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 are ready to provide a revised manuscript where the surrogate model will be constructed also for the final subset of simulations, and we hope that with the new addition, the message and merit of the study will be clarified.

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

Comment 1
“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 will upload all input files used to run the simulations (point A), which will also be 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 we are going to provide in addition along with their data sets will be published before submission of the revised manuscript.

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).

Comment 2

“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.

Comment 3

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.

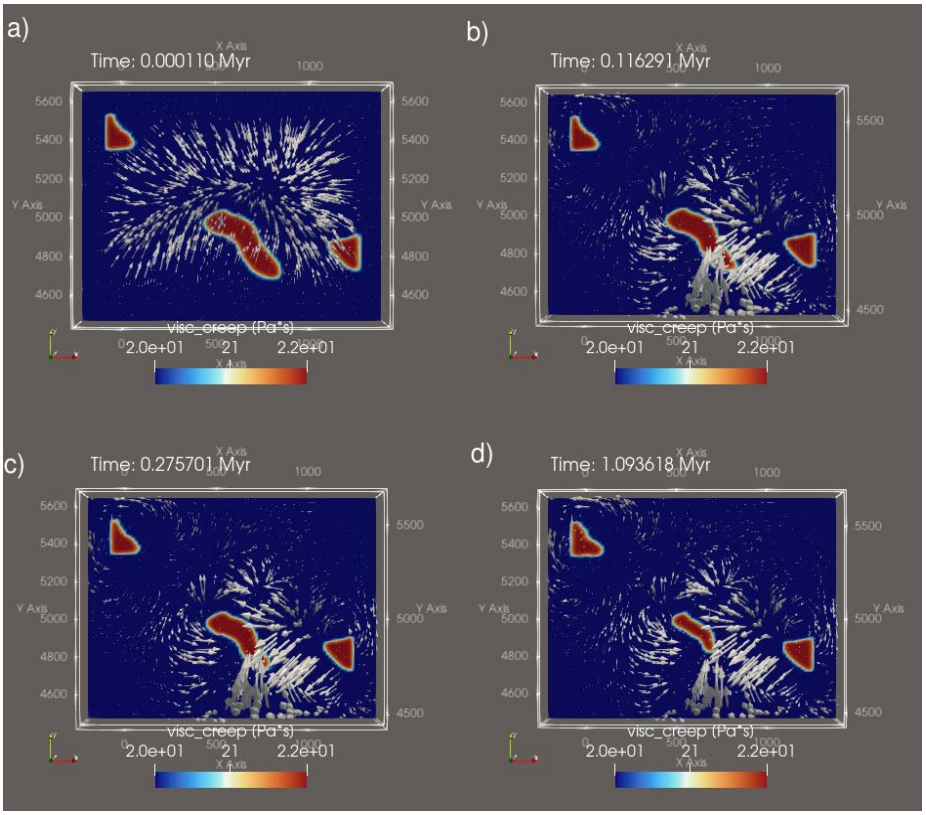

Figure 1: Depth slice of the viscosity at 220 km showing the geometry of the slab for the reference model (higher viscosity:red color) with time (a-d). Flow is shown by arrows scaled by magnitude. Note the decreasing size of the slab region with time.

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.

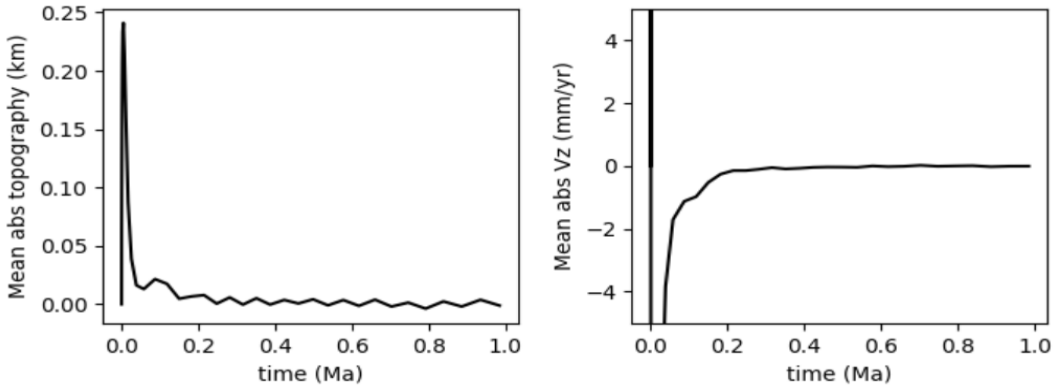

Figure 2: Time evolution of the mean absolute topography (left) and mean absolute vertical velocity (right) change for the reference model. Note that for the velocity panel y-axis is clipped for the initial time to visualise the later variation.

Comment 4
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.

Comment 5
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.

Comment 6
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.