Comments and responses

We thank the Reviewer for her/his insightful comments. We appreciate the time and effort invested in providing detailed suggestions. Below, we address each comment in detail and outline the corresponding actions we have taken.

1 General comments

Comment: What is the runtime of the evaluation tool for the given test case for a single time series?

Reply: For the specific case study of this paper, ML-AMPSIT takes about 10 seconds to generate a single time series using any of the tree-based algorithms, up to about 1 minute for the slowest algorithms, i.e., BRR and GPR. It should also be noted that the original WRF simulations took 4 hours for each run, i.e. about 400 hours to build the 100 members of the ensemble, but the GPR surrogate model took about 1 minute to generate the 5000 WRF-surrogate outputs used to implement the Sobol method, which would have taken $4 \times 5000 = 20000$ hours using standard WRF simulations. However, these speed benchmarks are only guaranteed for the specific setup used in this study and the specificity of the hardware used to run these programs, and should not be considered a general speed benchmark. It would be expected that for larger datasets and more difficult-to-achieve hyperparameter tuning, the runtimes could increase with respect to the ones observed in this study. Even in such cases, the computational time is expected to be substantially lower than high-fidelity simulations.

Action: To highlight these aspects, text was added at lines 340-343: "The low computational cost of these emulators allowed us to employ a surrogate sampling generated by `sobol.sample()`, with 5000 input values (the user can change this value by modifying the configuration parameter *Nsobol* in *loopconfig.ipynb*) with the overall Sobol method calculations performed in minutes against a single traditional WRF simulation typically taking several hours.", and at lines 638-643: "However, while actual runtimes depend on the specific dataset and hardware, the speed improvements observed in our case study highlight the potential of ML-AMPSIT to enable large-scale sensitivity analysis and ensemble generation with significantly lower computational requirements. The generation of surrogate outputs was observed to be sig-

nificantly faster than running high-fidelity WRF simulations, with runtimes reduced from hours to seconds or minutes, depending on the algorithm. This efficiency enabled the generation of thousands of surrogate results that would not have been possible by relying solely on traditional simulations."

Comment: Is the difference in the simulation results of the ensemble significant to evaluate feature importance or do they reflect the intrinsic uncertainty we have to expect in model simulations?

Reply: We thank the Reviewer for this pertinent question. If the ensemble spread were purely random or unrelated to the input parameters being tested, it would be unlikely for all algorithms to agree on parameter importance. In such a scenario, the regression task would either fail or overfit, resulting in unrealistic metrics and a lack of convergence despite an increase in sample size. However, the ability of our model to generate accurate predictions relative to simulated data suggests that most of the ensemble uncertainty is indeed attributable to variations in the selected parameters. This is highlighted in the text at lines 568-572: "It is worth noting that the MSE for GPR, LASSO, BRR and SVM does not show significant variations in the lowest 10 vertical levels both over land and over water (Figures 15 and 16), meaning that the observed variations in feature importance are related to changes in the input-output relation rather than to uncertainty issues. This is also supported by the fact that the metrics of these algorithms in Figure 9 show no deterioration associated with the changes in feature importance shown in Figure 11, and that these patterns are consistent across all the surrogate models."

Comment: Given WRF, which provides a large variety of parameterizations, is the investigation of model parameters a reasonable approach or may different parameterizations result in more ensemble spread and, thus, lead to more uncertainty? Also connected to the last point: Is the tool also applicable to other model uncertainties, e.g., the choice of different parameterizations or input data as land cover, SST, or emissions in the field of air quality. This may especially be important for the use of WRF, where it is more likely to first test different parameterization on their performance before evaluating the parameters within a single parameterization.

Reply: While this study focuses on a specific set of parameters within a particular land surface model, the ML-AMPSIT framework, in principle, is indeed adaptable to a variety of input-output scenarios. This flexibil-

ity means that ML-AMPSIT could be applied to different parameterization schemes, varied land cover types, alternative grid resolutions, and other simulation setups. Results of sensitivity analyses performed with ML-AMPSIT can also be compared to observations to evaluate the best model configuration.

Action: To emphasize this point, the following text has been added to the conclusions at lines 644-650: "Finally, it is worth noting that the application of the methods implemented in ML-AMPSIT is not only limited to the evaluation of land surface model parameters; these methods are inherently adaptable to any dataset containing input-output pairs, regardless of the data characteristics. This flexibility allows ML-AMPSIT to evaluate not only the influence of different input parameters, but also the effects of different simulation setups, such as physical schemes, subprocesses, land cover, numerical strategies, or geometric configurations. By using data-driven modelling, these tasks can be accomplished more quickly and with potentially less data. Moreover, since input-output frameworks are ubiquitous in scientific and statistical domains, the reach of a data-driven tool like ML-AMPSIT potentially extends far beyond the specific examples mentioned here."

Comment: By performing sensitivity and importance analyses, the overarching goal is to improve the models performance. For me, it is not clear how the tool can support this. Model performance is usually evaluated against observations, which seem not to be included in the described approach.
Do the surrogate models allow for testing further sets of model parameters to identify the set which best matches with the observations? Or does the surrogate model provide the best set of parameters by itself?
In this case, is it recommended to increase the assumed uncertainty in model parameters to ensure that the surrogate models include all possible solutions for the prediction of model behavior for other choices of parameter values?

Reply: The tool allows any set of parameters to be tested, and although the present paper is based on idealised simulations, the same can also be done with real-case simulations, thus allowing the comparison with observations. It is important to note, however, that the goal of the tool is to evaluate how much an input parameter affects an output variable, which is quite different from the task of finding the parameter values that best fit observations. For this reason, it finds the most important parameters that affect the variance in the data, but does not provide the best set of parameter values. Such an additional feature could in principle be implemented, perhaps in a future

version of ML-AMPSIT. However, once knowing which parameters cause most of the variance within a perturbed ensemble, the user can potentially concentrate on these parameters to improve model results. Indeed, knowing which parameters are most critical to the simulation output highlights which values should be estimated with more care to improve model results.
The user can arbitrarily change the uncertainty in the model parameters, but in this paper we chose to limit the ranges to realistic values to avoid unphysical situations.

Action: The main aim of ML-AMPSIT is summarised at lines 106-111: "ML-AMPSIT guides the user through the different steps of the sensitivity and importance analysis, allowing, on the one hand, for a simplification and automatisation of the process and, on the other hand, for extending the application of advanced sensitivity and importance analysis techniques to complex models, through the use of computationally inexpensive and non-linear interaction-aware methods. Once knowing which parameters cause most of the variance within a perturbed ensemble, the user can potentially concentrate on these parameters to improve model results. Indeed, knowing which parameters are most critical to the simulation output highlights which values should be estimated with more care to improve model results."

Comment: In the code repository, I'd recommend adding a readme file that details the workflow and basic principles of the ML-AMPSIT tool.

Reply: The README.md file in the code repository (https://dx.doi.org/10.5281/zenodo.10789930) explains the aims and principles of ML-AMPSIT and delves into the description of each file following a sequential order as intended in the workflow.

2 Minor comments

Comment: Line 57: ensemble perturbed parameter -> ensemble of perturbed parameters

Action: Text changed accordingly.

Comment: Line 67: Begin a new sentence "The study found..."

Action: Text changed accordingly.

Comment: Line 78: As I understand, the strength of ML-AMPSIT is

its applicability to different models in the field of atmospheric research. I suggest adding this information to the sentence to stand out against the studies presented previously in the introduction.

Action: We have clarified at lines 644-650 that ML-AMPSIT is designed to work with various input-output datasets, although this study focuses on land surface model parameters in WRF.

Comment: Line 108: the half-sentence "thanks to..." can be removed. It is a duplicate of the clause in line 105.

Action: Text changed accordingly.

Comment: Figure 1, step 4: ML-AMPSIT.ipybn $->$ ML-AMPSIT.ipynb ; looponfig.json $->$ loopconfig.json

Action: Typos in filenames corrected.

Comment: Figure 1, step 5: ConvergenceAnlys.ipybn $->$ ConvergenceAnlys.ipynb (This file is missing in the code uploaded to zenodo. Please add.)

Reply: The file ConvergenceAnlys.ipynb produces convergence plots (Figures 7 and 8 in the manuscript) from the file generated by ML-AMPSITloop.ipynb. However, its structure is not general enough to be used by any user without changes, because it depends on the details of the sensitivity analysis performed. Therefore, we decided not to include it in the main repository in the current version of the tool. However, users can easily perform convergence analyses, as presented in the manuscript, from the results provided by ML-AMPSIT, in particular by ML-AMPSITloop.ipynb. To avoid confusion, we have removed the convergence analysis step from Figure 1 and the workflow described in the paper. However, we have put the file ConvergenceAnlys.ipynb in the dataset repository (https://doi.org/10.5281/zenodo.14051616) as a reference for anyone interested in replicating the plots produced in the paper.

Action: The convergence analysis step has been removed from the workflow described in the paper and the file ConvergenceAnlys.ipynb has been put in the dataset repository (https://doi.org/10.5281/zenodo.14051616)

Comment: Line 188: Please give more information about the p-value (I suppose from a significance test) in relation to R2. How can it be interpreted and how is it related to R2? What is the hypothesis to be tested?

Action: Additional information about the p-value's relationship with R^2 has been added in the caption of Figure 6.

Comment: Line 225: What is the exact definition of Vi? Is it Vi = VAR(Y (X1, X2, ..., Xi + Δ, ..., Xk), Y (X1, X2, ..., Xi, ..., Xk))? But if this is true, how is the perturbation Δ accounted for considering that (at least in the linear case) larger perturbations lead to larger effects in Y?

Reply: We thank the Reviewer for pointing out this. The perturbations Δ must be prescribed as random values that uniformly probe the output response in an arbitrarily wide range. In the present study, we chose the Δ ranges such that the values remain realistic to avoid unphysical output. Larger Δ values imply, at least in the linear case, larger effects on the model output, but this does not necessarily translate to larger parameter importance. However, it is true that the results of a sensitivity analysis, regardless of the approach chosen, always depend on the range of exploration of the parameters, and that their transferability to arbitrary ranges of values is not guaranteed if the true sensitivity of the parameters in unexplored ranges is not known a priori.

Action: The effect of varying the range of variability of the parameters is commented on at lines 426-431: "It should be clear that the results of a sensitivity analysis, regardless of the approach chosen, always depend on the range of exploration of the parameters, and that their transferability to arbitrary ranges of values is not guaranteed if the true sensitivity of the parameters in unexplored ranges is not known a priori. The perturbation percentage in this work has been chosen to avoid unphysical values, but it must be noted that the aim of the present work is to introduce and test ML-AMPSIT functionalities in a simplified case study, whereas a more detailed analysis would require more attention to the choice of the parameter space."

Comment: In Eq. 4: Isn't summand S_{13} missing according to the rule in Eq. 3?

Reply: Yes, thanks for pointing out this error.

Action: Added the missing S_{13} term.

Comment: Line 261: "also known as a ridge-type regularization" can be removed as this is a duplicate of the statement in the previous sentence.

Action: Revised as suggested.

Comment: Line 389: change the beginning of the sentence to "The initial atmospheric potential temperature . . . "

Action: Text changed accordingly.

Comment: Figure 3: I suggest decreasing the size of the figure but increase the font size of the text. I also suggest zooming in to the center of the figure to highlight the area of investigation. Also, the "3 adjacent grid cells", which are also considered in the analysis can be included in the plot. In the caption, the land area is denoted as "red area". At least in my copy it appears green. Please revise.

Action: Figure 3 was revised according to the Reviewer's suggestions, highlighting the area of investigation and adjusting colours.

Comment: Line 505: "increasing and others decreasing": What is increasing/decreasing? Please clarify.

Action: Text changed to improve clarity: "Around these times, individual ensemble members exhibit divergent behaviour, some showing increases and some decreases in wind speed, which can complicate the prediction for the regression models".

Comment: Line 507: The authors highlight the self-validating feature of the ML-AMPSIT tool, where an agreement between the different approaches is assumed to be a measure of robustness of the results. As 3 out of 7 algorithms perform worse than the others, the question arises at which point the authors claim the results as "not robust". Also keeping in mind that "worse performance" does not automatically mean "bad performance" in general. How can the user discriminate between robustness of the results and the results being unstable.

Reply: The Reviewer is right that "worse performance" is different than "bad performance". The performance of the algorithms can be first assessed from the values of the evaluation metrics, as stated at lines 183-188. Then, the comparison between the results of the different algorithms helps to evaluate if a worse performance is also a bad performance or not. In the paper, although the metrics of the tree-based algorithms are quite lower than those of the other algorithms used, the robustness and stability of the results are inferred from the agreement among all the algorithms on the importance and ranking of the parameters, as well as from the convergence analysis of the results. This is stated, for example, at lines 485-490. Therefore, the user should

combine the information coming from the evaluation metrics and the comparison between the output of the different models to have a complete idea of their performance and thus to discriminate between "worse" and "bad" performance. The importance of comparing the results of the different models is highlighted at lines 500-502.

Comment: Line 523: In this sentence, Fig. 12 can be referenced for clarity.
Action: Reference to Fig. 12 added.

Comment: Line 565: Add "Figure" before 11.
Action: Text changed accordingly.

Comment: The citation of He and Ek (2023) needs to be revised. The other co-authors do not appear in the reference.
Action: The citation was revised.