Reply to reviewer comments

We thank the reviewers for their careful and constructive review of our manuscript. In the following we give a point-by-point reply where we detail what was changed and added in the revision of the manuscript in response to the reviewer comments. The reviewer comments are given in black color and our replies in blue color. Changes and amendments to the manuscript are given in red color. Two additional minor modifications to the text for more clarity are listed in the end of the document.

Reviewer #1 (Alicia Wilson)

This paper presents a very nicely executed groundwater flow model from a site with a remarkable set of field observations. This is an impressive model of an impressive field site, and I really enjoyed reading the paper. I made some notes in the pdf, but they are nothing more than minor word changes. Instead, my suggestions are ways to improve the focus of the paper, which should help improve its impact. This comes down to three suggestions:

Dear Dr. Wilson,
thank you very much for taking the time reading through the manuscript and making these valuable suggestions.

1. Improve the focus of the introduction. As written, the justification of the work is pretty much that no one has ever made a realistic model of such a well-characterized field site. In the authors' defense, it really is a remarkable model, and different people will take different things from it. As a coastal hydrogeologist, I am delighted to see such a detailed description of the model set-up, and I am looking forward to having my graduate students comb through the methods to make sure they understand them. As indicated more clearly in the abstract, there is a lot for biogeochemists in the mixing and variability at the fringes of the USP; for everyone, there is clear visualization of the flow paths that support observed temperatures, salinity and age dates; and the model provides a firm foundation for future modeling efforts that will include reactive transport. It needs to be published. So from some point of view "here is our model" is not a terrible justification for the paper. But still, it would be better (more citable, easier to get through review) if the research questions were clearer in the introduction.

To me, the obvious questions that this paper either answers or can answer are (1) where are the most active and transient zones of mixing surrounding the USP; (2) where and on what time scales should we sample in beach systems (possibly with added justification by posing this as an iterative, coupled investigation, where the initial observations were needed to create the model, and now the model (and its truly cool movie) can guide the location and timing of future field observations – with extra points for posing these locations in a way that is transferable to other sites, e.g. MLLW, MHHW, etc); (3) how much do major modifications of beach morphology change flow and mixing below the beach (the 2022 beach shortening event) and (4) how do the idealized (and typically steady-state) models that permeate the literature compare to a real field site (exactly how "bad" are they)?

We fully understand your points here, and in response we made explicitly clear that our major aims were to reconstruct the physical processes in the beach aquifer as a consequence of various superimposing hydro(geo)logical forces in a real high-energy beach setting and to identify potential hotspots of intense gradients and variability that may also be relevant for biogeochemical transformation processes. We have also picked-up on your suggestion below regarding the fresh and saline submarine groundwater (SGD) fluxes and groundwater ages of this SGD, and how they seem to be dependent on the hydrological forcing and beach morphodynamics.

In response to this point we updated the statement of our aims in the introduction:

"The aims of the present study are to (i) provide a numerical groundwater flow and transport modelling analysis for a high-energy beach site that can explain real-world field observations and, thereby, (ii) disentangle the complex system behaviour **and reconstruct the physical processes in the beach aquifer as a consequence** of various superimposing hydro(geo)logical forces in a real high-energy beach setting, **(iii) identify potential hotspots of intense gradients and variability, and (iv) estimate the fresh and saline SGD fluxes and respective mean subsurface residence times, as well as to illustrate their transience and the reasons for it.**"

To introduce the SGD topic we updated the second sentence in the introduction:

"Quantification of these elemental fluxes, however, is difficult, because **both (i) fresh and saline SGD water fluxes are not easy to measure under field scale conditions (e.g., Burnett et al., 2006; Grünenbaum et al., 2020a) and (ii)** complex hydrogeological and hydrobiogeochemical processes, and thus reactive transport within the STE, modify solute concentrations prior to discharge."

With respect to model-informed sampling as an objective: Given the rather complicated superposition of morphodynamics, spring-neap and daily tidal changes affecting the SGD fluxes, it will be rather hard, or even impossible to come up with an optimized, model-informed iteratively adapted strategy on sampling the different out-flowing SGD components, as conditions seem to quickly change in an unpredictable manner. In the beginning of the project this was not clear to us either. What can be said however, is that in such systems, dense monitoring at high resolution and over a few years are needed to come up with a reasonable conceptual model and parametrization that allows to visualize a more or less complete picture of the underground and which (i) can be used to identify specific locations for sampling where subsurface (reactive) hots-spots are likely prevailing on somewhat longer time-scales than just days, and (ii) serve as bases for further computing reactive transport.

We have added this conclusive elaboration to the Conclusion chapter as:

"Thus, the identified hotspots of temporal variability may also be reactive hots-spots. This means that our model can be used as a guide for field sampling campaigns targeting at subsurface biogeochemical transformation processes. Nevertheless, dense monitoring at high resolution and over a few years are needed to obtain a reasonably calibrated numerical model that allows to visualize a detailed picture of the flow and transport processes in the subsurface of a high-energy beach.
Further, the model results show that saline SGD fluxes vary considerable at daily and spring-neap time scales. The astronomical tide is always superimposed by meteoric effects which alter high water and low water levels from day to day. These are not easily predictable and therefore cannot be reflected in predictive modelling to guide field sampling campaigns on measuring these fluxes, e.g., by seepage meters. The model results suggest that a robust mean value can only be obtained when measurements capture these daily dynamics over a full spring-neap tide

cycle. On longer-term, i.e., monthly to yearly time scales these fluxes gradually change and appear to be mainly controlled by the spatial variability of the beach slope rather than its mean gradient. The large dynamics in groundwater age of the fresh SGD component seem to loosely correlate with the magnitude of the saline SGD flux. Higher saline flux also pushes out younger fresh water that originates from meteoric recharge at the upper beach."

Your suggestion on investigating the suitability of idealized models to capture the flow and transport patterns in the subsurface of a high-energy beach is an important point. Nevertheless, this would require more simulation runs, switching on and off the dynamics of individual boundary conditions, followed by a stringent analysis of the simulation results for the affected state-variables to be decided on, e.g., distribution, intensity and dynamics of hot-spot, fresh and saline SGD fluxes, residence time (distribution) etc. This requires careful work and planning, and is as we think, beyond the scope of this paper. It is certainly needed and will be worked on in the next phase of our project. The last two sentence in the original manuscript already pointed in this direction:

"*The present model will serve as tool for further analyses. For example, the importance of individual boundary effects (e.g., dynamics in beach morphology, storm events, or wave-set) on mixing and residence times, which are critically important for understanding and quantifying reactive transport in the present subterranean estuary, will be investigated.*"

2. Having just suggested the beach shortening event as a possible focus of the paper, it would be helpful to illustrate the changes that surrounded that event in a more focused way. As it is, the shortening event comes up a lot, but it is hard to see it in the figures, and it is dispersed among different sections. It is nicely visible in the movie, but I wonder if some kind of summary figure could be developed, with a focused discussion of the event all in one place. Part of this may also be that I thought I saw some interesting things presented as part of the calibration (e.g. illustrating the ability of the model to reproduce transients), when some of the findings could actually be interesting results in addition to calibration targets. (Perhaps sub-headings within the results section would help focus readers' attention on the most salient points?)

Although the beach shortening event was mentioned several times, it was not meant to be a focus of the paper, and we would like to keep it that way in the revised manuscript. One of the major aims was to reconstruct the physical processes in the beach aquifer as a consequence of the superposition of various impacting hydro(geo)logical forces, where the sudden beach erosion event is only one of out of many effects. We will make this aim clearer in the introduction (see our reply to point #1). Therefore, we would like to refrain from adding a summary figure on the beach erosion event.

3. Some of this may be intended for a later paper, but it would be very helpful to pull in the SGD crowd with some fluxes. The model is the perfect way to reconstruct volumetric fluxes through different parts of the model (the USP, the FDT, the total variation in fresh groundwater discharge from the dunes). This would actually be a great way to illustrate the potential importance of the 2022 beach shortening, where line graphs could show the width of the beach and the changes in the various fluxes over time. Perhaps with the addition of some kind of average age of water discharging from different flow systems (the USP, FDT), or the average area of each flow system (admittedly harder to automate than fluxes or ages)? Presenting information that is hard to collect

in the field (fluxes), expensive to analyze (age dates), or that happened before extensive monitoring began (beach shortening) is a fantastic justification for creating a model. Again, some of this might be for the next paper, but all of it would be interesting.

We agree, and have picked up on your suggestion with respect to the SGD fluxes. The analyses revealed that the sudden beach shortening and corresponding abrupt change in the mean beach slope appear to have little effect on the SGD fluxes. Rather, on the longer time-scales, the spatial variability of the beach slope (in terms of standard deviation) in the intertidal zone seem to control the SGD fluxes. Also, the spring-neap cycle play a role (as also previously shown in Heiss & Michael, 2014, WRR), and even on a daily time-scale, tide range changes due to meteoric effects seem to have a big impact.

In response to this comment we have added a new figure (Figure 7) showing SGD fluxes and mean groundwater ages over time:

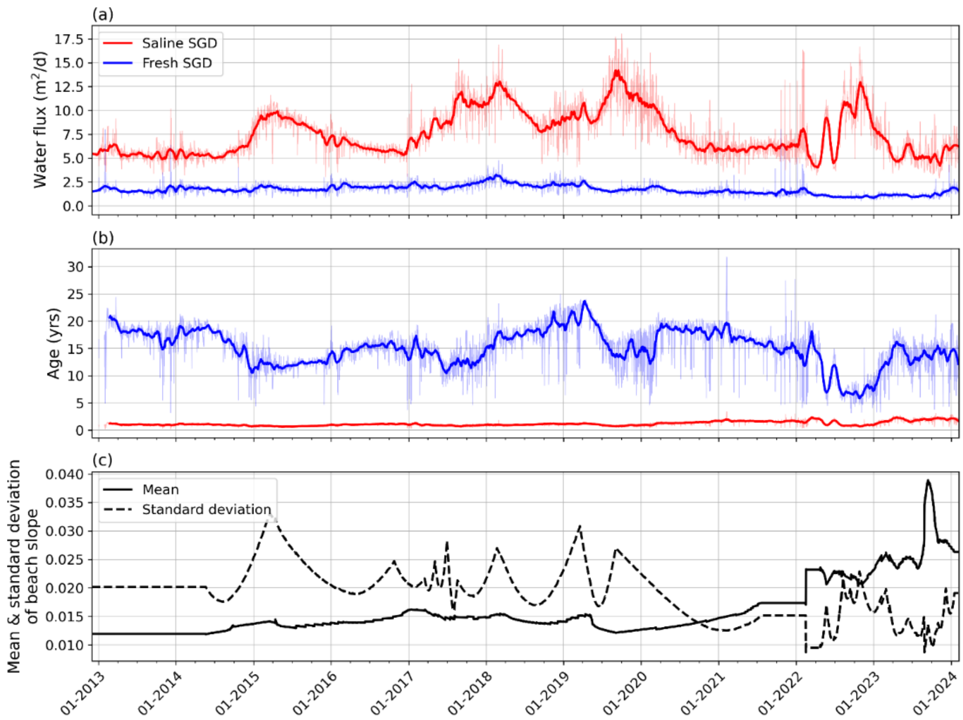

Figure 7: (a): Simulated time series of saline (red) and fresh (blue) SGD water flux. Thin lines represent the fluxes derived half-daily after each stress period, thick lines represent the signal smoothed with a Savitzky-Golay filter with a polynomial order of 2 and a window length encompassing 101 data points. (b): Simulated time series of flux weighted groundwater ages of the saline (red) and fresh (blue) SGD components. Thin lines represent the fluxes derived half-daily after each stress period, thick lines represent the signal smoothed with a Savitzky-Golay filter with a polynomial order of 2 and a window length encompassing 101 data points. (c): Mean and standard deviation of the beach slope in the intertidal zone. Only the time-period from Jan 2013 onwards are presented, i.e., the period were the dynamics of the sea side boundary condition is considered in the model.

We have also added a new figure (Figure S10) to the supporting information showing the correlation of the saline SGD flux and age with mean beach slope and its standard deviation:

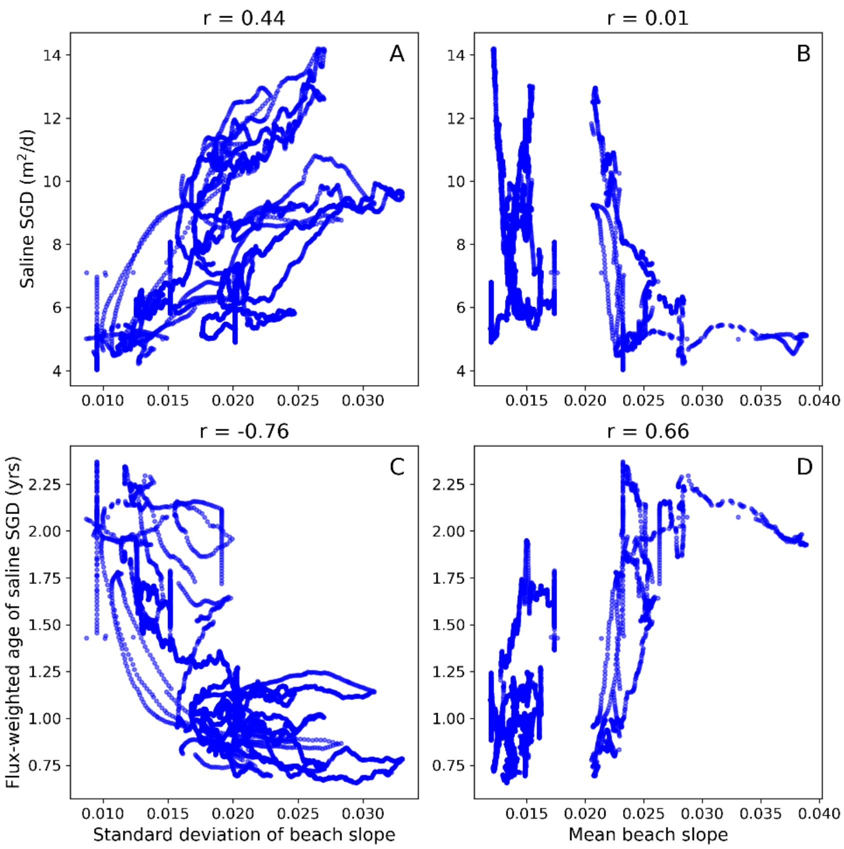

Figure S10: Scatter diagrams of simulated saline SGD versus (A) the standard deviation of beach slope and (B) the mean beach slope, and its flux weighted age versus (C) the standard deviation of beach slope and (D) the mean beach slope, with the corresponding Pearson correlation coefficient r.

And we added the following text to the Results and Discussion:

"The simulated fresh and saline SGD water fluxes vary on different time-scales and are in the range of 1.0 - 5 m^2/d and 2.5 – 17 m^2/d, respectively (Figure 7a). High-frequency variations can be seen on daily time-scales, as well as on spring-neap time-scales. Variation on daily time-scales are attributed to meteoric effects altering the astronomical tidal range. Changes between more than 3 m^2/d and 10 m^2/d for the fresh and saline component, respectively, can be observed (Figure 7a). High-frequency variations in SGD water flux stemming from semi-diurnal tides variation and spring-neap tides were also found by Heiss and Michael (2014) for a beach in Cape Henlopen, Delaware, USA. The longer-term variations for the saline component (smoothed red line in Figure 7a) appear to loosely correlate with the changes of spatial variation in beach morphology in terms of the beach slope's standard deviation, rather than the mean beach slope in the intertidal zone (Figure 7c, Figure S10A and Figure S10B). Thus, the sudden drop in the saline SGD water flux at and after the beach shorting event in February 2022 seemed to be attributed to the abrupt decrease in beach slope variability rather than the step-increase in mean beach slope. The flux recovered as the beach slope variability recovered. Previous model-based saline SGD flux estimates from May 2014 (Beck et al., 2017) were about 4.3 m^2/d. This is only a little bit lower than the here computed flux of about 5 m^2/d, but otherwise agrees very well with the current model results. The lower flux computed by Beck et al. (2017) is expected, as their model considered fixed beach morphology, mean tidal range and did neither account for dynamic wave-set up nor storm floods. The variation of the fresh SGD flux is mainly attributed to the seasonality of groundwater recharge on the island.

The simulated flux weighted mean groundwater age of the fresh SGD component range from 4 to 25 years, with an average of around 15 years (Figure 7b). Longer-term variation is also

superimposed by high-frequency variations on semi-diurnal and spring-neap time-scales. Correlation with the spatial variation of the beach slope is not visible. The longer-term variability appears to depend to some degree on the saline SGD flux (Figure 7a). It seems that predominantly younger freshwater originating from meteoric recharge at the upper beach is being pushed out during periods of elevated saline SGD fluxes. The flux weighted mean groundwater age of the saline SGD component varies between 0.75 and 2.3 years, and correlates with both, the spatial variation in beach morphology and the mean beach slope in the intertidal zone (Figure S10C and Figure S10D). The phenomenon of higher saline SGD flux and its younger mean age with increasing relief in the intertidal zone can be explained by so-called Tóthian flow. In his theoretically analysis, Tóth (1963) showed that topography-controlled flow with an increasing number of valleys and hills within a basin (and thus increased variability in the topographic gradient) leads to more local and shallower flow cells of higher discharge and shorter residence times."

These findings were also summarized in the Abstract as:

"Computed saline SGD water fluxes varied considerable on daily and spring-neap time scales, as well as on the longer term, i.e., monthly to yearly time scales. The rather gradual, longer-term changes in flux appear to be mainly controlled by changes in spatial variability of the beach slope. The simulated groundwater age of the fresh SGD component varied between 4 and 25 years, and predominantly depended on the magnitude of saline SGD flux."

In summary, I think this work is interesting for many reasons, and I hope that an improved focus will help a broader range of readers appreciate its contributions.

Thank you for your kind evaluation.

Minor comments from Alicia Wilson annotated in the pdf of our original submission

Line 59: "cross-shore transects": I may be the only one, but I can never tell what a "cross-shore" transect is. I strongly prefer "shore-perpendicular."

Yes, we agree, the first author of this manuscript also gets always confused about this. Thank you for this suggestion. We have replaced the term "cross-shore" with "shore-perpendicular" throughout the entire manuscript.

Figure 1: I'm certain this will be higher resolution in the final submission. (It's easier to understand the uppermost grid layer from the figure than the text, but only if you can see the grid!)
Yes, we have also added a Figure of the full model domain (Figure S1) in the supporting information (see our response to reviewer #2). In there, the grid and the top layer are better visible.

Lines 264-266: "*Preliminary calibration runs suggested that the applied 3rd type boundary condition could underestimate the seawater infiltration rate on the upper beach. Therefore, an additional seawater infiltration q_{sea} was implemented via a 2nd type boundary condition ...*": I gather that the discrepancy couldn't be fixed by altering the conductance instead? Is that because this zone was among the most significantly affected by wetting and drying? This is a bit in the weeds but since it's detailed so well here I am curious.

Yes, this is absolutely correct. We tested whether adaption of the conductance could improve the calibration results in this regard. But it didn't work out, likely because the conductance was already high due to the high hydraulic conductivity of the upper unit, which goes into the calculation of the conductance. As stated in the text, a rigorous description of the infiltration process needs a multi-phase modelling approach, but in SEAWAT we only have a 3rd-type boundary condition at hand, crudely trying to mimic the bulk behavior. On the other hand, in a multi-phase modelling approach, a lot of physical parameters (saturation dependent permeability's, air entry pressures, etc.) need to be known or calibrated in an otherwise numerically much more demanding procedure. That's why we introduced a defined saltwater infiltration flux during storm events, subject to calibration.

In response to this comment we introduced the following sentence (in bold):

"… the applied 3rd boundary condition could underestimate the seawater infiltration rate on the upper beach. **An adjustment of the hydraulic conductance C_b did not alleviate this behavior**. Therefore, an additional seawater infiltration q_{sea} was implemented via a 2nd type boundary condition …"

Other small linguistic improvements annotated on the original pdf were adopted but not listed in here.

Reviewer #2

The authors investigate the groundwater flow system, the heat and mass transport processes along a two-dimensional model of a high-energy subterranean estuary. During a year and a half of field observations — with sampling varying in time and space — they recorded the hydraulic head, temperature, salinity, and water age. A significant part of the model parameters (49 in total!) was determined by calibration, performing approximately 10,000 simulations and "selecting the best solution" using the particle swarm optimization (PSO) method. I consider the work performed to be very enormous, the techniques used (field measurements, numerical simulation, automated calibration) to be very comprehensive, and the results achieved and presented to be realistic. In light of all this, I consider the manuscript — after some clarification and completion — suitable for publication in the HESS journal.

Dear reviewer,
thank you very much for your careful evaluation of our manuscript and the kind suggestions for improvement. This is very much appreciated.

Major comments/concerns
The description of the numerical model is rather sporadic, which does not facilitate the work of modelers, a deeper understanding of the article, or its potential reproducibility.

We agree and we have now added detailed information in the manuscript and supporting information as specified below.

Therefore, I suggest:

a. the equations used in the simulations should be presented in the article if possible, or at least in the supplementary material, where it should be clarified that water density depends purely on concentration, while viscosity is constant.

We have included the governing equations in the supporting information.

b. Due to the complexity of the processes, the boundary conditions are also very complicated (Section 2.2) and difficult to follow. It would be useful for the reader if they were summarized in a figure, emphasizing their time dependence. This would also be useful because the total length of the model is 1450 m (Page5Line134), but the entire model area is never shown in the manuscript, only between 500–1300 m and 800–1300 m. This would also eliminate the shortcoming that I could not find any information on the salinity, temperature, and water age boundary conditions for the vertical and lower boundaries.

We agree, we have now included a figure (Figure S1) of the entire model domain, including the definition of all model boundaries and their time-dependence in the supporting information:

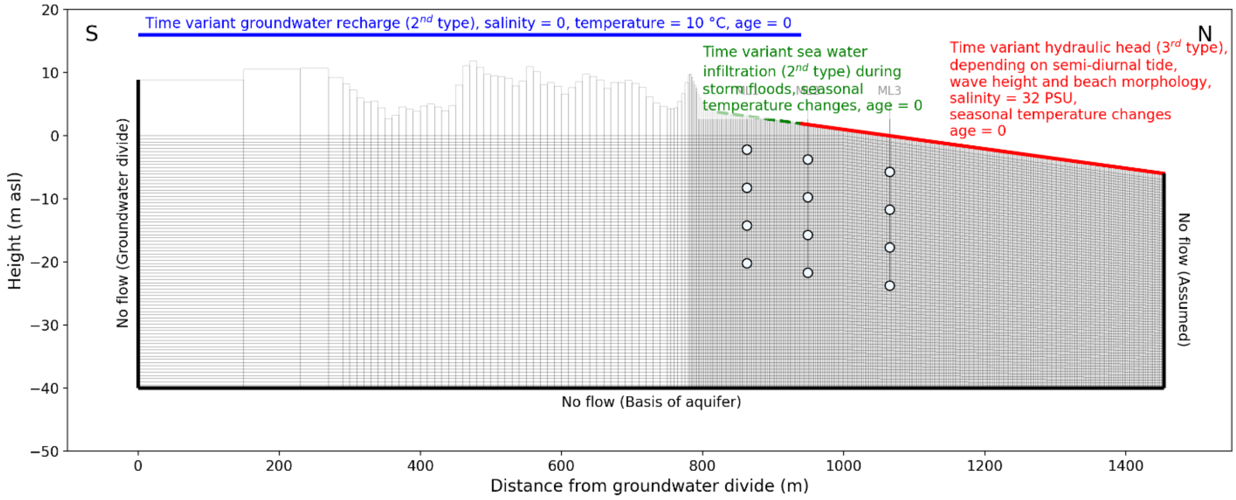

Figure S1: Model domain and defined boundary conditions. The multi-level observation wells ML1, ML2 and ML3 are presented for a better orientation.

c. Fixed model parameters (e.g. R, Ss, D) should be tabled too.

Yes, we have included them in Table A1.

2. The PSO algorithm is indeed suitable for finding the minimum of highly non-linear inversion problems, although this is a major challenge in the case of a model involving almost 50 unknowns, time-dependent, free topohaline convection. Could you show a figure that illustrates the convergence of the method? What object function did PSO use to quantify the individual simulations at the pilot points?

We have included the following text passages and figures in the supporting information that show how the parameters converge to their final values over the course of the optimization iterations:

Convergence of the PSO optimization

The preliminary optimization runs for finding optimal weights and an optimal number of pilot points by trial- and error indicated that 100 iterations were sufficient to converge to the global minimum (Figure S7). Though, more iterations than that could further bring down the objective function to some little extent, but virtually led to no visual improvement of the model fit.

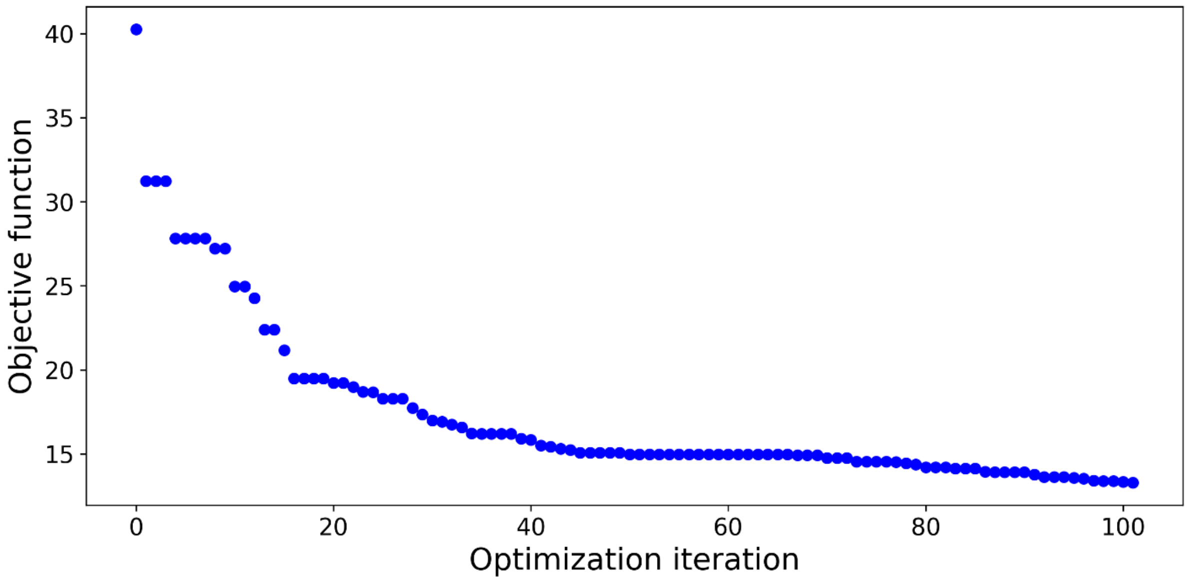

Figure S7: Objective function over the course of the optimization.

At the end of the optimisation run it is noticeable that for some parameters the spreading of the particles (i.e., parameter values associated with the particles) is broader than for others (Figure S8), suggesting their lower sensitivity on model results. A rigorous parameter sensitivity assessment, however, is not possible with the PSO approach.

Note that due to computational constrains of the High Performance Computer, the optimization was split in two subsequent batches of 50 iterations each. The second batch uses initial parameter values that came out at the end of the first batch as best particle values, but with re-set inertia of the particle velocities. This gives some perturbation at the beginning of the second batch, i.e., at optimization iteration 50 (Figure S8). While this leads to some delay in the convergence (Figure S7), it is still beneficial for further overcoming undesired local minima in the objective function.

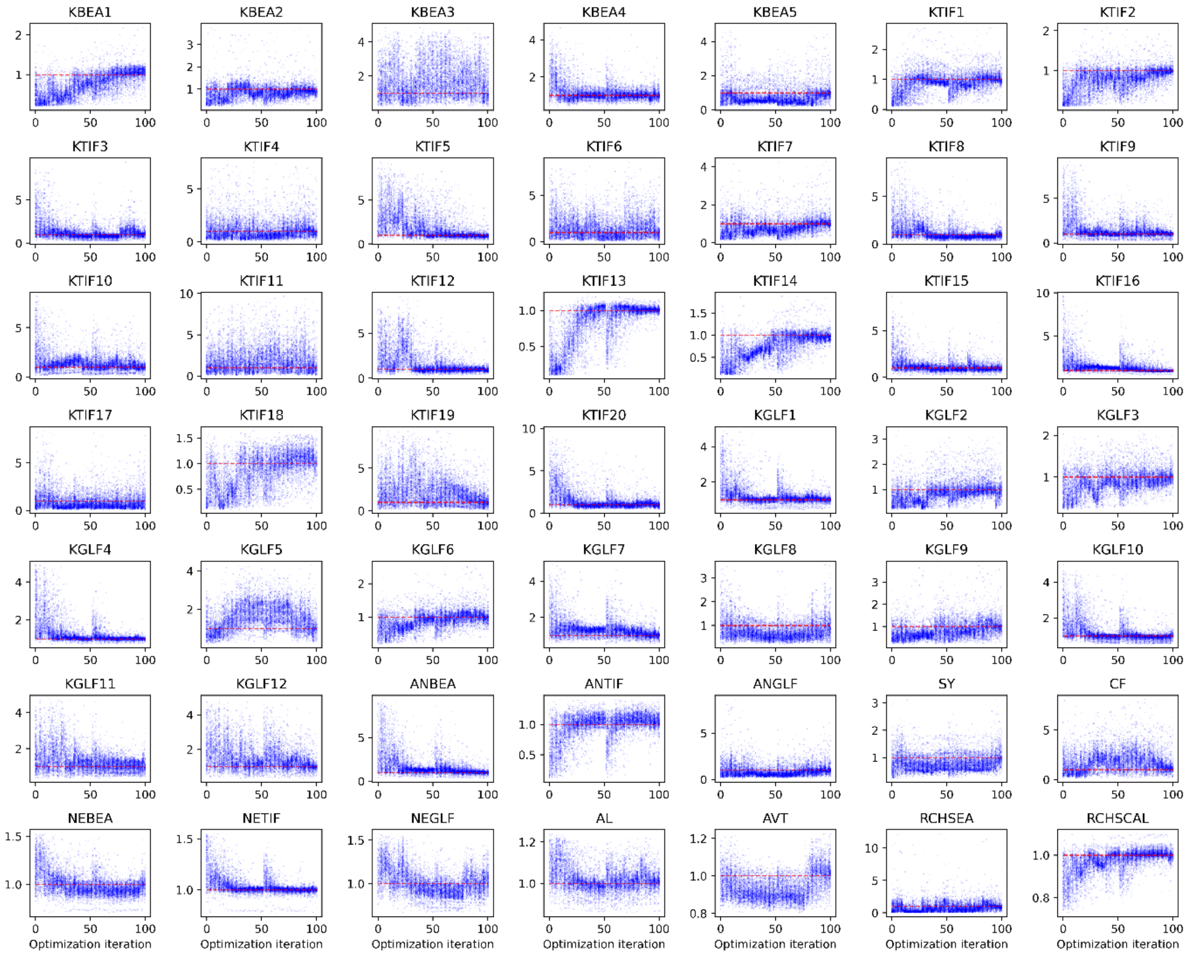

Figure S8: Evolution of the particle swarm (100 particles) for each parameter (normalized to the final value) over 100 optimization iterations (little blue dots). The dashed red line indicates the final normalized parameter value.

The objective function was the weighted sum of squared residuals, of all observations, i.e., one single value. We will elaborate on the weighting scheme in our reply to the next comment.
Regarding pilot points: From your comment we get the impression that our explanations on 'pilot points' were perhaps misleading. Pilot points are not used for evaluating the difference between observed and simulates state-variables, e.g., salinity, and thus they do not appear in the objective function. Instead, they are predefined locations where hydraulic conductivities are estimated and updated during each optimization iteration and which serve then as reference points to fill the grid-cells in between them via interpolation.
In response to this comment we now have extended the explanation from:

"Therein hydraulic conductivities were defined at a distributed set of locations within each geological unit, where they were subject to calibration and used to assign a spatially non-uniform hydraulic conductivity field via Kriging interpolation."

To:

"Therein hydraulic conductivities were defined at a distributed set of defined locations - the pilot points - within each geological unit. These serve as reference points where the hydraulic

conductivities were adjusted and used to assign a spatially non-uniform hydraulic conductivity field via Kriging interpolation over the entire model grid at each iteration step during the calibration procedure."

3. The cross plots in Figure 3 show the relation between the measured data and the simulated results for different calibration quantities. The figure is not discussed in the manuscript, which could answer some of the arising questions. Are the parameters of the best-fitting model shown on the y-axes? Were these diagrams used by PSO to qualify each simulation? Are the quantities at different pilot points and at different times shown here? Are all measurement points given equal weight? Could this explain why the root mean square error of the smaller number of water age data is much greater than that of the large number of e.g. head data? I would appreciate an explaining paragraph regarding Figure 3.

Yes, this figure was just to show how the model performs with respect to the different state-variables individually. All the displayed numerical performance measures were calculated by using the unweighted residuals between simulation and observation. We feel that the figure is not as important as it first appeared, as the model's performance is also very well visible in the follow-up figures in the manuscript. Figure 3 was merely meant as an add-on. Thus, we have decided to shift this figure into the supporting information (as Figure S9) and add some more explanations in the figure caption.

Weights: In the objective function, the weights are usually set simply as the inverse of the measurements standard deviation. However, in a multi-objective optimization framework, where different types of state-variables and with an uneven number of data points are used for calibration, weighting is not that straight forward any more. The weights need to balance the uneven amount of data points in the individual datasets of different state-variables and their often vastly different numerical values – And with that, their uneven contribution to the objective function. During the optimization process, these contributions likely change, as the goodness of fit for one or the other dataset may become superior. Thus, to compensate for that and keep the right balance, an iterative approach would be ideal to evaluate the objective function after each iteration in the optimization process and to adapt the weights 'on-the-fly'. However, there is still a debate on the weight problem in the scientific community when it comes to multi-objective optimization. It is difficult and we would refer to in-depth elaborations on the topic by e.g. Dai and Samper (2004) and to the PEST++ documentation (https://github.com/usgs/pestpp).
In the PSO framework of PEST++, however, there is no automatic on-the-fly update of weights. Instead, one has to stick with fixed weights. This leads to the problem that we do not end up with only one single optimization run, but with many trial runs, in which suitable weights have to be manually found via trial-and-error – A daunting task. While 10000 simulations were carried out for the 'final' optimization run, more than 10 of those runs had to be carried out prior to that in order to find suitable weights which could balance the different contributions of the individual datasets to the objective function in a way that finally a reasonable fit could be achieved.

In response to this comment we have added the following elaboration on how we assigned the weights in the Methods chapter where we explain the calibration approach:

"For the optimization, observation weights were applied for 5 individual observation groups, namely (a) hydraulic heads, (b) salinity, (c) temperature at the multi-level monitoring wells and

(d) temperature at SAMOS and (e) groundwater ages. In the objective function, the weights are usually set simply as the inverse of the measurement standard deviation. However, in a multi-objective optimization framework, where different types of state-variables with an uneven number of data points in each dataset are used for calibration, weighting is not straight forward anymore (Dai and Samper, 2004): The weights need to balance the uneven amount of data points in the individual datasets of different state-variables and their often vastly different numerical values. In addition, their contribution to the objective function, i.e., the weighted sum of squared residuals, change in the course of the optimization. Thus, an iterative approach is needed (Dai and Samper, 2004). In this work, preliminary optimization runs were carried out to find optimal weights for the different observation groups via trial and error. The final weights are presented in the supporting information. The convergence of the PSO optimization is also presented in the supporting information."

The computation of the objective function and the final weights are now presented in the supporting information.

Dai, Z., and J. Samper (2004), Inverse problem of multicomponent reactive chemical transport in porous media: Formulation and applications, Water Resour. Res., 40, W07407, doi:10.1029/2004WR003248.

4. The thermal retardation factor was fixed during the simulations, R=2. This value seems very high, although there is some ambiguity in the literature concerning the definition of the R value. What porosity, density and specific heat values give this factor?

Yes, in the literature different values of the thermal retardation factor R were reported. We have not calculated it from the porosity, density and specific heat capacity but simply used the average value of values reported in previous studies (referenced in the manuscript) that employed heat transport simulations in sandy aquifers. These ranged between 1.8 and 2.24.
With specific heat capacities of quartz and water of 710 J/kg/K and 4183 J/kg/K, respectively, and specific densities of 2650 kg/m^3 and 1025 kg/m^3, respectively, the porosity would need to be 0.32 when assuming and R of 2. The calibration results suggested effective porosities between 0.25 and 0.28. Given the uncertainty of those model-based estimates at field-scale, we think that the assumption of R=2 is acceptable.

Minor comments

Fig. 2 It would be useful to indicate the north-south direction.
Done.

P7L181 Dez → Dec
Done.

P15L326 3He → 3H
Done.

P15L332 3He → 3H
Done

P16L362 permeable material → permeable medium/sediment…
Done.

Fig. 7 caption I guess the figure illustrates the standard deviation of simulated salinity, groundwater age and temperature time series along…
Yes. We have added the term 'time series'.

P17L382 Figure 5 → Figure 5.b
Corrected.

P20L457 Reference is missing
We have now included appropriate references. The sentence now reads: "Here, mixing-controlled reactions, e.g., the oxidation of dissolved ferrous iron and precipitation of iron(hydr)oxides as so-called iron-curtain **(Charette and Sholkovitz, 2002)**, or nitrification due to ammonium oxidation **(Spiteri et al., 2008)** may be promoted."

Charette, M.A., Sholkovitz, E.R., 2002. Oxidative precipitation of groundwater-derived ferrous iron in the subterranean estuary of a coastal bay. Geophysical Research Letters 29, 85-1-85–4. https://doi.org/10.1029/2001GL014512

Spiteri, C., Slomp, C.P., Tuncay, K., Meile, C., 2008. Modeling biogeochemical processes in subterranean estuaries: Effect of flow dynamics and redox conditions on submarine groundwater discharge of nutrients. Water Resources Research 44. https://doi.org/10.1029/2007WR006071

P20L460 Reference is missing
We have now included appropriate reference. The sentence now reads: "In these zones, biogeochemical turnover and secondary reactions may be controlled by the varying degradability of dissolved organic matter as a function of groundwater age **(e.g., Waska et al., 2021)**."

Waska, H., Simon, H., Ahmerkamp, S., Greskowiak, J., Ahrens, J., Seibert, S.L., Schwalfenberg, K., Zielinski, O., Dittmar, T., 2021. Molecular Traits of Dissolved Organic Matter in the Subterranean Estuary of a High-Energy Beach: Indications of Sources and Sinks. Front. Mar. Sci. 8. https://doi.org/10.3389/fmars.2021.607083

It is somewhat surprising that the terms "biogeochemical" and "reactive" appear 16 and 11 times in the manuscript, resp., even though the simulation does not involve biogeochemical or reactive transport modeling…

Yes, there is a large scientific community that is interested in unravelling the hydrobiogeochemical transformation processes in subterranean estuaries (or beach aquifers) aiming at estimating element fluxes to the ocean via submarine groundwater discharge (SGD). Flow and transport dynamics critically affect biogeochemical turnover and thus reactive transport in the subsurface. As elaborated in our conclusion, the here identified hotspots of high temporal variability with respect to salinity and groundwater age conditions will likely be also hot spots for reactive processes. Thus, we think that our results are also important for hydro- and biogeochemists in this research field.

Nevertheless, maybe somewhat related to this comment, we will take up the suggestion of reviewer #1 and will extract and present fresh and saline SGD fluxes from the model in order to make this paper also potentially interesting for scientists that focus on the SGD water fluxes, rather than solely biogeochemistry.

Overall, the work presented in the manuscript is very substantial, multifaceted, and fits well with the subject matter of the HESS journal, so, after taking into account the above clarifications, I support its publication.

Thank you for your kind evaluation.

Additional minor modifications

In the Abstract we rephrased the sentence:

"*The calibrated model is able to replicate the principal behaviour of the highly transient system and enabled the identification of temporal variability hotspots.*"

To:

"The calibrated model is able to replicate the principal behaviour of the highly transient system and enabled the identification of hot spots of high temporal variability in the investigated state-variables"

Under 3.1 Calibration results we state:

"*In the observed data at ML2 it is noticeable that the temperature-based residence times (under consideration of a thermal retardation factor of R = 2) are considerably younger than the 3H/He groundwater ages.*"

In order to better understand this statement, we added:

"While between August 2022 and May 2023 the temperature peak shift suggests a seawater residence time of about 1.5 years to ML3-24m (Figure 3c), it is almost double according to the 3H/He age dating in November 2022 (Figure 5)."