Consolidated Replies to Reviewer Comments

Matthew J. Knowling[1], Jeremy T. White[2], Catherine R. Moore[2], Pawel Rakowski[3], and Kevin Hayley[4]

[1]Corresponding Author: GNS Science, New Zealand; m.knowling@gns.cri.nz
[2]GNS Science, New Zealand
[3]Hawke's Bay Regional Council, New Zealand
[4]Groundwater Solutions Ltd, Australia

December 16, 2019

Here we respond to each comment raised by the Associate Editor and the reviewers. Comments are shown in *italics* and are followed immediately by our response.

1 Reply to Associate Editor Dr Fabrizio Fenicia

The paper received an interesting mix of reviews, some enthusiastic, some very negative. I think the authors should seriously consider the comments of the negative reviewer, which they already started to do, and revise their paper accordingly. The paper will go to another round of review, to evaluate that the changes made are satisfactory.

We have carefully considered all of the reviewer comments, as per our detailed responses below. We have revised the manuscript accordingly. The manuscript has been improved as a result.

While Prof. Ferre suggested no changes to the manuscript, we have responded in detail to his thoughtful comments. We have addressed almost all

of the comments by Mr Turnadge through additions and modifications to the manuscript (and indicated otherwise where appropriate). We have addressed the anonymous reviewer's comments surrounding the level of methodological and case study details presented. We have also addressed their comments concerning the contribution of our paper with respect to previous studies by clarifying the text and adding references; this is notwithstanding that these comments were in contrast to the comments of Prof. Ferre and Mr Turnadge, who clearly valued the importance and timeliness of our paper.

We look forward to receiving the second round of comments from the reviewers and moving towards publication.

2 Reply to Short Comment by Ty P. A. Ferre

This is another excellent paper from this group. To me, it sits right between academic and applied hydrology. The group has such advanced modeling skills, that they have immediate credibility when they point out limitations based on model-based interpretations. In this paper, they carry their analysis through to the 'value' of data for decision support. They make a great case that data, even if it inherently carries important information, can be misleading if it is viewed through the lens of an imperfect model. Of course, they are all imperfect models. This is really important work and I feel that HESS is just the right target audience. I hope that it inspires continued careful examination of the interaction of data and models for decision support.

Well done Knowling et al!

Ty Ferre

We thank Prof. Ty Ferre for his positive and encouraging comments! We acknowledge his significant contribution to the field of decision-support modeling (the field in which this paper is cast). His comment that our study strikes a balance between academic and applied hydrology is pleasing as it reflects our endeavor to tackle problems that are relevant to both researchers and practitioners.

His comments also reiterate that our findings regarding the potential for assimilation of information-rich tracer data to cause ill-effects can be extended beyond just the (imperfect model) tracer data assimilation context (as covered in the Discussion and Conclusions section of the original mansucript).

Thanks again to Prof. Ferre!

3 Reply to Referee Comment by Ty P. A. Ferre

I provided an informal review previously, this serves as a somewhat more detailed formal review:

The authors continue to tackle one of the most important areas of applied hydrogeology with novel and insightful tools. Here, they challenge an accepted fact of hydrogeology that more data, and especially more diverse data will lead to better models for decision support. Their counterintuitive finding that isotopic data may have value little of no value for water resources is just the sort of result that could spark important conversations in our field.

As with their previous work, this group takes full advantage of their position as leading hydrogeologic modelers to offer constructive criticism for the field. There is no question that this group can build and calibrate large, complex models and that they are as able as anyone to extract information from hydrologic observations. This lends weight to what could otherwise be criticized as a finding based on lack of response. In this case, the group makes the point that increased complexity has a place in assimilating more and more varied data while recognizing that this increased complexity has limits for some applications. Again, the group takes advantage of its abilities to provide useful guidance for the community.

We thank Prof. Ty Ferre for his more detailed review and for his positive sentiments. Below we respond to each of his comments.

Prof. Ferre accurately summarizes our intention to challenge by rigorous investigation the widely accepted perspective among hydrologists and modellers that "more data and more diverse data lead to better model forecasts". Our experience extracting information from hydrologic observations through large, complex models—and the challenges associated with doing so—ultimately led us to undertake this study. We hope that our findings spark important conversations, and that these conversations ultimately lead to improved decision-support modeling practice.

Prof. Ferre echoes our position that increased model complexity may be required to appropriately assimilate both information-rich data, and data of various types, notwithstanding the additional challenges this added model complexity brings. Striking an appropriate balance between complexity and simplicity in the imperfect model-data assimilation context is an area that requires more attention in our opinion.

Personally, I appreciate the terse format of presenting the case studies. There may be call for providing more detail as supplemental information. I believe that the accepted White paper supplies these details. But, I will leave that to other reviewers and the editor to comment on whether it is appropriate to provide more detail in this manuscript.

We are pleased that Prof. Ferre finds the brevity of the case study details to be appropriate. We believe that this comment has been addressed by: (*i*) the availability of additional details regarding the second case study—the Hauraki Plains paired model analyses used for identifying tracer assimilation-induced bias—following publication of White et al. (ress), and (*ii*) the additional case study details added in response to the comments of the other reviewers (see below).

My only recommendation with regard to the authors approach reflects my own bias. As such, I completely understand if they do not address it in the paper! Regardless, I would like to hear the authors respond to the following question. From the perspective of a water manager or someone else tasked with assessing hydrologic risk, is a statistical reduction in the forecast the right measure of data value? Would the value of tritium, in this case, be viewed differently if decision-making were seen to be based on hypothesis testing of the plausibility of future high loading, for example? More generally, could the authors comment on the importance of considering the decision making context underlying the assessment of data worth?

Fantastic work I look forward to reading more in the series!

Ty Ferre

We thank Prof. Ferre for his interesting question on the contextualization and measurement of data worth (the assessment performed in the first case study). While there is probably no single "right" measure of data worth from the perspective of water resource managers across the board, we acknowledge that the assessment of data worth based on changes in the second moment (i.e., variance) of a forecast of management interest is unlikely to express the "full picture" from a decision maker's perspective. This is because decision makers need to evaluate forecast PDFs with respect to carefully defined decision thresholds governing management action. For example, if forecast variance is reduced through data acquisition, but this reduction in variance is of no consequence in terms of our ability to test hypotheses (e.g., if the entire PDF lies on one side of the decision threshold), were the collected data *really* worth it, in terms of the specific decision-support context in question? From the managers perspective, surely not, as the decision was not made any

easier following the acquisition of data.

A more decision maker-focused measure of data worth may therefore be to evaluate forecast PDFs (that are conditioned on different observations) with respect to a specific management decision threshold (or multiple decision thresholds as is more likely the case in practice; e.g., Vilhelmsen and Ferre (2018)). We feel that such a hypothesis testing approach to data worth assessment has significant merit and requires further investigation, notwithstanding that some excellent related works such as Nowak et al. (2012) and Wagner (1999) already exist in the literature. We also feel that identifying optimal monitoring data with respect to "decision difficulty" (e.g., Knowling et al. (2019)) has potential. For the data worth analysis presented in the first case study, it is difficult to anticipate how the specific outcomes of the analysis would be affected through adopting a hypothesis testing approach to data worth. This is due to the fundamental role that a specified decision threshold plays in the assessment of management action success/failure.

We also note that robust assessment of data worth in a hypothesis testing framework should ultimately involve non-linear uncertainty quantification such that not only the second moments but also the first moments of forecast PDFs can be assessed with respect to a decision threshold. Such an approach would also allow for the consequences of data assimilation-induced forecast bias in proximity to a decision threshold to be quantified (e.g., where the value of data in terms of variance reduction may be outweighed by its introduction of bias).

More generally, the decision-support context(s) in which the assessment of data worth can be framed is fundamentally important to the current study and more generally. Adopting a decision-support context allows for the prioritization of data collection towards improving the reliability of model forecasts that are used to support management decision-making. Exploring data worth in other contexts can also nevertheless be performed using similar approaches such as FOSM. For example, in the context of aquifer characterization, one can compute the value of, e.g., geophysics data in terms of the reduction in uncertainty associated with key aquifer properties, which may form the basis for improved system process understanding.

We welcome Prof. Ferre to contact us to discuss further some of the ideas above if he is interested.

4 Reply to Short Comment by Chris Turnadge

I would like to thank the authors for their invitation to provide comments on this manuscript. I believe this manuscript provides a valuable and timely contribution towards guidance in the use of subsurface environmental tracers to improve the predictive capability of groundwater flow and transport models. Many publications have implored researchers and practitioners to include a range of non-standard observation types such as temporal differences (Peeters et al., 2011), temperatures (Anderson, 2005), isotope concentrations and activities and inferred residence time (or "ages") (Schilling et al., 2019), and/or geophysical data (Hinnell et al., 2010) in model inversion and prediction uncertainty minimisation. However, investigations of when benefits may be obtained (or perhaps more importantly, when not) from these additional data types (and therefore specific guidance in their use) has been limited.

We thank Mr Chris Turnadge for his detailed and constructive comments. We agree that the guidance provided in the manuscript is timely for the reasons Mr Turnadge describes (and for reasons described in the Introduction and Discussion sections of the manuscript). It was one of our primary motivations to rigorously investigate the calls for increased use of diverse data (e.g., Schilling et al. (2019)) in the context of decision-support modeling. We thank Mr Turnadge for his positive sentiments.

On first reading of the manuscript I questioned whether the authors were "overreaching" in their conclusions. Having re-read the manuscript, I now believe that the authors have been careful to state that their conclusions regarding the applicability of environmental tracer observations are indicative, rather than comprehensive. More generally though, I do believe the manuscript will benefit from some revisions. Specifically, I would like to provide three major criticisms of the manuscript. These mainly relate to the suitability of the experimental design, in terms of the suitability of testing the hypotheses presented. These are followed by a number of minor criticisms that I believe the authors should also consider addressing. These minor criticisms mostly relate to the interpretation of environmental tracers, or to descriptions provided of the parameterisations of the numerical models used. I hope my comments are helpful to the authors in improving the manuscript and I am more than willing to provide further clarifications off-line, if they

are needed.

We agree that our conclusions are "indicative", and we are pleased Mr Turnadge believes our carefully formulated conclusions are appropriate. We feel that indicative conclusions are really all that can be drawn on the basis of two (or even more) real-world case study example demonstrations. We also consider empirical demonstrations to provide an important adjunct to theoretical demonstrations; we feel empiricism in the presence of inevitable site specifics are important to accompany theoretical investigation.

We believe that Mr Turnadge's comments can be addressed where appropriate through some minor yet important additions and modifications to the manuscript text. His comments also provide us with an opportunity to reiterate and expand on some of our current decision-support modeling perspectives and philosophies. We respond to each of his comments below.

Major comments:

1. Tracer observations are of limited additional value when direct observations of fluxes are already available

For the Heretaunga Plains example, I do not believe that the addition of environmental tracer observations (in this case, tritium) when flux observations (in this case, spring discharge) are already available provides an ideal (i.e. fair) test case. Tracer concentrations are proxies for fluxes (either recharge, lateral or discharge) so they are often measured (and subsequently included in model inversion) when direct observations of fluxes are not available. Assessing the value of tracer observations when flux observations are not available would provide a fairer test case and would be of greater interest. For the Heretaunga Plains example, this could be implemented by simply omitting spring discharge observations from all model inversion.

We consider the first case study to represent a fair and useful test case for the following reasons. First, our results do in fact already show the worth of MRT observations in the absence of spring discharge observations (albeit in the absence of other observations too—therefore representing the case where the maximum worth of MRT observations is apparent with an otherwise "empty" observation dataset). Please see the blue bars in the MRT column of Figure 2.

Second, the Heretaunga Plains case study presented reflects a real-world investigation, i.e., whereby tritium concentrations were measured after and in combination with discharge measurements. It is not the experience of the authors that tracer concentrations are typically only sampled where flux observations are lacking. This would imply that flux and tracer data contain

the same information and can therefore be substituted for one another. Contrast this with much literature suggesting the benefits of as many data types and as much data as possible (as acknowledged by Mr Turnadge above).

We also note that in the Discussion and Conclusions section, we provided a detailed description of circumstances by which the apparent worth of MRT observations in the first case study would likely have been higher (see Lines 246-254 of original manuscript).

2. Tracer observations are of limited additional value when collected upstream of prediction locations

For the Hauraki Plains example, subsurface tritium observations were recorded far upstream of nitrate discharge predictions. Since environmental tracers such as tritium integrate information along flow paths, it is intuitive that these tritium observations would contain limited information of value to predictions of nitrate concentration located far downstream. If subsurface tritium concentrations are available from locations in the vicinity of the predictions of interest, then I suggest that it would be more relevant to include these in the case study.

We regret that this comment appears to reflect a misunderstanding (and therefore a need to improve the communication of the manuscript). While we agree with Mr Turnadge's intuition regarding the integration of tracer-derived information along flow paths, we do not show that tritium concentration observations in the second case study are of limited value or "worth" (regardless of where the observations occurred within the basin). In fact, despite that the second case study does not actually explore the worth or information content of tritium observations, the results suggest that the tritium observations contain "too much" information (or, rather, misinformation when considered through the lens of an imperfect model that lacks parameter receptacles for the information contained in the tritium observations). That is, the sensitivity of the forecast of interest to uncertain parameters that were conditioned by tritium concentration observations led to forecast bias. This occurs due to two factors: (*i*) the "Firth" nitrate-load forecast aggregates flow paths across the entire domain (i.e., this forecast represents the only nitrate flux sink of the system) and in time; and (*ii*) the tritium observations provide insight into spatially and temporally averaged recharge and lateral flux rates in the upgradient portion of the domain, where most of the surfacewater/groundwater exchange occurs. In other words, the bias reflects the information content of the upgradient tritium observations related to averaged upgradient model parameters on which the forecast is sensitive.

To address this comment, we have added the text "This forecast aggregates flow paths across the entire model domain (i.e., it represents the only nitrate-flux sink of the system)." to the Second case study section of the revised manuscript.

We have also added the text "The biases identified reflect the sensitivity of the Firth forecast to uncertain parameters that were conditioned by tritium concentration observations. This occurs due to the spatially integrated nature of the Firth nitrate-load forecast, and because the tritium observations provide insight into spatially and temporally averaged recharge and lateral flux rates in the upgradient portion of the domain, where most of the surface-water/groundwater exchange occurs." to the Results section of the second case study.

3. Conservativeness and management thresholds when presenting prediction uncertainty

For the Hauraki Plains example, I believe that the authors criticise simple modelling approaches without providing any discussion of whether modelled predictions are conservative. Prediction histograms are presented in the absence of a management threshold; in this case, an upper permissible limit for nitrate discharge. I suggest that the authors present a relevant management threshold when presenting prediction uncertainty results. This would allow the authors the additional benefit of exploring whether predictions produced using simple and complex parameterisation approaches were conservative.

We agree that exploration of the conservativeness or otherwise of simplified model forecast PDFs (the former of which can be viewed as a metric for accepting such a simplified model) is an important undertaking, especially when performed with respect to a specific management decision threshold. We are pleased to inform Mr Turnadge that we have two papers that explicitly tackle this question—the first in terms of model parameterization (Knowling et al., 2019) and the second in terms of model vertical discretization (White et al., ress).

We therefore address this comment by adding an explicit reference to these manuscripts and how they cover a broader scope than that of the current manuscript in the Discussion and Conclusions section: "We refer the reader to Knowling et al. (2019) and White et al. (ress) for a broader exploration of the consequences of model simplification (in the form of parameterization reduction and vertical-discretization coarsening respectively) in terms of the decision-relevant forecast bias-variance trade-off and its implications for management decision making more generally.".

To further address this comment, we have added the text "(we refer the reader to White et al. (ress) for an exploration of the appropriateness of reduced-discretization models in decision support more generally)" to the Second case study section.

Minor comments

1. Mean residence times

Mean residence time (MRTs) values were used to quantify the age of groundwaters. However, MRTs are equivalent to groundwater ages only under very strict assumptions; specifically, requiring highly simplified conceptualisations. The latter are the basis of mixing functions used in lumped parameter models. In most (complex, real-world) cases, MRTs act as a fitting parameter. This is especially true if MRTs are derived from binary mixing models, which arbitrarily blend two mixing models, which generally undermines their physical bases. As a solution, and rather than the use of lumped parameter models to derived mean residence times, where possible I suggest simulating the reactive transport (or at least, non-reactive transport with first order decay) of environmental tracer concentrations (or activities. Admittedly, for some tracers (such as carbon-14) the complexity of reactions may make reactive transport simulation prohibitive. However, the examples presented by the authors already feature combined numerical flow and transport models. Additionally, the authors examples also feature a relatively simple tracer requiring only the simulation of first order decay, so I would assume that reactive transport simulation would be feasible.

We agree with Mr Turnadge regarding the simplicity and lack of physical basis of LPMs, and the potential benefits of full advective-dispersive (and reactive) transport numerical models for simulating tracer concentrations (as described comprehensively by Turnadge and Smerdon (2014)), notwithstanding their practical limitations, which are amplified in formal decision-support modeling contexts.

Importantly, however, we reiterate that LPM-derived MRT observations are only used in the first case study (in combination with advective-transport simulations); the second study employs full advective-dispersive transport modelling together with a first-order reaction rate to simulate radioactive decay of tritium (see also comment below). Our intention here was to employ both of these "standard practice" tracer modeling techniques as a basis for exploring the ramifications of model tracer-data assimilation, such that the findings can be as useful as possible to industry. The literature reflects the common use of both advective-only particle-tracking simulations (combined

with LPM-based "age" observations) and advective-dispersive simulations (combined with tracer concentrations) (e.g., Gusyev et al. (2014)).

We have addressed this comment by more explicitly stating in the Introduction section (where the two case studies are introduced) the different modeling approaches undertaken for the two case studies.

2. Binary mixing models

The authors state that, as part of lumped parameter modelling, binary mixing models (BMMs) were used to derive mean residence times from subsurface tritium concentrations. BMMs are a linear combination of two other (ideally) physically-based mixing models. I suggest that the authors describe which mixing models were combined using the BMMs, the relative contribution of each, and the physical meaning of the combined result.

The MRT observations that are considered in the first case study only, were derived using a combination of exponential piston flow models (EPMs) and BMMs (comprising two "parallel" EPMs), as described in detail in Morgenstern et al. (2018).

We have addressed this comment by adding the following text to the First case study section of the manuscript: "Specifically, a combination of exponential piston-flow models (EPMs) and binary-mixing models (BMMs) (that comprise two EPMs) were used. BMMs were employed for wells where long time-series data are available for multiple tracers, and where an adequate fit to different tracer signals could not be obtained on the basis of a single EPM. Relative EPM mixing fractions were specified on the basis of aquifer confinement conditions and well-screen length (mixing fractions of 80-95% were applied for wells with a long screen in unconfined conditions, whereas mixing fractions of 50-60% were applied for wells with shorter screens in confined conditions). The reader is referred to Morgenstern et al. (2018) for more details. ".

3. Atmospheric tritium concentrations

The authors state that the historic record of atmospheric tritium concentration features a "shape [that allows] for unique interpretation" (lines 69 - 70). It is unclear what the authors mean by the term "unique" in this context. Unlike SF6, it would not be true to state that the historical atmospheric tritium record is consistently monotonic. Tritium values increased initially due to nuclear weapons testing and have declined ever since. In some cases, the historical atmospheric tritium record features more than one peak value after the cessation of nuclear weapons testing. In addition, historic records of atmospheric tritium concentrations at various locations are typically quite

noisy, unlike SF6. See Figure 1 in McCallum et al. (2014) for an example of a noisy, two-peak example of a historical atmospheric tritium record. For this example, and following correction for radiometric decay in the subsurface, a tritium observation of 20 TU would correspond to any of four different recharge times (and therefore ages). In comparison, a historical atmospheric SF6 record is also shown by McCallum et al. (2014), which increases monotonically and would therefore permit unique interpretations of ages from measured concentrations. I suggest that the authors could state instead that tritium is a popular tracer for the identification of young age groundwaters (i.e. <70 years old) for the following reasons. Unlike CFCs, it is not affected by microbial degradation or contamination and, unlike SF6, it is not affected by potential subsurface sources. The authors may wish to cite Beyer et al. (2014), who provided a comparison of traditional (e.g. 3H, CFCs, SF6) and emerging (e.g. Halon-1301, SF5CF3) young age tracers.

We agree with the need to be more specific regarding the reasons why tritium is a often-favoured tracer and indeed why its consideration herein is relevant. We thank Mr Turnadge for his suggested revision to the text. We have revised the manuscript directly following his suggestion.

4. Non-reactive modelling of contaminant transport

The Hauraki Plains model simulated nitrate as a non-reactive constituent. In practice, nitrates in the subsurface are subject to a range of processes: assimilation, nitrification/denitrification, volatilisation, sorption/desorption and retardation (Kendall and Aravena, 2000). For this reason, I suggest that the authors explain why nitrate was not simulated as a reactive constituent in the forward model.

The second case study does in fact use a first-order decay rate to simulate the process of denitrification reactions. We consider this approach to be "standard" modeling practice.

We have addressed this comment by adding the text "Denitrification and radioactive tritium decay processes are simulated using first-order reaction rates." to the Second case study section of the revised manuscript.

5. Screen lengths of wells sampled for environmental tracers

When interpreting subsurface concentrations of environmental tracers, knowledge of the length of screened sections in sampled groundwater wells is crucial. If lumped parameter models are used, this affects the choice of mixing model. For example, if sampled wells are open holes or fully screened then the exponential mixing, exponential-piston flow, or dispersion models may be appropriate. Alternatively, if sampled wells are partially screened then

the partial exponential model may be appropriate. Given the importance of this information to tracer interpretation, I suggest that the authors describe the screen extents of each sampled well. The authors could also state how this information was used to select an appropriate mixing model for lumped parameter modelling, from which mean residence times were calculated.

We agree with Mr Turnadge regarding the importance of the well screen length for tracer interpretation—especially when using LPMs for interpretation (as is the case for the first case study).

We have addressed this comment by indicating the role that well screen lengths had on the deployment of LPMs to infer MRT in the First case study section, as per our response to minor comment (2). A full description of well details such as depth, screen interval and the corresponding LPM used for MRT interpretation can be found in Morgenstern et al. (2018) (publicly available online, as referenced in the original manuscript). We therefore prefer not to repeat these details in the current manuscript.

6. Pilot point parameterisation

Pilot point parameterisations of the Heretaunga Plains and Hauraki Plains models are not described explicitly in the manuscript. Specifically, it is not clear which parameters were parameterised using this method. I suggest that the authors describe explicitly which model parameters were implemented on a cell-by-cell basis, or using pilot points, zonation or using spatially uniform values, including horizontal and vertical hydraulic conductivity, specific yield/storage, recharge and, for transport models, porosity.

We have addressed this comment by the following changes.

First, we have added the following text to the First case study section: "Spatially-distributed parameterization of hydraulic conductivity (horizontal and horizontal/vertical anisotropy ratio), effective porosity, specific storage and specific yield is achieved using pilot points (e.g., Doherty, 2003). Spatially-distributed river-bed and boundary conductance parameters are defined on a reach and zone basis, respectively. We refer the reader to the Supplementary Information for more information".

Second, we have added the following text to the Second case study section: "Spatially-distributed parameterization of (horizontal and vertical) hydraulic conductivity, effective porosity, recharge rate, first-order denitrification rate, initial concentration and dispersivity is achieved using a combination of cell-based and zone-based multipliers. Nitrate-loading rate and abstraction well rate is parameterized using cell-by-cell and well-based mulitipliers, respectively. Streamflow-routing (SFR) elements are parameterized on a stream-

segment basis.".

7. Variogram definitions

It is not clear whether, for a given model, the same variogram was used to implement pilot point parameterisation for one or many parameter types. For example, I would not expect that the spatial correlation between hydraulic conductivity values to be the same as for recharge rates. I suggest that the authors state explicitly which variogram parameter values (e.g. correlation length, range, sill, nugget) were used to define which model parameter values, and describe the spatial analyses used to quantify spatial correlation between parameter values. Given that the degree of model complexity (particularly in relation to the ability of a model to assimilate observed data) is a key focus of the manuscript, I believe that detailed descriptions of the model parameterisation used are relevant.

Briefly, for the Heretaunga Plains case study, the same variogram (variogram parameters already defined in manuscript—see Lines 136-139) is used for pilot-point based distributed parameters; no spatial correlation is assumed otherwise. We have addressed this comment by adding "pilot-point based" to Line 137 of the original manuscript. This addition, in combination with the details on parameterization devices added in response to minor comment (6), addresses this comment for the first case study.

For the Hauraki Plains case study, variogram details regarding the various different parameter types have already been described in both White (2018) and White et al. (ress). We therefore address this comment by adding a reference to the additions made in response to minor comment (6): "We refer the reader to White (2018) and White et al. (ress) for more information on model parameterization and construction of prior parameter covariance matrices".

On a more general note, we agree that variograms could theoretically be defined on a parameter type-specific basis from a physically-based (or perhaps more appropriately a "physically-motivated") parameter standpoint. However, given our recent experience and findings regarding the significant potential for ill-effects (e.g., forecast bias) in real-world decision-support modeling (e.g., Knowling et al. (2019); White et al. (ress)), we tend to consider spatially distributed parameters employed by regional-scale models to be significant "abstractions" (i.e., from their intended property representation; e.g., Watson et al. (2013)). It follows that questions such as "what is the variogram for spatially distributed recharge bias-correction parameters?" and "how do we represent uncertainty in the variogram model used to describe prior param-

eter correlation and heterogeneity (i.e., a "hyper-parameter")?" arise when trying to rigorously deal with real-world model error.

8. Bias and underestimation

The authors state that the assimilation of tritium can induce "biased first moments or underestimated second moments" (line 55). I suggest that the authors could unpack this statement by providing simple examples to support this statement, for both bias and underestimation. The authors could also state explicitly the nature of the bias; i.e. whether prediction mean values were under- or overestimated.

On Line 55 (of the original manuscript), we are making only a general statement that data assimilation through history matching an imperfect model can result in forecast bias (either mean over- or under-estimation) and/or uncertainty underestimation; these ill effects occur as a result of both model simplification and history matching. We cite literature that demonstrate these phenomena. This statement does not relate to tracers or tritium specifically, or any other data in particular. Therefore, no comments can be made at this point as to the nature of bias or variance underestimation. The "direction" of tritium assimilation-induced bias (i.e., under- or over-estimation) is covered in the Results section of the second case study (although we note that the direction of bias may not be very generalizable between different forecasts and between different sites).

9. Ensemble size representativeness

The authors state that their implementation of the Iterative Ensemble Smoother featured ensemble sizes of 100 (lines 194-195). This value appears to have been selected arbitrarily, likely based on logistical constraints (e.g. forward model and inversion computing times). Was bootstrapping or other representativeness/convergence testing methods used to assess whether an ensemble size of 100 representative, and/or whether ensemble statistics converged as the ensemble size approached 100? I suggest that the authors demonstrate that the ensemble size used was representative.

The ensemble size was in fact selected on the basis of an approximation of the solution-space dimensionality. This approximation was obtained through a subspace analysis of predictive error variance (Moore and Doherty, 2005). We refer Mr Turnadge to the Supplementary Information of Knowling et al. (2019) for more information on this, including a plot of the singular-value spectrum.

To address this comment, we have added the above-mentioned singular-value spectrum plot to the Supplementary Information, and a supporting

sentence to the Second case study section.

10. Vertical coarsening of model grid

The authors provide limited explanation of why vertical coarsening of the model grid led to fewer relatively long flow paths. Since this observation is crucial to the interpretation of the authors results, I suggest that the authors expand their discussion of this key point.

Fewer relatively long flow paths occur when vertically coarsening the model grid simply due to the aggregation of numerical discretization effects—the flow paths of a coarser-layer model will be a smoother and averaged representation of those derived from a finer-layer model. As described above, the ramifications of model simplification in terms of reduced vertical discretization in the uncertainty quantification and data assimilation context more generally is covered by the separate manuscript White et al. (ress). Nevertheless, we agree that this is an important explanation to support the current findings.

We have therefore added the following explanation to the Results section of the second case study in the revised manuscript: "Briefly, this occurs due to the aggregation of numerical discretization effects—the flow paths of a coarser-layer model will be a smoother and averaged representation of those derived from a finer-layer model.".

Many thanks again to Mr Turnadge for his helpful comments.

5 Reply to Anonymous Referee #2

In the present manuscript, Knowling et al. aim to demonstrate that environmental tracer observations in general are not as informative for groundwater model data assimilation as previously thought because, in their eyes, flow models are typically too wrong for adequate physical representation of tracer behavior. The authors base their conclusions on only two case studies involving groundwater model calibration against only one environmental tracer (i.e. tritium, in one case study using tritium-derived groundwater residence times and in a second case study using tritium concentrations directly). The authors specifically identify errors in groundwater model vertical discretization as a reason for why data assimilation of groundwater model with tritium concentrations is prone to result in biased model predictions.

While a systematic study on this topic is potentially interesting and useful, the present study lacks the necessary rigor in experimental design and

standard in scientific reporting to be able to demonstrate what the study aims to demonstrate and to be a valid contribution to HESS. Shortcomings include: Failure to properly describe (1) the model calibration procedures, (2) the observation data, and (3) the models and assumptions used to derive residence times from tritium concentrations. The authors also fail in properly referencing scientific literature which already demonstrated aspects of the present study. Moreover, misleading statements are made about existing studies, and the general conclusions that were drawn on the value of environmental tracer observations for groundwater model calibration in general are not justified from the results of the simple experimental setup and use of tritium alone. Due to the lack in reporting, it isnt even possible to fully understand, assess or reproduce the findings. Below I elaborate on some of the shortcomings of the study which I see as reasons for rejecting of this paper.

We thank the anonymous reviewer for their comments. We believe that their comments can be addressed where appropriate through some minor, yet important additions and clarifications to the manuscript text. The comments also provide us with an opportunity to revisit and expand on some of our findings and recommendations.

First we wish to clarify the following points related to the reviewer's summary of our work:

1. Not only do we assess the apparent or "theoretical" information content of environmental tracer observations for decision-support groundwater model data assimilation, as judged by rigorous data worth exploration, we also assess the potential for the assimilation of these data to cause unwanted effects such as forecast bias, by considering model error and using paired complex/simple models. To our knowledge, no other work has examined the assimilation of environmental tracers in the context of the bias-variance trade-off relevant to the use of imperfect models. We feel that this central aspect of our paper has been over-looked by the reviewer. Framing our findings and recommendations in the context of real-world decision-support modeling is fundamentally important to the purpose of our paper. The importance of this aspect was acknowledged by the other reviewers, e.g., "They make a great case that data, even if it inherently carries important information, can be misleading if it is viewed through the lens of an imperfect model" and "This manuscript provides a valuable and timely contribution towards guidance in the use of subsurface environmental tracers to improve the predictive capability of groundwater flow and transport models. Many publications have implored researchers and practitioners to include a range

of non-standard observation types ... However, investigations of when benefits may be obtained (or perhaps more importantly, when not) from these additional data types (and therefore specific guidance in their use) has been limited".

2. We wish to follow-up on the reviewers comment that "in (our) eyes, flow models are typically too wrong for adequate physical representation of tracer behavior". Recent studies have showed that even seemingly minor model defects can cause significant ill-effects such as bias and uncertainty under-estimation. However, even more importantly, the outcome of these ill-effects depends on the purpose of the modeling analysis. The challenge is therefore to try to avoid these ill-effects in the context of the given modeling analysis and its purpose. We are advocating for careful and forecast-specific model design, to ensure that the rich information contained within tracer data can be properly assimilated. Potential use of more abstract means to assimilate these data into simpler models, is a promising model design option as it alleviates the need for increased model complexity and the costs associated with it. We suggest that this is an area of future work, as discussed in the Discussion and Conclusions section of the manuscript. These recommendations were explicitly valued by the other reviewers, e.g., "In this case, the group makes the point that increased complexity has a place in assimilating more and more varied data while recognizing that this increased complexity has limits for some applications."

The reviewer's summary suggests that our conclusions are not warranted on the basis of two case studies and the consideration of one environmental tracer. We wish to point out that an exhaustive exploration of how and when environmental tracers can be most usefully assimilated into models represents a research field in itself. The purpose of our paper is to raise and illustrate the following two points: (*i*) the assimilation of environmental tracer observations may not always be worthwhile, depending on the forecast being made, and (*ii*) careful model design is central to the ability to assimilate environmental tracer observations into models.

We now address each of the reviewer's comments in detail below.

- *The manuscript lacks key information on model calibration:*

  The present manuscript doesn't sufficiently explain the observation data, models which were used to derive the different observation types or calibration procedures. In the first case study, the value of observations of tritium-derived groundwater residence times are compared to

the value of groundwater levels and spring discharge observations for the reduction of the predictive uncertainty of spring discharge predictions. However, information about the calibration procedure is not provided, i.e., it isnt clear whether an ensemble-based data assimilation procedure (i.e., the iterative ensemble smoother as mentioned in the abstract), or whether a classic history matching calibration procedure (i.e., based on a weighted, multivariate maximum likelihood estimation procedure as described in a referenced modelling report) is used. Even though in the abstract it is stated that iterative ensemble smoother was used in the present study, the method isnt explained in the methods section of model study 1.

One can either assume that it was the same as for model study 2, i.e. Iterative Ensemble Smoother. This is suggested by the wording of the abstract and the term data assimilation via history matching (line 117). An Iterative Ensemble Smoother approach, and ensemble-based data assimilation procedures in general, would however make the direct application of linear predictive uncertainty analysis based on FOSM impossible because to the jacobi matrix isn't calculated by these approaches. Or, one could assume that data assimilation was not conducted but instead classic history matching after reading a referenced modelling report (however, Rakowski and Knowling, 2018, is not referenced in the respective model calibration and uncertainty quantification methodology section (2.4)). Using classic history matching would be a contrast to what was stated in the abstract and make the data worth assessment difficult to compare to the findings of modelling case study 2. The authors should also explain in detail what they mean by how the jacobi matrix was populated. For the second modelling case study, in section 3 after the description of methods and results of model study 1, it is explained that an Iterative Ensemble Smoother with 100 realisations was used. While for model study 1 it was stated that 882 parameters were calibrated, for model study 2 one does not learn how many parameters were calibrated. While for model study 1 there is a referenced modelling report available, the report referenced for model study 2 was not accepted or published at the time of the article submission and therefore not available for checking (on lines 184-185 it is stated: 'The model, and the vertical-discretization simplification analysis, is described in detail in White et al. (forthcoming)' and the said study is listed in the

bibliography as 'accepted, subject to minor revisions'). Key information in the calibration procedure is essential when the purpose of the study is to demonstrate the value of different observation types, as the calibration procedure strongly influence the data worth results.

The reviewer states that "information about the calibration procedure is not provided" for the first case study. This comment reflects a misunderstanding, and therefore a need to improve the presentation of this portion of manuscript. We employ FOSM techniques to quantitatively assess the worth of tritium-derived MRT along with other hydrologic observations by comparing forecast uncertainty changes following the notional data assimilation of different observations (e.g., see Lines 71-72, 75-78 and 128-130). FOSM analyses do not rely on formal history matching or on the pre-existence of a "calibrated model". To be clear, no actual parameter estimation is undertaken as part of the first case study. FOSM techniques have been widely employed for data worth assessment in this notional context in many settings as it enables rapid exploration of the worth of many different combinations of conditional forecast variances in a computationally efficient manner (e.g., Wallis et al. (2014); Zell et al. (2018)).

We have addressed this comment by revising Line 117 (i.e., add "notionally", remove "via history matching") and by replacing the "History matching" sub-section heading with "Observations for assimilation".

We feel that these changes will address the reviewer's confusion regarding whether the iterative ensemble smoother was used for the first case study. The ensemble smoother was used only for the second case study, where we explore the potential for model simplification-induced forecast bias and how this may be exacerbated when assimilating tracer concentration observations.

To further address this comment, we have made the distinction between the two different data assimilation approaches undertaken for the two case studies more explicit throughout the revised manuscript.

The paired complex/simple model analysis undertaken for the second case study involves formal history matching and non-linear uncertainty quantification (via the iterative ensemble smoother) for various models with varying vertical discretizations. Each of these analyses are performed twice, once with and once without tritium concentration data

for assimilation (i.e., Figure 4 is the result of six ensemble smoother experiments). This first-of-its-kind analysis for environmental tracer assimilation allows us to identify biases arising directly from the assimilation of these information-rich observations with imperfect, real-world groundwater models in a decision-support setting. This approach was necessary to explore otherwise invisible biases induced through assimilating tritium data. As discussed above, we feel this critical and novel part of the paper has been over-looked by the reviewer.

While the number of uncertain model parameters used in the second case study (for each of the 7-, 2- and 1-layer models) is provided in the now-published article White et al. (ress), we agree with the reviewer that this constitutes an important detail that, when absent, may obscure some details regarding the second case study. We have therefore added the following text to Section 3.2 (as well as other data assimilation details; see responses below): "This parameterization approach gives rise to a problem dimensionality of 141268, 50180 and 29050 for the 7-layer, 2-layer and 1-layer model history-matching experiments, respectively."

- *The manuscript lacks key information about the used observation data:*

  Observation data which were used for the modelling study are not provided, even though this is critical information to understand and reproduce the reported findings. While for model study 1 at least the different observation types which were used are mentioned, for model study 2 it is completely unclear what observations were used alongside tritium. It isn't clear how many observations of tritium, what uncertainty these observations are associated with, and the study which probably contains such information was not accepted at the time of submission and is not available.

  The reviewer is correct that the second case study does not contain information regarding observations used for history matching aside from tritium concentration observations. While the other hydrologic observations used for history matching are described in detail in White (2018) and White et al. (ress), we note that this omission was purposeful in the original manuscript: the second case study does not compare the relative value of different data including tritium observations (as is the case in the first case study). Instead, the second case study investi-

gates an additional and equally-relevant aspect of environmental tracer assimilation concerning differences in the posterior forecast distribution in terms of first-moment (i.e., bias) and second-moment (i.e., variance) characteristics. These differences arise directly from the assimilation of tritium concentration observations using models that are progressively less equipped to assimilate this information.

Nevertheless, to address this comment for completeness, we have added details on the observation data used for history matching in the second case study (other than tritium concentration observations), including plots of observation locations in the Supplementary Information: "Other observations such as long-term averaged groundwater levels and surface-water flows, and transient surface-water and groundwater nitrate conentrations were also used for history matching (see the Supplementary Information for observation locations)".

While more information regarding the tritium concentration observation data used for history matching in the second case study are also presented in White (2018) and White et al. (ress), we agree with the reviewer that these details are warranted here. We have therefore added the following: "The history-matching experiments included 20 tritium concentration observations from the groundwater system (Figure **??**) (see also Supplementary Information for observation locations per model layer)" to the Second case study section.

Key questions that should be addressed before data worth can be objectively assessed are: What data were used alongside tritium? Is tritium an informative tracer for each of the two given systems, i.e., is the groundwater residence time in both catchments sensitive to tritium? How was tritium analyzed and which equations were used to postprocess tritium concentrations into groundwater residence times? How were flux measurements obtained? What is the uncertainty of spring discharge observations? Are the uncertainties comparable to tritium-based residence time uncertainties? What are the weights that were used during calibration and do they reflect the uncertainty of the different observations? None of this is described in the manuscript. This information is needed for the readers to assess whether the results of the present study are correct and meaningful.

Our response to each question above are as follows:

- As described above, we have added details to the Supplementary Information regarding the other observations used alongside tritium for assimilation in the second case study.

- Tritium is indeed an "informative tracer" for the hydrologic settings in both case studies. We agree it is important to state this more explicitly. We therefore address this comment by adding ".... in hydrological environments where young groundwater components are decision relevant" to "we focus specifically on the ramifications of assimilating the information contained within tritium concentration observations and tritium-derived mean residence time (MRT) observations for decision support concerning low flow and nutrient transport at the regional scale" in the Introduction section. We have also added the following text to the Introduction (in response to Mr Turnadge's comments): "Tritium is a popular tracer for the identification of relatively young age groundwaters (i.e., <70 years old), for the following reasons: (*i*) unlike CFCs, tritium is not affected by microbial degradation or contamination; and (*ii*) unlike SF6, it is not affected by potential subsurface sources (e.g., Morgenstern and Daughney, 2012; Cartwright and Morgenstern, 2012; Beyer et al., 2014)".

- We agree with the reviewer that details regarding how tritium measurements were interpreted (for the first case study) are warranted. This was also raised by Mr Chris Turnadge in his review comments. We have added details on the interpretation of MRT from tritium measurements to the description of the first case study (i.e., where MRT observations were used)—see responses to Mr Turnadge above.

- We have added the following text to the revised manuscript regarding the field techniques used to measure fluxes: "..., obtained using a range of techniques including flow gauging, electrical conductivity and temperature surveys, water isotopic analyses, etc. (Wilding, 2017)".

- The spring discharge measurement uncertainty is assumed to be proportional to the absolute flux magnitude (i.e., a heteroscedastic error model; e.g., Sorooshian (1981)). The assumed measurement uncertainty for MRT observations are also assumed to be proportional to the MRT magnitude, and are generally lower than those of spring discharge observations (primarily reflecting the difference in magnitudes). However, as described on Lines 144-146 and 300-301, in order to ap-

proximately account for the role of model error in reducing a model's ability to fit observations (i.e., we should not be fitting observations to a level commensurate with measurement noise given the presence of model error), we use the model-to-measurement residuals as a basis to adjust the uncertainty surrounding observations (see, e.g., Doherty (2015)). We address this comment by adding an explicit reference to Appendix A (where this is discussed in more detail) on Line 143: "(see Appendix A)".

- Observation weights assigned to different observations used for assimilation indeed reflect uncertainty—specifically epistemic uncertainty (accounting for both measurement and structural sources of uncertainty)—as described above. This is stated on Line 143.

- *The relevance of the authors findings is over-stated:*

  It is unclear why it is concluded that tritium is representative of environmental tracers in general. The manuscript lacks an important number of references which have already published similar results on the value of spring discharge or tritium or which have shown, in much more systematic and rigorous experimental approaches, that environmental tracers are highly valuable for groundwater model calibration. While the title is very broad, i.e.: 'On the assimilation of environmental tracer observations for model-based decision support', the present study does not generally assess the value of environmental tracer data in a data assimilation context. It appears as if only for one of the two modelling studies formal data assimilation has been conducted (however, as outlined in the previous comment, it is not entirely clear what calibration approach was used in the first modelling case study). Furthermore, only one single environmental tracer is used: tritium. Tritium is certainly not reflective of all environmental tracers and for many groundwater systems, tritium is not a useful tracer because groundwater residence times are of an order on which tritium isnt sensitive. The wording of abstract, introduction, discussion and conclusions strongly suggests that the authors believe that their two case studies of tritium are representative of the wider worth of environmental tracer data for groundwater model calibration (e.g., Lines 268-271) : "We consider this recommendation to be in stark contrast to the common belief that calibrating to more data improves the model and its predictions. We therefore also consider this

recommendation to be of significant implication to decision-support environmental modeling practitioners. It is expected that this finding can be extended to the general approach of assimilating diverse observation types in environmental modeling."

All comments are addressed in detail in the following responses. First we respond in more general terms.

The purpose of this paper is to show, through two real-world decision-support models, that the assimilation of environmental tracers is not a panacea to all the ills of environmental simulation, and that care in the model design is essential to ensure that the information contained within these tracers is not squandered. While we have endeavored to make this purpose more clear in the text to avoid misinterpretation, we note that the comment that "the relevance of our findings is overstated" is in direct contrast to the technically detailed comments by Mr Turnadge, who explicitly stated the carefulness with which our conclusions were drawn: "Having re-read the manuscript, I now believe that the authors have been careful to state that their conclusions regarding the applicability of environmental tracer observations are indicative, rather than comprehensive" (see our response above). We have taken care to ensure that all general statements are accompanied by appropriate caveats. Prof. Ferre's comments also reflect the more general implications of our findings (i.e., beyond tritium data assimilation): "They make a great case that data, even if it inherently carries important information, can be misleading if it is viewed through the lens of an imperfect model".

While we agree that a "systematic study on this topic is potentially interesting and useful" (the reviewer's words), a comprehensive analysis into the value or otherwise of assimilating environmental tracers, and the models required to do so, can never be systematic because it will always be context specific. This context reflects a combination of important factors, such as the decision-support quantities of interest, the hydrologic setting, the complexity (or otherwise) of the model, the other available observations, among others.

Tritium is not representative of environmental tracers in general, as it requires more complex mathematical simulation procedures to do its complex decay and production pathways justice. Showing that a sim-

ple one-layer model cannot properly represent tritium transport and therefore calibrating it against tritium results in biased predictions is not generating insights representative for environmental tracer value in general. Numerous previous studies have much more systematically analysed and identified the large benefits of environmental tracers for groundwater model calibration in general, but the large majority are not referenced in the present manuscript. Here are a few examples:

Carniato et al. (2015), Highly parameterized inversion of groundwater reactive transport for a complex field site. DOI: 10.1016/j.jconhyd.2014.12.001

Delsmann et al. (2016), Global sampling to assess the value of diverse observations in conditioning a real-world groundwater flow and transport model. DOI: 10.1002/2014WR016476

Hunt et al. (2006), The importance of diverse data types to calibrate a watershed model of the Trout Lake Basin, Northern Wisconsin, USA. DOI: 10.1016/j.jhydrol.2005.08.005 (cited in the present manuscript)

Rasa et al. (2013), Effect of different transport observations on inverse modeling results: case study of a long-term groundwater tracer test monitored at high resolution. DOI: 10.1007/s10040-013-1026-8 (cited in the present manuscript)

Xu and Gomez-Hernandez (2016): Characterization of non-Gaussian conductivities and porosities with hydraulic heads, solute concentrations, and water temperatures. DOI: 10.1002/2016WR019011

Oehlmann et al. (2015), Reducing the ambiguity of karst aquifer models by pattern matching of flow and transport on catchment scale. DOI: 10.5194/hess-19-893-2015

Masbruch et al. (2014), Hydrology and numerical simulation of groundwater movement and heat transport in Snake Valley and surrounding areas, Juab, Millard, and Beaver Counties, Utah, and White Pine and Lincoln Counties, Nevada. DOI: 10.3133/sir20145103

We do not "conclude" that tritium is representative of tracers in general. Instead, we consider tritium as a representative environmental tracer in the context of the potential outcomes of its assimilation into imperfect models in general. Our consideration of tritium simply reflects that it is one of the most widely used environmental tracers in

younger groundwater systems (as is the case for our two case studies—see comments above).

The purpose of the paper is not to assess the representativeness of tritium relative to other tracers. Instead, it is to demonstrate that the usefulness or otherwise of information-rich data, including environmental tracers, is related to the forecasts being made, and that such data may induce (undetectable) forecast bias, when assimilated into imperfect models. We believe that these issues are relevant and transferrable to the assimilation of environmental tracers in general. This reflects that we consider the primary barrier to appropriate assimilation of tracer data to be the difficulties associated with extracting information from spatially-discrete concentration observations when using upscaled or simplified representations of hydraulic properties within a model that simulates tracer concentrations using the advection-dispersion equation. To the extent that the simulated output corresponding to observed tracer concentration(s) are sensitive to model details or parameters that are "missing" in a simplified model (e.g., White et al. (2014)), inappropriate parameter compensation will occur. Then, to the extent that the forecast of management interest is dependent on these biased parameter estimates, the forecast will be biased, potentially leading to resource mismanagement. Therefore, we believe that these factors, which all real-world decision-support model analyses share, will also cause issues for tracer data assimilation and assimilation of diverse data in imperfect models more generally. Despite this, we believe that the larger challenge for the transferability of our work is the specificity of different decision-support contexts, and the infinite spectrum of model design, as we discuss in the Discussion and Conclusions section.

To address comments related to how our findings can be applied to other tracers, for example, we have added a paragraph to the Discussion and Conclusions section of the revised manuscript based on the response above.

What is demonstrated in the first modeling case study, i.e., the complicated nature of using residence/travel time observations derived from tritium for groundwater model calibration, is very well known and was already subject of multiple much more systematic and thorough comparisons and reviews, some of which are even referenced in the present manuscript (e.g., Turnadge and Smerdon 2014 (DOI: 10.1016/j.jhydrol.2014.10.056),

McCallum et al. 2014 (DOI: 10.1111/gwat.12052) and 2015 (DOI: doi:10.1111/gwat.12237), Schilling et al. 2019 (DOI: 10.1029/2018RG000619), Sanford 2011 (DOI: 10.1007/s10040-010-0637-6)). All these studies concluded already that it is better to calibrate a flow model against environmental tracer concentrations, or yet even better, direct flux observations, rather than against residence times due to the fact that the simulation of residence times is often faulty due to structural inaccuracies in the numerical groundwater model.

We agree with the reviewer that it has been reported in the literature that it is a preferred approach to simulate tracer concentrations (involving solution of the advective-dispersive equation) and history match to tracer concentrations directly, rather than simulate residence times (involving advective-only particle-tracking schemes) and history match to derived quantities such as MRT. We state this on Line 40-43 (original manuscript). However, in real-world modeling, the compounding challenges associated with simulation of tracer concentrations, e.g., computational demands, can complicate the assimilation of tracers. These challenges are why the latter approach is still popular in the industry (e.g., Turnadge and Smerdon (2014); Gusyev et al. (2014)), as stated in the manuscript.

It is interesting that the reviewer states that model "structural inaccuracies" generally explain why residence times cannot be reliably simulated. We contend that a significant degree of "structural inaccuracy" will persist, even when dispersive and decay processes are simulated explicitly (i.e., rather than simplified advective-only simulations), as is demonstrated explicitly in the second case study. The difficulty in simulating spatially and temporally distributed tracer concentrations (or more generally, simulating advective-dispersive transport) has been discussed by many (e.g., Zheng and Gorelick (2003); Riva et al. (2008))—see also our response to the previous comment. Our second case study demonstrates how discretization-related model error combined with assimilation of discrete-point concentration observations can induce considerable biases and ultimately corrupt resource management.

Specifically, the fact that spring discharge observations contain the largest amount of information for spring discharge predictions is neither surprising nor new. Exchange fluxes in general, be it groundwater discharging as spring water or into a surface water body, or surface water

infiltrating into the subsurface, have been demonstrated to not only be more valuable data for groundwater model calibration than travel/residence times observations, but also to be much less prone to bias due to straightforward implementation into flow model calibration compared to the more complex physical underpinnings required for groundwater residence times simulations. The authors even reference one study which has demonstrated this systematically in comparison to groundwater residence time observations: Hunt et al. (2006, DOI: 10.1016/j.jhydrol.2005.08.005, already cited in the manuscript) compared the worth of several different flux observations to the worth of hydraulic heads, environmental tracer concentrations and travel time information, and found that groundwater exfiltration onto the surface (providing baseflow of a stream) was the most information rich data type overall, and that many other flux observation types were also more informative than travel time observations.

The authors failed to reference studies which have already demonstrated the high importance of spring discharge more specifically: Masbruch et al. (2014, DOI: 10.3133/sir20145103, not cited in the manuscript) systematically compared the information content of spring discharge to observations of groundwater levels, temperature and environmental tracers, and found that spring discharge observations were the most informative overall data type. A similarly high importance of spring discharge observations was identified by La Vigna et al. (2006, DOI: 10.1007/s10040-016-1393-z, not cited in the manuscript), who systematically elaborated the worth of spring discharge observations for the calibration of groundwater flow models in comparison to hydraulic head observations. Oehlmann et al. (2015, DOI: 10.5194/hess-19-893-2015, not cited in the manuscript) systematically analysed the calibration of karst groundwater models against observations of spring discharge, groundwater residence times and groundwater levels. They identified that spring discharge observations provide indispensable information for karst groundwater model calibration, but also showed the large information content of residence time observations. The use of all three observation types together was the most beneficial approach for groundwater model parameterisation.

The authors' literature review is unbalanced, misses many key references, and makes incorrect statements about findings of key studies.

We agree that the specific finding from the first case study that the spring discharge observations are of most "worth" when considering spring discharge forecasts is not surprising. We state this explicitly on Lines 243-245: "The worth of MRT observations relative to various hydraulic potential and discharge observations across the different forecasts are, in general terms, similar to those reported by Zell et al. (2018)".

To address this comment, we will add the Hunt et al. (2006), Masbruch et al. (2014) and Oehlmann et al. (2015) reference to further support this sentence.

We note that the reviewer states that only "aspects" of our study have been demonstrated in other studies. We agree. However, as discussed in detail above, the aspects that the reviewer is referring to do not encapsulate the thrust of our paper. We agree with the reviewer that numerous studies have explored the value in environmental tracer data for history matching groundwater models, and we accept that we have not cited every paper on this in our literature review. However, we do not feel that it is necessary to do so, as the literature we cite in the Introduction collectively expresses the state of the science: that studies have "identified the large benefits of environmental tracers for groundwater model calibration in general" (to use the reviewer's words).

In contrast, our study provides an important demonstration of how "variable" the worth of these data may actually be, e.g., in the presence of other data and when making water quantity-related forecasts. We provide a series of detailed explanations for this—we refer the reviewer to Lines 235-254 of the Discussion and Conclusions sections. We feel that such "worked examples" are important for the community to see—this perspective was strongly supported by the other reviewers, e.g., "This is really important work and I feel that HESS is just the right target audience", "this is just the sort of result that could spark important conversations in our field" and "the group takes advantage of its abilities to provide useful guidance for the community" (Prof. Ferre), and "this manuscript provides a valuable and timely contribution towards guidance in the use of subsurface environmental tracers" and "investigations of when benefits may be obtained (or perhaps more importantly, when not) from these additional data types (and therefore specific guidance in their use) has been limited" (Mr Turnadge).

Furthermore, we note that the references suggested by the reviewer do not tackle the entangled issues of model error, data assimilation and predictive reliability—in contrast to our study. For example, following consideration of the references suggested by the reviewer, we consider questions such as: (*i*) how could these studies explore the potential for forecast bias given that history matching was undertaken using only a single model?; and (*ii*) how can the observation data responsible for inducing forecast bias through assimilation be identified? These questions illuminate how our study differs from previous studies. This also provides an insight into the importance of the context is which our paper is framed—and how this differs to groundwater modeling practice in general terms. We were very careful to make this point clear, e.g., "Modeling for the purpose of decision support is the context in which the remainder of this paper is framed" (Line 29), and "The benefit or otherwise of direct assimilation of tritium concentration data in other decision contexts, or for more general system understanding and conceptual model development, is therefore not the focus of the current study—this study is concerned with a models ability to "predict" (in two decision-support contexts) rather than "explain" (observed system behavior), as contrasted by Shmueli (2010)." (a revised version of Line 231).

References

Beyer, M., Morgenstern, U., and Jackson, B. (2014). Review of techniques for dating young groundwater (<100 years) in new zealand. *Journal of Hydrology (New Zealand)*, 53(2):93–111.

Cartwright, I. and Morgenstern, U. (2012). Constraining groundwater recharge and the rate of geochemical processes using tritium and major ion geochemistry: Ovens catchment, southeast australia. *Journal of Hydrology*, 475:137 – 149.

Doherty, J. (2015). *Calibration and uncertainty analysis for complex environmental models - PEST: complete theory and what it means for modelling the real world*. Watermark Numerical Computing.

Doherty, J. E. (2003). Ground water model calibration using pilot points and regularization. *Ground Water*, 41(2):170–177.

Gusyev, M., Abrams, D., Toews, M., Morgenstern, U., and Stewart, M. (2014). A comparison of particle-tracking and solute transport methods for simulation of tritium concentrations and groundwater transit times in river water. *Hydrology and Earth System Sciences*, 18(8):3109.

Hunt, R. J., Feinstein, D. T., Pint, C. D., and Anderson, M. P. (2006). The importance of diverse data types to calibrate a watershed model of the trout lake basin, northern wisconsin, usa. *Journal of Hydrology*, 321(1):286 – 296.

Knowling, M. J., White, J. T., and Moore, C. R. (2019). Role of model parameterization in risk-based decision support: An empirical exploration. *Advances in Water Resources*, 128:59 – 73.

Masbruch, M., Gardner, P., and Brooks, L. (2014). Hydrology and numerical simulation of groundwater movement and heat transport in snake valley and surrounding areas, juab, millard, and beaver counties, utah, and white pine and lincoln counties, nevada.

Moore, C. and Doherty, J. E. (2005). Role of the calibration process in reducing model predictive error. *Water Resources Research*, 41(5):1–14.

Morgenstern, U., Begg, J., van der Raaij, R., Moreau, M., Martindale, H., Daughney, C., Franzblau, R., Stewart, M., Knowling, M., Toews, M., Trompetter, V., Kaiser, J., and Gordon, D. (2018). Heretaunga plains aquifers : groundwater dynamics, source and hydrochemical processes as inferred from age, chemistry, and stable isotope tracer data.

Morgenstern, U. and Daughney, C. J. (2012). Groundwater age for identification of baseline groundwater quality and impacts of land-use intensification the national groundwater monitoring programme of new zealand. *Journal of Hydrology*, 456-457:79 – 93.

Nowak, W., Rubin, Y., and de Barros, F. P. J. (2012). A hypothesis-driven approach to optimize field campaigns. *Water Resources Research*, 48(6).

Oehlmann, S., Geyer, T., Licha, T., and Sauter, M. (2015). Reducing the ambiguity of karst aquifer models by pattern matching of flow and transport on catchment scale. *Hydrology and Earth System Sciences*, 19(2):893–912.

Riva, M., Guadagnini, A., Fernandez-Garcia, D., Sanchez-Vila, X., and Ptak, T. (2008). Relative importance of geostatistical and transport models in describing heavily tailed breakthrough curves at the lauswiesen site. *Journal of Contaminant Hydrology*, 101(1):1 – 13.

Schilling, O. S., Cook, P. G., and Brunner, P. (2019). Beyond classical observations in hydrogeology: The advantages of including exchange flux, temperature, tracer concentration, residence time, and soil moisture observations in groundwater model calibration. *Reviews of Geophysics*, 57(1):146–182.

Sorooshian, S. (1981). Parameter estimation of rainfall-runoff models with heteroscedastic streamflow errors the noninformative data case. *Journal of Hydrology*, 52(1):127 – 138.

Turnadge, C. and Smerdon, B. D. (2014). A review of methods for modelling environmental tracers in groundwater: advantages of tracer concentration simulation. *Journal of Hydrology*, 519:3674–3689.

Vilhelmsen, T. N. and Ferre, T. P. (2018). Extending data worth analyses to select multiple observations targeting multiple forecasts. *Groundwater*, 56(3):399–412.

Wagner, B. J. (1999). Evaluating data worth for ground-water management under uncertainty. *Journal of water resources planning and management*, 125(5):281–288.

Wallis, I., Moore, C., Post, V., Wolf, L., Martens, E., and Prommer, H. (2014). Using predictive uncertainty analysis to optimise tracer test design and data acquisition. *Journal of Hydrology*, 515:191–204.

Watson, T. A., Doherty, J. E., and Christensen, S. (2013). Parameter and predictive outcomes of model simplification. *Water Resources Research*.

White, J. T. (2018). A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions. *Environmental Modelling & Software*.

White, J. T., Doherty, J. E., and Hughes, J. D. (2014). Quantifying the predictive consequences of model error with linear subspace analysis. *Water Resources Research*, 50(2):1152–1173.

White, J. T., Knowling, M. J., and Moore, C. R. (in press). Consequences of model simplification in risk-based decision making: An analysis of groundwater-model vertical discretization. *Groundwater, doi: 10.1111/gwat.12957*.

Wilding, T. K. (2017). Heretaunga springs: Gains and losses of stream flow to groundwater on the heretaunga plains.

Zell, W. O., Culver, T. B., and Sanford, W. E. (2018). Prediction uncertainty and data worth assessment for groundwater transport times in an agricultural catchment. *Journal of Hydrology*, 561:1019 – 1036.

Zheng, C. and Gorelick, S. M. (2003). Analysis of solute transport in flow fields influenced by preferential flowpaths at the decimeter scale. *Groundwater*, 41(2):142–155.

On the assimilation of environmental tracer observations for model-based decision support

Matthew J. Knowling[1], Jeremy T. White[1], Catherine R. Moore[1], Pawel Rakowski[2], and Kevin Hayley[3]

[1]GNS Science, New Zealand
[2]Hawke's Bay Regional Council, New Zealand
[3]Groundwater Solutions Ltd, Australia

Correspondence: Matthew J. Knowling (m.knowling@gns.cri.nz)

Abstract. It has been advocated that history-matching numerical models to a diverse range of observation data types, particularly including environmental tracer concentrations and their interpretations/derivatives (e.g., mean age), constitutes an effective and appropriate means to improve model forecast reliability. This study presents two regional-scale modeling case studies that directly and rigorously assess the value of discrete tritium concentration observations and tritium-derived mean residence time (MRT) estimates in two decision-support contexts; "value" ~~herein is measured as~~ is measured herein as both the improvement (or otherwise) in the reliability of forecasts through uncertainty variance reduction and bias minimization as a result of assimilating tritium or tritium-derived MRT observations. The first case study (Heretaunga Plains, New Zealand) utilizes a suite of steady-state and transient flow models and an advection-only particle-tracking model to evaluate the worth of tritium-derived MRT estimates relative to hydraulic potential, spring discharge and river/aquifer exchange flux observations. The worth of MRT observations is quantified in terms of the change in the uncertainty surrounding ecologically-sensitive spring discharge forecasts via first-order second-moment (FOSM) analyses. The second case study (Hauraki Plains, New Zealand) employs paired simple/complex transient flow and transport models to evaluate the potential for assimilation-induced bias in simulated surface-water nitrate discharge to an ecologically-sensitive estuary system; formal data assimilation of tritium observations is undertaken using an iterative ensemble smoother. The results of these case studies indicate that, for the decision-relevant forecasts considered, tritium observations are of variable benefit and may induce damaging bias in forecasts; these biases are a result of an imperfect model's inability to properly and directly assimilate the rich information content of the tritium observations. The findings of this study challenge the unqualified advocacy of the increasing use of tracers, and diverse data types more generally, whenever environmental model data assimilation is undertaken with imperfect models. This study also highlights the need for improved imperfect-model data assimilation strategies. While these strategies will likely require increased model complexity (including advanced discretization, processes and parameterization) to allow for appropriate assimilation of rich and diverse data types that operate across a range of spatial and temporal scales commensurate with a forecast of management interest, it is critical that increased model complexity does not preclude the application of formal data assimilation and uncertainty quantification techniques due to model instability and excessive run times.

1 Introduction

Numerical models used to provide water resources management decision support are often subjected to data assimilation through history matching (or "calibration"). This is due to the large information deficit accompanying the development of these models, and the potential for the history matching process to lead to an increased reliability of simulated outputs of management interest (herein referred to as "forecasts") through variance reduction. Modeling for the purpose of decision-support is the context in which the remainder of the paper is framed.

It is widely advocated that the assimilation of multiple types of state observations (i.e., "diverse data") is of benefit in "constraining" models. In other words, as more data are used for history matching, and the more diverse those data are, the reliability of the forecasts increases. This is an intuitive stance arising from direct application of Bayes equation and from the recognized rich information content of diverse data types; this intuition is supported by many studies, ~~e. g., Sanford et al. (2004); Michael and Voss (200~~ (e.g., Sanford et al., 2004; Michael and Voss, 2009; Ginn et al., 2009; Li et al., 2009; Gusyev et al., 2013; Hansen et al., 2013). For example, Hunt et al. (2006) demonstrated the importance of unconventional observations including lake/aquifer exchange fluxes, depth of lake isotope plume and groundwater travel times in achieving "well-constrained parameter values" (e.g., acceptable posterior variance) through history matching a regional-scale groundwater model.

History-matching to environmental tracer observations, in particular, is widely a regarded mechanism to improve the reliability of forecasts. In a review of approaches for modeling environmental tracers in groundwater systems, Turnadge and Smerdon (2014) state that age data have been useful for constraining models; in particular, "simulation of environmental tracer transport that explicitly accounts for the accumulation and decay of tracer mass, has proven to be highly beneficial in constraining numerical models". Zell et al. (2018) showed the relative importance of water-level, stream discharge and environmental tracers (including tritium, CFCs, SF6) in the conditioning of groundwater travel time forecasts. They reported that, overall, tracer data were of considerable benefit in terms of forecast uncertainty reduction. In a recent review paper, Schilling et al. (2019) state that assimilation of concentration observations through surface water/groundwater flow model history matching "harbors huge potential", based on the findings of previous studies, while assimilation of tracer-derived residence time observations in these models also often help significantly (where an appropriate approach is adopted, ~~e. g., Sanford (2011); Zuber et al. (2011)~~ (e.g., Sanford, 2011; Zuber et al., 2011).

However, the notion that unabated assimilation of diverse data types (including environmental tracers) is always of benefit holds only from a theoretical standpoint. Direct evaluation of the likelihood term of Bayes theorem is predicated on a "perfect" simulator to appropriately condition uncertain model parameters through data assimilation. In real-world modeling contexts, however, the presence of model error can invalidate even the most rigorous data assimilation techniques ~~(e. g., Doherty and Welter (2010); White et al. (2014); Oliver and Alfonzo (2018)).~~ (e.g., Doherty and Welter, 2010; White et al., 2014; Oliver an. Therefore, when an imperfect simulator is used in a data assimilation framework, extreme care must be taken to assure that the model imperfections do not corrupt (through biased first moments, or under-estimated second moments) the forecast posterior distributions. A number of recent works have shown that the failure to appropriately frame the imperfect-model data assimilation problem can result in severely biased results ~~(e. g., Doherty and Christensen (2011); Knowling et al. (2019); White et al. (in press)~~

~~).~~ (e.g., Doherty and Christensen, 2011; Knowling et al., 2019; White et al., in press). The largely unknown ability of an imperfect regional-scale model to simultaneously assimilate diverse data types that operate over different spatial and temporal scales—and how these imperfections may affect model-based decision support in some contexts—serves as motivation for the current study.

A subtle, yet very important distinction should be made at this point. There is no doubting that diverse data types, and in particular environmental tracers, have contributed significantly to the understanding of catchment processes and properties ~~(e. g., Kirchner et al. (2001); André et al. (2005); Stewart and Thomas (2008); McDonnell et al. (2010); Morgenstern et al. (2010); Han et al. (20~~ ~~).~~ (e.g., Kirchner et al., 2001; André et al., 2005; Stewart and Thomas, 2008; McDonnell et al., 2010; Morgenstern et al., 2010; Han et al., . However, as discussed, this study focuses instead on the role of (imperfect) models in two selected decision-support contexts, and how the assimilation of environmental tracers in particular affects their utility in these contexts, i.e., by increasing (or otherwise) the reliability of forecasts.

Herein, we focus specifically on the ramifications of assimilating the information contained within tritium concentration observations and tritium-derived mean residence time (MRT) observations for model-based decision support concerning low flow and nutrient transport at the regional scale in hydrological environments where young groundwater components are decision relevant. Tritium is ~~an often favored environmental tracer due to its half-life and atmospheric signal shape allowing for unique interpretation (e~~a popular tracer for the identification of relatively young age groundwaters (i.e., <70 years old), for the following reasons: (*i*) unlike CFCs, tritium is not affected by microbial degradation or contamination; and (*ii*) unlike SF6, it is not affected by potential subsurface sources (e.g., Morgenstern and Daughney, 2012; Cartwright and Morgenstern, 2012; Beyer et al., 2014 . ~~g., Morgenstern and Daughney (2012); Cartwright and Morgenstern (2012)).~~

The ~~objectives~~ objective of this study ~~are~~ is two-fold. First, we ~~evaluate~~ investigate the theoretical worth of tritium-derived MRT observations ~~(as quantified by forecast variance reduction)~~ relative to other observation data types. This investigation is performed using a case study (Heretaunga Plains, New Zealand) that adopts first-order second-moment (FOSM) techniques; our analysis focuses on the relative worth of MRT observations in terms of changes in the uncertainty associated with spring discharge forecasts at various locations that are of management interest due to their ecological significance. This first case study employs advective-only particle-tracking modeling approach to simulate MRT.

Second, we ~~investigate~~ explore the use of discrete tritium concentration observations in data assimilation in the context of a controlled model simplification experiment as a means to understand what, if any, ill-effects may be induced by using these ~~rich~~ information-rich data types in a simplified (i.e., imperfect) model.

~~The remainder of the paper is structured as follows. We first present a case study that adopts first-order second-moment (FOSM) techniques to explore the theoretical worth of tritium-derived MRT observations compared to steady-state and transient hydraulic potential and flux data in terms of reducing the uncertainty associated with spring discharge forecasts at various locations that are of management interest due to their ecological significance. We then present a~~ This exploration is performed using a second case study that employs a recently-presented paired simple/complex model analysis ~~to rigorously investigate the consequences of assimilating tritium concentration data in terms of the potential for assimilation-induced~~ (White et al., in press) . The paired model analysis is used herein to allow for the identification of possible (and otherwise undetectable) bias and uncer-

tainty under-estimation surrounding ~~nutrient load forecasts. A discussion and some concluding remarks are then provided~~forecasts of nutrient load to an ecologically-sensitive estuary system. This second case study simulates (tritium and nitrate) tracer concentrations directly—using a full advective-dispersive modeling approach that also accounts for first-order reaction rates.

2 First case study

The first case study serves to investigate the ability of tritium-derived MRT observations to constrain ecologically-sensitive spring discharge forecasts (i.e., the "worth" of these observations) using a model of the groundwater system of the Heretaunga Plains (New Zealand) (Figure 1). The model was constructed primarily for the purposes of groundwater allocation management decision-support.

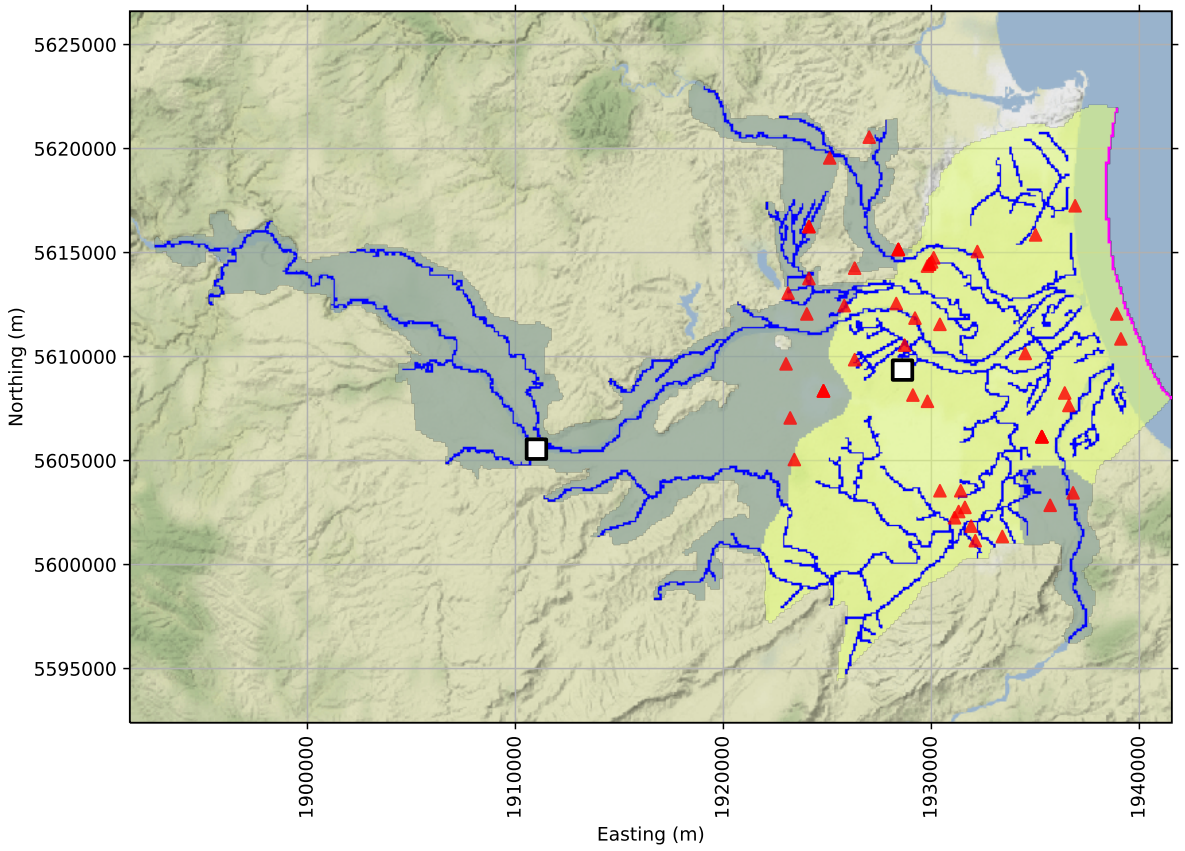

Figure 1. Heretaunga Plains model schematic, including river boundary conditions (blue lines), top-layer confinement status (unconfined areas shaded blue and confined areas shaded yellow), coastal general-head boundary (magenta line). The location of groundwater tritium-derived MRT observations are shown as red triangles. The location of forecasts—spring discharge rates during summer and winter—are shown as white markers.

2.1 The model

The model comprises 302 rows and 501 columns (uniform 100 m by 100 m horizontal grid discretization). Two layers are used for flow simulations, whereas six layers are used to generate more vertically detailed cell-by-cell flow budgets for particle tracking simulations. MODFLOW-2005 (Harbaugh, 2005) is used to simulate groundwater flow under steady-state and transient conditions. Separate simulations are conducted for data assimilation and forecasting purposes spanning different time periods (and temporal resolutions) of interest (e.g., separate transient flow simulations are conducted using annual stress periods for the period 1980—2015, and using monthly stress periods for the periods 1997—1999 and 2011—2015). MODPATH (Pollock, 2012) is used to simulate advection-only (i.e., neglecting diffusion, dispersion and retardation) reverse particle tracking, thereby providing a basis for assimilating tritium-derived MRT estimates (Figure 1). Specifically, the mean particle exit time corresponding to each observation location is compared with tritium-derived MRT estimates ~~(e. g., Sanford (2011); Gusyev et al. (2014)).~~ (e.g., Sanford, 2011; Gusyev et al., 2014).

Relevant aspects of the model are as follows:

- Land-surface recharge estimates, derived from a daily soil water balance modeling assessment (Rajanayaka and Fisk, 2018), are specified using the (specified flux) recharge package.
- The interaction between groundwater and surface water (including rivers, streams and springs) is simulated using the (head-dependent flux) river package. Time-varying river stage values are specified for the three main rivers in the region based on observed values. River-bed conductance values are varied seasonally to reflect in an approximate manner the non-linear relationship between field observations of spring discharge and groundwater levels.
- The coastal boundary condition is represented using the (head-dependent flux) general-head boundary package. The general-head stage is specified using a density-corrected mean sea-level ~~(e. g., Morgan et al. (2012)).~~ (e.g., Morgan et al., 2012) .
- Groundwater abstraction rates, based on observed and estimated data, are represented using the (specified flux) well package.

For a more detailed description of the Heretaunga Plains models, the reader is referred to Rakowski and Knowling (2018).

2.2 Forecasts

We focus on the following forecasts—due to their ecological significance and their potential to be impacted by groundwater abstraction:

- ~~spring~~ Spring discharge rate during summer at two locations (one in the central Heretaunga Plains, and one in the upper reaches of the catchment) (Figure 1)
- ~~spring~~ Spring discharge rate during winter at the central Heretaunga Plains location (Figure 1)

2.3 ~~History matching~~Observations for assimilation

Data assimilation ~~via history matching is undertaken~~ is undertaken notionally via FOSM techniques using the following observations:

- 6,167 groundwater levels (comprising time-averaged water-levels, absolute and deviation-from-mean annual, monthly and daily water-levels, long-term differences in water-level, and vertical head differences);
- 92 surface-water/groundwater fluxes (time-averaged and transient river gain and loss fluxes and spring discharge fluxes~~)~~ ~~(including time-averaged and transient~~, obtained using a range of techniques including flow gauging, electrical conductivity and temperature surveys, water isotopic analyses, etc. (Wilding, 2017)); and
- 52 groundwater MRT estimates derived from tritium concentrations using ~~binary mixing analytical models(Morgenstern et al., 2018)~~ lumped-parameter models. Specifically, a combination of exponential piston-flow models (EPMs) and binary-mixing models (BMMs) (that comprise two EPMs) were used. BMMs were employed for wells where long time-series data are available for multiple tracers, and where an adequate fit to different tracer signals could not be obtained on the basis of a single EPM. Relative EPM mixing fractions were specified on the basis of aquifer confinement conditions and well-screen length (mixing fractions of 80-95% were applied for wells with a long screen in unconfined conditions, whereas mixing fractions of 50-60% were applied for wells with shorter screens in confined conditions). The reader is referred to Morgenstern et al. (2018) for more details.

A highly parameterized approach was adopted ~~(e.g., Hunt et al. (2007); Knowling et al. (2019)),~~ (e.g., Hunt et al., 2007; Knowling et al., , involving a total of 822 uncertain parameters~~(see Supplementary Material)~~. Spatially-distributed parameterization of hydraulic conductivity (horizontal and horizontal/vertical anisotropy ratio), effective porosity, specific storage and specific yield is achieved using pilot points ~~(e. g., Doherty (2003)).~~ (e.g., Doherty, 2003). Spatially-distributed river-bed and boundary conductance parameters are defined on a reach and zone basis, respectively. We refer the reader to the Supplementary Information for more information.

2.4 Uncertainty quantification and data-worth exploration

Here we employ FOSM techniques ~~(e.g., Tarantola (2005); Doherty (2015))~~ (e.g., Tarantola, 2005; Doherty, 2015) to investigate the theoretical worth of various observation data types in terms of the their influence on the uncertainty variance surrounding forecasts following data assimilation. Application of FOSM in this context requires only consideration of the relative differences in estimated forecast variance as a result of conditioning on different observation data types. Use of FOSM in relative contexts has been shown to be especially robust ~~(e. g., Dausman et al. (2010); Herckenrath et al. (2011); Knowling et al. (2019)).~~ (e.g., Dausman et al., 2010; Herckenrath et al., 2011; Knowling et al., 2019).

The theoretical underpinnings of FOSM-based uncertainty quantification and data-worth assessment and details related to its application herein are presented in Appendix A.

Aspects that are relevant to the application of FOSM herein include:

- The prior parameter covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ was specified as a block-diagonal matrix whereby geostatistical correlation between pilot-point based spatially-distributed ~~parameter~~ parameters is represented through use of an exponential variogram with a range of approximately 10,000 m, and a sill proportional to the expected prior variance (the range of the square-root of the diagonal elements of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, i.e., the standard deviation of prior parameter uncertainty, is given in the Supplementary ~~Material~~Information). Non-spatially- and temporally-distributed parameters are assumed to be uncorrelated and therefore occupy diagonal matrix elements only.

- The Jacobian matrix $\mathbf{J}$ was populated using 1% two-point derivative increments.

- The diagonal elements of the epistemic noise covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ (see Appendix A) was specified on the basis of observation "weights", adjusted in such a way that the measurement objective function equals the number of non-zero weighted observations, in order to approximate epistemic noise (i.e., the combined impact of random measurement errors and model simplification errors) based on model residuals ~~(e. g., Doherty (2015)).~~ (e.g., Doherty, 2015).

2.5 Results

For the summer spring discharge forecast in the central Heretaunga Plains, MRT observations display a worth that is considerably less than that of spring discharge observations during the summer months (i.e., when lower flows persist) and transient head observations (Figure 2 A). This is not surprising given that the forecast and the summer spring discharge observations are of the same type and represent the same temporal condition, and transient head observations are plentiful (5,704), spanning different time periods at annual, monthly and daily resolutions. The worth of MRT observations is greater than winter spring discharge observations, indicating a higher relevance of the spatially and temporally integrated information contained within MRT observations for this low-flow related prediction compared to the higher frequency and magnitude signals captured within spring discharge observations during winter.

Similar results from a relative perspective are apparent for the summer spring discharge forecast in the upper portion of the Heretaunga Plains. That is, transient head observations and spring discharge observations during summer are of highest worth, followed by observations of time-averaged heads, MRT and winter spring discharge (Figure 2 B)—for reasons described above. The greater worth of MRT observations for this forecast compared to the summer spring discharge forecast located downgradient indicates that this forecast is more senstive to (uncertain) model parameters that are conditioned through assimilating MRT observations. This is due to the fact that the forecast is located where the aquifer is unconfined and receives rainfall and river recharge—these recharge rates are informed by MRT observations and have a large influence on the forecast.

For the winter spring discharge forecast, the worth of MRT observations is lower than that of other observations (Figure 2 C). This indicates a low relevance of the spatially and temporally integrated information contained in MRT observations with respect to a forecast concerning higher frequency and magnitude signals. This is also supported by the relatively low worth of the time-averaged head observations due to the temporally integrated nature of these quantities. As expected, a significantly greater worth of spring discharge observations during winter is evident for this forecast due to the unique and directly relevant information content associated with discharge observations that capture high-flow transience signals.

Across the three forecasts, a significantly larger worth is evident when MRT observations are added to the observation dataset compared to when MRT observations are removed from the observation dataset (red versus blue; Figure 2). This indicates that correlation occurs between the information contained within MRT observations and other observations. This is generally in contrast to the more unique information contained within spring discharge observations.

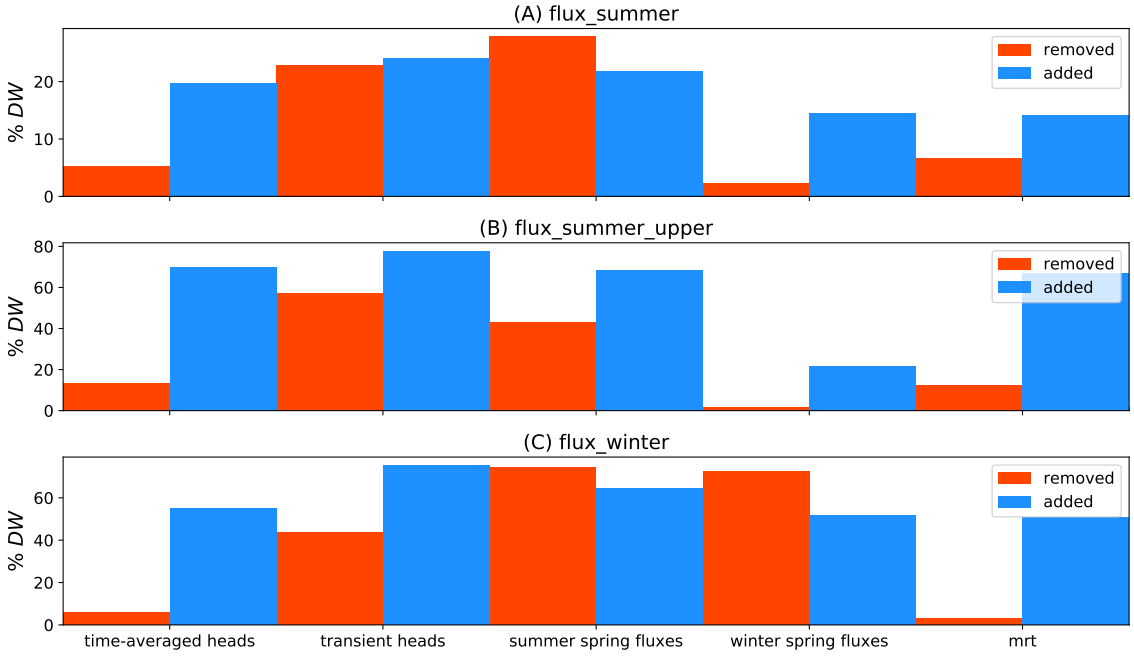

Figure 2. Worth of different observation groups (%DW) with respect to forecasts: (A) spring discharge flux during summer in central Heretaunga Plains; (B) spring discharge flux during summer in upper portion of Heretaunga Plains; and (C) spring discharge flux during winter in central Heretaunga Plains (see Figure 1 for locations). %DW is quantified as both the increase in forecast uncertainty variance following the removal of an observation group available for conditioning (red), and the decrease in forecast uncertainty variance following the addition of an observation group available for conditioning (blue) (see Appendix A). Note the different scales on the y-axes.

3 Second case study

The second case study serves to evaluate how assimilating discrete groundwater tritium concentration observations may affect the robustness of forecasts in the context of a controlled model simplification experiment, where the simplification is related to model vertical discretization ~~. Compared~~ (we refer the reader to White et al. (in press) for an exploration of the appropriateness of reduced-discretization models in decision support more generally). In contrast to the first case study, which focused on the theoretical worth of derived tritium observations in terms of changes in forecast variance, this case study proceeds with repeated data assimilation in a paired simple/complex model analysis both with and without assimilating tritium observations. Through these paired-model analyses, any potential biases or under-estimation of variances arising from the assimilation of

tritium observations with a simplified model ~~will~~ can be exposed. A linked hydrologic-nutrient transport model of the Hauraki Plains (New Zealand) (Figure 3) is used as a basis for the model simplification experiment.

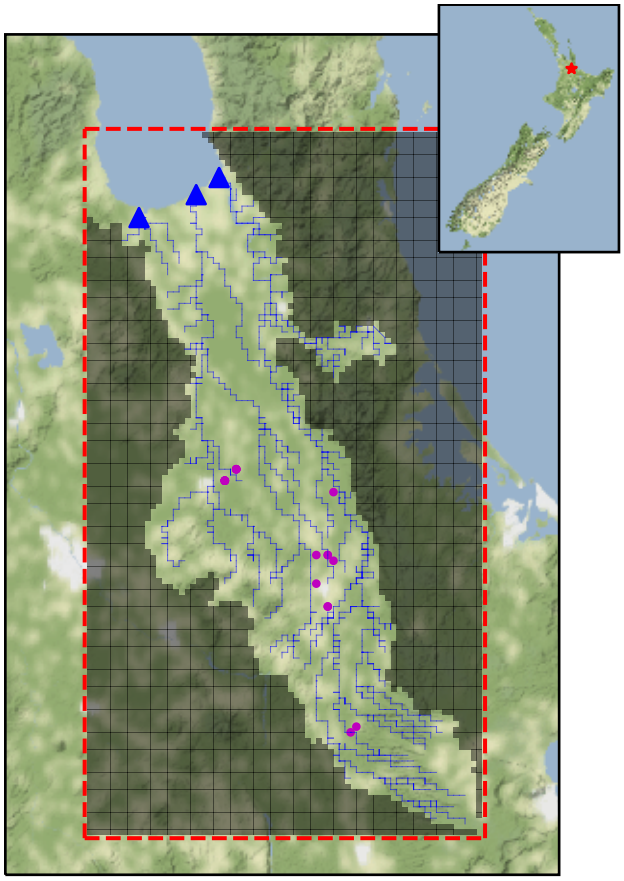

Figure 3. Hauraki Plains model extent (red dashed line), layer-1 inactive area (shaded), and surface-water network (blue lines). The terminal surface-water reaches that discharge to the Firth of Thames are shown as blue triangles. The location of groundwater tritium measurements are shown as magenta dots.

3.1 The model

The linked hydrologic-nutrient transport model simulates groundwater and surface-water flow using MODFLOW-NWT (Niswonger et al., 2011); advective and dispersive transport of nitrate and tritium in the groundwater and surface water system is simulated using MT3D-USGS (Bedekar et al., 2016). ~~The model~~Denitrification and radioactive tritium decay processes are simulated using first-order reaction rates. The model is described in detail in White (2018), and the vertical-discretization simplification analysis ~~,~~is described in detail in White et al. (in press).

Herein, we focus on a single forecast: the cumulative load of nitrate discharging from the surface-water system to the Firth of Thames—an ecologically-sensitive estuary system—over a 10-year projection scenario involving present-day (2018) flow and transport model forcing conditions. This forecast aggregates flow paths across the entire model domain (i.e., it represents the only nitrate-flux sink of the system). This forecast is referred to herein as the "Firth forecast".

3.2 ~~History-matching~~ Data assimilation and ~~uncertainty-quantification methodology~~uncertainty quantification

As described in White et al. (in press), data assimilation was undertaken via history matching three versions of the model, each with a different vertical discretization scheme~~, were history-matched~~; history matching was performed using the iterative ensemble smoother PESTPP-IES (White, 2018). ~~Each history matching experiment employed a dense, grid-scale parameterization approach, and involved parameter conditioning to several different types of flow and transport state observations. The data assimilation process~~

History matching was conducted using 100 stochastic parameter realizations~~; following~~. An ensemble size of 100 was deemed sufficient to avoid under-utilization of observation data (i.e., "under-fitting") based on an exploration of the solution-space dimensionality using a subspace analysis (Moore and Doherty, 2005) (see the Supplementary Information and Knowling et al. (2019) for more details). Following history matching, the 10-year projection scenario was evaluated with the 100 history-matched realizations (effectively a 100-member sample of the posterior distribution). From the resulting 100 scenario evaluations, a posterior probability density function (PDF) of the First forecast was constructed~~. Each of these~~.

The reader is referred to White (2018) and White et al. (in press) for a full description fo the Hauraki Plains model data assimilation process; a brief overview is nevertheless provided as follows:

- *Model parameterization.* Spatially-distributed parameterization of (horizontal and vertical) hydraulic conductivity, effective porosity, recharge rate, first-order denitrification rate, initial concentration and dispersivity is achieved using a combination of cell-based and zone-based multipliers. Nitrate-loading rate and abstraction well rate is parameterized using cell-by-cell and well-based mulitipliers, respectively. Streamflow-routing (SFR) elements are parameterized on a stream-segment basis. This parameterization approach gives rise to a problem dimensionality of 141268, 50180 and 29050 for the 7-layer, 2-layer and 1-layer model history-matching experiments~~included available discrete tritium~~, respectively. We refer the reader to White (2018) and White et al. (in press) for more information on parameterization and construction of prior parameter covariance matrices.

- *Observation data for assimilation.* The history-matching experiments included 20 tritium concentration observations from the groundwater system (Figure 3) ~~.~~ (see also Supplementary Information for observation locations per model layer). Other observations such as long-term averaged groundwater levels and surface-water flows, and transient surface-water and groundwater nitrate conentrations were also used for history matching (see the Supplementary Information for observation locations).

As shown in White et al. (in press), the reduced-discretization (1-layer and 2-layer) model posterior PDFs for the Firth forecast display significant bias compared to the corresponding 7-layer model posterior PDF (Figure 4 A,D,G). In White et al. (in

press), it was hypothesized that the tritium observations were giving rise to the apparent bias in the 1-layer and 2-layer posterior PDFs through the phenomenon of (inappropriate) parameter compensation ~~(e.g., Clark and Vrugt (2006); White et al. (2014))~~ (e.g., Clark and Vrugt, 2006; White et al., 2014) arising from history matching models with simplified model vertical discretization. Herein, we test this hypothesis by conditioning all three uniquely-discretized models again, but without using the discrete tritium observations, and then comparing the resulting posterior PDFs to the corresponding PDFs in White et al. (in press). Any apparent difference in the posterior PDFs for the Firth forecast is therefore directly attributable to the exclusion of the tritium observations during history matching.

3.3 Results

The process of history-matching with and without available groundwater tritium concentration observations yields substantial differences in the posterior PDFs of the Firth forecast (Figure 4). In the case of the 7-layer "complex model" (Figure 4 A,B), excluding the tritium observations results in a posterior PDF with a larger second moment and a slightly larger first moment compared to including tritium observations for history matching; the difference between the Firth forecast posterior PDFs with and without assimilating tritium observations is between 0 and ~~$2 \times 10^7 kg$~~ 2×10^7 kg of nitrate (Figure 4 C). The larger second moment of the posterior PDF when excluding tritium observations represents an intuitive and expected outcome: using fewer observations for parameter conditioning through history matching should (theoretically) result in a larger posterior variance for the forecasts that depend on those parameters.

Herein, for the purposes of identifying bias, the 7-layer model is considered to represent the best-available estimate of the Firth forecast. Using this construct, we see that there are significant differences in posterior PDFs across the uniquely-discretized models arising from data assimilation that included the tritium observations (Figure 4 A,D,G). This is largely in contrast to the case where data assimilation is undertaken without the tritium observations, which leads to much more subtle differences in posterior PDFs across the uniquely-discretized models (Figure 4 B,E,H).

The bias apparent in the posterior difference PDFs for the reduced-layer models relative to the 7-layer model (Figure 4 C,F,I) are directly attributable to the use of tritium observations in the data assimilation process. The difference between the Firth forecast PDFs resulting from data assimilation with and without tritium is most pronounced for the 1-layer model (Figure 4 I). In this case, excluding tritium observations from the history matching results in a decrease in simulated nitrate discharge of 2×10^7 to ~~$4 \times 10^7 kg$~~ ~~approximately~~ 4×10^7 kg—approximately a 40% decrease in simulated mean nitrate discharge. We attribute the apparent 1-layer PDF bias to the loss of simulated vertical flow and associated deeper groundwater flow paths. ~~While these~~ Briefly, this occurs due to the aggregation of numerical discretization effects—the flow paths of a coarser-layer model will be a smoother and averaged representation of those derived from a finer-layer model. While these deeper flow paths are not important for simulating the nitrate transport cycle (given the relatively high denitrification rates in the Hauraki system), it is apparently important for assimilating the tritium concentration observations.

The biases identified reflect the sensitivity of the Firth forecast to uncertain parameters that were conditioned by tritium concentration observations. This occurs due to the spatially integrated nature of the Firth nitrate-load forecast, and because

the tritium observations provide insight into spatially and temporally averaged recharge and lateral flux rates in the upgradient portion of the domain, where most of the surface-water/groundwater exchange occurs.

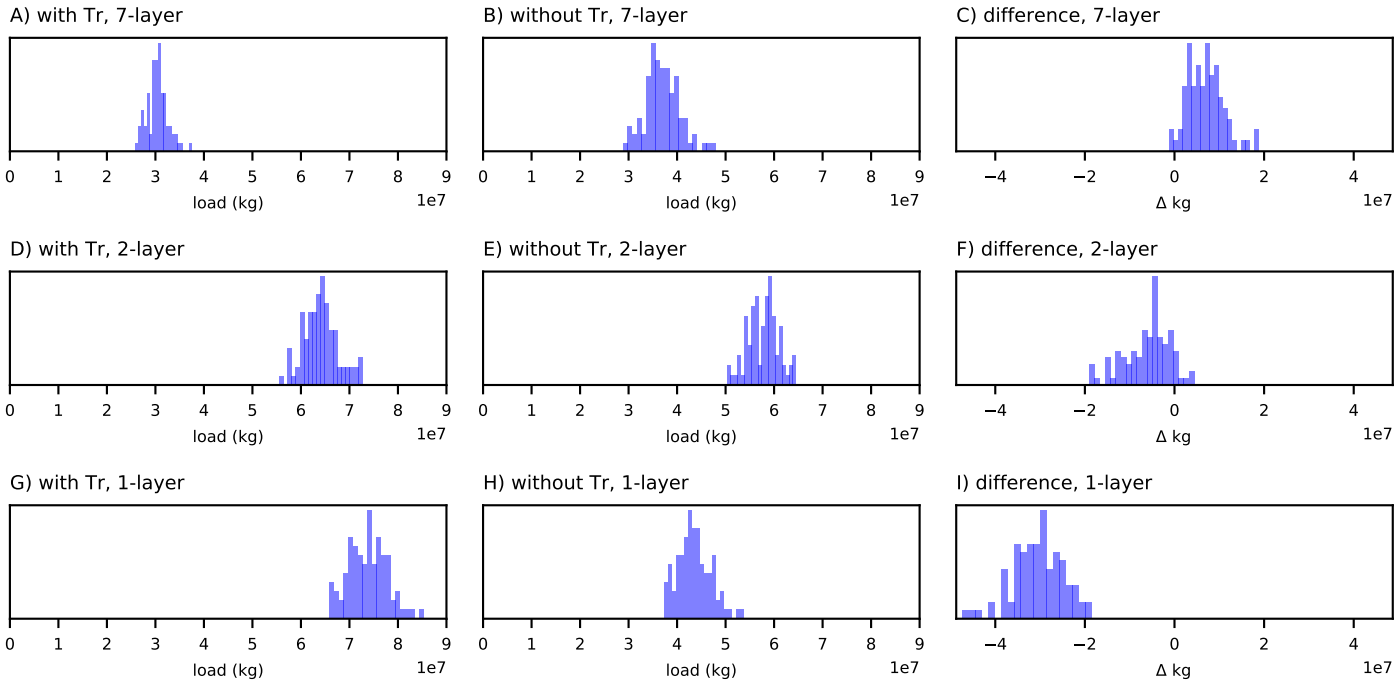

Figure 4. Comparison of posterior probability density functions (PDFs) for the Firth forecast. The left column (A,D,G) allows identification of bias as a result of both model simplification and tritium assimilation. By comparing to the middle column (B,E,H), model simplification-induced bias can be separated from that induced by assimilating tritium observations. The isolation of tritium assimilation-induced bias evident with different simplified models is shown in the right column (C,F,I). Including tritium observations in the conditioning of the 1-layer and 2-layer models (G,D) yields significant bias compared to the 7-layer PDF (A). However, if tritium observations are excluded from conditioning, the 1-layer and 2-layer PDFs (H,E) have considerably less bias compared to the corresponding 7-layer PDF (B). The differences in the PDFs (C,F,I) show that the tritium observations have the greatest biasing effect on the Firth forecast for the 1-layer model.

4 Discussion and Conclusions

285 This study explores the ramifications of assimilating tritium concentration and tritium-derived interpretation observations, specifically in the context of two examples of decision-support modeling. The benefit or otherwise of tritium data in other contexts such as site system characterization/understanding and conceptual model development is therefore not the focus of the current study—this study is concerned with a model's ability to "predict" (in two decision-support contexts) rather than "explain" (observed system behavior), as contrasted by Shmueli (2010).

The first case study presented herein serves to demonstrate that assimilating the rich information contained within tritium-derived MRT observations may be of variable worth in terms of improving the reliability of forecasts, especially where MRT observations are correlated with other available state observations (e.g., where hydraulic data are widespread, given the apparent spatially and temporally integrated information content of MRT observations, as supported by Ginn et al. (2009)). Moreover, the worth of MRT observations is shown to vary between forecasts in such a way that reflects the underlying physics represented by the model (e.g., the MRT observations are of greatest worth for forecasts that are located where the aquifer is receiving recharge)—these physics dictate the "information flow" rather than the spatial proximity of the MRT observations and the forecast. The forecast-specific nature of observation worth has also been reported previously ~~(e. g., Dausman et al. (2010); Fienen et al. (2010); White et al. (2016)).~~ (e.g., Dausman et al., 2010; Fienen et al., 2010; White et al., 2016). The worth of MRT observations relative to various hydraulic potential and discharge observations across the different forecasts are, in general terms, similar to those reported by Hunt et al. (2006), Masbruch et al. (2014), Oehlmann et al. (2015) and Zell et al. (2018) (especially when considering the discussion point in the following paragraph).

While the particle-tracking model used in the first case study provides a mechanism for MRT observations to inform uncertain model parameters, including aquifer porosity (which is otherwise uninformed by other historical field observations), it is important to note that the forecasts are insensitive to porosity. That is, the information contained within MRT observations is spread between parameters that both do and do not play a role in constraining forecasts—effectively "diluting" the information available for conditioning. It is therefore expected that the worth of MRT observations presented herein would generally be larger for forecasts that are dependent on both uncertain hydraulic and transport parameters (e.g., particle travel times)~~(~~. This is notwithstanding that the uncertainty variance for such forecasts may be larger given the additional source of uncertainty associated with porosity~~)~~. These findings are nevertheless highly relevant in that MRT observations are widely regarded to be of benefit in constraining uncertain model parameters more generally (Schilling et al., 2019)—regardless of the forecast.

The second case study serves to demonstrate that assimilating tritium concentration observations with simplified (i.e., imperfect) numerical models may induce significant bias in forecasts—bias that is undetectable without a complex/simple model pair ~~(e. g., Doherty and Christensen (2011); White et al. (2014); Knowling et al. (2019)).~~ (e.g., Doherty and Christensen, 2011; White et al., 20. The forecast bias revealed in the second case study occurs as a result of the vertical discretization-simplified model's inability to appropriately assimilate the rich information content of the tritium observations. Generally, the observed pattern of simplification and resulting forecast bias implies that as the simplification of the model increases, the dangers of assimilating rich and diverse data types also grows. This result is highly relevant to decision-support modeling practitioners since all numerical models are gross simplifications of real environmental systems that they attempt to simulate. We refer the reader to Knowling et al. (2019) and White et al. (in press) for a broader exploration of the consequences of model simplification (in the form of parameterization reduction and vertical-discretization coarsening respectively) in terms of the decision-relevant forecast bias-variance trade-off and its implications for management decision making more generally.

Collectively, these results suggest that the assimilation of tritium and tritium-derived observations through history matching with an imperfect model should be strategic and approached with caution. It is recommended that these information-rich observations should not indiscriminately be incorporated in a data assimilation framework, given that this study has shown

that such an approach (i) may ~~only~~ be of variable ~~benefit,~~ apparent benefit, depending on the forecast being made, and (ii) when using imperfect models, may produce far worse forecast outcomes than those that would have been arrived at without assimilating these observations at ~~all—an outcome similar to that postulated by Brynjarsdóttir and O'Hagan (2014)~~ all. This recommendation is similar to those by Brynjarsdóttir and O'Hagan (2014) and He et al. (2018). We consider this recommendation to be in stark contrast to the common belief that "calibrating to more data improves the model and its predictions". ~~We therefore also consider this recommendation to be~~ , and therefore of significant implication to decision-support environmental modeling practitioners. ~~It is expected that this finding can be extended to the general approach of assimilating diverse observation types in environmental modeling.~~

Furthermore, we expect the above-mentioned issues associated with imperfect-model data assimilation to be relevant and largely transferrable to the assimilation of other environmental tracers, other information-rich observations and diverse data types more generally. This is because we consider the primary barrier to appropriate assimilation of tritium observation data encountered in the second case study to be fundamental challenges associated with extracting appropriate information from spatially-discrete concentration observations when using upscaled or simplified representations of hydraulic properties within a regional-scale model that simulates tracer concentrations using the advection-dispersion equation (e.g., Zheng and Gorelick, 2003; Riva et a . To the extent that simulated outputs corresponding to observed tracer concentrations are sensitive to model details or parameters that are "missing" in a simplified model (e.g., White et al., 2014), parameter compensation will occur (e.g., Clark and Vrugt, 2006) . To the extent that the forecast of management interest is dependent on these biased parameter estimates, the forecast will also become biased, potentially leading to resource mismanagement. The ubiquitous nature of model error and the challenges in appropriately accounting for differences in, e.g., representative spatial scales between field observations and model-derived quantities, suggests that the ill-effects identified in this study such as history matching-induced bias are not unique to the specifics of our study (e.g., consideration of tritium as a tracer). The similar findings and recommendations of Brynjarsdóttir and O'Hagan (2 and He et al. (2018) in the statistics and petroleum reservoir disciplines, respectively, also supports the potential for the transferability in our findings and recommendations to data assimilation in other envrionmental modelling contexts.

If diverse and information-rich data such as tritium and MRT observations are available, and data assimilation through history matching is deemed necessary and/or appropriate, then a targeted modeling approach is needed that identifies which of these data are relevant to the forecast. This is critical to avoiding the ill-effects of model error in the context of decision support modeling ~~(e.g., White et al. (2014); Knowling et al. (2019)),~~ (e.g., White et al., 2014; Knowling et al., 2019), as well as to avoid adding unnecessary complexity (through processes and parameters) needed to simulate the equivalent values of the diverse data for assimilation purposes, which may greatly increase the computational cost of the modeling analysis.

It should be noted, however, that even when the forecast is well "aligned" with observation data (i.e., the forecast is solution-space dependent), some degree of parameter compensation will inevitably occur—all models are gross simplifications and therefore model parameters do not perfectly represent real-world properties ~~(e. g., Clark and Vrugt (2006); White et al. (2014)).~~ (e.g., Clark and Vrugt, 2006; White et al., 2014). However, if the data used for assimilation are commensurate with the forecasts, then the ill-effects of model error may be expected to be negligible ~~(e. g., Doherty and Christensen (2011); Watson et al. (2013)).~~ (e.g., Doherty and Christensen, 2011; Watson et al., 2013).

The above findings and recommendations suggest that there is a significant need to identify better ways to assimilate diverse observation types including tracer concentration and tracer interpretation observations in numerical models for decision support. An enhanced ability to assimilate tracer observations, for example, will likely require increased model complexity (including advanced discretization, process representation and parameterization) to provide appropriate assimilation of rich and diverse data types that operate across a range of spatial and temporal scales commensurate with a given forecast.

However, an important and challenging compromise will be encountered: the need for enough model complexity to appropriately assimilate rich and diverse observations, while simultaneously ensuring that this level of complexity does not preclude the application of formal data assimilation and uncertainty quantification techniques due to the associated numerical instability and excessive run times. The navagation of this trade-off is central to effective and efficient decision-support modeling practice. In the meantime, tracer data model assimilation should involve processing or transforming of concentrations into quantities that may be more useful and may guard against ill-effects of history matching imperfect models (e.g., by integrating observations in space and time~~; Rasa et al. (2013); Knowling et al. (2019); White et al. (in press))~~) (e.g., Rasa et al., 2013; Knowling et al., 2019; White et al., in press).

Appendix A: First-order second-moment (FOSM) ~~uncertainty quantification and data-worth assessment~~methodology

This section provides a description of the FOSM approach used in the first case study to quantify uncertainty variance and assess data worth.

The posterior ~~(i.e., post-history matching)~~ covariance matrix of uncertain model parameters ~~,~~ $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ ~~,~~ can be approximated using the Schur complement ~~(Golub and Van Loan, 1996)~~(Golub and Van Loan, 1996; Tarantola, 2005):

$$\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\mathbf{J}^T\left[\mathbf{J}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\mathbf{J}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\right]^{-1}\mathbf{J}\boldsymbol{\Sigma}_{\boldsymbol{\theta}} \tag{A1}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is the prior ~~(i.e., pre-history matching)~~ parameter covariance matrix, which is specified based on expert knowledge pertaining to site system characteristics, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ is the epistemic observation noise covariance matrix (often assumed to have non-zero diagonal elements only), which includes the effects of model structural errors and measurement errors, and $\mathbf{J}$ is the Jacobian matrix of partial first derivatives (i.e., sensitivities) of simulated model outputs with respect to parameters. The Schur complement can be considered a linearized form of Bayes equation to estimate the second moment of the parameter and forecast posterior distribution ~~(e. g., Goldstein and Wooff (2007); Christensen and Doherty (2008); Dausman et al. (2010)).~~ (e.g., Goldstein and Wooff, 2007; Christensen and Doherty, 2008; Dausman et al., 2010).

Equation A1 assumes a linear relation between model parameters and simulated outputs (i.e., the sensitivities encapsulated within the $\mathbf{J}$ matrix is independent of the parameter values $\boldsymbol{\theta}$). It also assumes that parameter and epistemic uncertainty distributions are Gaussian (i.e., normal).

While the posterior parameter and forecast uncertainty variances yielded by FOSM may only be approximate (depending on the validity of the linear assumption), the computational efficiency with which a large number of different number of

conditioning "experiments" can be performed is unparalleled—these experiments facilitate rapid evaluation of the worth of different types of observations to reduce forecast variance. In addition, a number of studies have shown support for its usage especially in a relative second-moment sense ~~(e. g., Dausman et al. (2010); Herckenrath et al. (2011); Knowling et al. (2019)).~~ (e.g., Dausman et al., 2010; Herckenrath et al., 2011; Knowling et al., 2019).

The prior and posterior uncertainty variance surrounding a forecast σ_s^2 can be expressed by mapping uncertainty from parameter to forecast "space". This is achieved by computing the sensitivity of the forecast to model parameters, comprising the vector $\mathbf{y}$ (i.e., a row of $\mathbf{J}$). That is:

$$\sigma_s^2 = \mathbf{y}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \mathbf{y} \tag{A2}$$

and

$$\overline{\sigma}_s^2 = \mathbf{y}^T \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} \mathbf{y} \tag{A3}$$

The worth of data, expressed as a percentage, is given by:

$$\%DW = \frac{|\sigma_{\pm obs}^2 - \sigma_{base}^2|}{min\{\sigma_{base}^2, \sigma_{\pm obs}^2\}} \times 100 \tag{A4}$$

where $\sigma_{\pm obs}^2$ is the increase/decrease in forecast uncertainty variance as a result of the removal/addition of one or more observations or observation groups used for parameter conditioning, respectively, and σ_{base}^2 is either the forecast uncertainty calculated on the basis of all observation data/zero observation data, depending on whether data worth is being quantified by adding or removing observations.

Herein, we quantify %DW as a result of both the removal and addition of observation groups. We primarily focus on %DW values based on the removal of an observation group from an otherwise full observation dataset available for assimilation, given that these values reflect the unique (i.e., uncorrelated) information content of observations. However, the difference between %DW values arising from these different data-worth quantification approaches is used herein to comment on the level of information uniqueness/redundancy within observation groups.

It is important to note that each FOSM-based data worth assessment is conducted with respect to a single forecast (notwithstanding that we evaluate the worth of different observation data with respect to a number of different forecasts). We consider this to be a side-benefit of this approach, especially given the need for decision-support modeling to be undertaken in a forecast-targeted manner, as discussed recently by White (2017).

Author contributions. MJK, JTW and CRM contributed to the concept. MJK and JTW undertook the modeling analyses. MJK prepared the manuscript with input from JTW. JTW and CRM contributed to the manuscript preparation. PR and KH contributed to the underlying Heretaunga Plains models.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This research was performed as part of both the Te Whakaheke o te Wai and Smart Models for Aquifer Management Programmes, funded by the Ministry of Business, Innovation and Employment (New Zealand), with co-funding from Hawke's Bay Regional Council and Waikato Regional Council. The authors wish to thank Ty Ferre, Chris Turnadge and the anonymous reviewer for their helpful comments.

References

André, L., Franceschi, M., Pouchan, P., and Atteia, O.: Using geochemical data and modelling to enhance the understanding of groundwater flow in a regional deep aquifer, Aquitaine Basin, south-west of France, Journal of Hydrology, 305, 40 – 62, https://doi.org/https://doi.org/10.1016/j.jhydrol.2004.08.027, http://www.sciencedirect.com/science/article/pii/S002216940400407X, 2005.

Bedekar, V., Morway, E., Langevin, C., and Tonkin, M.: MT3D-USGS version 1: A U.S. Geological Survey release of MT3DMS updated with new and expanded transport capabilities for use with MODFLOW, U.S. Geological Survey Techniques and Methods 6-A53, p. 69, 2016.

Beyer, M., Morgenstern, U., and Jackson, B.: Review of techniques for dating young groundwater (<100 years) in New Zealand, Journal of Hydrology (New Zealand), 53, 93–111, https://doi.org/10.2307/43945058, http://www.jstor.org/stable/43945058, 2014.

Brynjarsdóttir, J. and O'Hagan, A.: Learning about physical parameters: The importance of model discrepancy, Inverse problems, 30, 114 007, 2014.

Cartwright, I. and Morgenstern, U.: Constraining groundwater recharge and the rate of geochemical processes using tritium and major ion geochemistry: Ovens catchment, southeast Australia, Journal of Hydrology, 475, 137 – 149, https://doi.org/https://doi.org/10.1016/j.jhydrol.2012.09.037, http://www.sciencedirect.com/science/article/pii/S0022169412008529, 2012.

Christensen, S. and Doherty, J.: Predictive error dependencies when using pilot points and singular value decomposition in groundwater model calibration, Advances in Water Resources, 31, 674 – 700, https://doi.org/https://doi.org/10.1016/j.advwatres.2008.01.003, 2008.

Clark, M. P. and Vrugt, J. A.: Unraveling uncertainties in hydrologic model calibration: Addressing the problem of compensatory parameters, Geophysical Research Letters, 33, https://doi.org/10.1029/2005GL025604, 2006.

Dausman, A., Doherty, J., Langevin, C., and Sukop, M.: Quantifying data worth toward reducing predictive uncertainty, Ground Water, 48, 729–740, 2010.

Doherty, J. and Christensen, S.: Use of paired simple and complex models to reduce predictive bias and quantify uncertainty, Water Resources Research, 47, 2011.

Doherty, J. and Welter, D.: A Short Exploration of Structural Noise, Water Resources Research, 46, 2010.

Doherty, J. E.: Ground water model calibration using pilot points and regularization, Ground Water, 41, 170–177, 2003.

Doherty, J. E.: PEST and its utility support software, Theory, Watermark Numerical Publishing, 2015.

Fienen, M. N., Doherty, J. E., Hunt, R. J., and Reeves, H. W.: Using prediction uncertainty analysis to design hydrologic monitoring networks: Example applications from the Great Lakes water availability pilot project, U.S. Geological Survey Scientific Investigation Report 2010-5159, p. 44, 2010.

Ginn, T. R., Haeri, H., Massoudieh, A., and Foglia, L.: Notes on Groundwater Age in Forward and Inverse Modeling, Transport in Porous Media, 79, 117–134, https://doi.org/10.1007/s11242-009-9406-1, https://doi.org/10.1007/s11242-009-9406-1, 2009.

Goldstein, M. and Wooff, D.: Bayes linear statistics, theory and methods, John Wiley & Sons, 2007.

Golub, G. and Van Loan, C.: Matrix Computations, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, http://books.google.com/books?id=mlOa7wPX6OYC, 1996.

Gusyev, M., Abrams, D., Toews, M., Morgenstern, U., and Stewart, M.: A comparison of particle-tracking and solute transport methods for simulation of tritium concentrations and groundwater transit times in river water, Hydrology and Earth System Sciences, 18, 3109, 2014.

Gusyev, M. A., Toews, M., Morgenstern, U., Stewart, M., White, P., Daughney, C., and Hadfield, J.: Calibration of a transient transport model to tritium data in streams and simulation of groundwater ages in the western Lake Taupo catchment, New Zealand, Hydrology and Earth System Sciences, 17, 1217–1227, https://doi.org/10.5194/hess-17-1217-2013, 2013.

Han, D. M., Song, X. F., Currell, M. J., and Tsujimura, M.: Using chlorofluorocarbons (CFCs) and tritium to improve conceptual model of groundwater flow in the South Coast Aquifers of Laizhou Bay, China, Hydrological Processes, 26, 3614–3629, https://doi.org/10.1002/hyp.8450, https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.8450, 2012.

Hansen, A. L., Refsgaard, J. C., Christensen, B. S. B., and Jensen, K. H.: Importance of including small-scale tile drain discharge in the calibration of a coupled groundwater-surface water catchment model, Water Resources Research, 49, 585–603, https://doi.org/10.1029/2011WR011783, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011WR011783, 2013.

Harbaugh, A. W.: MODFLOW-2005, the U.S. Geological Survey modular ground-water model – the Ground-Water Flow Process, vol. 6, 2005.

He, J., Reynolds, A. C., Tanaka, S., Wen, X.-H., and Kamath, J.: Calibrating Global Uncertainties to Local Data: Is the Learning Being Over-Generalized?, Society of Petroleum Engineers, https://doi.org/10.2118/191480-MS, 2018.

Herckenrath, D., Langevin, C. D., and Doherty, J. E.: Predictive uncertainty analysis of a saltwater intrusion model using null-space Monte Carlo, Water Resources Research, 47, n/a–n/a, https://doi.org/10.1029/2010WR009342, http://dx.doi.org/10.1029/2010WR009342, 2011.

Hunt, R. J., Feinstein, D. T., Pint, C. D., and Anderson, M. P.: The importance of diverse data types to calibrate a watershed model of the Trout Lake Basin, Northern Wisconsin, USA, Journal of Hydrology, 321, 286 – 296, https://doi.org/https://doi.org/10.1016/j.jhydrol.2005.08.005, http://www.sciencedirect.com/science/article/pii/S0022169405003926, 2006.

Hunt, R. J., Doherty, J., and Tonkin, M. J.: Are Models Too Simple? Arguments for Increased Parameterization, Groundwater, 45, 254–262, https://doi.org/10.1111/j.1745-6584.2007.00316.x, 2007.

Kirchner, J. W., Feng, X., and Neal, C.: Catchment-scale advection and dispersion as a mechanism for fractal scaling in stream tracer concentrations, Journal of Hydrology, 254, 82 – 101, https://doi.org/https://doi.org/10.1016/S0022-1694(01)00487-5, http://www.sciencedirect.com/science/article/pii/S0022169401004875, 2001.

Knowling, M. J., White, J. T., and Moore, C. R.: Role of model parameterization in risk-based decision support: An empirical exploration, Advances in Water Resources, 128, 59 – 73, https://doi.org/https://doi.org/10.1016/j.advwatres.2019.04.010, http://www.sciencedirect.com/science/article/pii/S0309170819300909, 2019.

Leray, S., de Dreuzy, J.-R., Bour, O., Labasque, T., and Aquilina, L.: Contribution of age data to the characterization of complex aquifers, Journal of Hydrology, 464-465, 54 – 68, https://doi.org/https://doi.org/10.1016/j.jhydrol.2012.06.052, http://www.sciencedirect.com/science/article/pii/S0022169412005537, 2012.

Li, H., Brunner, P., Kinzelbach, W., Li, W., and Dong, X.: Calibration of a groundwater model using pattern information from remote sensing data, Journal of Hydrology, 377, 120 – 130, https://doi.org/https://doi.org/10.1016/j.jhydrol.2009.08.012, http://www.sciencedirect.com/science/article/pii/S0022169409004880, 2009.

Masbruch, M., Gardner, P., and Brooks, L.: Hydrology and numerical simulation of groundwater movement and heat transport in Snake Valley and surrounding areas, Juab, Millard, and Beaver Counties, Utah, and White Pine and Lincoln Counties, Nevada, https://doi.org/10.3133/sir20145103, 2014.

McDonnell, J. J., McGuire, K., Aggarwal, P., Beven, K. J., Biondi, D., Destouni, G., Dunn, S., James, A., Kirchner, J., Kraft, P., Lyon, S., Maloszewski, P., Newman, B., Pfister, L., Rinaldo, A., Rodhe, A., Sayama, T., Seibert, J., Solomon, K., Soulsby, C., Stewart, M., Tetzlaff,

D., Tobin, C., Troch, P., Weiler, M., Western, A., Wörman, A., and Wrede, S.: How old is streamwater? Open questions in catchment transit time conceptualization, modelling and analysis, Hydrological Processes, 24, 1745–1754, https://doi.org/10.1002/hyp.7796, https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.7796, 2010.

Michael, H. A. and Voss, C. I.: Estimation of regional-scale groundwater flow properties in the Bengal Basin of India and Bangladesh, Hydrogeology Journal, 17, 1329–1346, https://doi.org/10.1007/s10040-009-0443-1, https://doi.org/10.1007/s10040-009-0443-1, 2009.

Moore, C. and Doherty, J. E.: Role of the calibration process in reducing model predictive error, Water Resources Research, 41, 1–14, 2005.

Morgan, L. K., Werner, A. D., and Simmons, C. T.: On the interpretation of coastal aquifer water level trends and water balances: A precautionary note, Journal of Hydrology, 470-471, 280 – 288, https://doi.org/https://doi.org/10.1016/j.jhydrol.2012.09.001, http://www.sciencedirect.com/science/article/pii/S0022169412007494, 2012.

Morgenstern, U. and Daughney, C. J.: Groundwater age for identification of baseline groundwater quality and impacts of land-use intensification – The National Groundwater Monitoring Programme of New Zealand, Journal of Hydrology, 456-457, 79 – 93, https://doi.org/https://doi.org/10.1016/j.jhydrol.2012.06.010, http://www.sciencedirect.com/science/article/pii/S0022169412004933, 2012.

Morgenstern, U., Stewart, M. K., and Stenger, R.: Dating of streamwater using tritium in a post nuclear bomb pulse world: continuous variation of mean transit time with streamflow, Hydrology and Earth System Sciences, 14, 2289–2301, https://doi.org/10.5194/hess-14-2289-2010, https://www.hydrol-earth-syst-sci.net/14/2289/2010/, 2010.

Morgenstern, U., Begg, J., van der Raaij, R., Moreau, M., Martindale, H., Daughney, C., Franzblau, R., Stewart, M., Knowling, M., Toews, M., Trompetter, V., Kaiser, J., and Gordon, D.: Heretaunga Plains aquifers : groundwater dynamics, source and hydrochemical processes as inferred from age, chemistry, and stable isotope tracer data, https://doi.org/10.21420/G2Q92G, 2018.

Niswonger, R., Panday, S., and Ibaraki, M.: MODFLOW-NWT, A Newton formulation for MODFLOW-2005, U.S. Geological Survey Techniques and Methods 6-A37, p. 44, 2011.

Oehlmann, S., Geyer, T., Licha, T., and Sauter, M.: Reducing the ambiguity of karst aquifer models by pattern matching of flow and transport on catchment scale, Hydrology and Earth System Sciences, 19, 893–912, https://doi.org/10.5194/hess-19-893-2015, https://www.hydrol-earth-syst-sci.net/19/893/2015/, 2015.

Oliver, D. S. and Alfonzo, M.: Calibration of imperfect models to biased observations, Computational Geosciences, 22, 145–161, https://doi.org/10.1007/s10596-017-9678-4, https://doi.org/10.1007/s10596-017-9678-4, 2018.

Pollock, D. W.: User guide for MODPATH version 6—A particle-tracking model for MODFLOW: U.S. Geological Survey Techniques and Methods, U.S. Dept. of the Interior, U.S. Geological Survey Reston, Va, version 6. edn., 2012.

Rajanayaka, C. and Fisk, L.: IRRIGATION WATER DEMAND & LAND SURFACE RECHARGE ASSESSMENT FOR HERETAUNGA PLAINS, https://www.hbrc.govt.nz/assets/Document-Library/Publications-Database/Aqualinc-Irrigation-water-demand-land-surface-recharge-assessment-Heretaunga-Plains-20180713.pdf, 2018.

Rakowski, P. and Knowling, M.: Heretaunga aquifer system groundwater model development report, https://www.hbrc.govt.nz/assets/Document-Library/Publications-Database/4997-Heretaunga-Model-Groundwater-Development-Report.pdf, 2018.

Rasa, E., Foglia, L., Mackay, D. M., and Scow, K. M.: Effect of different transport observations on inverse modeling results: case study of a long-term groundwater tracer test monitored at high resolution, Hydrogeology Journal, 21, 1539–1554, https://doi.org/10.1007/s10040-013-1026-8, https://doi.org/10.1007/s10040-013-1026-8, 2013.

Riva, M., Guadagnini, A., Fernandez-Garcia, D., Sanchez-Vila, X., and Ptak, T.: Relative importance of geostatistical and transport models in describing heavily tailed breakthrough curves at the Lauswiesen site, Journal of Contaminant Hydrology, 101, 1 –

13, https://doi.org/https://doi.org/10.1016/j.jconhyd.2008.07.004, http://www.sciencedirect.com/science/article/pii/S016977220800106X, 2008.

Sanford, W.: Calibration of models using groundwater age, Hydrogeology Journal, 19, 13–16, 2011.

Sanford, W. E., Plummer, L. N., McAda, D. P., Bexfield, L. M., and Anderholm, S. K.: Hydrochemical tracers in the middle Rio Grande Basin, USA: 2. Calibration of a groundwater-flow model, Hydrogeology Journal, 12, 389–407, https://doi.org/10.1007/s10040-004-0326-4, https://doi.org/10.1007/s10040-004-0326-4, 2004.

Schilling, O. S., Cook, P. G., and Brunner, P.: Beyond Classical Observations in Hydrogeology: The Advantages of Including Exchange Flux, Temperature, Tracer Concentration, Residence Time, and Soil Moisture Observations in Groundwater Model Calibration, Reviews of Geophysics, 57, 146–182, https://doi.org/10.1029/2018RG000619, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018RG000619, 2019.

Shmueli, G.: To Explain or to Predict?, Statist. Sci., 25, 289–310, https://doi.org/10.1214/10-STS330, https://doi.org/10.1214/10-STS330, 2010.

Siade, A., Prommer, H., Suckow, A., and Raiber, M.: Using Numerical Groundwater Modelling to Constrain Flow Rates and Flow Paths in the Surat Basin through Environmental Tracer Data, https://doi.org/10.25919/5b8055bfe3ea5, 2018.

Stewart, M. K. and Thomas, J. T.: A conceptual model of flow to the Waikoropupu Springs, NW Nelson, New Zealand, based on hydrometric and tracer 18 0, Cl, 3 H and CFC) evidence, Hydrology and Earth System Sciences, 12, 1–19, https://doi.org/10.5194/hess-12-1-2008, 2008.

Tarantola, A.: Inverse problem theory and methods for model parameter estimation, SIAM, 2005.

Turnadge, C. and Smerdon, B. D.: A review of methods for modelling environmental tracers in groundwater: advantages of tracer concentration simulation, Journal of Hydrology, 519, 3674–3689, 2014.

Watson, T. A., Doherty, J. E., and Christensen, S.: Parameter and predictive outcomes of model simplification, Water Resources Research, https://doi.org/10.1002/wrcr.20145, http://dx.doi.org/10.1002/wrcr.20145, 2013.

White, J. T.: Forecast First: An Argument for Groundwater Modeling in Reverse, Groundwater, 55, 660–664, https://doi.org/10.1111/gwat.12558, http://dx.doi.org/10.1111/gwat.12558, 2017.

White, J. T.: A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions, Environmental Modelling & Software, https://doi.org/https://doi.org/10.1016/j.envsoft.2018.06.009, http://www.sciencedirect.com/science/article/pii/S1364815218302676, 2018.

White, J. T., Doherty, J. E., and Hughes, J. D.: Quantifying the predictive consequences of model error with linear subspace analysis, Water Resources Research, 50, 1152–1173, https://doi.org/10.1002/2013WR014767, http://dx.doi.org/10.1002/2013WR014767, 2014.

White, J. T., Fienen, M. N., and Doherty, J. E.: A python framework for environmental model uncertainty analysis, Environmental Modelling and Software, 85, 217 – 228, https://doi.org/http://dx.doi.org/10.1016/j.envsoft.2016.08.017, 2016.

White, J. T., Knowling, M. J., and Moore, C. R.: Consequences of model simplification in risk-based decision making: An analysis of groundwater-model vertical discretization, Groundwater, doi: 10.1111/gwat.12957, https://doi.org/10.1111/gwat.12957, http://dx.doi.org/10.1111/gwat.12957, in press.

Wilding, T. K.: Heretaunga Springs: Gains and losses of stream flow to groundwater on the Heretaunga Plains, 2017.

Zell, W. O., Culver, T. B., and Sanford, W. E.: Prediction uncertainty and data worth assessment for groundwater transport times in an agricultural catchment, Journal of Hydrology, 561, 1019 – 1036, https://doi.org/https://doi.org/10.1016/j.jhydrol.2018.02.006, 2018.

Zheng, C. and Gorelick, S. M.: Analysis of Solute Transport in Flow Fields Influenced by Preferential Flowpaths at the Decimeter Scale, Groundwater, 41, 142–155, https://doi.org/10.1111/j.1745-6584.2003.tb02578.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-6584.2003.tb02578.x, 2003.

Zuber, A., Różański, K., Kania, J., and Purtschert, R.: On some methodological problems in the use of environmental tracers to estimate hydrogeologic parameters and to calibrate flow and transport models, Hydrogeology journal, 19, 53–69, 2011.