Reviewer 1

We thank reviewer 1 for his/her thoughtful review that greatly helped to improve our paper. The feedback and suggestions were very motivating for us to improve our paper. In this document, we have entered your comments in italics, added our response as regular text, and subsequently added suggested changes to the paper in red. After our responses you can find the suggested changes made to the text in red. We also like to add that we added supplementary information to this document, as it spares you the effort of looking for an appendix of a previous paper we wrote. Line numbers are written as "**[P… L…]**", indicating page number, line number in the no markup document.

General comments

First of all though, with regards to the Introduction section, you can build a stronger justification for this study by first discussing the studies that have been conducted in other areas (presently starting on page 3, line 25), which demonstrate that over-simplified models of large-scale aquifer systems are conceptually flawed. Then you can bring in the Nile delta system, and argue that there is also a need there to analyze the palaeo-geographical evolution in order to understand the present-day conditions. By starting off with a focus on the Nile straight away, you present it too much as a regional problem, not a scientific analysis that yields outcomes that can be transferred to other study regions.

This is a good suggestion, as it shifts the focus of the paper to the scientific problem instead of the regional problem. This will also help with Reviewer 2's main concern. We have added sentences of introduction to the broader problem **[P2 L1-7]** and consequently followed your suggestions. Given the severity of the local problems, we subsequently moved the former first paragraph to a dedicated section under "Area description" **[P4 L2-21]** .

The paper now starts as follows:

Palaeohydrogeological conditions have influenced groundwater quality in the majority of large-scale groundwater systems, since groundwater below 250 m depth is dominated by groundwater with an age of over 12 ka (Jasechko et al., 2017). These conditions can especially be found in deltaic areas, where the effects of marine transgressions are often still observed in groundwater salinities (Larsen et al., 2017). More specifically, their low elevation allowed for far reaching marine transgressions, leading to a large vertical influx of sea water, and hampered subsequent flushing with fresh water after the marine regression. This hypothesis is supported by hydrogeochemical research in several deltas (e.g. Colombani et al., 2017; Fass et al., 2007; Faye et al., 2005; Manzano et al., 2001; Wang and Jiao, 2012). The physical justification for this hypothesis, however, often still has to be tested, with a few notable exceptions (Delsman et al., 2014; Larsen et al., 2017; Van Pham et al., 2019), and can be provided by palaeohydrogeologic modelling, which has recently provided important insights for several cases. Gossel et al. (2010) created a large-scale 3D variable-density groundwater model of the Nubian Aquifer System and showed that seawater intrusion has occurred since the Pleistocene Lowstand towards the Qattara Depression (North-West Egypt). Later, Voss and Soliman (2013) showed with a parsimonious 3D model of the same groundwater system that groundwater tables are naturally lowering during the Holocene, since they receive limited recharge and are drained into oases or sabkhas. Moreover, the authors used an inventive validation method by comparing the position of discharge areas in the model with a dataset of oases or sabkha locations.

Delsman et al. (2014) conducted a detailed palaeohydrogeological reconstruction of the last 8.5 ka over a cross-section in the Netherlands to show that the system has never reached a steady state. They showed that the Holocene transgression caused substantial seawater intrusion, from which the system is still recovering. Using a combination of geophysics and 2D numerical models, Larsen et al. (2017) showed that during the Holocene transgression sea water preferentially intruded in former river branches in the Red River Delta, Vietnam. Van Pham et al. (2019) showed that most of the fresh groundwater in the Mekong Delta (Vietnam) was likely recharged during the Pleistocene and preserved by the Holocene clay cap. Despite being in a humid climate, recharge to the deeper Mekong Delta groundwater system is very limited and freshwater volumes are still declining naturally.

Moreover, the paper as a whole, but the Discussion section in particular, is a bit of a confusing mix of a number of problems. There is (A) the scientific problem of understanding the evolution of the groundwater salinity distribution in large delta systems over long timescales. This is mixed with (B) the local management problem/question of how much freshwater there is in the Nile delta. And then there is (C) the problem that previous models have assumed steady-state conditions. For a publication in HESS, the local management problem (B) is not the most important. The scientific problem (A) is, and it should be made clear from the onset and throughout the paper that this is the main focus. The implications for the local management problem can be mentioned toward the end (it is especially interesting to note that depending on the conceptualization, the locations where freshwater occurs will differ), but should not feature prominently anywhere else.

Following one of your suggestions in the specific comments (on P14L2), we discuss the scientific problem at the start of our discussion **[P15 L1-30]** and the local management issues to a later part of the discussion **[P16 L19-30]**. Following the suggestion in the first general comment, the scientific problem is now also discussed at the start in the introduction **[P2 L1-7]** . Furthermore, the start of the abstract is changed. This has aided a lot in changing the focus of this paper **[P1 L11-12]** .

With regards to the assumption of the system being in steady-state (C), a considerable portion of the paper is devoted to a comparison between dynamic (i.e., models considering the palaeo-geographical evolution) and steady-state conditions (equivalent simulations for the same geology). It is concluded that the dynamic models are a better fit to the data. But given that the data set has severe limitations (there are relatively few data points, and the quality of the available data is also not assured, at least the paper does not describe the QC/QA procedures), one could wonder if that is really a such strong criterion. I think a much stronger argument could be made by looking at the time required to reach steady state. This information is not presented for all model scenarios, but in line 28 on page 11 it is mentioned that it took 60 ka for the B-model scenarios. Doesn't this automatically invalidate the steady-state assumption, without having to perform anymore detailed comparisons between the dynamic models and their steady-state equivalents? I am not sure what time is required for the other simulations, but I am guessing it will be on the same order, except maybe for the most unrealistic representations of the delta's lithology. I would encourage the authors to present the timescale aspect in more detail and use it to build the argument against steady-state being a realistic assumption. Much of the detailed comparisons such as those presented in figures 9, 10 and 11 could then be omitted.

This would also have the benefit that the paper becomes more easy to follow, because as it stands, one quickly gets lost in the many different scenarios.

We think the suggestion to discuss the time to the steady state is good. This, in combination with changing the order in which we discuss the results and discussion, will improve the readability of the paper. Omitting much of the detailed comparisons though, would negate one of the strong points of this study though, namely that we ran a lot of scenarios for a complex 3D model. These scenarios provide a lot of unique information that is relevant to our discussion, as it makes this study more applicable to other deltas. So here we don't fully agree. Still, we have removed what was Fig. 10 in the initial submission, as we barely discussed it in the text. Furthermore, we removed Fig. 5 and accompanying text as the data is limited.

We have added the time to reach a steady state to "3.7 model evaluation", as this section appeared to be the most fitting. We added the following lines to the methods section **[P11 L10-12]:**

To assess the validity of the steady-state assumption, we checked for all equivalent steady-state models the time they reached a steady state. This time was determined by calculating the derivative of the fresh water volume over time. If this did not change more than 1E-04% of the total volume, we considered the model to have reached a steady state.

To the results section we added **[P13 L4-8]:**

All equivalent steady-state scenarios required at least several thousands of years to reach a steady state (Table 4) from an initial Pleistocene steady state. Most notable are the B-scenarios, where the hypersaline groundwater caused the system to respond very slowly, over tens of thousands of years, thus exceeding the extent of the Holocene. The shortest scenarios were the N-T scenarios, as they did not include HGw and due to the lack of clay layers the salt water did not experience any resistance during its flow upwards from its initial Pleistocene state to the Holocene steady state.

To the discussion we added **[P15 L6-9]:**
Our equivalent steady-state scenarios required at least 5500 years to reach a steady state (Table 4), a period in which already considerable changes occurred to the boundary conditions (Fig. 3). This increased to tens of thousands of years for the more complex models. We therefore doubt that using a steady-state approach with current boundary conditions results in a reasonable estimate of the current fresh-salt groundwater distribution for such a complex, large-scale system.

To the conclusion we added **[P18 L2-L7]**:
It was found that large timescales are involved, as steady-state model scenarios required at least 5500 years to reach equilibrium. Hence, none of the evaluated paleohydrogeological scenarios reached a steady state over the last 9000 years, meaning that the transient boundary conditions definitely had an influence on current groundwater salinity. Given the large range variation in delta-architectures analyses, we can conclude that steady-state models are not likely to result in realistic FGw distributions in deltaic areas.

To the abstract **[P1 L24-L25]**::
Furthermore, the time required to reach a steady state under current boundary conditions exceeded 5500 years for all scenarios.

The Results section
could start with what is now subsection 4.2, which could be expanded and/or merged with subsection 4.4. This would give the reader a much better overview of the actual processes before diving into the more detailed analysis of model performance and freshwater volume.

We followed this suggestion as follows: We have started the results sections with a statement that for readability we first discuss the results of a few scenarios that were selected based on a later discussed model performance assessment **[P12 L2-L5]**. Then we start with the former section 4.2 **[P12 L6]**, which describes the spatial distribution. We continue with the former section 4.4, describing the contribution of several boundary conditions through time **[P12 L14]**. After that, we discuss the model performance **[P12 L23]**.

So the outline changed to:

4 Results

Based on a comparison with observations, five scenarios were selected that showed the best match with the observations. These are called the "acceptable" scenarios. To keep the results comprehensible, we start with discussing the results of these five selected scenarios through space (section 4.1) and time (section 4.2), before discussing this actual selection procedure (section 4.3).

4.1 Current spatial TDS distribution of acceptable model scenarios

4.2 Salt sources over time

4.3 Model evaluation

4.4 Freshwater volume dynamics and sensitivity analyis

4.5 Fresh-salt distribution: the palaeohydrogeological reconstruction against its equivalent steady-state

Specific comments

Page 1

lines 15-16, "observed by hydrogeochemists": No need to suggest a disciplinary bias

We have removed this in **[P1 L10-L14]**.

line 17, "palaeo-reconstruction": This suggests your reconstruction itself is ancient. Choose better wording

This is indeed confusing. We changed this to "palaeohydrogeological reconstruction" (just as there are for example palaeoclimate reconstructions) throughout the complete document.

line 18: Insert "a" between "using" and "state-of-the-art model"

We have corrected this in **[P1 L17]**

line 23: You use both "sea water" and "seawater" in your manuscript. Pick one, and check for consistency.

We were indeed inconsistent and moreover grammatically incorrect in the original text. However, according to the Cambridge Dictionary “saltwater” should be used as adjective (saltwater intrusion) and “salt water” as a noun.

Noun:
https://dictionary.cambridge.org/dictionary/english/salt-water
Adjective:
https://dictionary.cambridge.org/dictionary/english/saltwater

We assumed here that sea water and seawater should be used the same as saltwater and salt water, and hence changed “sea water intrusion” to “seawater intrusion” and use “sea water” as a noun. Thus, to stay grammatically correct, we have to stay inconsistent. The same holds for “fresh water” and “freshwater”.

line 29: add "s" to "distribution"

We corrected this in **[P1 L29]**.

Page 2

line 6: insert "that they were" between "indicated" and "pumping". More generally, the language usage needs some attention, the paper is generally well written, but every now and then it lacks some attention to detail. I will not focus on these issues from this point onward, but the authors should do a careful proofread of their revised manuscript to resolve them.

We have corrected this specific point **[P4 L7]** and conducted extra proofreading (by a native speaker). We hope this improves our attention to detail sufficiently.

Page 4

line 31: delete "hypersaline and", the fact that they are hypersaline should not be a reason to discard them. On the next page you talk about hypersaline groundwaters as well (or at least, salinities greater than seawater)

This is indeed confusing, we have deleted these two words in **[P5 L9-10]**.

Page 5

line 1: I am not sure if this reasoning holds true. A 1000y old groundwater can still move appreciable distances over a couple of decades if it is near a large well field

Good point, our reasoning is flawed here. We have added the different measurement dates as an extra source of the spatial variation in observed salinities.

Thus changing the list of potential causes to **[P5 L11-13]**:

This variability in measured salinity can be explained with 1) the different measurement depths, 2) different dates of the measurement campaigns, 3) heterogeneity in the hydraulic conductivity of the subsoil resulting in heterogeneous salt transport, and 4) heterogeneous evapoconcentration.

line 26: Some additional information is required here to explain the choices and rationale behind these 9 different lithological models

We have extended this section in **[P6 L11-12].**

The height of the clayey sediments determines how disconnected the deeper groundwater system is from the sea, and thus the ability of the system to preserve denser hypersaline groundwater in its aquifers (van Engelen et al., 2018). The hydraulic conductivity of the onshore-reaching clay layers is varied to get a first-order approximation of the effect of clay layers on regional groundwater flow. We assigned a continuous hydraulic conductivity to these clay layers, based on three different lithologies (in order of decreasing hydraulic conductivity): sand, fluvial clay and marine clay (Table 3). The rationale behind this is that small clay lenses have negligible effect on regional groundwater flow, thus are assigned a hydraulic conductivity of sand. Fluvial clay layers are assigned a hydraulic conductivity of the current confining Holocene clay layer, as this was deposited under fluvial conditions (Pennington et al., 2017). Marine clay layers present continuous layers of low conductive with a big influence on the regional groundwater flow and thus have the lowest hydraulic conductivity.

Page 6

line 5: Replace "its'" with "its"

We corrected this in line **[P6 L24]**

lines 10, 11: Not sure what you mean by "keeping memory locally".

We meant here that each subdomain is assigned its own private memory at the computational node, instead of fitting all subdomains into one big shared memory. We changed the sentence, to clear this up. **[P6 L29-30]**

using the Message Passing Interface (The MPI Forum, 1993) to exchange data between subdomains, where each subdomain has its own private memory assigned at a computational node.

line 15 a.f.: You need to include a map with the model area.

We have added this map as Appendix A1. **[P40 L1-5]**

lines 23, 24, "as the hills above this height have no important contribution to the groundwater flow": Without the aforementioned map it is hard to assess the validity of this statement. Where are these hills? How high are they? More importantly, what is the recharge and the water table elevation. The elevation of the hill itself is not so important, its hydrogeological characteristics much more so, of course...

This was indeed an offhand remark. We do not have actual groundwater tables of this specific area, but we know that the amount of rainfall is very low and that there is no surface water located there. This is supported by Geirnaert and Laeven (1992), who found that shallow groundwater in these locations is >5000 years, meaning very limited recent recharge. We assumed that this leads to groundwater tables well below surface level such that loss through evaporation is negligible.

We added extended this sentence to provide support **[P7 L12-14]**:

The top of the NDA was clipped off above 20 m AMSL, as the hills above this height have no important contribution to the groundwater flow (Geirnaert and Laeven, 1992), because there is very limited rainfall in the south and no surface water here.

Page 7

line 3, "time slices": This sounds like what we would normally call a stress period in groundwater modelling. Why the confusing terminology? And why with a space in the title, and without in this sentence. Please pay more attention to detail.

This is indeed what we call a stress period in groundwater modelling. Still, we decided to stick with the term "time slices", following Delsman et al. (2014). "Stress period" makes sense from a modelling perspective, but can be confusing for non-modellers (like paleogeographers, hydrogeochemists), hence why we prefer this term. We added a sentence of explanation in **[P7 L23-25]**

A time slice is also known as "stress period" by groundwater modellers (Harbaugh, Arlen, 2005)

line 12: Replace "announced" with "at"

We corrected this. **[P7 L32]**

Next, time slice 3 covered the marine transgression, which was a period of rapid sea-level rise at the start of the Holocene

line 22, "100 day resistance" (better would be "a resistance of 100 days" BTW): How sensitive are the modelling results to this assumption? A single value is of course highly unrealistic, and citing studies from the Rhine-Meuse Delta does not provide any justification, because this parameter is just as uncertain and spatially (and even temporally at this timescale) variable there. But you've got to work with what you have, I understand that, but in the end, some sensitivity analysis is required to test to what extent the study outcomes might be affected by this modelling choice.

We have conducted a local sensitivity analysis of this resistance on one scenario. We initially wanted to vary the resistance with a factor 10 or 5 but this led to numerical convergence issues for the scenarios with a decreased resistance. Therefore, we stayed at a factor 2. To not repeat ourselves in this answer, the text we added as Appendix B should explain the rest **[P18 L26]**:

To assess the effects of our assumption of the boundary resistance, we ran two alternative versions of the "H-N-T-P" scenario with a different resistance. This scenario was chosen as it presumably is the "acceptable" scenario that would be affected the most by the resistance value, as it has no horizontal clay layers that resist changes in boundary conditions and the sea boundary has the most open connection with the sea. We multiplied the resistance with a factor 0.5 and 2. Lowering the resistance more than with a factor 0.5 lead to numerical convergence issues. Fig. B1 shows that throughout the Pleistocene the resistance influences the groundwater types, as the lower resistance allows more river water to be replaced with hypersaline groundwater. The groundwater types of the different models quickly converge, however, through the Holocene. We therefore think that the choice of boundary resistance has limited effect on our results and conclusions, despite that we only varied the resistance to a limited extent in this sensitivity analysis.

line 31: replace "acquiring" with "achieving"

Corrected in **[P8 L20]**.

Page 8

lines 3-6: Did you do this via SEAWAT's density options in the CHD package?
A range of 2-18 g TDS/L gives quite a range in density. What value did you adopt? And again, how sensitive is it?

We used the RIV package for the lagoons. We estimated the mean salinities from the Flaux et al. (2013) *A 7500-year strontium isotope record from the northwestern Nile delta (Maryut lagoon, Egypt)* for each time slice. Of the total of seven time slices, the last four have lagoons, since lagoons started to form from 7.5 ka. To be specific, the salinities assigned to time slices 3 to 7 are respectively: 18, 9, 4.5, 2.5 g TDS/l. Despite these decreasing salinities, we can still observe quite high salinities at the (former) locations in lagoons that fall in between 5-15 g TDS/l (see Figure 4), so past lagoonal salinities are still locally present in our model.
We do not, however, think that different salinities will have a large effect on our conclusions. Since the mixing zones are relatively small, changing the lagoonal salinity will only have a small effect on the fresh groundwater volumes, as long as this salinity does not get lower than 1 g TDS/l. This is presumably more controlled by the location of the lagoons. Which model scenarios were deemed "acceptable" was in the end mainly assessed based on the location of hypersaline groundwater, so this will also not change strongly, since the density of HGw is much higher (up to 120 g/l) than that of the lagoons (up to 18 g/l). This brackish groundwater will exert limited force on the HGw.

We added the following text that is more specific here **[P8 L23-27]**:

Of the total of seven time slices, the last four have lagoons, since lagoons started to form from 7.5 ka. The lagoon stage was set such that it was in hydrostatic equilibrium with the sea, so that its pressure is corrected for salinity (Post et al., 2007). Lagoonal palaeo-salinities were estimated from the published strontium isotope ratios from the Maryut lagoon (Flaux et al., 2013) for each time slice. To be specific, the salinities assigned to time slices 3 to 7 are respectively: 18, 9, 4.5, 2.5 g TDS/l. These salinities show strong variation through time, because the inflow of the Nile varied through time.

Page 10

line 7: You could cite the following article here: Sanford, W.E. & Pope, J.P.
Hydrogeol J (2010) 18: 73. https://doi.org/10.1007/s10040-009-0513-4

This is a very fitting suggestion, thanks. Also relevant for the discussion.

We added the reference to method section 3.7 **[P10 L29]** and added an extra sentence to the discussion **[P16 L1-2]**:

Similar conclusions were drawn by Sanford and Pope (2010) for the Eastern Shore of Virginia (USA), an area with a similar observation density.

line 10 a.f.:
This is somewhat hard to follow and it might be worth adding a sketch that illustrates

the principle (could be in a supplementary document). Other authors may choose to adopt this methodology, hence it could be worthwhile doing this.

We have added a sketch as Fig A2 **[P41 L1-5]**.

Page 11

line 19: what do you mean with "behavioural"? And what is the justification for using 0.07?

With the term “behavioural” we mean the ability of models to reproduce certain patterns observed, following Beven and Binley (1992) *The future of distributed models: Model calibration and uncertainty prediction.* Whatever these patterns specifically are, is up to the researcher to decide. So, on second thought, perhaps the word “acceptable” captures the inherent arbitrariness of this decision better. The value of 0.07 was based on a visual inspection of the figure, as around that value there is a separation visible. The scenarios with $\mathrm{Md}|\Lambda| < 0.07$, predict the location of the HGw with similar skill as to with which they predict the location of saline groundwater. These scenarios we call “acceptable” .

In the text **[P12 L29 – P13 L3]**:

More striking differences are observed in the hypersaline zone, where we observe a division around $\Lambda = 0.07$ into two groups. There are the scenarios with $\mathrm{Md}|\Lambda| < 0.07$, that predict the location of the HGw with similar skill as to with which they predict the location of saline groundwater. We call these scenarios “acceptable”. Specifically, these are the following five model scenarios: C-M-B-P, C-N-T-P, H-M-T-P, H-F-T-P, H-N-T-P. The other scenarios perform considerably worse in predicting the location of the HGw.

lines 19-21: Sentence does not seem to flow well due to a grammar error, not sure what you are trying to say here.

This should be clear now **[P12 L29 – P13 L3]**.

Page 12

line 3: Not clear what you mean here with "behavioural"

Hopefully cleared up now with the changes made to the description of the model evaluation **[P12 L29 – P13 L3]**. See response comments of P11L19.

line 4: Up until this point the difference between fluvial and marine clay layers has not been explicitly discussed.

Now expanded on in section 3.1 **[P6 L11-12]**. So see answer to comment for P5L26.

lines 7, 8: replace nondescript terms like "more 3D patterns are visible in the salinities" and "partly has a conical" with more accurate descriptions

We are more specific here now **[P12 L10-13]**

Regardless of the differences between scenarios, in all realizations the fresh-salt interface roughly follows the coastline, except in the west where there is far extending seawater intrusion visible towards Wadi El Natrun (Fig. 1). Next to this depression, (former) lagoons are visible as shallow brackish zones and (former) dune areas are visible as freshwater lenses.

lines 24 a.f., "This table also shows...": I could not follow what you are trying to say here

This sentence could be more to the point. Changed the sentence to **[P13 L22-L24]**:

This table also shows that these parts of the model are also the most uncertain, since disregarding potential deep and offshore fresh groundwater volumes decreases the spread from 74% to 32%.

Page 14

line 2 a.f.: I think the point you want to make here is that you can come up with multiple models that fit the observations equally well and yet, the volume of freshwater varies a lot. What do you mean with "The variance in the results should also affect management decisions."? I think the management decisions will not be based on the total volume of freshwater, but on the possibility to be able to extract groundwater in a particular region. See also general remark made before about the relevance of the freshwater volume issue for this paper. I would not start the Discussion section with this paragraph (see next point)

We have removed these sentences now, since it shifts the focus too much to the regional problem. We tried to say that policy makers and managers should take into mind the large uncertainty of groundwater models in these complex aquifers, as shown in this research, but since we do not provide any suggestion to do so this is a bit of a weak statement.

However, even though it has no effect on the changes made to this paper, we would like to express our different view on how management decisions are made in the Nile Delta though. Currently, there is an ongoing political discussion between Egypt and the upstream countries on the effects of the large dams that are being built. Discharge of the Nile will decrease and thus the volume of fresh groundwater that can serve as a “strategic stock” during periods of low-flow is becoming of increasing interest, from the perspective of delta-scale management.

Page 15

line 6: Start the Discussion section with this paragraph, it is much more relevant to a broad readership than local aspects such as the discussion of the total freshwater volume (also see general comment) and flow to Wadi El Natrun.

Good suggestion, we have brought this paragraph to the start of the Discussion section. **[P15 L2-30]**

line 32: Also include a discussion of the representation of free convection phenomena in your model. The large grid size is prohibitive for an accurate process representation. How confident are you that this does not harm the general conclusions drawn from the model outcomes?

We are confident that the general conclusions are not harmed, since we have investigated the errors made in modelling with a crude resolution for the Nile Delta in a previous paper. This was published as an appendix, so easily overlooked. We have added a deliberation on this in the discussion of this paper in lines **[P17 L10-18]**. For your convenience we have added this appendix of our previous paper as supplementary information to this response to authors.

To the discussion we added:

In addition, for a proper physical representation of free convection, a finer grid is required. A coarse horizontal cell size results in a delay in the onset of free convection, while a coarse vertical cell size results in an onset of free convection even for situations that are expected to be stable (Kooi et al., 2000). van Engelen et al. (2018, Appendix D) investigated the errors caused by coarse model cells for the Nile Delta and found that especially the crude horizontal grid size had an influence. They found that this resulted in similar downward fluxes, but a delay in the onset of free convection. This effect, however, was negligible after ~50 years and thus dwarfed by the timescale of our time slices. The coarse vertical grid size was not an issue, since the marine transgression occurred over sand with a high hydraulic conductivity, meaning there is a very instable situation and free convection has to occur. We thus think that the errors made in modelling free convection do not influence our conclusions.

Figure 1: It could just be because of the pdf, but the resolution is very poor. Figure needs an inset showing the location of the area within Egypt/the Mediterranean region. Add north arrow and scale bar! In this figure all areas outside the delta are white, better to make the Mediterranean blue and the desert yellow(-ish).

Our maps indeed lacked these crucial map features, so therefore we added them to this figure **[P30 L1-2]**. The resolution presumably deteriorated in converting the vector file first to a .png to add in MS Word and consequently to a pdf. We plan on submitting the original vector files as pdfs for the final version, which should solve this.

Figure 2: Somewhere we will need a map with the model boundary. On this map you will need to indicate which part of the model is shown here.

We added a dedicated figure to this (Fig. A1), since adding these features to Fig. 1 resulted in a too complicated, cluttered map.

Figures 4, 5: Take the reader by the hand here as the figures are complex. Explicitly mention in the caption that a, b, c and d reflect the salinity classes, and explain the n value.

We changed the caption to **[P35 L1-8]**:

Figure 6: Goodness of fit boxplots for all palaeohydrogeological reconstructions, binned into four salinity classes. The higher the value of Λ, the worse the fit. Codes indicate model scenarios (Table 1). TDS values are binned in the classes [0, 1], [1, 5], [5, 35], [35, 100] g TDS/L for respectively “fresh”, “brackish”, “saline”, “hypersaline”, these are respectively plotted in panels A, B, C, D. “n” indicates the amount of observations available in the TDS bin to evaluate model results with. Diamonds indicate outliers, defined as values separated from the first or third quartile at 1.5 times the interquartile range. When no box is plotted for a scenario, 75% of the measurements equal zero, which consequently causes the interquartile range to be zero, rendering every non-zero value an outlier.

Supplementary material

Effects grid size on free convection

Context:
In this Appendix we conducted a grid convergence test for a simple sandbox model. “NDA-f model” means “Nile Delta Aquifer free convection model”. There was also a model without free convection, and for this model we conducted a grid convergence test on the original model. Therefore, the text below starts with the words “different approach”. In addition, we removed all less relevant parameters to this research from Table 1.

The text below is taken from Appendix D of :
van Engelen et al. (2018) ***On the origins of hypersaline groundwater in the Nile Delta aquifer***

For the NDA-f model, a different approach was taken, to assess the effect of a wider range of Δx, since this dimension had the coarsest resolution (1 km). We modeled a simple 800 m thick sandbox with a hypersaline lake on top. This allowed us to stretch the domain as much as necessary in the horizontal direction. The chosen parameters for this model are the same as in Table 1. The amount of cells was kept at 100 in the horizontal direction. Ξ is the normalized fluid mass increase, which is calculated by dividing the increase in fluid mass across the top boundary by the maximum mass storage increase:

$$\text{(D.1)} \qquad \Xi = \frac{\iint \frac{Q}{\Delta x}\rho \,\mathrm{d}x\mathrm{d}t}{V_{tot}\left((\rho_{max}-\rho_0)n+\frac{1}{2}S_f\frac{\rho_{max}-\rho_0}{\rho_0}z_{max}\rho_{max}\right)}$$

where Q is the volumetric flux through a boundary cell (m^2/d), V_{tot} is the total volume of the model domain (m^2), ρ_{max} is the concentration at the top boundary which was set at 1078 kg/m^3, ρ_0 was the initial density of the groundwater in the domain (1000 kg/m^3), S_f is the specific storage in terms of fresh water head (d^{-1}), z_{max} the depth of the domain (m), and n the porosity (-). For the test with Δx, Δz was kept at 10 m; in testing Δz, Δx was kept at 10 m. As the density at the top boundary of the model was perturbed randomly, 10 simulations were started for each discretization, of which the minimum, maximum and mean of Ξ are plotted. It can be seen in figure D.2 that there is some spread in Ξ, which is caused both by integration errors and differences between realizations. Furthermore, a larger Δx causes a delay in the onset of free convection and causes a slower downward movement of fingers. However, all errors are unimportant on timescales larger than ~100 years, and thus also on our timescales of interest, which is over 1000 years.

Table 1. Parameters for the NDA-f model. We are purposely inconsistent with units, as it allows for easier comparison with values found in the literature.

Parameter	**Description**	**Value**	**Unit**
$K_{h,\,sand}$	Horizontal hydraulic conductivity sand. Compaction case and free convection case	75	m/d
$\frac{K_h}{K_v}$	Anisotropy	10	-
α_l, α_t, α_v	Longitudinal, transversal and vertical dispersion length	10, 1, 0.1	m
n_e	Effective porosity	0.10	-
$\frac{\Delta h_{riv}}{\Delta x}$	River gradient	9.375e-2	m/km
S_f	Specific storage in terms of fresh water head	1e-5	1/m
$\frac{\partial \rho}{\partial C}$	Slope linear equation of state	0.71	(kg/m^3)/(g/l)

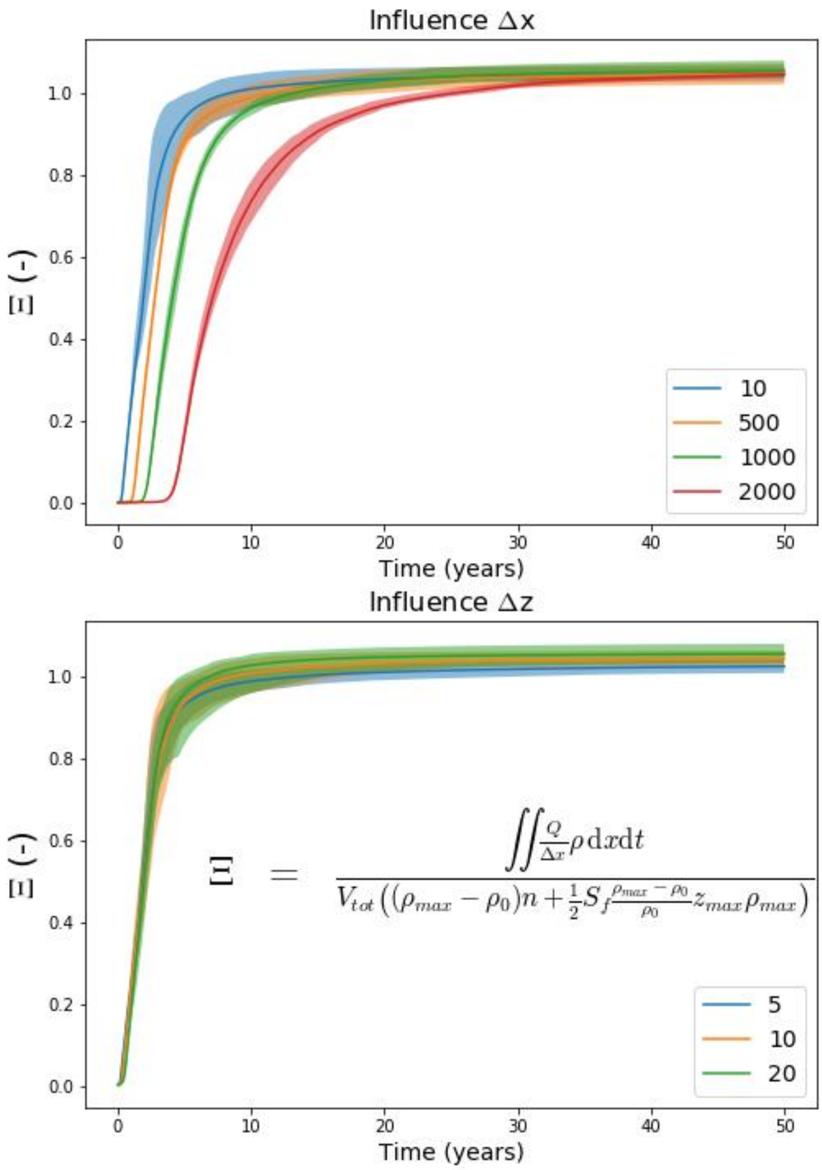

Figure D.2: Results of the test of the errors that are introduced in the free convection model by the large grid size: a. shows the influence of Δx and b. shows the influence of Δz.