We thank the reviewer for their insightful comments which have touched upon topics that have been discussion points between authors. The reviewer's comments have helped us refine and clarify the narrative of this manuscript. Our responses to the reviewer's comments (in black) are in blue below.

General comment

The manuscript presents the use of two machine learning algorithms to estimate groundwater transit time distributions as complement to or in the absence of appropriate age dating tracers.

The central research question, the state of the art and the local setting are all very competently presented.

The analysis itself however seems to me still incomplete and rather unfocused. While the authors discuss at length the effect of different processes and parameter choice on estimate results, they only pay lip service to what should be the central step, namely quantifying the deviation and bias of the estimates obtained from the metamodels compared to the reference transit time distributions. I think that simply showing the transit time distributions and then declaring that the fit is overall satisfactory is not convincing enough. Furthermore, the discussion on the results of the case study are too detailed and specific, and dilute the manuscript instead of keeping it compact and to the point.

We thank the reviewer for this comment, which provided an opportunity to better clarify the purpose of the analysis. The purpose of the analysis was to explore in a 'proof of concept' sense the effectiveness of data-driven or machine learning algorithms to the estimation of groundwater age on the basis of water chemistry data. Because the groundwater age data was itself an estimate based on a very simple representation of the physical system (e.g. the Lumped Parameter Model 'LPM') we decided that defining the deviation and bias from this estimate of groundwater age was of questionable value in these experiments. We agree that this output would be useful if the physical system was simulated in greater detail, and this is the focus of our current work.

The authors' suggestions as to the further use that could be made of the metamodel approach presented are interesting and seem potentially useful indeed, but I would have liked to see some concrete examples.

We agree that concrete examples of the use of metamodel approaches in water management will provide better context for the utility of this work. As the reviewer notes we had some examples in the results and discussion section. We have added some additional text to add further conceptual examples to the manuscript to provide further context.

This includes discussion of the insights that groundwater age can provide into how groundwater systems function, and how they may be changing over time. These estimated ages can also provide history matching targets for numerical models used to inform groundwater allocation decisions.

We have not included any numerical examples due to the length of the paper, and the intention to fully describe concrete examples in future publications.

All in all, the approach seems interesting, but the analysis itself would gain in depth by (i) refocusing it on the central question of how well the metamodels perform, (ii) reorganising and cutting down the discussion and (iii) potentially illustrating what is only hinted at in the concluding section.

We thank the reviewer again for raising these issues. In terms of the three points raised above we have done the following:

(i) Performance: We have replaced concepts of 'fit' with 'correspondence' to avoid ambiguity around any claims to fit a 'truth'. This makes the proof of concept focus of the paper more clear. This also avoids erroneous extrapolations given the fact we are fitting to a simple model which is also accompanied by simplification errors that have not been explored in this work.

(ii) We have reorganised and streamlined the discussion. We have retained sections that we believe are critical to highlight the additional insights and potential value of this type of meta modelling approach. We are hesitant to remove more of the discussion at this stage of the review process, as four other reviewers have not raised this concern.

(iii) We have added a brief discussion of how meta modelling outputs can be used to inform environmental management decisions.

Specific comments

L36: The parameters of a lumped parameter model are sometimes estimated from hydrogeological information rather than from tracer data (see for instance Abrams and Haitjema, Ground Water 56 (3), or Bailleux et al., Hydrogeology Journal 23 (7)). I think this point should also be taken up in the discussion (L616).

Amended text. We agree with the reviewer and have added a comment on this into the introduction. However, we did not revisit this in the discussion as that may distract from the 'proof of concept' focus of the paper (e.g. to estimate age on the basis of water chemistry).

L209: Do you mean median of MEAN residence times, or the median of the distribution of individual flow lines' residence times ?

Amended text. We meant median of the mean residence times and have adjusted the manuscript to clarify this.

L284: I do not understand why you write that the cause for the lack of fit of a single EPM to tracer data indicates that the transit time distribution has "changed over time". Isn't it rather because the real transit time distribution deviates significantly from the exponential model, as could be expected for more complex hydrogeological settings ?

We amended the text to clarify our meaning and address any confusion.

In this case, we are not talking about e.g., EPM vs BMM, but rather this is referring to different samples taken over time at the same site. The multi-age tracer long-term data appear to fit two different age distribution modes, one with a younger, and one with a slightly older age distribution. Such bi-modal age distributions are plausible, with the potential reason for this being seasonal changes, for example increased pump rates during summer, and increased recharge rates during winter. Therefore, we fitted two different age distributions to the data,

one matching the data indicative of younger water, and one matching the data indicative of older water.

L325: By “appropriate”, do you mean that the choice of these criteria has been tested in any way, or is it simply that this is what everyone in that field does, hoping for the best ?

Amended text. We agree the word ‘appropriate’ is ambiguous and so we have removed it. It is indeed standard practice, but there are issues with it as discussed in Schöniger et al. (2014) as cited.

L365: The matches shown in figure 6 and in the supplementary figures S1 to S4 are sometimes close and sometimes not. The widths of some estimated ensemble percentiles are also huge in some cases. So overall, unless you find a way to quantify the deviations between LPM transit time distributions used as reference, and estimated percentiles, I would be much more guarded in the assessment of goodness of fit.

Amended text. We agree with the reviewer. We are focussing on the broad ability of the metamodels to represent a LPM derived distribution. To better convey the ‘proof of concept’ focus of the paper we have amended the discussion, adopting the term ‘correspondence’ rather than ‘fit’ between LPM transit time distributions and those derived from the metamodels.

We also acknowledge the heterogeneity in the system that cannot be presented by an LPM, which may compromise the age estimation, and have also mentioned this in the text.

L384-391: The paragraph starting with “This finding […]” may be moved to the discussion, I think.

We are unsure of the reviewers meaning here as this section is already in the Results and Discussion section.

L415: Since the core of the manuscript is to test whether machine learning can be used to estimate transit time distributions from hydrochemical datasets, I think you need to include a way to quantify goodness of fit and deviations from the reference transit time distributions. Simply relying on a graphical comparison, and then declaring that the fits are good enough seems very unsatisfactory to me.

Amended text. We agree with the reviewer. Goodness of fit information was provided in the Supplementary material Table S2, and we have amended the manuscript text to make this clearer.

We also acknowledge that the goodness of fit metrics are less appropriate in this study, where the LPM estimates are also accompanied by a model simplification error. We note that a paired simple-complex model intercomparison could be undertaken to better estimate model structural errors associated with the LPM in this context (Doherty and Christensen 2011 discuss this method).

We have amended the text further, so that we discuss correspondence between the metamodel and LPM outputs, rather than ‘fit’. However we retain the standard goodness of fit metrics in the SI, as a metric of similarity rather than ability, in a heuristic sense.

L424: The phrasing “can successfully be used to predict groundwater age distribution” is too much of a statement to be taken at face value instead of the result of a thorough analysis of goodness of fit (see preceding comment).

Amended Text. We agree and have amended the sentence. This also links to the response above to L415.

L427: The sensitivity analysis is nice, but should not replace the much more central step of quantifying the deviation and the bias between SR and GBR models and the reference distribution. In my opinion, this step is still largely missing from the overall analysis.

We thank the reviewer for this comment. In future work we are looking at characterising bias and deviation of LPM models using numerical experiments where a synthetic truth can be used as an objective reference (this would address the fact that we don’t have ‘real’ age data, only interpreted age). At this point we believe that analysis of metamodel bias and deviation would be useful. However, we believe that this analysis goes beyond the scope of this paper which is to explore whether or not chemistry-age relationships are sufficiently strong to support metamodels.

L570: I am surprised that seasonal variations in groundwater heads would be so large as to affect the calculation of mean residence times in such a wet environment as New Zealand, at least for aquifers that are not largely fed from infiltrating streams.

L583: Same remark as above concerning water quality. How come water quality is so variable and dependent on sampling season for groundwater environments ?

This response is in regard to the reviewer’s comments on L570 and L583.

Most aquifers in New Zealand receive a significant proportion of recharge from infiltrating streams. For the particular case study, this is estimated to be more than 66% of the mass balance (as discussed in section 2.2 of the paper).

Rainfall, evapotranspiration, as well as water use, vary a lot between seasons in New Zealand. As a result, the rainfall and river recharge patterns and amounts may vary, which can affect groundwater chemistry and groundwater age distributions in the aquifer.

While this is generally the exemption, usually, shallow unconfined wells are affected, which, in winter, tap into the fresh local recharge pulse (young water). In contrast in summer, the fresh winter recharge pulse becomes depleted and, as a result, the discharge from the well at lower water levels reflects only the deeper older groundwater. This seasonal variability is clearly indicated by different age tracer concentrations at different seasons, and if the older and shallower water contrast in hydrochemistry, also hydrochemistry would vary.

Not much is known yet internationally about such seasonal variabilities of groundwater age in wells. It is especially in the southern hemisphere (and particular in NZ with its high-resolution tritium input available), where due to the absence of ‘bomb-tritium’ young and older water show a large contrast in tritium concentrations, that such effects can easily be studied. However, this is not the subject of this paper.

L598: I would use the word “reference” rather than “truth”, even in brackets.

Good point. Amended.

L607-609: I completely agree.