

Reviewer's comment:

General comment

The authors have significantly reworked their manuscript, clarifying many points, and streamlining some of the analysis. I still think that the manuscript is too long, especially the introduction (that reads like a review), and that the case study chosen is too complex and fraught with specificities, but this may be a question of taste.

There is however one aspect of the analysis I disagree with. The authors present the metamodelling as successfully reproducing the reference transit time distributions obtained from tracer data using lumped-parameter models. But I do not think that figure 6 conveys this at all. On the top two figures, the range of transit times estimated for the 90th and the 95th percentiles is enormous, and hence obviously not constrained well by the chemical data. The bottom left figure on the contrary clearly shows that the metamodel manages to approach the reference distribution. But the bottom right figure illustrates another problem, namely that the metamodels can systematically bias the entire transit time distribution to lower values. So maybe the metamodelling works well on average, as seems to be indicated by the statistics presented, but there are still serious problems in some cases, and I feel this is being slightly swept under the carpet. Further below in the discussion (L688), the authors suggest that this deviation might in fact indicate that the calibrated transit time distribution is inappropriate, and that using groundwater chemistry data together with a metamodel could guide model choice for the lpm. This is interesting, but still a bit too hypothetical.

I understand that the authors see this manuscript as a proof of concept, but I cannot help thinking that exploring first what is only been hinted at in the discussion, and then presenting a series of simple case studies illustrating clearly the advantages and the problems of the proposed metamodelling instead of suggesting what one could do with it would have made the method the authors propose more concise, much clearer and to the point.

Authors' response:

We are pleased to have the opportunity to further improve our manuscript.

As the reviewer points out, it is complex! Unfortunately, this complexity is a reality for applications in dynamic real-world systems.

The remaining concerns of both the reviewer and the editor appear to relate to the presentation of the age distributions in Figure 6 (and S1-S4), and the associated text. We want to clarify that our choice of sites to show, as examples, in Figure 6 was designed to highlight the variation in model results; these sites definitely do not represent the site with the "best" correspondence. We refute the notion that the poorer results have been "swept under the carpet". However, we acknowledge that the text around these results may not have highlighted the variation in performance as we intended. We have reworded and slightly restructured Section 4.1 to try and make this clearer.

We hope that this reworking provided more clarity to the presentation of the predicted age distributions, including the coherency (or otherwise) of the prediction median absolute deviations.