Referee report for "Validation of climate mitigation pathways"

The manuscript introduces and details the *piamValidation* package, developed to "enable systematic comparison of variables" from IAM datasets. The package is presented as an instrument that can improve reliability and transparency of IAMs, and two uses are highlighted: the comparison and harmonization of scenarios across IAMs, and the comparison of near-term dynamics in IAM scenarios with external, historical datasets. Two applications, corresponding to these two use cases, are then presented: one on NGFS scenarios, highlighting the different types of comparisons/harmonization that the tool can perform, and one for the improvement of the REMIND model through the validation of near-term technology dynamics.

The manuscript is clearly written and provides a good overview, description and discussion of the package. As someone who has some knowledge of the IAM landscape but is not a modeller, I found it easy to follow and understand, which suggests it is suitable for the introduction of a package that is presented as user friendly.

Below are a few points that deserve clarification and/or discussion.

- Section 3.2 "REMIND Validate short-term technology trends" presents an application of the *piamValidation* package to REMIND scenario data and discusses subesquent changes between REMiND 3.2.0 and REMIND 3.5.0 to improve the accuracy of the near-term technological dynamics in the model, but it is not always clear how exactly the use of *piamValidation* informed the changes. This is especially the case in subsections 3.3 and 3.4 where the outcomes from the use of *piamValidation* are not mentioned explicitly, so it does not come out clearly how the changes implemented relate to the diagnosis (as opposed to updating/refining in the light of recent literature or data only). Section 3.5 is much clearer in that respect.
- Very minor point on this section, but for consistency the subsections should be labelled "3.2.1/3.2.2/3.2.3" rather than "3.3/3.4/3.5".
- The introduction refers to "the key evaluation criterion outlined by Wilson et al. (2021)". It would need to briefly state what these criteria are and briefly explain how the *piamValidation* tool is aligned with them, or with which of them (was it explicitly designed with these criteria in mind?). This is also important because Wilson et al. present and compare a range of evaluation methods for IAMs, so it would be necessary to state explicitly how and to what extent the tool contributes to the programme that Wilson et al. call for.
- The introduction also refers to the "mapmaker and navigator" approach to IAM-generated scenarios, which was central in the IPCC AR5. On this note, the quote from Beck and Oomen (line 20) should be introduced in a way that is less ambiguous, because as of now it suggests that the metaphor is Beck and Oomen's, while they are in fact explaining the vision outlined in Edenhofer and Minx (2014) and how it understands the relation between the COP and the IPCC; so it would be better to rephrase as "In this approach, the role of the COP is understood through a metaphor : 'the COP operates as a navigator…'" or something along this line.

This approach was also specific to AR5, and indeed was formulated by the co-chair of WGIII in that cycle. It would be relevant to mention developments in the AR6 cycle as well, especially the discussions about the design, scope and use of the Scenarios Database (which was arguably designed as the organising device for the use of scenarios in the report of WGIII, and informed by feedback on and criticism of the approach to IAM scenarios in AR5).
This seems especially relevant here because it would make it possible to discuss how *piamValidation* relates to other initiatives for harmonizing/comparing/validating IAM scenarios, including within the IAMC, in relations to Scenario Databases/explorers, and, more focused on the AR6, to the discussions about the way the vetting of scenarios was carried out in the AR6 and how to work with scenario ensembles (e.g. having a tool that can help compare across scenarios and compare scenarios to historical data could be presented as potential a contribution to future vetting processes). The following reference (published after the initial submission of the manuscript) maybe relevant for that discussion:
Sognnaes, I., Peters, G.P. Influence of individual models and studies on quantitative mitigation findings in the IPCC Sixth Assessment Report. *Nat Commun* **16**, 8343 (2025). https://doi.org/10.1038/s41467-025-64091-w

- *piamValidation* is presented as a tool that can improve the realism and reliability of IAMs on the one hand, and their transparency on the other hand. These are two often related but distinct requirements. The abstract is quite clear on this distinction, but the bulk of the paper is more focused on realism (with the application to REMIND as a clear example of how the tool can help improve realism) and the distinction and articulation between the two objectives (which can of course be combined) could be made even clearer. The NGFS application shows that the tool is more versatile and can have broader uses, oriented more towards intercomparison and handling of large scenario datasets, which is a strength that could be highlighted more. The value of scenarios is not only/not necessarily always in their realism – in some cases exploratory, idealized or extreme scenarios can be useful, and in that case a tool that can make comparison more robust and transparent is valuable; the capacity of the tool to adapt to a range of conceptions, approaches and purpose of scenarios development (instead of just driving harmonization/convergence towards one vision of realism or usefulness of scenarios) is then important.
- In terms of use for model improvement, judging from authors' affiliations, it seems that the tool was developed by the same team which develops REMIND, and so it is presumably particularly suited to REMIND. Would such use for model validation and improvement be replicable with other IAMs, i.e. would the tool be suitable and/or adaptable to other model architectures? This does not need to be tested for this paper, but the question would deserve to be opened in the conclusion.
- The question of the reference data, its availability, selection and processing is not discussed extensively. The conclusion states that "The REMIND case study demonstrates that the usefulness of the tool for historical and benchmark validation critically depends on the chosen validation settings and the availability of reliable reference data" but I did not find that this so was clearly explained in

the relevant section. The selection of reference data does raise some questions. Could the tool also be used to compare different sources of reference data?
In the REMIND case study, it is explained that the tool was used to compare with "historical data" but the comparison extends to 2030 and includes projected trends. Why are IEA trends to 2030 deemed more reliable than REMIND near-terms trends and why calibrate against them? How does this relate to the ambition to improve transparency, considering that IEA data is not open and IEA models are less transparently documented than most IAMs?

- The conclusion (line 386) mentions "the ability to identify high-quality scenarios from an extensive scenario database". The reference to "high quality" begs the question of what defines the "quality" of a scenario: is it realism? fitness for purpose and ability to help address/clarify a specific question? replicability and transparency? Is the quality of scenario an absolute attribute or conditional on context and purpose of the scenario exercise? etc… To avoid opening these epistemological questions in the conclusion, it might be better to use a more neutral phrase e.g. "the ability to handle, compare and vet scenarios from an extensive scenario database".