We thank all four reviewers for their insightful and constructive comments, shown in black text below, to which we respond below using blue text.

In our responses (blue text), all references to line numbers refer to positions in the revised manuscript with tracked changes active. On the other hand, all of the reviewer's references to line numbers (black text) refer to positions in the original manuscript. We have addressed all concerns and actioned most suggestions in the revised manuscript, as detailed in our responses.

We have combined the review comments, and our responses, from all reviewers into one document and introduced numbered subheadings in the remainder of this document so that we can more easily cross reference our responses between the different reviewers who raised similar concerns. For each reviewer, we have identified the comments that we consider to be more substantive versus those that are relatively minor.

In our view, the most substantive comments from the reviewers pertain to:

- Whether the chemistry-based metamodels should have been trained by matching to LPM-derived age distributions as opposed to some other information source (Reviewer RC1, RC2 and RC4)
- Whether the sites should have been segregated into different chemical clusters, then metamodels developed for each cluster separately (Reviewer RC1)
- Our explanations of the drivers for the inferred age-hydrochemistry relationships (Reviewer RC3)
- The appropriateness of the chaining approach we have employed (Reviewer RC4).

1.1 RC1 – Scott Wilson

This paper makes an excellent and novel contribution to predicting transit times using a wider dataset than isotopic age tracers. I found the text enjoyable and easy to read, and the perspective is fairly balanced. I have two main comments on the approach taken, which have some bearing on the conclusions that can be derived from this work.

1.1.1 Training chemistry-based metamodels on LPM-derived age distributions

The main drawback of the approach taken is that the chemically based models use the lumped model age estimates as a response variable, ie training a model on another model which is acknowledged as having shortcomings, although this is the motivation for the paper. The difficulty is that this creates an ambiguity as to whether mismatches in the trained models are due to poor lumped model estimates, or poor local performance of the trained models, or both, or something else (eg parameter or model selection). As a suggestion, an alternative or complementary approach would be to firstly train models to predict the isotopic tracer con\centration. This would provide some prior information on the mismatch between the lumped model predictions, and ensemble predictions. This approach could perhaps inform how the lumped parameter estimates could be improved, which was suggested in 553.This is

not a step necessary for this paper, but perhaps something that could be carried out in future work.

This is a very good comment, which we interpret may have arisen in part because our original manuscript neglected to explicitly make the point that groundwater age and groundwater age distributions cannot be measured directly – *they must be inferred using some form of model*. Our original manuscript also laid out the two main options for modelling groundwater age – LPMs vs. numerical flow/transport models – and discussed their strengths and weaknesses.

- We now make this point that groundwater ages must be derived from models, at **Line 30**.

The reviewer commented that LPM-derived age estimates carry some uncertainty which would then be brought over into the metamodels. We agree, and this is of course also true if the age estimates had been derived from a numerical flow/transport model instead.

- We now emphasise that metamodels will inherit the uncertainties of the models they were trained on, at **Line 76**.
- We have also made an additional point about potential spatiotemporal correlation and bias in the age estimates from LPMs vs. numerical flow/transport models at **Line 37-39** and **Line 46-48.** The point is that both types of models contain uncertainties but the LPM-derived ages may be less subject to spatially correlated biases. This can be a strength or weakness depending on the objectives of any given study, but we consider it to be helpful for our present investigation, as noted at **Line 104-107**.

We acknowledge but disagree with the reviewer's suggestion that our metamodels could have been trained directly on the measured concentrations of the individual age tracers, rather than on a single LPM-based age model based on a combination of all age tracer data. This is for several reasons. Firstly, the different age tracers have different age ranges over which they are valid. Secondly, some of the age tracers are gases whereas tritium is part of the water molecule, so are subject to different processes in the aquifer system (e.g. gas exchange). Thirdly, some of the age tracers can be subject to degradation or alteration in the aquifer, so the measured concentrations in the sample may not represent their concentrations in the original groundwater recharge. For all of these reasons, best practice in our lab and elsewhere is to carefully compare and evaluate the time-series measurements from all available age tracers and derive the most robust LPM interpretation consistent with them all. As a side note, we do agree with using the measured concentrations of the age tracers when calibrating a numerical flow/transport model, but that is not directly relevant to the present manuscript.

- We now explain why the LPMs are developed by fitting to all available age tracer data for the relevant site, at **Line 261-263**.

The reviewer is correct though, that we have considered that errors in the underlying LPM interpretation may be one reason why the metamodels fit more poorly at some sites. Indeed, a whole section of the original manuscript was dedicated to this idea (Section 4.3.2).

- We have not modified Section 4.3.2, because we consider that it already encapsulated the point being raised here by the reviewer.

1.1.2 Segregating sites based on chemistry prior to developing metamodels

The modelling approach applied here is to generate a global model from a subset of individual models, and it is assumed that the input data are spatially and temporally independent. However, the hydrochemical clustering results do indicate that the chemistry data have a predictable spatial and temporal variability. Some evidence of this influence is apparent in the results (eg Fig 6, T2_34), and hence in the applications section some spatial and temporal discrepancies are acknowledged. To overcome this, it may have been beneficial to train models on each hydrochemical cluster, although it has to be acknowledged that there is little data available for clusters 4 and 5. An alternative approach would be to introduce some additional predictive parameters in the model to account for spatial and temporal variability, eg elevation, depth, position, hydrochemical cluster. Some of these parameters have been used for validation, but they could also have been training parameters, or tested to see if they do inform model predictions. In doing so, one could have more confidence in the application of the models to areas with no age data.

We agree that in some applications it can be helpful to pre-segregate observations and develop metamodels on the populations separately. Indeed, we had considered this approach as we were developing our own methodology. However, we opted to attempt to develop a single metamodel for the entire dataset, without pre-segregation, primarily to determine whether the SR and GBR algorithms could themselves account for any inherent differences in the age-chemistry relationships between sites. As shown, our results demonstrate that the models did perform well across all hydrochemical clusters.

- At the start of the section in which the SR and GBR methodologies are described, we have inserted a short paragraph to explain why we did not use pre-segregation of sites, at **Line 292-297**.
- We have added a single sentence in the results section at **Line 415-417** to explain that site-specific biases in model fit were not systematically related to the site's cluster assignments shown in Figure 1, which justifies our approach of developing metamodels for all hydrochemical clusters simultaneously,

We appreciate the reviewer's suggestion that other datasets such as elevation, well depth, well location could also be used as input data, along with the site-specific hydrochemistry data. Indeed, assessing the information content of different input datasets is one of our research interests, and we are currently looking into using hydrophysical parameters in addition to the hydrochemistry. This is planned for a future paper.

- We have added a sentence about future research in this direction to Section 5 at **Line 641-642s**.

1.1.3 Minor comments

The paper would also benefit from some corections and the clarification of some points listed below.

Title and line 60: SR and GBR methods are not metamodels per se. They are applied in this paper as metamodels because they are trained on the LPMs rather than raw observation data

We agree and have changed the introduction to clarify how we use the term metamodel in this paper at **Line 67-69**.

Line 97: Should be Heretaunga Plain not Plains (also elsewhere)

No change has been made to the manuscript because "Heretaunga Plains" is the most common usage and is also being used by Hawke's Bay Regional Council, the regional governing authority for this area (e.g., https://www.hbrc.govt.nz/environment/aquifers/, https://www.hbrc.govt.nz/home/article/851/the-plan-for-healthier-heretaunga-plains-waterways?t=featured&s=1).

Line 125: There are red lines on Fig 1 which are unreferenced. Are these flow barriers? It seems odd that there is a flow path towards a flow barrier (centre top)

Good find. The red lines identify areas where Morgenstern et al. (2018) found indication that there is no surface water flow contributing to the main aquifers. We have added an explanation to the figure caption.

Line 154: The clustering detailed in the hydrochemistry section provides some background context, but is not used in the modelling or subsequent analysis.

See our reply in subsection 1.1.2 above. As noted, we now explain why the hydrochemical clusters are not used as input to the metamodeling, and we discuss the implication of this approach in our results section.

Line 219: It's good practice to state that this is the response variable for the statistical modelling, and the hydrochemistry data are the predictor variables

Good point. We have now added these at **Lines 229–230 and Line 240**.

Line 247-249: How much error is the distance to these input signal datasets likely to introduce to the age estimates, and how would that compare to the error introduced by the EPM?

We agree that if a different model was used, that spatial correlations within their error functions can be explored, however in this project we adopted an average error term for simplicity. Spatial correlation within error functions could be investigated in future work. However, we believe that this choice doesn't undermine the general conclusions of this work. Generally, the model simplification error introduced by the LPM would tend to create larger errors than the spatial distance to the input signals (see e.g. Doherty and Moore 2019; White et al. 2014). As discussed in 1.2.1, a detailed analysis of the possible errors and their propagation through to the age distribution estimates is a complex topic and beyond the scope of this paper.

Doherty, J. and Moore, C.: Decision support modelling: data assimilation, uncertainty quantification and strategic abstraction. Groundwater, 58(3), 327-337 doi: 10.1111/gwat.12969, 2019.

White, J.T., Doherty, J. and Hughes, J.D.: Quantifying the predictive consequences of model error with linear subspace analysis. Water Resour. Res., 50(2):1152-1173. DOI: 10.1002/2013WR014767, 2014.

Line 269: Were not was

This seems to refer to this sentence: "*SR grammar rules permitted arithmetic, exponential and logarithmic functions; permission of conditionals (e.g. if/then statements) was also assessed in terms of ability to improve model fits*." We consider that the original grammar was correct because the word 'was' refers to the word 'permission', which is singular.

Line 273: The primary aim of tuning is to improve model performance, not assist convergence

Agree. Have amended text accordingly (now at **Line 304**), as follows: 'The hyperparameters were tuned to find the optimal parameters (tree depth = 4, sample split = 2 and learning rate = 0.05) that result in the best performance of the models.'

Line 278: The terms 'chained' and 'unchained' models is unorthodox, and perhaps not an apt description of what the models represent. Perhaps these would be better referred to as 'independent' (see line 276) or 'individual models', and the chained models as 'ensemble models'

No change has been made in response to this reviewer comment, though we do acknowledge that the terminology is tricky, given the complexity of the methods we have used.

The terms for 'unchained' and 'chained' models are based on the "chained regression" modelling approach, which is for example used in the RegressorChain scikit.learn class that we are employing in the GBR models. This terminology is widely used.

We do not favour the reviewer's suggestion to replace the term 'unchained' with a term such as 'independent' because for clarity we repeatedly use the latter term to emphasise that unchained models are constructed individually (i.e. independently) for each of the modelled percentiles (the chained models are too). To use the term 'independent' in more than one way would make the manuscript more difficult to follow.

We also do not favour referring to the chained models as 'ensemble models' because we in fact applied an ensemble model approach with both the 'unchained' and the 'chained' models. We believe this is already adequately explained in the text and depicted in Figure 4.

Line 286: Why do the train/test splits differ for the two models? This approach doesn't enable a clear comparison of modelling performance between the two models to be made

We compared a range of test/train split ratios based on typical approaches used by practitioners of these modelling methods. We now point this out on **Line 317-318**. Later in the same paragraph we have modified the wording to explain that the final selection of test/train split was based on these tests that we had conducted.

For GBR, we found that a 10/90 % test/train split (i.e. K = 10-folds) achieved a better, more consistent performance with our smaller data set (76 inputs) compared to larger split ratios

(e.g. 20/80% and 30/70%). When we decreased K (e.g., K=5 etc.), we provide a smaller training set for GBR models to learn from, and the performance of the GBR models is limited by the amount of data.

For SR, a test/train ratio of 33/66 provided good stability. Due to the random sampling procedure in the SR technique, we understand that even if we had followed the same split ratio, there could be some method specific small discrepancies in the test/train sample selection procedures between SR and GBR procedures.

286: As a comment, a 10/90 split is quite heavy-handed and could lead to overfitting. The unchained GBR R2 values are very high, although this is also true for the SR R2 values

No change has been made in response to this comment. Here, the split (K) is defined by the kfold cross-validation procedure, and this 10-fold cross-validation resulted in 10/90 % test-train split. 10/90 % test-train split in GBR achieved a better, more consistent performance with our small data set (76 inputs) compared to larger split ratios (e.g. 20/80% and 30/70%). Scikit learn cross-validation usually uses 10-fold as the default value and it is known that this 10-fold work well with smaller data sets as the models are trained with 90% of data, resulting in lower bias in estimates. Also, by using a strict model selection criterion we have minimised the effects of overfitting. Also see response to previous comment.

Line 290: There seems to be an error in the Pearson formulas

Good catch. Symbols must have been converted between versions. Have corrected this (now **Line 326**).

Line 375: Last Glacial (is a noun)

Good point. Changed in the text at **Line 139** and **Line 411-412**.

Line 399: The third value is 1.7 (ie >1)

Also a good catch. Corrected with 0.17 (now **Line 439**).

Line 405: Perhaps the models could achieve good age distributions with substantially less parameters?

We agree that our results demonstrate that good estimates of the age distributions can be achieved with fewer chemical parameters as input. While this is implicit in the discussion in our original manuscript, we didn't say so directly so have now added this point at **Line 448-450**.

Line 410: It might have been more informative to plot the cluster results here rather than the ensemble weights, since the most informative parameters are already described in the text. As a reader, I'm intrigued by the relationship between the model performance and the clusters.

No additional change has been made in response to this comment. We opted to present the results according to parameter weights in order to demonstrate the hydrochemical variables that exert greatest influence in the model fits across all sites. As per subsection 1.1.2 above, we have already added a sentence to explain that there were no strong differences in model fit

for the different clusters, at **Line 417-420.** Indeed, some of the clusters contain very few sites so we would not likely expect to distinguish statistically different parameter weights if metamodels were developed for clusters individually.

Line 434: Perhaps water chemistry has some influence of the source rock, which wouldn't necessarily be reflected in the age estimates

No change has been made in response to this comment because we consider that it is consistent with what the original manuscript already reported. We agree that the chemistry at any point reflects the water-rock interaction both from the point of recharge and along the whole flow path to the sampling point. There are other factors that would also affect the evolution of chemistry over time, such as microbial processes. As noted in subsection 1.1.2 above, our approach in this study was to attempt to develop metamodels for all sites simultaneously, to determine whether they could identify these age-chemistry relationships themselves. Our methods were shown to be effective in this context, but metamodels would need to be retrained if these same methods are to be applied in a different catchment.

Line 517: It's ambiguous how these parameters were treated. Were their values set to the detection limit?

As the parameters B, F, oxygen-18 and deuterium are not routinely monitored, we used dummy values equal to the average across all samples in the predictor dataset. Due to the generally low weighting of these parameters, there should be very limited impact from this on the results. We have clarified this at **Line 569-570**.

Line 522: I think this claim is a bit of a stretch since there are no spatial aspects to this study. The model is aspatial, and global, and appears to generalise well to most, but not all the data. The model has the potential to be applied to other areas with confidence if the successful or unsuccessful predictions could be identified as having an association with something eg a particular cluster. NB this comment also applies to the last sentence of the abstract.

We accept this comment. We have removed the term 'spatial extrapolation' (now **Line 574**), but we maintain that the SR and GBR models can produce estimates of age distribution in areas of similar hydrogeological regimes even where no age tracer measurements have been made. The application described in Section 4.3.1 illustrates that doing so provides rich information about age distributions in areas where they would not have otherwise been available. We have also added the caveat 'in a similar hydrogeological regime' to the abstract.

Line 543-547: I don't think these statements are valid, particularly in light of the preceding sentences. There is no spatial aspect to the modelling to this modelling approach, it only uses age and chemistry data.

We feel that the reviewer has slightly misinterpreted the intent of statements made in the original manuscript. Our view is that the *metamodeling approaches* demonstrated in this paper can be applied in other catchments where sufficient age and chemistry data are available. While we already note that it is not reasonable to apply a metamodel from one catchment to another catchment, we state that within a single catchment the metamodels can provide estimates of age distribution where no age tracer concentrations have been measured, as long as chemistry data are available. We have slightly modified the sentence on **Line 596**

to indicate that applications within a single catchment must pertain to a single hydrogeological regime.

Line 578: Which of these models would you have the most confidence to apply elsewhere?

The performance of these models in other catchments will likely be comparable if the catchment has a similar hydrogeological regime. The similarity or not of the hydrogeological systems and the key hydrochemical processes are likely going to be the most significant considerations for applying either of these modelling approaches, rather than any differences between these modelling approaches. As mentioned above, the purpose of the paper was not to compare the model performance but rather to demonstrate the utility of metamodels in this context using two different models and their typical model set up (mimicking practitioners use of these modelling methods). As stated in the original manuscript (from Line 383 onwards), we believe that either method would be suitable for similar application elsewhere, depending on user requirements, experiences and understanding.

1.2 RC2 – Camille Bouchez

This work explores the use of metamodelling techniques to predict groundwater age distributions from hydrochemistry. It is a novel and interesting contribution aiming at increasing the availability of groundwater age information from easily available hydrochemical data in catchments. The knowledge gap is convincing and the paper is nicely written. However, I have some comments that should be addressed before publication.

1.2.1 Training chemistry-based metamodels on LPM-derived age distributions

My main concern comes from considering the LPM-derived age distributions as the true representation of groundwater age distribution, which is later used as the metamodel prediction target. I understand the interest of this choice, but I think it is a strong assumption that should be further discussed in the paper. In particular, the following points are missing:

- Where are the age tracer data? They are not in the Supplementary Material as indicated l. 219, and I could not easily find them in Morgenstern et al. 2018. There is an extensive description of how these data where acquired (l. 228-238) and how they are used to fit LPM (l. 238-264) but results are never presented in the paper while they are very important. Age tracer data fitted by the LPM must appear in Supplementary Material, to evaluate the confidence in the LPM predictions later used.

We thank the reviewer for pointing this out. The majority of, and currently publicly available age tracer data, together with information on tracer inputs, are provided in Morgenstern & van der Raaij (2019). It was an oversight from us to not include the reference to this report, and we have amended **Line 241** accordingly.

- Without this, it is hard to evaluate uncertainties associated with the LPM-derived age distributions. Would it be possible to estimate the uncertainties? How much are the trained models sensitive to the LPM? Could uncertainties in LPMs explain part of the errors?

We agree that this would be an interesting topic to study, but this is a separate study in itself. Fitting LPMs to age tracer datasets is subject to several error sources, including the number of different age tracers used, their age and error ranges, the number of measurements at the same XYZ location over time, and the use of binary or single mixing models (dependent on the hydrogeologic system). Therefore, the combination of the errors and how they propagate into the metamodels is quite complex. A paired simple-complex model comparison would be one such methodology that could be used (Doherty & Christenson 2011). We also did touch on in the paper that potentially the metamodelling approaches could be used to identify issues with the LPM interpretations.

Doherty, J. and Christensen, S.: Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. Water Resour. Res., 47, W12534, https://doi.org/10.1029/2011WR010763, 2011.

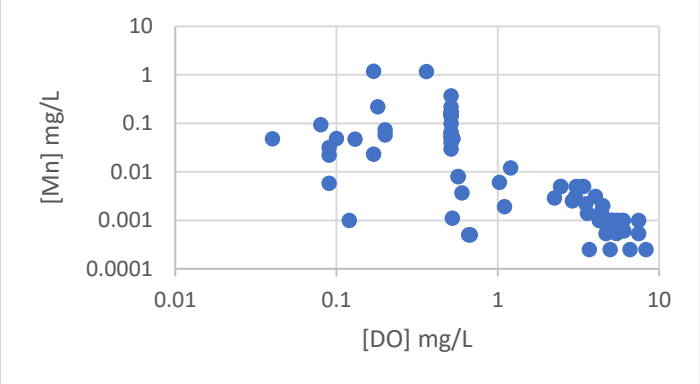
1.2.2 Explaining the age-hydrochemistry relationships

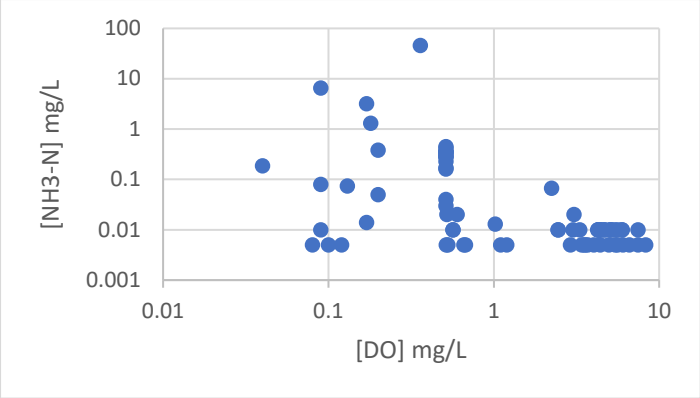
My second main concern comes from the relationships obtained between hydrochemical data and groundwater age distribution and the processes that could explain them.

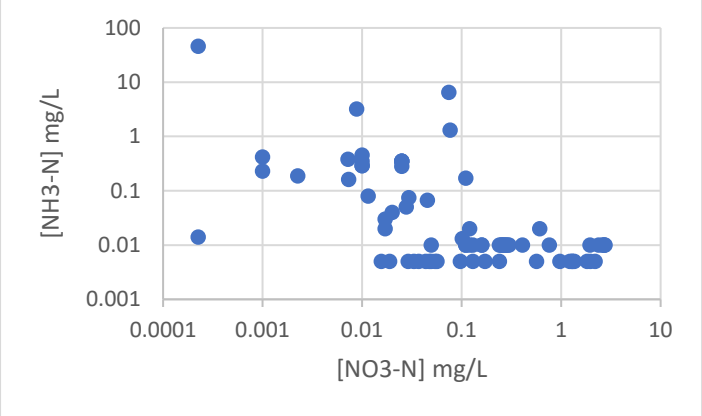
- Based on which argument and figure can you tell that "NH3-N, Fe and Mn all tend to increase with groundwater age, whereas concentrations of DO and NO3-N tend to decrease" (l. 424)? This affirmation does not appear clearly on Figure 8 and it does not appear clearly either in the correlation matrix Figure 2.

Figure 8 displays the relative weightings of the hydrochemical parameters in the metamodels. It does not, nor is it intended to, display the correlations between the variables.

Figure 2 does indeed show the magnitude and direction of correlations among hydrochemical variables across the whole dataset. For example, DO is seen to be positively correlated with NO_3-N (r = 0.2) but negatively correlated with Fe (r = -0.2), Mn (r = -0.3), NH_3-N (r = -0.1) and PO_4-P (r = -0.4, mistakenly labelled as DRP, which will be changed). While some of these appear weak, it's because Pearson's r is a measure of linear correlation whereas several of the hydrochemical relationships are known to be non-linear, as shown for the Heretaunga Plains data below (note log scales). The non-linearity in relationships between redox-sensitive parameters is expected given that they tend to be consumed through microbial respiration in a step-wise sequence; for example, NO_3-N is usually has to be largely depleted through denitrification before appreciable concentrations of NH_3-N build up. We have added a short comment to this effect at **Line 219-221**.







- I found interesting to try to quantify the consumption of DO in the catchment, by assuming that the organic matter oxidation is only related to DO. However, no explanations are given on how the average rate constant was derived and additional information are required. A first-order kinetics on the DOM concentration was considered, therefore not accounting for the DO concentrations (if I understood correctly from the reference given). Is it correct? It should be specified. Which groundwater age percentile was considered for the calculation? How were the DOM concentrations averaged?

We appreciate this interest from the reviewer. Our aim here is just to provide a general indication of the sorts of insights that could be generated on DO consumption rates if better data were available. We caution that our original manuscript acknowledged that the rate of DO consumption can only be evaluated semi-qualitatively at best from the data we have available. We have modified the wording on **Lines 481** to make this clearer.

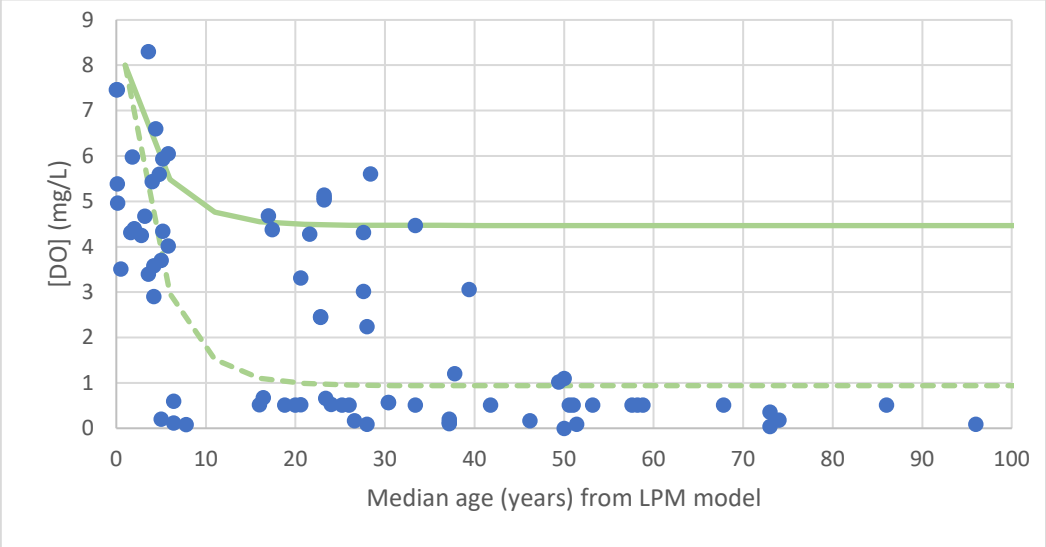
We also have corrected an error at **Line 491** in the reporting of our estimated rate constant (k). The original manuscript stated that k = -0.6, but actually this should have been log k = -0.6.

To make the calculations easier to follow we have added the first-order kinetic rate expression to the text at **Line 488**.

An important point that we didn't make clearly in the original manuscript is that the decline in DO concentration over time depends on the organic matter concentration in the groundwater, either introduced via recharge or acquired during groundwater passage through the aquifer. We have clarified this at **Line 489**.

An illustrative plot is shown below. It displays the measured concentrations of DO vs the median of the LPM-derived age distribution for each site. Two model curves are shown based on the rate equation now given in the text. The models assume that the governing reaction is $CH_2O + O_2 = H_2O + CO_2$. Both model curves assume the same rate constant (log k = -0.6) and the same initial concentration of DO (8 mg/L), but the top and bottom curves assume different initial organic matter concentrations of 3.3 and 6.6 mg/L, respectively. Note that organic matter concentrations are not typically measured in Heretaunga Plains groundwater or other aquifer systems in New Zealand, so these values were just selected to bracket the dataset – but they are potentially reasonable based on overseas studies.

We welcome advice from the Editor as to whether inclusion of such a plot would be valuable for the manuscript. Our initial sense is that it is based on so many assumptions that it would be better to exclude, but we would welcome feedback.



- The inverse relationship between age and temperature is not expected as we would expect that older groundwater shows higher temperature. But this relationship is really strong and I think this paper would highly benefit from a close look at this relationship and clarifications in the explanations given. I do not understand the calculation of the activation energy made and I doubt the interpretation that is made from it. First, it somehow considers an aggregation of all reaction types. Secondly, where does the k1/k2=0.8 come from? Here, the age ratio is 0.8. But why would the kinetic rate ratio be equal to the age ratio? I agree that an increase in T would increase the reaction rates. However, how do you relate this to the effect of T on modelled age? Please clarify the process that is presented here to explain the inverse relationship between age and temperature. I would be more convinced by a hydrological explanation. The paper would benefit from a more convincing explanation of the relationship obtained between temperature and age.

We appreciate the reviewer's interest in this finding. Indeed, the strong inverse correlation between T and the modelled groundwater age was a surprising result for us as well.

At **Line 527-529,** we have clarified that the available data from this study do not permit elucidation of the cause(s) of the strong inverse relationship between T and estimated groundwater age, and that the following paragraphs simply present two concepts that could be explored through further investigations.

We feel that the Arrhenius Equation is quite well established and so shouldn't require further explanation in the text, but we have added a reference to Langmuir (1997) at **Line 535**, which contains these same equations.

In the application to the present study, the Arrhenius equation has to be aggregated across all reaction types because there is no available means of identifying specific types of reactions that may be more important than others, or applying the equation to any particular type of reaction. We clarify this on **Line 538-539**.

The reviewer asks why the reaction rates at two temperatures should be tied to the ratio of ages. This is because we assume that there is a fixed reaction rate constant for each temperature, so a given reaction will proceed at a different rate for the two temperatures being compared. For the reaction to have progressed to the same degree from the same initial chemical condition, our results suggest that the warmer system will take less time. Assuming that the form of the kinetic reaction equation doesn't change (for example it remains first-order), then the ratio of times for the reactions to progress to the same point should be equal to the ratio of their reaction rate constants.

- Relationships between Ca, Mg, Na, K and SiO2 and age would highly depend on the aquifer lithology. Would these elements be better predictors of groundwater age if an *a priori* classification based on the rock lithology was made?

This is a suggestion made by Reviewer RC1, to which we reply in subsection 1.1.2 above. As noted there, we have added an explanation of why we did not apply an a priori segregation of sites based on rock lithology at **Line 292-297**, and we have added a sentence to our results section at **Line 415-417** that justifies our approach.

1.2.3 Minor comments

Fig.1: What are the red lines?

Good find. The red lines identify areas where Morgenstern et al. (2018) found indication that there is no surface water flow contributing to the main aquifers. We have added an explanation to the figure caption.

l. 136: it would be interesting to give the value of the recharge rate of the area

Rakowski & Knowling (2018) estimate the total recharge to the aquifer to be approximately 264 M m3/year, of which losing rivers contribute about 185 Mm3/year and 79 Mm3/year come from rainfall recharge. We have added the total recharge to the updated manuscript at **Line 150**.

Rakowski, P. and Knowling, M. J.: Heretaunga aquifer groundwater model: development report, 182 pp., 2018.

Line 140: what is the confined aquifer zone near the coast? Maybe worth showing on the map?

We agree that it could be useful to show the confined aquifer zone on the map. However, the confined aquifer boundary in the Heretaunga Plains is currently in the process of being updated with new data. As an alternative, we have added the currently mapped extents of fine (sand, silt, clay) terrestrial and estuary deposits at the ground surface to Figure 1 and Figure 3.

Line 195: there is a confusion between the text and Fig. 3, one refer to mean residence time and the other to the 50th percentile, please correct.

Good catch. We have corrected this in **Lines 211-212**.

Figure 6: At least for the example given in Figure 6, the lumped parameter model should be described in the main text (singular or binary EPMs? Which values of the parameters?)

We have added this information to the figure.

Line 337: MAE : Mean Absolute Error?

Good catch. We must have missed explaining this in the paper. In this paper, MAE stands for Median Absolute Error. We have added this to the text at **Line 372**.

Figure 8: change DRP for PO4-P as this is how it is referred to in the main text

We have corrected this in Figure 2 and Figure 8 accordingly.

Line 540: I wonder of the generalization of the approach and on the application of the trained model elsewhere. The obtained hydrochemistry-age relationships are not easy to explain (at least for temperature), and therefore it is difficult to tell if they are applicable elsewhere or if they are only related to some local effects. Would other predictive parameters such as depth, distance to the river, or elevation inform on water age predictions?

We agree. We are currently looking into using hydrophysical parameters in addition to the hydrochemistry. This is the plan for the next paper. We have added a sentence about future research in this direction to Section 5 at **Line 641-642**.

The authors acknowledge that the work might be only applicable to the selected catchment. Is there another similar catchment, where age data are available and where the models could be applied to determine groundwater age distributions from hydrochemistry, in order to validate the method?

Testing and validating of the models in other catchments is planned for further work.

1.3 RC3 – Anonymous referee

Overall, this is an interesting metamodeling application using water quality information to emulate a lumped-parameter model and make forecasts of groundwater age. Two methods

were used (gradient boosted regression and symbolic regression) with advantages to each and with generally similar performance. The authors also make a detailed interpretation of the parameter and model behavior.

This is a fine contribution and I have just a few minor comments to consider.

1.3.1 Minor comments

Line 61: There is some ambiguity to how the model is described here. It's not really trained on data, but rather is trained on the LPM model that, in turn, is trained on data. Being super clear here is important, particularly for readers less familiar with metamodeling

This is a good suggestion. As noted in our response to RC1, we have clarified this on **Lines 66-69**.

Figure 1 and in the text: The clusters from previous work are both identified on the figure and in the text, but no context is provided beyond a reference to previous work. A sentence or two would be key to explain this.

We use the clustering from previous work for two purposes. The first purpose is just to simplify the description of hydrochemical variations across the Heretaunga Plains aquifer system. We have added a sentence to explain this rationale at **Lines 182-184**.

The second purpose we use the clustering for is to test whether the age models are able to perform adequately on all groundwater chemical categories, but without have pre-segregated the dataset and training separate machine learning models for each cluster. Please note our responses to similar comments from Reviewer RC1 in subsection 1.1.2 above. As noted there, we have added an explanation of why we did not apply an a priori segregation of sites based on rock lithology at **Lines 292-297**, and we have added a sentence to our results section at **Lines 313-315** that justifies our approach.

Figure 2 and elsewhere: Many of these water quality constituents are obviously identified by their chemical formulae, but some of not defined. Even if it's in supplemental material, a table defining the quantities would be helpful.

We agree that this is an oversight from us. We have added a table to the supplemental material (new Table 1 in the pdf) as suggested and given a reference to the table in the text at **Line 174**.

Line 290: There seems to be a formatting glitch here – hard to understand what the equation is meaning to explain.

Good catch. Symbols must have been converted between versions. As already noted in our response to RC1, we have corrected this (now **Line 326**).

Line 327: more formatting glitches

Another good catch. Have corrected this (now **Line 362**).

Lines 359-362: This is a great point and I appreciate the context because it's true that the extrema of the distribution would be of interest to many users.

We thank the reviewer for support of this point made in the original manuscript.

1.4 RC4 – Anonymous referee

This manuscript aims at assessing the validity of using two machine leaning techniques to extrapolate beyond available groundwater age data and infer the lumped RTD from hydrochemistry.

This contribution is novel and appears quite appealing to complement tracers dataset which are costly and time consuming.

The manuscript is nicely written and easy to follow.

1.4.1 Training chemistry-based metamodels on LPM-derived age distributions

I have reservations about the choice of the LPM models as calibration targets. I understand that the study is closer to the reality in which the age distribution is unknown. Still, I consider that it would have been much stronger to test the validity of the methodology on a pure synthetic case controlling every aspect of the problem: data and associated uncertainty, full shape of the age distribution, etc. An important aspect as well is that, without a priori information about the age distribution, a few LPM differing in their hydrogeological conceptual representation can equally fit. My point is that it is difficult to evaluate the validity of a calibration or inference methodology on real largely under-constrained cases. One way to tackle this would be I think to highlight the fact that the system studied here is a "not so complex" system (a textbook system?) and have been widely studied so that the target LPM is a more than reasonable estimation (see my minor comment below).

Please refer to our response to Reviewer RC1, given in subsection 1.1.1 above. As listed there, we have made several modifications to the manuscript to clarify and justify our approaches.

The reviewer suggests that we could have conducted a purely synthetic study based on models in which all parameters and processes were fully constrained. We did consider this idea and may pursue it in the future, but we ruled it out for the current investigation because we considered the chemical process datasets and models to be too uncertain to be useful. More detail is on this below.

In order to undertake a purely synthetic study we would need to develop a model of the groundwater flow and transport system. Much of this modelling has already been undertaken. Simulations are already possible for tritium transport and particle tracking or direct age simulation to enable the age distributions of groundwater to be evaluated spatiotemporally across the model domain. Further work is underway to improve the existing groundwater flow and transport models but is not yet published.

In order to undertake a purely synthetic case study we would also need to implement a model of the chemical evolution of groundwater over time. One option would be to use a forward

simulation based on lab-derived reaction rates, e.g. using a programme such as PHREEQC, Geochemist's Workbench, etc. The other option would be to apply a reaction rate model based on field observations, such as PROFILE or ForSAFE (see Sverdrup et al. 2019, cited in the original manuscript). In either case, we would need to be certain that these models contain geochemical processes that are relevant to our particular field location.

But just as important, regardless of which type of geochemical reaction model was selected, we would have the major challenge that we do not have data including but not limited to: 1) the mineralogical composition of the aquifer materials or how they vary spatially; 2) the key factors such as reactive surface area, which control water-rock reaction rates; 3) microbial processes and their rates and spatial variability; or 4) dissolved and solid phase organic matter concentrations and their reactivities.

The result is that we felt that we could conceivably develop a groundwater flow and transport model with accompanying estimates of age distributions, but our ability to model the geochemistry would be too underconstrained to be useful.

1.4.2 Chaining approach

I have reservations as well about the independence for the percentiles and the further chaining approach. It appears to me that it goes again physics and flow mechanics to consider percentiles as separate entities, and not the age distribution as a whole. My point is that a LPM or numerically-generated distribution lies on a hydrogeological conceptual representation which describes the functioning of the system. It has been shown (Leray et al, 2019: https://doi.org/10.1016/j.jhydrol.2019.04.032) that local modification of the system properties affects not only local flow lines and mass balance locally but the overall response and functioning of the system and consequently the age distribution. So it is confusing to me that the distribution is considered by part (even if the chaining approach intends to reconstruct the puzzle)

We thank the reviewer for this insightful comment. We agree that the percentiles in a single age distribution must have a mathematical relationship driven by the groundwater flow regime. The approach that we took in the manuscript should not be taken as a disagreement with this statement – rather, our approach was followed to test the appropriateness of using LPM-derived age distributions as our modelling objective.

We initially constructed unchained models for the *individual* percentiles as a means of testing the validity of the shapes of the age distributions produced by the LPMs. By modelling one percentile at a time, we aimed to determine whether there were any sites for which the SR or GBR models produced misfit, which may have demonstrated that the shape of the LPM age distribution was inappropriate at those sites. Then by comparing the age estimates derived for different percentiles (i.e. from different unchained SR or GBR models) at single site, we could diagnose whether the LPM mean age was erroneous (as shown by systematically incorrect age estimates for all percentiles), or the LPM's age distribution had the wrong shape (as shown by different misfit for different percentiles), or a combination of both. We have clarified this at **Lines 313-315**.

The reason for subsequently constructing the chained models was already described in the original manuscript (now at **Lines 340-342**): "[Chaining] was done to ensure that the separately simulated percentiles had an appropriate relationship to each other, e.g., that the

value for the 10th percentile in the age distribution for any sample had to be greater than or equal to the 5th percentile in the age distribution at the same sample."

Our results section already discussed the insights that could be gained by comparing the quality of SR and GBR model fits across sites and percentiles. We have added a sentence at **Lines 417-420** to explain that "there were few sites for which clear errors in the shape of the LPM-derived age distribution could be identified based on differences in the quality of fit of unchained model fits across different percentiles, so we conclude that the LPMs applied in this investigation are generally appropriate to represent the age distributions in the study area."

1.4.3 Minor comments

Line 26: I would write the age as plural ("understanding the ages of water") to reinforce the fact that natural groundwater systems are made of a wide variety of flow paths and consequently of residence times (or ages). If it is correct grammatically of course.

Agreed. Text changed at **Line 26** as recommended.

Line 53: "most such previous studies". Revise

No change has been made in response to this comment. Perhaps we are being slightly pedantic, but we have opted to keep the original wording. The use of the term 'such previous studies' is to show that we are referring to the 'various less time and cost-intensive methods have previously been trialled to increase the amount of available groundwater age data in areas where no age tracers have been sampled' – in other words, we are not referring to all studies about groundwater age.

Line 218: I am not an expert but should it not be half of the detection limit?

We have clarified the text at **Line 236-237** to say "Censored and uncensored results below the highest censoring threshold for each parameter were replaced with the corresponding analytical detection limit (Helsel et al., 2020)". The reason for this approach is that it isn't possible to tell the difference between concentrations reported (for example) as <0.05, 0.03, <0.06 so they should all be considered equivalent and set at 0.06 for SR and GBR model training.

Lines 252 to 254: It is argued that the EPM provided good matches for a wide range of New Zealand systems. A fault-bounded, local, relatively homogeneous and thick system with uniform recharge rate upstream and zero recharge rate downstream looks like an EPM to me. So, I think the validity of the EPM should be argued considering specific aspects of the system (that may be quite similar to other sites in New Zealand)

We agree that this is an important aspect to include. We have added a sentence to clarify that the EPM does indeed match the hydrogeological system, at **Lines 278-280**.

Line 278: to differentiate.

We have amended the text at **Line 310**.