The below documents the major changes to the manuscript which were implemented as a direct result of reviewer feedback. For some of the below responses, the additional or modified text is directly reproduced here. However, several reviewer comments below, as well as the annotated PDF responses, are only viewable in the latex-diff document. As well, there are additional changes beyond the below, largely for clarity of text, which are viewable in the provided latex-diff document.

Reviewer 1

“This is one of my biggest concerns with this manuscript: Why is this needed? According to lines 104/105 NMSS and Seakon-SS are equivalent.”
&
“Coming back to my point above (at lines 127/128): According to lines 104/105, the NM and Seakon model outputs are equivalent for SS structures, so why is there a difference?”
On lines 104/105 the manuscript states “ When lateral variations are not included, Seakon model output is equivalent to the NMSS model (assuming parameter values corresponding to the 3-layer viscosity structure are the same).”

This statement is a simplification which may be the source of confusion for the reviewer. When Seakon is used in the spherically symmetric (SS) configuration, it is functionally equivalent to the NMSS numerical model. Given the significant differences in numerical methods, and subtleties such as the differences in time-stepping and grid-discretization, the model output is not bit-for-bit identical as some readers may interpret the statement on lines 104/105. As such, to avoid introducing any potential source of structural error that may arise from these differences, and also for the ease and expediency of data processing during the investigation, we elected to use Seakon in the SS configuration rather than rely on it solely for the calculations which incorporate lateral variability. We will incorporate a brief summary of the above (as well as the comments made to the other reviewer with respect to this subject) into the text and remove the statement on lines 104/105 to avoid this potential source of confusion.

Implementation Action: None taken as with the statement removed (as discussed above and in other comments by Reviewer #2)

'My concern is mainly that ROC of RSL is not used when comparing to observations. ROC of the radial displacement is used, so there it is okay to use the ROC, but I would not use it for RSL.'
The reviewer indicates that they have an issue with the use of the rate of change (ROC, i.e., the first derivative with respect to time) of relative sea level (RSL) to produce predictions of RSL itself. The reasons for this choice are twofold: Firstly we can easily convert between the two due to the definition of RSL, and secondly, we found reduced misfits using this

approach. With respect to the conversion aspect, the reason we can use the ROC RSL in lieu of RSL itself in the emulator is that, by definition of RSL, we always know that RSL at present day is equal to zero. I.e., RSL(t=0) = 0. As such, it is a simple matter to integrate the provided ROC of RSL from either the emulator, or actual RSL data itself, to reproduce the RSL timeseries. With respect to the second part, early sensitivity tests revealed nearly an order-of-magnitude reduction in the prediction misfits from the neural networks when using the rate-of-change rather than trying to predict the RSL value directly. As such, we chose to incorporate the ROC RSL into the emulator rather than the direct prediction.
As well, part of the goals of the emulator is to eventually be able to provide input to other numerical models, in which case the ROC of the field is more useful for inclusion into numerical solvers than an absolute value. We will incorporate a brief summary of the above into the manuscript.

Implementation Action:
Text now reads: "Given RSL and RAD are both defined to be zero at present day, we can readily recover the full timeseries of the field by integrating the ROC from present day."

"You discuss the ice model effect of the trained dataset at the end, but how much is the trained dataset depending on the tomography and lithosphere model? Please discuss this as well."
&
"Why these? Why 2/3/5/7 factors of 9? What about 36 and 54? You could test it with N being anything between 1 and 330. Why not testing it for values between 27 and 63 with steps of 1? Of course, this is is time consuming, but could identify where the misfit line deceases most and reaches a "plateau". How is this number also ice, tomography and lithosphere model depending?"

These values are a result of the initial experimental setup. With the limited compute resources available at the outset of the project we budgeted ~100 3D and SS model parameter vectors (PVs) to explore with Seakon and a limited number of Artificial Neural Networks (ANNs) to train. The subsequent numbers of PVs incorporated into the training datasets of the ANNs were a result of combining two sets of PVs: a core set of PVs and the supplemental PV sets (which have increasing numbers of PVs). The core PVs consist of the 8 most extreme PVs from this initial sampling of the LT/UMV/LMV space (i.e. minimum and maximum of each of the 3 dimensions, resulting in 23 PVs) as well as the center-most PV. The supplemental sets had 9, 18, 36, and 54 members. The lowest bound of 9 PVs was used to determine if we could accurately represent the 3D-SS difference as a function of LT/UMV/LMV with only the core PV set and a single intermediate PV between each of these extreme PVs. From there we doubled the supplemental PV set to 18 and subsequently 36. Finally, the set of 54 was chosen to be half-way between a subsequent doubling and not evaluating the impact of adding another supplemental PV set. The reasoning was that requiring so many members in the training set would indicate this fast-surrogate approach was not sufficiently resource-effective.
This latter reasoning is also why we do not test ANNs with >63 (i.e. 9+54) PVs included in

the training dataset. For the purposes of the examples presented, requiring the same order-of-magnitude of computational resources to train the neural networks would indicate that the method employed in this study is ineffective.

As for training and subsequent analysis within the values of 27 and 63 with steps of 1, such an undertaking would be an order of magnitude more expensive to train than what is presented in the investigation. We could potentially add several more intermediate steps, but preliminary work with other tomography models indicates that there is a relationship between the complexity of the 3D Earth structure and the quality of the ANN predictions for a given training dataset size. As such, we argue this scale of investigation would be better suited to a study which explores this approach more fully (i.e., one which varies not only the SS Earth structure, but also ice and tomography). We did do some preliminary investigation work exploring a simpler 3D Earth model, one which uses the same tomography but assumes a spherically symmetric elastic lithosphere, and found that for the same number of PVs included in the training dataset we obtained reduced ANN:model misfits.
We will condense the above reasoning regarding the sizes of the PV datasets and add additional context around line 210. As well, we will include some details regarding the preliminary results of the impact of different 3D Earth structures.

Implementation Action: Text now reads:

"Additional supplemental PVs were added to this baseline set, with 9, 18, 36, and 54 members each. The lowest bound of 9 PVs was used to determine if we could accurately represent the 3D-SS difference as a function of LT/UMV/LMV with only the core PV set and a single intermediate PV between each of these extreme PVs. From there we doubled the supplemental PV set to 18 and subsequently 36. Finally, the set of 54 was chosen to be half-way between a subsequent doubling and not evaluating the impact of adding another supplemental PV set. The reasoning being that requiring so many members in the training set would indicate this fast-surrogate approach was not sufficiently resource-effective. Combining the core and supplemental sets resulted in 4 trained networks with $\text{N}=18$, 27, 45, and 63 members, respectively."

As well some discussion of an alternative fully 3D Earth model is now incorporated

---

"Why do you use the NMSS model output here? I understand that the goal is to use it in the future, but here you describe the method and the most obvious would be to compare the emulated 3D GIA-SS on SS Seakon to 3D GIA Seakon as these would show you the real differences between emulated and calculated 3D models."

The choice of using the NMSS results is largely a result of being the eventual workflow with this tool, as pointed out by the reviewer. Previous iterations of the manuscript used the approach (i.e. using Seakon data only) the reviewer specified, but it made no appreciable difference to our results or conclusions. We will clarify this point in the manuscript.
Implementation Action:

Text now reads:

“... and the NMSS model combined with explicit 3D-SS Seakon output. The choice of combining NMSS model output with the emulated 3D-SS output is made to reflect the end-goal of this workflow, and we note that the SS Seakon output is, for the purposes of calculating the misfit, equivalent to that of the NMSS model. The results of these calculations…”

“A calculation of NMSS+emulated 3D-SS GIA Seakon is okay to have here, but the purpose of doing this is for future application studies. It would be first interesting to see what the differences are between emulated and calculated Seakon models.”

Given the functional interchangeability of the Seakon SS and NMSS output, the differences the reviewer requests are already available in the supplemental materials by comparing the bottom two rows of Figures S10 through to S12 or by comparing Figure S7 to S8.

Implementation Action: None to take

“I do not understand why you calculate the emulator:NMSS. For what? To find out the difference between 1D and 3D? But this is not the aim of this paper. It can be a side result, but shouldn't cover that much in the manuscript.”

Calculating the mean-square-error between the emulator and the NMSS is done to demonstrate that, while there are still non-negligible misfits between the emulator and the explicit 3D data, the output of the emulator are significantly more like the 3D output than the NMSS output. We will re-evaluate the amount of discussion dedicated to this comparison.

Implementation Action: Text already motivates the emulator:NMSS MSE, and represents less than 5% of the manuscript line count. We feel this is acceptable given it provides a quantitative measure of emulator performance.

Reviewer 2

Main comments
1. The first application of neural network to emulate a GIA model as far as I know is Lin et al. (2023), which has been published since September 2023 and presented at AGU Fall meeting in 2022. That does not take away any of the scientific value of the current manuscript but Lin et al. (2023) should be discussed in some detail, e.g. , in terms of type of neural network selected, time steps, loss function, but also because it uses many different ice histories which the current manuscript states as important future work (of course Lin et al did not use 3D GIA models).

We were aware Yucheng Lin and colleagues at Durham University were working on the application of Machine Learning to glacial isostatic adjustment (GIA) modelling. In fact, we had some discussions with these colleagues to share ideas and make sure our contributions would be complementary. However, at the time we submitted of our manuscript to GMD, we were not aware that the Lin et al. paper had been published, hence the reason it is not mentioned. We will revise our manuscript to mention the Lin et al study in the Introduction and describe how it is complementary to our work. The different goals of the Lin et al. study to ours is also reflected in the type of emulator used (graphical vs non-graphical) as well as other methodological aspects that will be noted in the Introduction. Finally, we will compare our emulation results to those of Lin et al. where appropriate in Section 3.1 (Results).

Implementation Action: Text now reads:
" … limiting factor in exploring the parameter space (Tarasov et al., 2012; Sellevold and Vizcaino, 2021; Williams et al., 2023). A recent study (Lin et al., 2023) applied a graph-based spherical convolutional neural network algorithm to the GIA problem. Their focus was on emulating RSL on a single 1D Earth model for a wide range of ice history models, and so very different to the aims of this study. Given the high computational efficiency of the 1D GIA model, a relatively large training set of 1200 simulations was used in their analysis. In general, good results were obtained indicating the potential utility of ANN methods in GIA applications. A primary challenge for the 3D case considered here is the much reduced computational efficiency, limiting the number of simulations that can be performed to generate a training set."

Among other related textual changes noted in the latex-diff document.

---

2. 335-355: this part of the paper is separate from the main goal of the paper and the emulator does not seem necessary for it. I think it can be removed without any loss for the main objective. If the authors have a strong wish to keep this part, the results and conclusions need to be placed in context with a long list of earlier studies that have reached similar conclusions as in line 338, 348 and 353. The extra research goal should be introduced in more detail compared to what is now in line 70-72, including previous work and what the current paper adds to it, and the conclusion in line 389 should also state that the conclusions are in agreement with many earlier Studies.

We agree that this aspect of the analysis is secondary to the primary aims. However, a key motivation for better searching the 3D Earth model parameter space is to (eventually) demonstrate that 3D models do produce improved fits compared to the 1D Earth models. This is the rationale for including this component of the analysis. Although the results are disappointing, in that the more thorough exploration of parameter space did not result in markedly improved fits relative to the 1D case, we prefer to keep this aspect of our study. Therefore, the text will be revised by: (1) expanding the Introduction to make this research goal more explicit and to provide a short review of past work that has compared data-model fits with 3D and 1D Earth models; and (2) referring to past work to ensure appropriate credit is given in Section 3.2 when discussing the results. Prior to submission of the revised manuscript, we will inform Dr. van der Wal of the detailed edits to ensure no important studies have been Overlooked.

Implementation Action: Text has been updated as per repsonse-to-reviewers document, textual changes are noted in the latex-diff.

---

3. I found the description of the method sometimes lacking in detail. As the first application of neural networks to a 3D GIA model the it will be followed up by other studies. For that reason it is important know what has been tried and why certain choices are made. Especially I would say in a journal such as GMD. Below are specific comments. I think none of them requires extra modelling.

We agree that more detail on these aspects is warranted and so will significantly expand the Methods section (particularly 2.2 and 2.3). More details on these revisions are outlined below.
Implementation Action: Addressed via additional edits discussed below

---

104: this difference should be quantified because it is a source of error that will contribute to the difference between emulated and explicit results, since the emulation is based on the 3D model but the result is added to the NMSS model.

In fact, we chose to calculate the 3D minus 1D signal using the Seakon code only (run in 1D and 3D configurations) to avoid any potential errors due to differences between these two codes. This choice meant a greater computational requirement of generating output for 330 (1D) model parameter sets using the Seakon code instead of the NMSS code. However, it negates the need to benchmark the NMSS and Seakon codes. We will clarify this aspect of the analysis in Sections 2.1 and 2.2 and remove the sentence “When lateral variations are not included, Seakon model output is equivalent to the NMSS model (assuming parameter values corresponding to the 3-layer viscosity structure are the same).” as it is not relevant to the analysis and led to confusion for both reviewers.

Implementation Action: Highlighted statement has been removed

---

134 and further: The text here is hard to grasp. Testing and comparison is not specific enough, and the proxy-data: model comparison is not explained yet. With neural networks it is good to be clear about what tests and validation are done. In line 232 it is now not clear what the validation subensemble is.

We have modified the specified section of text in the PDF for clarity. We will add text to extend the explanation with regards to the generation of the testing and validation datasets to further increase clarity.

Implementation Action: PDF changes have been implemented for clarity as well as other minor textual additions for clarity.

145-147: This is an important finding, and for future development of ANN it is very useful to know why RSL itself can not be accurately emulated, and whether this is worthy of further research or not. Other questions: How is the ROC computed, the difference between two consecutive timesteps? How is RSL later reconstructed from this, by integrating the emulated ROC over time? It would be helpful to provide an indication of some of the preliminary results.

We agree this is an interesting and important question worthy of further investigation. Our work does not address why the 3D-SS ROC of RSL is able to be more accurately emulated compared to 3D-SS RSL itself. We were disappointed in the results for RSL and so considered ROC of RSL and found the results to be significantly improved and, therefore, worthy of publication. As well, we note that there is no easy way to directly ensure that RSL(PD) = 0 as per the definition of RSL, and therefor a data transformation, such as was used in the manuscript, is appropriate.

We will revisit our preliminary results that focused on RSL (rather than ROC RSL) and, if they are compatible with the current structure of the experiments described in the manuscript, we will incorporate additional information to at least document the improvement in going from RSL to ROC of RSL. Future work will expand on the results presented in this manuscript in several respects and we hope to explain the relative success of the ROC results as part of that analysis and potentially explore other functional transformations. With respect to how ROC was calculated, it is as the reviewer suggests and is simply the difference between RSL over consecutive timesteps divided by the time interval, though we do note that the size of the timestep varies throughout the glacial cycle (from 500 to 2000 years). The reviewer is also correct with regards to the reconstruction of RSL via integration.

Implementation Action: Initial investigations are not compatible with the current iteration of the manuscript. As such there are no text modifications made.

154: It is not clear to me how the probability density function is created, could you clarify this? Should this be seen as a histogram of all ROC for all timesteps or per timestep?

We will add additional details regarding the construction of the probability density function (PDF). The PDF can thought of as the histogram of the ROC of RSL across all variables and for all timesteps.

Implementation Action: Text now reads:
“...probability density function of the ROC of RSL \& RAD.
This probability density function is calculated by using the ROC of RAD \& RSL from files constructed for model performance evaluation (or training without any filtering). From each of

these files ROC of RAD \& RSL is extracted, then concatenated across all PVs, and then evaluated using the numpy histogram function.
This distribution is used …”

164: Can you provide some insight on why you selected this library?

The Tensorflow library/framework was chosen for several reasons. However, a primary one was the support for this specific library at the Digital Research Alliance of Canada (formerly ComputeCanada) platform where the training and model execution was conducted. The Keras application programming interface also made the implementation of the neural networks themselves simple (both from the perspective of the authors as ‘developers’ and readers as ‘users’). We expect that comparable results can be obtained from other libraries/frameworks such as PyTorch.

Implementation Action: Text now reads:
“The choice of framework/library was motivated by the available support at the high performance computing center where the network training was conducted, we anticipate comparable results can be achieved with other frameworks.”

165: “we train separately”. Doesn’t this result in a different ANN (different weights) for each ice history, viscosity, lithosphere thickness? I might miss something obvious but it would be good to clarify.

It results in a different set of weights for each 3D configuration (i.e., lithosphere model and seismic velocity model) so, in this study, three sets of weights were determined (for each of the 3D model configurations defined in Section 2.2). This will be mentioned explicitly in the text in Section 2.2.

Implementation Action: Added clarifying text, text now reads:
“... variable earth structure (i.e., S40RTS and S40RTS-LR18) and ice sheet history (i.e., ICE6G). As a result, each combination of LV and ice sheet history in this approach results in a separate set of ANN weights to be used with the emulator.
The inputs to the ANNs can be grouped into 4 aspects…”

169: Can you explain why 4 time steps is a good choice? It is not intuitive as the “memory” of GIA would go back further than that, and the paper later concludes that it could be a reason for the worse performance for present-day uplift rates.

As with the structure of the neural networks, this choice was the result of preliminary testing and evaluating trade-offs with respect to hardware limitations and quality of predictions. Generally, providing more previous timesteps to the networks resulted in reduced misfits, but there were swiftly diminishing returns and technical issues (largely due to memory and storage constraints on the hardware used for training) with adding significantly more past time-step data. Four previous timesteps were found to be a useful balance between model expense and useful predictions.

With regards to this being a potential source for poor performance regarding present-day uplift rates: this statement in the conclusions would benefit from additional clarity. While technically true, we would have to incorporate at least ~10,000 years of timesteps (~20) before providing multiple non-zero ice thickness changes between timesteps for many locations. This is beyond what we are able to evaluate with the hardware accessible to us, and would also still only provide a small subset of ice thickness data which is not constant. As such, a different approach for providing ice history for predictions of 3D-SS differences in present-day uplift rates needs to be explored in future work.
We will incorporate a brief summary the above with respect to the choice of 4-timesteps and rephrase the text in the Conclusions to increase clarity.

Implementation Action: Text now reads:
“... to emphasize regions with greater amplitude signals.
The choice of using the previous 4-time steps was motivated by preliminary testing and evaluating trade-offs with respect to hardware limitations and quality of predictions.
Generally, providing more previous timesteps to the networks resulted in reduced misfits, but there were swiftly diminishing returns and technical issues (largely due to memory and storage constraints on the hardware used for training) with adding significantly more past time-step data.
Four previous timesteps were found to be a useful balance between model expense and useful predictions.”

And

… respect to ice history (e.g., maximum ice thickness at that location within the last 10 ka). An important extension…”

174: Can you specify it here? This is now done in line 214. From my experience the choice of stopping condition can be important. Please explain if you have tried other stopping criteria, for example averaging only over locations with significant signal (and compare to Lin et al).

We will add details regarding the stopping criteria tested in the initial stages of the investigation. We did not investigate stopping criteria with a spatial dependence as the reviewer suggests. We will review the Lin et. al. manuscript for methodological comparisons on this aspect of the analysis and include relevant aspects in our discussion.

Implementation Action: Additional details have been provided, text now reads:
“...against the training dataset, is activated.
This early stopping condition for RSL is set such that, if there is no improvement of the MSE by at least $0.01\:(\text{mm}/{yr})^{-2}$ within 100 training epochs, the training stops and the set of weights which results in the lowest MSE is chosen.
This approach is used to prevent, or at least minimize, over-fitting of the trained ANNs \citep{Chollet_2021}.

We note there was no evidence of over-fitting in the training diagnostics.
As discussed, the training data is already filtered at this stage to emphasize regions with greater amplitude signals. "

183: Because the training expense and performance are the main results of the paper, can you give some insight in how these vary?

We will add an overview of the initial results of our investigation into the impact of artificial neural network model structure.

Implementation Action: Text now reads:

"... network as network depth increased.
A variety of ANN structures (layer counts from 2 to 20 layers and widths from 8-1024, in steps of 2^n, were varied using an initial test dataset) were evaluated.
Some results of initial explorations are summarized in Fig. \ref{sfig:hyperparameterEvaluation} which shows ROC RSL MSE for ANNs with network widths of 64 to 1024 nodes, and fully connected hidden layers (with normalization layers) between 2 and 10.
Optimal results were generally for network widths of 512 and 1024, and depths of 8 or 10.
The configuration outlined above provides a good balance between performance and training expense.
The Python scripts used…"

224: This is the first time the weights are mentioned. It would be good to mention these weights already in section 2.1 as I think these are the 'output' of the training.

We will add in a description of the weights, specifically the interpretation of the first layer (which effectively maps the relative importance of each of the inputs) in Section 2. The reviewer specifies Section 2.1 (GIA/RSL Models), but given the material discussed in 2.2 (Generation of Model Training Inputs) and 2.3 (Training of the ANNs), these sections seem more appropriate for this subject matter and will be used instead.

Implementation Action: Text updated noted in the latex-diff document.

324: Could you add a conclusion or implication from the result in this paragraph? Do you think the ANN should not be used for intermediate field, or should N be increased?

The intermediate field (globally not just for North America) is a difficult region to produce useful predictions of 3D-SS ROC RSL or ROC RAD via the ANNs employed. For the current work, and given the scale of misfits involved, it suggests that this approach (i.e., this specific selection of inputs to an ANN to produce estimates of 3D-SS ROC RSL or ROC RAD) should not be used to explore the intermediate field alone, but rather only as part of a spatially larger dataset as was done in the manuscript. Increasing N did not greatly improve fits in the intermediate field any more so than the near or far field. We will add a concluding sentence summarizing the above to the specified paragraph.

Implementation Action: Text updated and noted in the latex-diff document.

328: “within 2 parameter value increments” In table 1 this does not appear to be the case for delta_total for USGC: 0.05 for EMU and 0.8 for EXP. Can you check this?

We have validated the data in Table 1 and the reviewer is correct: as such the statement is not valid given the results in Table 1, and so we will modify the text to reflect this.

Implementation Action: Qualified such that the text specifies that it it is ‘generally within two increments’; e.g.,
“…such that generally the region of minimum misfit either overlaps the 3D GIA model results, or is within two increments in the parameter space”

Miscellaneous comments
Caption figure 1. The lithosphere values shown are scaled, but the values in North America are above 200 km which is thick compared to other GIA studies. It would be good to comment on that in the text

The lithosphere thickness values shown in Fig. 1 are not scaled. As stated in the caption, the values shown are those of the LithoRef18 model (Afonso et al., 2019). The actual scaling will vary depending on the target global value (for the ‘reference’ 1D viscosity profile). For most cases, the scaling is less than 1 and so the values shown in Fig. 1 are significantly reduced (by about 10-40%). Caption will be revised to improve clarity.

Implementation Action: Caption text now reads:
“... is equivalent to the value specified in the SS configurations.
The scaling depends on the target global value for the `reference’ 1D viscosity profile.
For most cases, the scaling is less than 1 and so the values shown are significantly reduced (by $\approx 10-40\%$). “

Figure 2: There are a few apparent outliers for LT = 96 km, could you comment on those?

We expect the outliers to come from one of two sources: (1) Potential issues with individual runs which comprise the datasets used to either train or validate the neural networks or (2) some unknown issue from the training of the ANNs themselves. With respect to (1), the Seakon code was not designed around ensemble scale work and requires at least 2 iterations produce a converged solution within tolerance. There have been multiple instances where model runs have failed unexpectedly (either due to internal software, or external hardware/computing-platform issues) but still produced output up to the point of failure, thus resulting in a mixed-iteration dataset. While we have quality-checked these datasets to prune (i.e., re-run) those parameter vectors which stood out previously or produced errors during model execution, we will redo the parameter vectors the reviewer highlights in Figure 2. If the re-runs do not remove the outliers, then they must be due to (2) and we will document them as such.

Implementation Action: We have examined and rectified those outliers identified within the time constraints of the publishing process. Where anomalous results remain we have replaced the values by NaN and noted in the caption with the additional text:
“Note that results which are anomalous and suspected to be affected by technical issues for specific ensemble members are set to NaN and rendered in gray.”

262: If this is correct, does that mean that the RSL anomaly is also relatively large for present? If you have these results it would be good to report on that to support your tentative conclusion

The 3D-SS ROC of RSL emulator:model misfit would be large relative to the signal, as with the ROC RAD. However, these errors are only applicable over a relatively short duration and do not significantly impact the RSL predictions themselves. Indeed, this is also the case for the RAD predictions in the past (prior to ice disappearance in North America and Fennoscandia). However, as the text notes, this does render the emulator, as applied here, inaccurate for estimating contemporary land motion and also ROC RSL. Improving this is a target of ongoing work.

Implementation Action: No action to implement

332: It is a nice idea to use the emulator to find a larger area in the parameter space that can be searched with the explicit method, but what do you mean exactly by the “parameter space that provides the optimal fits”. Is it the best fit parameters with a confidence region? How does it differ between EMU and EXP?

This statement means the sub-area or (sub-areas) of the total LT-UMV-LMV parameter space that produce the lowest misfit values. For example, the approximate range UMV (0.1-03 x 1021 Pas) and LMV (1-10 x 1021 Pas) for the emulated model output in Fig. 6 (middle row). One could define this range to some degree of confidence using a statistical test (e.g., F-test), but we have not done this. As evident in Fig. 6, there are some differences between the emulated and explicitly modelling delta values (compare results in middle vs bottom frame). Given this, when using the results based on the emulator, it would be best to expand the boundaries of the sub-area by one or two increments in UMV and LMV to have greater confidence in finding the optimum parameter set via scaled-down search using the model (Seakon in this case).
Text will be revised and expanded to clarify these aspects.

Implementation Action: Text updated and noted in the latex-diff document.