Answer to Reviewer 1 (Sergio Zlotnik)

Thank you very much for taking the time to review our paper and helping us to improve its current state. Below, we outline how we will address the comments in the manuscript itself.

- Section 2.1: is the forward model linear? or the viscosity \eta is a function of velocities \bf{u}? This is very relevant to the discussion of section 4.1.1 (errors for eta are smaller that for rho)
    - We solve for a linear forward Stokes Flow problem, that is, we do not consider any functional behavior of the effective viscosity parameter either in terms of solid velocities nor strain rates. This information has been added to section 2.1 in line 67 and lines 80-81.
- Section 2.2: Are the snapshots centered before the svd? Number of snapshots? Size of the test set? This is explained later, but maybe it can be repeated in 2.2.
    - Reviewer 1 is right: the snapshots are centered before executing the svd. We added all required information to section 2.2 in lines 104-107 and line 110.
- Section 4.1.1: How local and global errors are measured? It is stated that errors are of order of 1m (tol 1E-4) although in Fig 2 errors are 1 to 5%. How should I interpret these results?
    - Global errors are calculated as the MSE of the back-projected solution. We added the exact formula in lines 112-114. The problem with global errors is that, locally, the error can exceed these values. Therefore, we additionally investigated the local error behavior (by taking the difference between the full- and reduced order solution) for designated solutions. This we discussed in Fig. 2, where we demonstrate how the maximum local difference is an order of magnitude higher than the global error. We added a more detailed description in lines 205-210 to better highlight the differences between these two error metrics and their importance for the evaluation of the approximation quality.
- I could not follow the discussion on the viscosity parameterization. Are there 6 or 2 viscosity parameters?
    - The forward model has six viscosity parameters. However, during the construction of the surrogate model, we decided to combine some layers in terms of their viscosity parameterization, and consider only the viscosity of the lithosphere (including slabs) and asthenosphere as free parameters. We add some clarifying text in lines 219-222.
- "However, as we will detail in subsection 4.3, this surrogate model is currently not constructible because of challenges related to the data set." why?
    - The reason behind our statement is that, during the construction of the surrogate model for the case where we combined both density and viscosity, we did find that some of the full order forward solutions showcased velocities and topographies that vary by several orders of magnitude with respect to the other family of solutions. Worth to mention is that these were solutions to fully-solved forward simulations (that is we did encounter any crash or premature ending of the forward model) and that they appear to follow a random pattern within the parameter space. The presence of random variations, that is, solutions that do not follow the expected physical behavior, hindered to construct a surrogate model. While addressing the comments received during the first round of review, community post and reviewer

    - 2, we did carry out an in-depth revision of the workflow to extract the results from the full order solutions to construct the surrogate model and we realized that the appearance of these random „unexpected“ solutions was due to an error in the shell script we use to identify the timestep from where to extract the full order solutions (for those cases it was picking an earlier timestep before the system was fully isostatically balanced). We would like to stress here that we double check all three scenarios discussed in the manuscript and we found that this error only occurred for scenario 3, and only for few realizations within this scenario, being the reason, we did not spot the problem before. Given that we identified the source of the problem, we carried out the construction of the surrogate model for this last scenario as well, which is presented in section 4.3 and in Figure 2.
- The use of POD is fairly clear. But for reproducibility it would be necessary that the configuration of the NN is included in the manuscript or in the public data.
    - The main hyperparameter of the NN are listed in the tables of Figure 2, additionally we provided a data folder containing the scripts for the surrogate model construction of all surrogate models used in the paper. The "SurrogateModelScripts" folder has always a file which contains the configuration of the NN. To better highlight this, we added a description in lines 178-180.

Answer to Reviewer 2 – Frist Round

We would like to thank reviewer 2 for his/her comments.

We would like to emphasize, that the main objective of our study is rather on describing a novel methodology to construct surrogate models for geodynamic simulations and to allow for probabilistic inverse studies and not on the respective geodynamic simulation methods that the surrogate model will be applied to. However, the comments from reviewer 2 are about aspects related to some of the forward simulations performed with the open-source software LaMEM and only applied to the last scenario presented and discussed, scenario 3. Thus, the comments of reviewer 2 exclusively focus on aspects that make up only a small portion of the entire work. We thank reviewer 2 for these comments that helped together with the community post to identify the issues with the last scenario and want to elaborate with some detail to those: As a matter of fact, the methodology presented in the manuscript is not dependent on the forward dynamic solver adopted (as illustrated in section 5). The problems encountered while constructing the surrogate for this last scenario do not impact the scientific merit of the results in principle and we added the surrogate models for the combined scenario. As detailed in our answer, we confirm that there was an error in the extraction of the timesteps and we acknowledge all comments raised from reviewer 2, whose expertise with the specific software has helped improving parts of the applications presented. Having identified this error in the script we used to extract the corrected timesteps, and having realized that the error only affected a few simulations for the last scenario, we have now constructed a surrogate model for this last case in the revised manuscript. Let us disclaim here that the problems encountered, wrong timestep extraction, are solely related to the combined scenario.

Please find below a description of how we addressed the comments in the paper.

- The development of surrogate models represents a noteworthy and valuable contribution. However, it is important to express concerns regarding the reliability of the underlying forward models. Several simulations terminated prematurely due to a crash in a third-party library (details are provided in Appendix A). These failed simulations were not correctly identified and were instead interpreted as successful, albeit exhibiting "instabilities" such as low topography and high vertical surface velocities. In reality, these anomalies are attributable to the premature termination of the simulation, which prevented the system from reaching isostatic equilibrium. Substituting the crashing solver component with an alternative resolves the issue entirely.

  - First of all, we would like to clarify that all our simulations did not „terminate prematurely" and we did not face any problem with any specific „third-party library". The source of the problem stems from an error in the automatic extraction of the results through a shell script. All results were supposed to be extracted at ~0.27 Myr (a clarification has been added in lines 148-151), a time when the system is in isostatic balance. However, the error in the shell script leads to an occasional extraction after 0.027 and 0.0027 Myr. Because of the random nature of this occurrence, we did not spot the problem before. Therefore, we cannot agree with the conclusive remark from reviewer 2 stating that the source of errors stems from solvers failure. As already stated in our general comment above, all results from the forward simulations have been already stored, so there is no need to carry out additional „full-order simulations".
  - A general note of caution here: the problem rarely lies on the library (MUMPS), rather on the way it is used. MUMPS is a well-known and widely used distributed-memory parallel solver (the same apply to superLU_DIST). The reason for us to prefer to rely on MUMPS instead of a more frontal solver as superLU_DIST stems mainly from the former being designed for a relatively broader range of matrices and relying on a dynamic pivoting over hybrid parallelization. This said, it is relatively hard for us to pin-point a likely potential reason for the crashing simulations experienced by reviewer 2: we had too few details on the configuration, architecture, library versions and similar aspect on his/her local installation (a configuration log would help us in this regard) to be able to reproduce the crashing simulation.

- Although the affected simulations are explicitly reported only for the combined training set, it cannot be definitively ruled out that similar problems may exist in the other sets. Therefore, it is strongly recommended that all forward simulations be recomputed to ensure consistency and reliability. As this task exceeds the scope of a standard major revision, the manuscript should be reconsidered for resubmission after the complete regeneration of the training datasets.
  - We double checked the extraction regarding all data sets and can confirm that the issue is solely related to the combined scenario 3, where we changed both the density and viscosity. Given that the error is not related to the forward solution, but rather to a wrong extraction of the results, no further forward simulations are required to generate the training datasets for the last scenario. The training data sets for scenario 1 and 2 do not require any modification either.
  - The newly generated surrogate models are presented in section 4.3 and Figure 2 has been updated to account for the additional surrogate models. We also updated Figure 7 correcting for the previous wrongly extracted forward solves. In light of these changes also the Discussion has been updated, especially in lines 364-368 and 424-444 and the conclusion was adapted in lines 464-472.
- To reduce the computational cost associated with regenerating these datasets, an optimized set of input parameters and solver configurations may be employed. Appendix B summarizes a series of such optimizations, which together yield a reduction in run time by more than two orders of magnitude.
  - Thank you for the suggestions. However, as stated above, all simulation results for all scenarios, including scenario 3, have been previously stored. Therefore, there is no need

    - to rerun any simulations, but rather to redo the extraction step. A reference to the revised version of the data set is provided in the revised manuscript.
- Rather than including the datasets themselves, the supplementary material should provide scripts used to extract relevant information from the simulation outputs. It is also recommended to include automation scripts for input file generation and forward model execution. Appendix C presents an example Python implementation of this functionality.
    - As pointed out in the general comment, the main objective of the paper is to leverage surrogate model construction for geodynamical simulations to enable probabilistic inverse methods in this field. Therefore, the data folder concerning the surrogate models was constructed with this main aim in mind. To ease the surrogate model construction, the simulations results are provided in form of data sets. Our decision was motivated mainly by the wish to enhance ease, from the user side, to reproduce the surrogate model construction and its applications (the main subject of the manuscript) without involving any installation of the forward solver used for generating the training set. Nevertheless, to ease the reproducibility of the forward model, we included the mentioned shell script in the data folder containing the information about the forward model (see second round of comments).
- Detailed comments
    - These detailed comments relate to specific aspects regarding the combined scenario. Because of the above detailed reasons, we completely reworked section 4.3 and parts of the discussion in the revised manuscript. Hence, the detailed comments refer to text parts that are no longer existing in the document.

Answer to Reviewer 2 – Second Round

We would like to thank reviewer 2 for his/her comments.

On a general note, we would like to stress again the focus of the study, that is, to showcase a workflow to construct surrogate models that conform to the actual physics at play in order to enable setting up and solving inverse problems as applied to typical geodynamic simulations.

This said, we notice that, unfortunately, the comments raised by reviewer 2 again do focus on a particular aspect related to a subset of forward full-order geodynamic simulations, but do not address the other scientific aspects of the study. While we already stated that, thanks to the community post as well as the first round of comments raised by reviewer 2, we have been able to identify the source of errors for some forward simulations (due to a wrong timestep selection during the postprocessing stage), we would like to clarify that the error made relates to a minor part of the study (a single sub-set of the simulations) and should not be considered as downgrading the scientific merit of the study, which stems from the surrogate model construction and its usability for geodynamic applications. As stated in our first round of replies, we provide a revised manuscript where the surrogate model is constructed also for the final subset of simulations, and we hope that with the new addition, the message and merit of the study is clarified.

In what follows, we provide a point-by-point answer to all comments raised during the second round of posting.

- "First of all, we would like to clarify that all our simulations did not „terminate prematurely" and we did not face any problem with any specific „third-party library."

To allow an independent assessment of the correctness of this statement, please upload the log files of all forward simulations.
In addition to the log files, please provide the complete software infrastructure used to accomplish the following tasks:
A) Preparation of the input files for each realization
B) Submission of each realization using a workload manager (e.g., slurm)
C) Error checking for forward model runs (this part is absolutely critical).
D) Extraction of input data for surrogate models from forward simulation outputs
According to GMD policy, all of these critical software components must be made publicly available.

- We can agree only partly with the point raised by reviewer 2. We entirely agree with the „FAIR" principles of code and data as per the journal policy, but found some of the requests from reviewer 2 not consistent with the GMD data policy. To be more specific, we agree with the majority of requests, but find point (B) puzzling, which relates to bash scripting for the slurm scheduler (or any other scheduler) used. We personally do not see how such additional information would adhere to the „FAIR" principles, since job submission depends on the particular architecture of the cluster used (and this also entails available libraries and modules) as well as the running scheduler. How would it help to judge the scientific merit of the manuscript?
- Concerning all other points, we uploaded all input files used to run the simulations (point A), which are also informative of the error criteria adopted in each run (point C). In addition, the input data used to construct the surrogate models (currently present in the paper) has already been stored in a dedicated repository. The surrogate models that are provided in addition along with their data sets have been added to the data repository.
- To clarify the working procedure, we rely on bash shell script "make_models.sh" to prepare the forward full order model [**https://doi.org/10.5281/zenodo.16640814**], with a combination of "awk" to read the input parameters from the comma delimited files, and a stream editor "sed", to edit the material input properties in the LaMEM input file. Both of these, awk and sed, are part of Unix/Linux distributions. The shell script produces directories for each model containing a LaMEM input file and a slurm script to submit the job, which are provided here [**https://doi.org/10.5281/zenodo.16640814**].
- We ran the simulations until a minimum of 0.3 Myr since we fixed the time step at which data is extracted for the surrogate model (see below). For this, we relied on another shell script, "time_step_use_find.sh, which lists the time step within a given range of interest (0.2-0.3 Myr). This script produces an output text file with the model name (first column) and the associated time-step directory names (following columns). We then used this output file to read the simulation results in Paraview and write them out as .csv files using a customised Python script ("save_time_step.py" [**https://doi.org/10.5281/zenodo.16640814**]). For the models for which those timesteps were missing, we checked the simulation status in the slurm log file, and, in case of premature failure, non-convergence of the solvers, we adjusted the solver settings related to the number of iterations and the multigrid solver options [solver_type_2.dat,**https://doi.org/10.5281/zenodo.16640814**]. All this information is provided within the respective input files [**https://doi.org/10.5281/zenodo.16640814**]. To avoid any confusion, we have now

added all input files to the data repository instead of providing only one exemplary input file (as done during the initial submission).

- *"The source of the problem stems from an error in the automatic extraction of the results through a shell script. All results were supposed to be extracted at 0.27 Myr, a time when the system is in isostatic balance. However, the error in the shell script leads to an occasional extraction after 0.027 and 0.0027 Myr."*
  This error appears rather puzzling, especially considering that your simulations do not output data at any of the mentioned time steps (0.27, 0.027, or 0.0027 Myr). The closest timestamps I found in the output logs are: 2.796436e-01, 2.964359e-02, and 2.843589e-03. Furthermore, there is no apparent pattern in the problematic realization numbers (82, 89, and 95), which suggests the issue occurs randomly. Please provide a detailed explanation of this bug, along with the actual extraction script.
  - We realised that this point requires further clarification. The pattern in the time step we do reference (0.27-0.0027 Myr) is to illustrate the issue. Indeed, the actual timesteps from the output list might follow a similar order, as for example extracted from a random solution:
    Timestep_00000100_2.28830780e-02
    Timestep_00000620_2.82883078e-01
  - As shown in the example, the naming convention lists the time step number as the first number in the file name. Afterwards, the value of the timestep is printed, followed by the exponent power as the last number. Hence, this exponent power dictates the extraction, whether we are at 0.28 or 0.028 Myr.
  - To illustrate the issue we encountered, we show below the previously used script for the extraction (in bold the relevant lines):
    ```
    #!/bin/bash
    rm -f use_time_Step # file to store model name and time step directory; remove if it
    #exists
    touch use_time_Step # create the file
    for dir in Train_* Validate_* # iterate over all the models
    do
    cd ./$dir # go inside the directory and run following
    echo $dir `ls -d Timestep_0000*_2.[6-9]*`>>./../use_time_Step # print model name
    #and listed directory with the prescribed pattern
    cd ./../
    done
    ```
    This script listed all the timesteps with the above pattern, from the lowest exponent (e-03) to the highest in ascending order. The issue arises from the circumstance that the first entry in the time-step column is read by the Python script in Paraview. Meaning that for some realisations, the timestep with the wrong exponent has been considered. As stated, we noticed an error in the script, which we have now corrected to:
    ```
    echo $dir `ls -d Timestep_0000*_2.[6-9]*e-01`>>./../use_time_Step
    ```
    This ensures that the data is picked at the desired time step.
- Why was 0.27 Myr chosen? This duration is not sufficient to reach isostatic balance. Based on your typical model results, I estimate that at least 0.5 Myr is required. To remain on the safe side, 1 Myr should be used. Alternatively, I suggest implementing a custom termination criterion in LaMEM based on a surface velocity threshold. Therefore, I strongly recommend

that all forward models be recalculated for a longer simulation period (at least 0.5 Myr, preferably 1 Myr).

- The aim of the paper is the construction of surrogate models to enable a better constraint of the present-day dynamic state of the lithosphere and mantle in a manner which is consistent with available observations (e.g, Topography and GNSS velocities), and to investigate in future studies the sensitivity of the results, via global sensitivity analysis, to parameters variations due to inherent uncertainties. Hence, we are interested in the quasi-instantaneous dynamic response of the system to the internal deep mass distribution, mantle and slabs.
- Our choice of considering an elapsed time of 0.27 Myr and not 1 Myr is based on the following criteria:
  - 1) The Stoke flow induced by the considered slabs is established in the model and does not deform significantly (Fig. 1);
  - 2) Changes in the topography and surface velocities do not vary significantly. This threshold time is decided by observing the change in variation of topography and velocities for the reference model (Fig. 1).
- After ~0.2 Myr change in mean topography is in the order of ~10 m and < 1mm/yr, respectively. These changes are well within the resolution of digital elevation models and GNSS-derived velocities. Hence, we chose a threshold time of ~0.27 Myr, to also account for variations in physical properties to be on the safe side. An explanation has been added in lines 148-151.

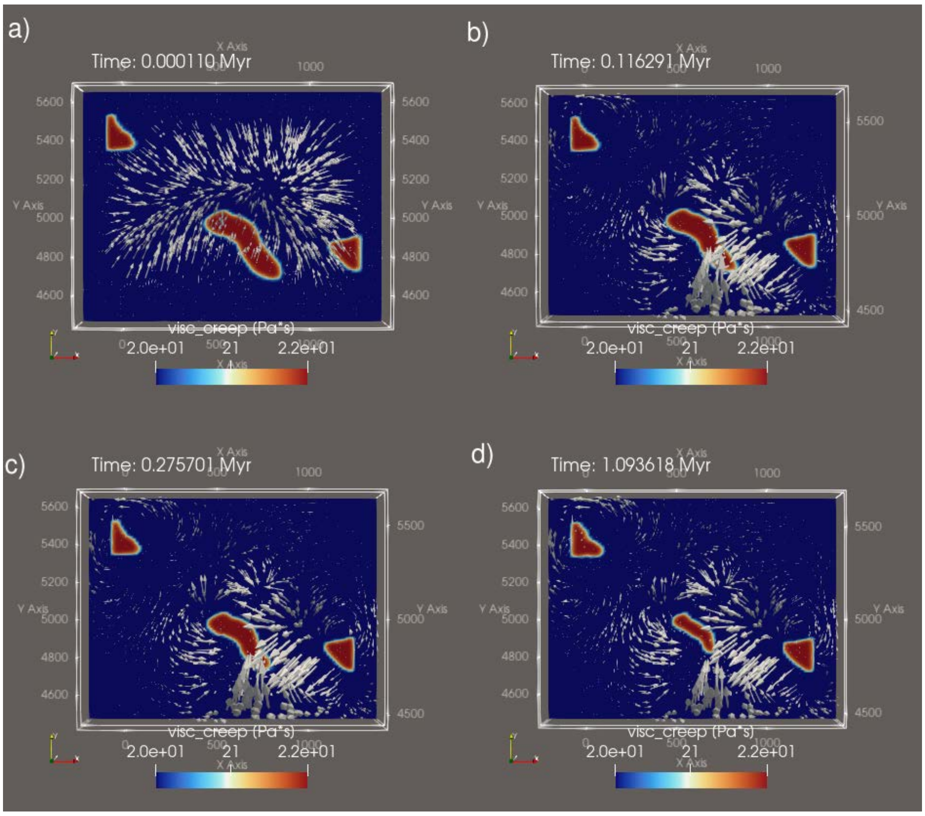

Figure 1: Depth slice of the viscosity at 220 km showing the geometry of the slab for the reference model (higher viscosity:red color) with time (a-d). Flow is shown by arrows scaled by magnitude. Note the decreasing size of the slab region with time.

- Worth to note is that considering a longer time-period, would have induced active deformation within the slabs, leading to a system geometry (and mass distribution) no longer representative of the present-day architecture of the mantle (the one we imposed at the beginning of the model), which would enter systematic (epistemic) bias in the later sensitivity analysis where the actual configuration would have been dependent on the applied material properties, thus hindering an objective comparison of the different response only in terms of the varying properties.

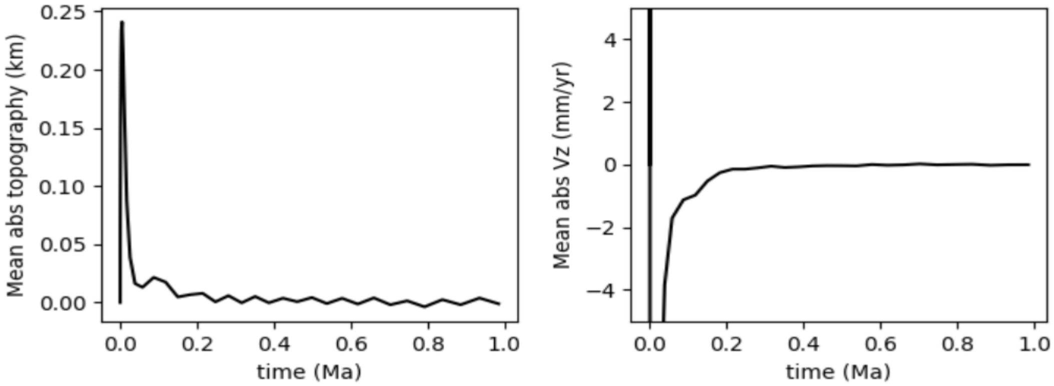

Figure 2: Time evolution of the mean absolute topography (left) and mean absolute vertical velocity (right) change for the reference model. Note that for the velocity panel y-axis is clipped for the initial time to visualise the later variation.

- Additionally, I recommend increasing the spatial resolution, as some geological units appear to be underrepresented. Ideally, a resolution sensitivity test should be included.
  - Regarding the resolution tests and increasing the resolution of the models, we do not agree with the comments raised with respect to the specific objective of the study. In our study, we target the deeper mantle configuration only, and therefore, we consider that the current resolution (~13 km in E-W, ~17 km in N-S, and ~3 km along depth) is enough to reflect the typical wavelength of the expected response within constraints from geophysical/geological data. For example, at this resolution, the influence of the lithospheric thickness, upper-mantle architecture derived from the different tomography models (i.e., effects of slab), and the local sensitivity analysis of their physical properties (density and viscosity) are well captured as discussed in Kumar et al 2022 (https://doi. org/10.1029/2022GL099476). We would agree with the comment raised if our study had attempted to further investigate the source of the smaller wavelength response, as due to internal variations in the crustal configuration, which is fixed. Furthermore, the main focus of the paper is the construction of the surrogate models, and the structure of the model is not changed; rather, the physical properties are varied.

- There is no need to upload gigabytes of extracted data, as this can easily be regenerated. Of course, uploading it remains optional.
  - That seems to be a misunderstanding of our previous answer. We wanted to confirm that the data sets are still on our internal servers and thus allow us to redo the extraction step. However, we see no added value in uploading all output files, since, as pointed out by the reviewer, they can be easily regenerated.

- In any case, the complete software infrastructure related to the forward modeling process must be shared in accordance with GMD policy.

- We have shared the scripts used to generate the input files and the data extraction. In addition, we also share the input files for all the models such that they can be reproduced. Please also refer to our previous answer to Comment 1 by the same reviewer.