Review of “ML-AMPSIT: Machine Learing-based Automated Multi-method Parameter Sensitivity and Importance analysis Tool”

The manuscript details the workflow and application of the newly developed ML-AMPSIT. The tool provides an automated framework to conduct sensitivity and importance analysis for model parameters using seven different machine learning algorithms. It was developed for the WRF model but is applicable to all models dealing with model parameters. The aim of the manuscript is the introduction of the new tool and the description of its capabilities to support scientists conducting their sensitivity analysis with respect to model parameters. The tool was applied to the WRF model coupled with the NOAH-MP parameterization to analyze the dependence of the model results on the input parameters. With this, the capabilities of the ML-AMPSIT have been demonstrated.

The manuscript is well written and of good quality. All reviewer comments have been sufficiently addressed during the first review process. Also, the manuscript and the code fit well into the scope of GMD. However, I have stated a few comments below that remain unclear for me after reading the script. I recommend the publication of the manuscript after addressing my below posted comments.

General comments:

As this manuscript is a description of ML-AMPSIT, I would have expected a stronger discussion about the performance of the tool and the interpretation of the results. For example, the authors may address questions like: What is the runtime of the evaluation tool for the given test case for a single time series? 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? 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.

The idea of ML-AMPSIT is to construct surrogate models to evaluate the sensitivity of the model to certain parameters as well as the importance of these parameters. 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?

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

Minor comments:

- Line 57: ensemble perturbed parameter -> ensemble of perturbed parameters
- Line 67: Begin a new sentence “The study found…”
- 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.
- Line 108: the half-sentence “thanks to…” can be removed. It is a duplicate of the clause in line 105.
- Figure 1, step 4:
    - ML-AMPSIT.ipybn -> ML-AMPSIT.ipynb
    - looponfig.json -> loopconfig.json
- Figure 1, step 5:
    - ConvergenceAnlys.ipybn -> ConvergenceAnlys.ipynb (This file is missing in the code uploaded to zenodo. Please add.)
- Line 188: Please give more information about the p-value (I suppose from a significance test) in relation to $R^2$. How can it be interpreted and how is it related to $R^2$? What is the hypothesis to be tested?
- Line 225: What is the exact definition of $V_i$? Is it $V_i = VAR(Y (X_1, X_2, ..., X_i + \Delta, ..., X_k), Y (X_1, X_2, ..., X_i, ..., X_k))$? 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?
- In Eq. 4: Isn’t summand $S_{13}$ missing according to the rule in Eq. 3?
- Line 261: “also known as a ridge-type regularization” can be removed as this is a duplicate of the statement in the previous sentence.
- Line 389: change the beginning of the sentence to “The initial atmospheric potential temperature …”
- 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.
- Line 505: “increasing and others decreasing”: What is increasing/decreasing? Please clarify.
- 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.
- Line 523: In this sentence, Fig. 12 can be referenced for clarity.
- Line 565: Add “Figure” before 11.
- The citation of He and Ek (2023) needs to be revised. The other co-authors do not appear in the reference.