geological features: "large-scale geological features, such as faults, fractures, and stratigraphic variations" and their standard scales compared to domain extension.

We will modify the text as suggested at the revision of the manuscript.

- **Introduction:**

[42-43] The paper should clearly state how the methodology " improves the consistency for subsurface flow". The author should provide a more precise explanation of why faults are relevant for contaminant transport in porous media. The manuscript should provide a deeper analysis of the role of heterogeneity within the graph-based approach.

The previous work from Rizzo and de Barros (2017) is limited to 2D multi-Gaussian heterogeneous medium and compares the graph approximation with results from particle tracking. Here we go one step further by integrating general flow direction information and by doing a comparison with flow and transport simulations (thus aiming at improving the consistency with subsurface flow). We will add these precisions at the revision of the manuscript.

[47] Consider addressing the role of heterogeneity in the main body of the manuscript.

I am not sure what this comment means. I agree that we should talk more about the role of heterogeneity in the first paragraph of the introduction. We will add these precisions at the revision of the manuscript and make sure that it is discussed properly with respect to application of the proposed approach when mentioning scenario selection.

- **Method:**

[60] Figure 1. There are no dimensions indicated in the figure. Is there a reason for the orientation of the scheme?

We will reorient the scheme and add dimensions as in figure 2 at the revision of the manuscript.

[70-73] The description of the experimental setting should be more specific about the position of the source points relative to the grid size. The authors indicate only one coordinate point; it is unclear where the random 10 positions fall on the modeling grid.

The 10 positions are displayed on Figure 2. We will add a reference to the figure in the revised manuscript.

[75-80] This section should also address how the authors evaluate the role of heterogeneity for the simulation domain for the different subsurface properties, as this section indicates a variability in the behavior of the faults but does not answer the effect of the hydraulic conductivity or porosity for this approach. Appendix A should be referenced here.

For each scenario, each geological unit is a stochastic multi-Gaussian SRF realization whose parameters are described lines 78 to 81. We will add this precision at the revision of the manuscript.

[98] Figure 2 shows the hydraulic conductivity values of one scenario. The color bar should be properly labeled, and the formatting of the relative position of the two plots needs to be adjusted.

The label of the colorbar will be added at the revision of the manuscript.

[100] Equation 2. This equation needs to be properly referenced and described in the text. The variables are not defined.

R_γ is the hydraulic resistance along the path gamma. For each point l on gamma (where l is a dummy variable), we calculate the absolute value of the scalar product between the inverse of the conductivity tensor at point l, $K^{-1}(l)$, and the infinitesimal distance dl. We then compute the integral of this value over the path gamma.

This will be added at the revision of the manuscript.

[105] Equation 3. This equation needs to be properly referenced and described in the text.

It is the scalar product between the inverse of the hydraulic conductivity simplified tensor [k_{xx}, k_{yy}, k_{zz}] and the oriented edge [e_x,e_y,e_z]. This will be added at the revision of the manuscript.

[126] is the function “get_shortest_paths” the same as the Dijkstra algorithm?

Yes, it is the implementation of Dijkstra algorithm used here.

[140] Figure 3. At this stage of the reading, it is still not clear what s32 is. The figure needs quality improvement. Include units for the color bars. Figures c and d should be moved further down as it is not clear at this point what they mean, and they are not formatted properly. Labels for figures c and d should indicate the modeling framework

used (MODFLOW, GRAPHFLOW). Furthermore, the choice of histogram plot to compare the output of 80 simulations using the new methodology compared to one single scenario using MODFLOW is confusing as it does not indicate the performance of each simulation against its corresponding physics-based.

S32 is scenario 32 (numbering is explained lines 73 & 74). Cumulative mass, computed from MODFLOW outputs, is in units of mass per cubic meter (as define line 78) and distances, computed from GRAPHFLOW outputs, are homogeneous to a length in meters, but are not meters as the length through the graph is the product of weighted lengths, so we do not prefer to add a unit to the colorbar of this subfigure, as it could be misleading for the readeran d) display the correlation coefficients between cumulative mass and graph distance.

- **Metrics**

[148] Figure 4 needs to improve its quality. Some recommendations: use the same font size of the plots and add labels to the color bars and units of measure. Adjust formatting. Since this is a workflow of the proposed metric, use more descriptive texts next to the figures.

We agree with these formatting recommendations and will implement them at the revision of the manuscript.

[178] Variables have different formatting than the previous equation. 2-Wassertein Distance (W2) needs to be numbered.

We have checked the formatting and all ‘b’s should be bold and italic. The equations formatting, as well as the equations numbering, will be updated at the revision of the manuscript.

- **Method of scenario selection**

[205-214] This section seems to address a different problem: the uncertainty of uncharacterized faults. However, the proposed methodology to validate the graph model has not been discussed up to this point. Consider including the evaluation of the model with the proposed metric first. This analysis should reflect the desirable range of the metric and its limitations.

We justify the use of the graph model as a proxy by analysing the ranking correlation between the graph distances and the cumulated mass (Figure 3d). We will precise this around line 145 in the revised manuscript as well as clarify how the proposed metric contribute to assess the performance of the proxy.

- **Results**

[265] In this section, the author should provide a thorough justification of why a metric of 0.3 is considered valid. Based on the plots presented in Figure 5, for a validation coefficient of 0.31, the cumulative mass and the shortest distances seem to differ.

It depends on what we consider a "good" approximation. In line 265, we refer to an "acceptable" threshold. In the case of Figure 5.f, we observe that our method captures two out of the three cumulative mass patches present in the MODFLOW simulation. Indeed, if the user is more demanding, they can choose a higher threshold, such as 0.4 or 0.5. These explanations will be added at the revision of the manuscript.

[272] How does the discretization of the domain affect the binary maps and, consequently, its validation?

Here, we just tested the computing time scalability. We did not compare binary maps of different resolutions. This is something that we can add to the discussion for further work to potentially increase the computing efficiency of the approach.

Figure 5. This figure needs to improve its quality. Consider including the name of the scenario presented in each plot.

The scenario numbers will be added at the revision of the manuscript.

We are providing them now.

Fig a: scenario 0, Fig b: scenario 76, Fig c: scenario 36,

Fig d: scenario 8, Fig e: scenario 27, Fig f: scenario 10

[276] There is no reference to what position 5 is.

Although present in the code, the coordinates of the different positions are indeed not included in the paper. This will be added in the revised version of the paper. We are providing these coordinates for 10 positions, indexed from 0 to 9 :

ID	X	Y	Z
0	2011.8216247	2950.46369633	512.5
1	1644.15961272	2948.64944714	512.5
2	1811.83145201	2423.32644897	512.5
3	2327.70259382	2409.19913637	512.5
4	2049.59368767	2027.55911324	512.5
5	2253.51310867	2538.14331322	512.5

6	1829.7317165	2788.42870343	512.5
7	1803.19482929	2453.49788948	512.5
8	1634.04169725	2403.11298645	512.5
9	1703.45524068	2262.31334044	512.5

We will add such a table at the revision of the manuscript.

Figure 7. This plot references 8 different scenarios from the graph method against one single scenario solved using a physics-based model. In the following paragraph, the author should provide an explanation of why two different scenarios lead to similar or equal validation metrics. This is misleading as it could mean that the proposed validation metric is not robust.

When fault 1 act as a preferential path and fault 2 as a barrier, most of the flow goes through fault 1, which reaches the model outlet independently of fault 3 (that could act either as a barrier or a preferential path). It means that fault 3 does not influence the shortest path through the graph. These explanations will be added at the revision of the manuscript.

Table 2. The caption and names of the scenarios don't match.

Thanks for pointing this out. It is about scenarios 12 and 65 and the caption will be modified accordingly at the revision of the manuscript.

Technical corrections

The figures in the manuscript could be significantly improved in terms of clarity and readability. To enhance the visual appeal and understanding of the results. The font size for labels, axis titles, and legends should be increased to improve visibility. Clear and concise labels should be used to identify different components of the figures. Avoid using abbreviations or overly technical terms. Employ distinct color bars for different variables to facilitate comparison and interpretation. Consider the overall layout of the figures, ensuring that the elements are well-organized and easy to follow.

We will update the figures to improve their readability as suggested and such that the font size are increased and are consistent through the different figures of the manuscript.