Response to Anonymous Referee #4

*Throughout the document, the original comments of the anonymous referee are presented sequentially in **black** and italic font.*

The authors’ responses are provided in **blue** font to ensure clear distinction.

***Comment**: 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. I recommend its publication.

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

4 Response: We are grateful to Reviewer #4 for the comprehensive and constructive feedback. The insightful remarks helped refine the manuscript and strengthen its clarity.

***Comment**: 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.*

4.1 Response: Thank you for this comment. In the second application case, the *piamValidation* tool is used to identify deviations from near-term trends and to document improvements that result from model adjustments. We have added clarifying remarks to the respective subsections which complement the general explanation on the validation and subsequent improvement process in section 3.2.

***Comment**: 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”.*

4.2 Response: We agree and have adjusted the manuscript accordingly.

Comment: *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.*

4.3 Response: Thank you for this important comment. We now mention the evaluation criteria by Wilson et al. (2021) and clarify how the *piamValidation* tool aligns with their framework. Specifically, we highlight the evaluation methods that the tool supports and thus can help to strengthen the evaluation criteria, with a particular focus on credibility.

Comment:*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 and analyses of scenarios databases). The following reference (published after the initial submission of the manuscript) may be 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

4.4 Response: We acknowledge that the original phrasing could have been interpreted ambiguously and have revised the manuscript accordingly. We also appreciate the reference to climate model validation in AR6 WGIII Chapter 3 and have incorporated this point into the revised manuscript.

Comment*: 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.*

4.5 Response: Thank you for this insightful comment. In Section 3.1, the first application case using NGFS scenarios illustrates different types of application cases of the *piamValidation* tool. This application case is not focused on near-term realism, but instead highlights the versatility of the tool. In contrast, Section 3.2 demonstrates the use of *piamValidation* tool to spot near-term deviations in the REMIND model and document subsequent improvements.

The observation that scenarios should be allowed to span a wide range of possible pathways instead of focusing on one most likely scenario is also the reason why we limit the validation checks to the near-term. Depending on technology, limits to scale-up, technology readiness and installation speeds allow the definition of feasibility boundaries which translates to the definition of thresholds up to 2030.

Finally, overall transparency is promoted through open distribution of configuration files and by supporting community-driven discussion on threshold choices and validation settings.

Comment*: 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.*

4.6 Response: The model is agnostic in respect to which IAM is used and more generally could theoretically be applied to other types of models as well. The only restriction is on the structure of the data which is required to follow the IAMC format as described in section 2.2. This is also reflected in section 3.1 where the tool is applied to model output from MESSAGE and GCAM besides REMIND.

***Comment**: 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?

4.7 Response: The selection of observational data and projected estimates is discussed in section 3.2 with the most detailed explanations described in the GitHub discussions linked in the manuscript (https://github.com/pik-piam/mrremind/discussions/544). We agree that care is needed when using other models to validate REMIND. We do think the comparison to specialized sector models with high granularity can yield valuable insights for a full-system model such as REMIND. (In the interest of transparency, open data and open-source models should be preferred where available.)

We see the presented validation exercise as a starting point of an ongoing community effort to identify reliable reference data and welcome discussions that help curate a collection of thresholds. We acknowledge uncertainty in external data sources by adding additional tolerances to reference values while also aiming to validate against the data range spanned by multiple sources where possible (see section 2.2.2).

The tool could also be used to compare different sources of reference data by treating them as different models and performing a model-intercomparison akin to the one presented in section 3.1.2.

***Comment**: 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".*

4.8 Response: Thank you for this comment. We agree and have changed the text accordingly.

Literature

Wilson, C., Guivarch, C., Kriegler, E., Van Ruijven, B., Van Vuuren, D.P., Krey, V., Schwanitz, V.J., Thompson, E.L., 2021. Evaluating process-based integrated assessment models of climate change mitigation. Climatic Change 166, 3. https://doi.org/10.1007/s10584-021-03099-9