Response to Anonymous Referee #1

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: *This reviewer's understanding of piamValidation and the study presented in 'Validation of climate mitigation pathways' is that the package serves as a crucial tool for improving the reliability of Integrated Assessment Models (IAMs). While IAMs are essential for shaping climate policy, they often face criticism for their lack of transparency and their limited ability to account for real-world technological advancements. piamValidation aims to address these concerns by systematically comparing IAM scenario data against historical observations, feasibility limits, and other model results. The package is designed for ease of use, requiring minimal coding to generate interactive HTML reports with heat maps, which encourages broader adoption and the development of more realistic near-term scenarios. Its effectiveness is demonstrated using the REMIND model, highlighting its ability to detect emerging technological trends that diverge from expected patterns—such as developments in carbon dioxide transport and storage, electric vehicles, and offshore wind power. Clear visual feedback, including 'traffic light' evaluations, helps model developers implement meaningful improvements.*

1 Response: We would like to thank reviewer #1 for the comprehensive and insightful review. The level of involvement and depth of feedback has substantially helped us enhance the quality of our manuscript.

Comment: *If this interpretation is correct, then this reviewer has identified weaknesses and several areas for improvement.*

First, the scope of validation variables and case studies appears limited. The current application primarily focuses on select technologies and the REMIND model. To enhance the tool's applicability and robustness, this reviewer suggests expanding its scope to include other sectors, variables and case studies with different IAMs (e.g. MESSAGE, GCAM; https://www.ngfs.net/ngfs-scenarios-portal/glossary/#IAM). A broader range of applications would help demonstrate the versatility of piamValidation and strengthen its reliability across diverse modeling frameworks.

1.1 Response: Thank you for this valuable comment. We agree that additional application cases can further demonstrate the tool's applicability and strengthen its robustness. Therefore, in the revised manuscript, we will add a section on the general applicability of the tool. In addition, we support the reviewer's suggestion to apply the tool to open-source NGFS scenarios, and we show an example below.

NGFS Application

The following paragraph demonstrates the versatility of piamValidation by performing four types of validation checks on the NGFS scenarios v5.0 (https://zenodo.org/records/13989530). This validation exercise is a qualitative one, focusing on the type of checks performed rather than the exact selection of threshold values. This implies that threshold violations do not indicate limitations in the scenario data but rather illustrate how the tool can be used to identify specific patterns. Furthermore, the plots use additional colors to indicate whether upper or lower thresholds are violated via the function argument "***extraColors = TRUE***" when calling "***validateScenarios***".

The corresponding validation configuration file for these application cases and the markdown file to create the plots below are available on GitHub piamValidation.

1. Validation overview for multiple models.

The heat maps of the validation report are able to represent four dimensions: in Figure 1 below for instance, the dimensions are scenarios, years, variables and models. The model's dimension is shown when another dimension has only one value, either the region (here: "World") or the variable. One conclusion that can be drawn from this visualization is that the near-term dynamics of CCS are consistently flagged across all models: most tend to underestimate the 2020 data point while overestimating the 2030 value in many scenarios.

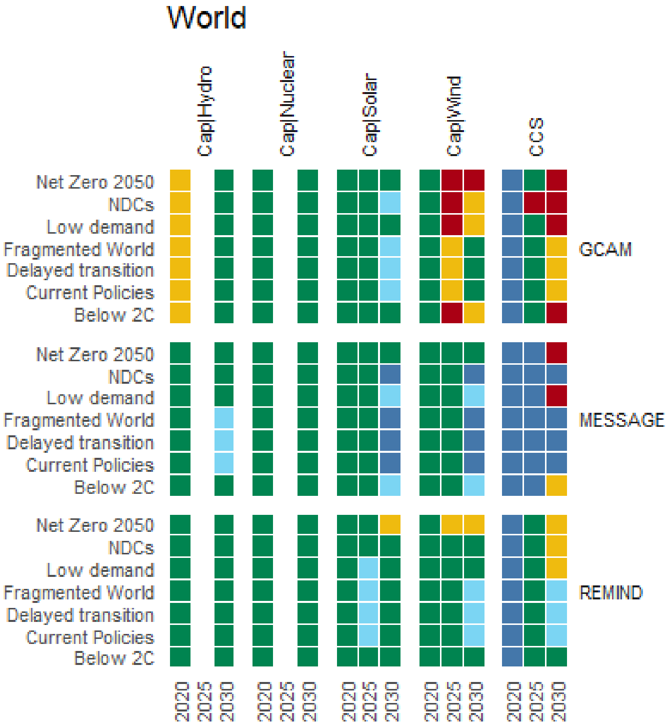

Figure 1: NGFS model overview

2. Model intercomparison

The piamValidation package allows for model intercomparison exercises by selecting one model as the reference (here: MESSAGEix-GLOBIOM). In this example, we examine global CO2 emissions and identify occurrences of REMIND or GCAM deviating more than 20% (weak threshold) or 40% (strong threshold) from MESSAGE within each scenario.

The heat map in Figure 2 a) reveals that the strongest deviations appear after 2050, with REMIND and GCAM showing lower emissions than MESSAGE. However, as emissions drop closer to zero, the relative differences being used as thresholds make up smaller absolute values. This becomes clearer when looking at a line plot of a specific scenario (here: “Delayed transition”) and seeing a “closing” funnel (see Figure 2 b).

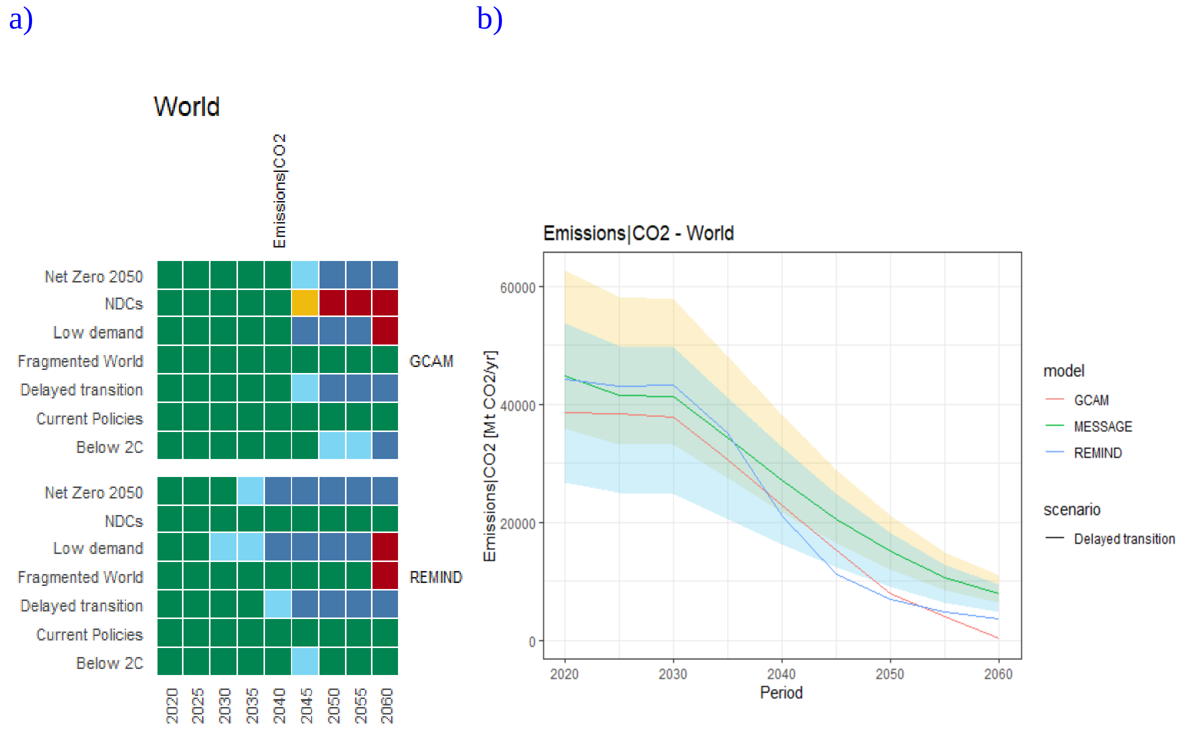

Figure 2 NGFS relative model intercomparison in relation to MESSAGEix-GLOBIOM

Users who want to avoid this case can choose the “difference” metric instead of the “relative” one to define constant thresholds around the reference model. Applying a buffer of +/- 5/10 Gt CO2/yr results in a validation outcome as shown in Figure 3 a) and b).

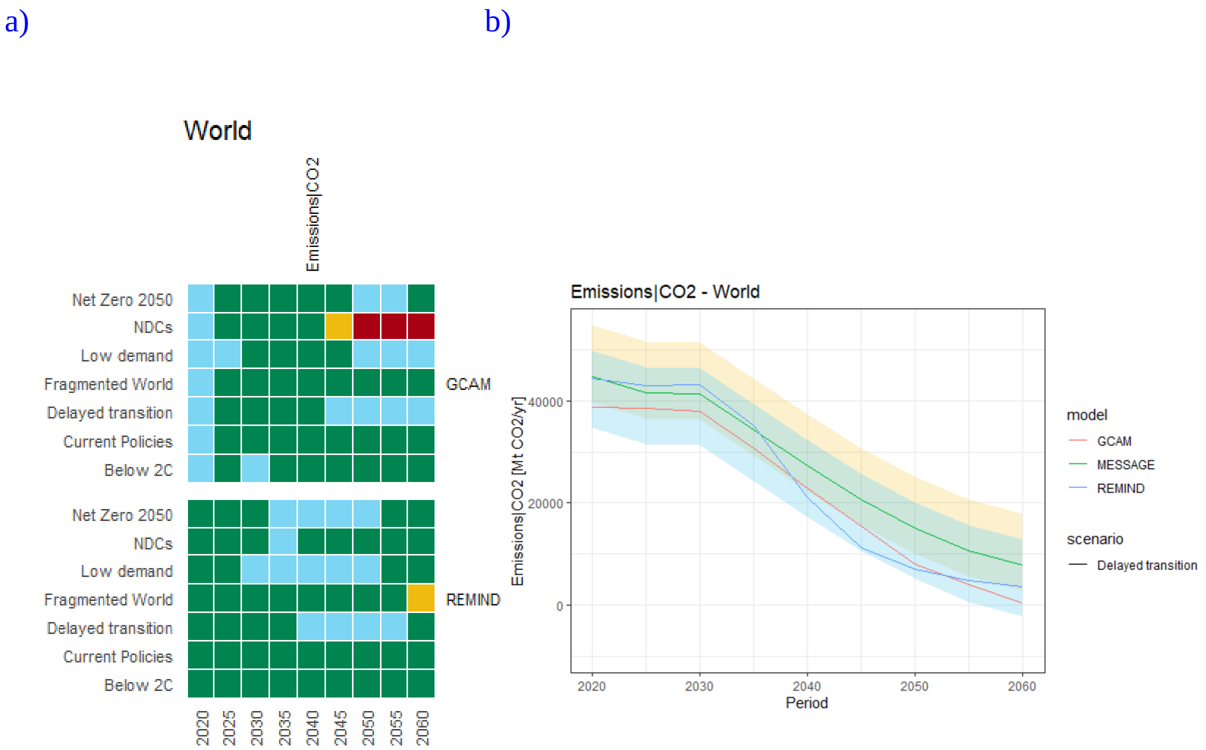

Figure 3 NGFS absolute model intercomparison in relation to MESSAGEix-GLOBIOM

3. Scenario intercomparison

In a similar fashion, scenarios can also be compared with each other (here: the reference is the “Below 2C” scenario). Consistent with the underlying scenario narratives, more ambitious scenarios such as “Net Zero 2050” and “Low Demand” are characterized by lower CO_2 emissions, whereas less ambitious scenarios such as “Current Policies” and “Fragmented World” exhibit higher CO_2 emissions. This application case can serve as a straightforward means of conducting a preliminary plausibility check of scenario narratives.

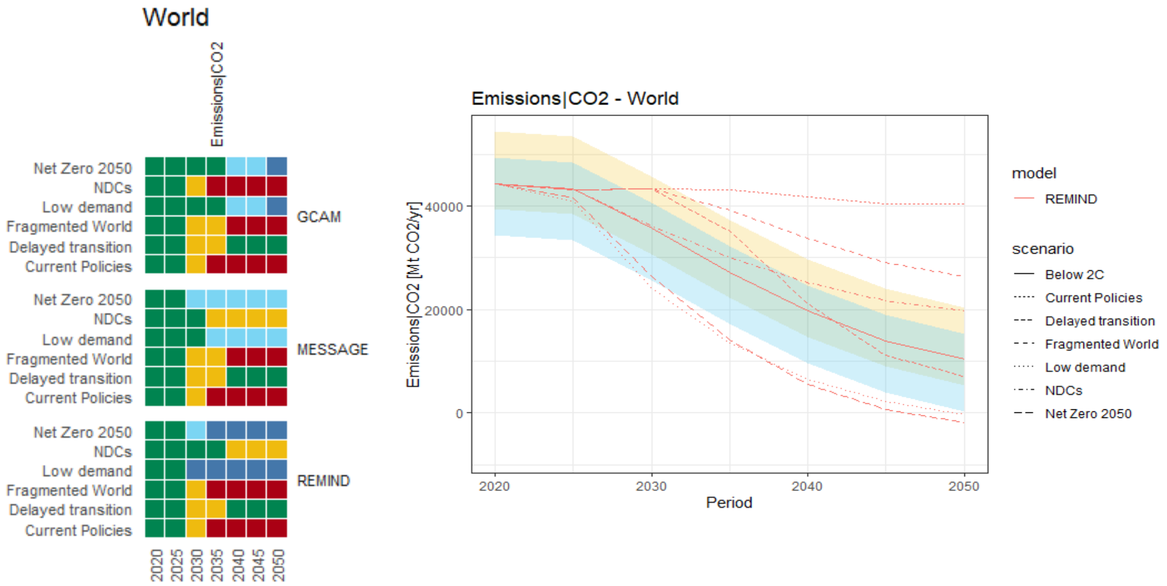

Figure 4 NGFS scenario intercomparisopn

4. Period intercomparison

Finally, periods can also be selected in relation to one another. This example checks whether the periods 2025 and 2030 compared to the period 2020 are between -20% and 0% (weak thresholds) or -40% and +10% (strong thresholds). Note that this case also demonstrates the option of choosing asymmetrical thresholds.

a) b)

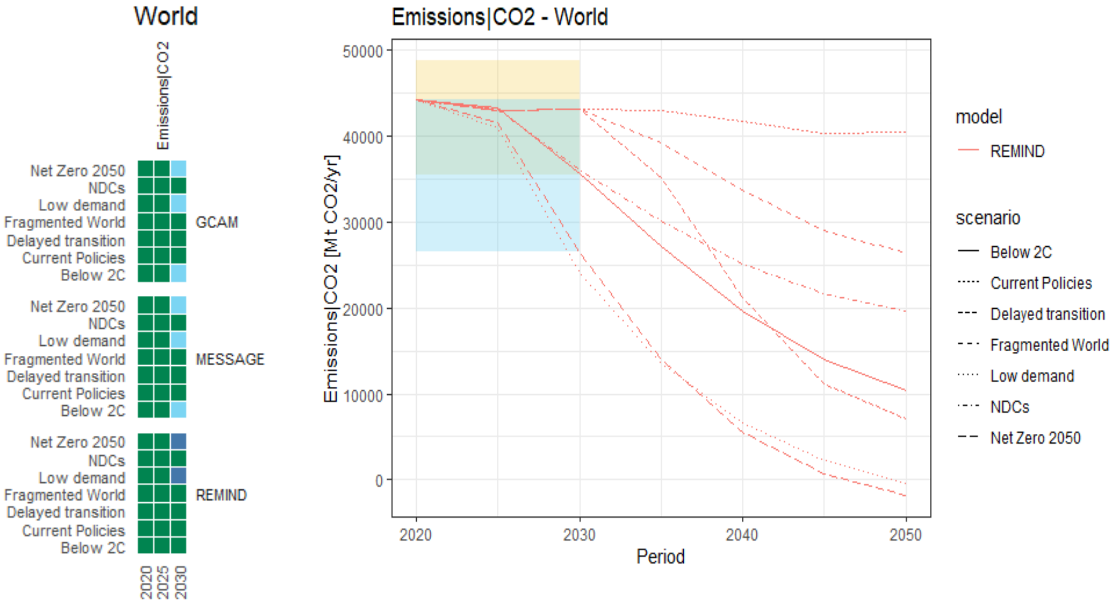

Figure 5 NGSF period intercomparison

Comment: *Second, the effectiveness of the validation is inherently dependent on data quality and methodological transparency. The reliability of the process hinges on the accuracy and robustness of observational and benchmark data. This reviewer calls for a more detailed discussion on managing uncertainties in reference datasets, along with a deeper technical explanation of how validation thresholds are determined, particularly for complex or uncertain data. In addition, this reviewer suggests that the authors incorporate metadata quality indicators for input reference datasets and establish a shared, moderated repository of standard validation thresholds to enhance transparency and reproducibility.*

1.2 Response: Thank you for this valuable remark. In the revised version of the manuscript, we will place greater emphasis on how the usefulness of the tool and the quality of its results depend on the reference data employed and the thresholds selected.

Through the tool's configuration file, the applied thresholds and data source references are published and ensure reproducibility. An open discussion with other IAM teams and interested stakeholders regarding the selection of data sources and the determination of thresholds has already been initiated for the REMIND application case, as documented here: https://github.com/pik-piam/mrremind/discussions.

To tackle the uncertainty around specific sources, the tool is able to take several sources for the same threshold. With the configuration ref_model "***range(sourceA, sourceB)***", the flag will appear only when the data is beyond the threshold for sourceA *and* for sourceB. Use-cases include allowing different models to rely on different sources or smoothing yearly data so that they ignore short-term disruptions (like the 2020 variations due to the pandemic).

In addition, the platform https://pik-piam.github.io/piamValidation/ provides systematic procedures and standardized benchmarks to ensure transparency and reproducibility of the results.

Comment: *Third, technical barriers and user accessibility warrant further consideration. While piamValidation is designed for ease of use, its reliance on R and the IAMC data format may present challenge*s for users unfamiliar with these tools. To broaden accessibility, this reviewer suggests providing more guidance for non-R users or exploring interfaces for alternative platforms, such as Python. Expanding compatibility across multiple programming environments would help ensure that a wider audience - including researchers and policymakers with varying technical backgrounds - can effectively use the tool.*

1.3 Response: Yes, we agree that the tool should also be used for non-R users. As the purpose of the piamValidation tool is to validate IAM data, the validated data should be in the standard IAM format, as this is the format of open IAM scenario data. The manuscript can be used as a step-by-step tutorial for non- R users as it explains the steps for:

1. Installing R environment: install R and the integrated development environment RStudio 4. Download the freeware here: R https://www.r-project.org/ and RStudio https://posit.co/products/open-source/rstudio/

2. Installing piamValidation:
   install.packages("piamValidation",repo="https://rse.pik−potsdam.de/r/packages")
3. Single execution command:
   validationReport(c("path_to_IAM_data","path_to_ref_data"),"path_to_your_config")

This procedure enables users with no prior R experience to install the necessary environment, run *piamValidation*, and obtain a full validation report with minimal interaction.

Comment: *Fourth, the future directions are not entirely clear. While ongoing development is mentioned, a more structured roadmap outlining planned enhancements would be beneficial. This reviewer suggests specifying future improvements, such as incorporating machine learning techniques, expanding the range of validation variables and integrating the tool with additional modelling frameworks. In addition, extending validation metrics to assess long-term feasibility and policy robustness would strengthen the tool's relevance for decision-making in climate policy.*

1.4 Response: We appreciate the reviewer's thoughtful comment and will restructure the outlook in the revised manuscript. In terms of future developments, the highest priority is the expansion of validation variables and refinement of existing thresholds. This requires regularly updating the data sources that we currently use and monitoring and evaluating new data sources to integrate them into the tool.

Rather than focusing on expanding the validation tool towards other languages or data frameworks, the development direction is set on making the tool more stable, precise, and intuitive to use. At this stage, we do not see a clear benefit from incorporating machine learning methods and prefer to rely on established data sources. Although highly desirable, extending the validation period to enable a long-term feasibility assessment in this application is currently constrained by the limited availability of reliable reference data.

First attempts were started to compare IAM scenarios to other energy-focused projections, such as the IEA "Net Zero by 2050" scenario. However, due to methodological differences, e.g., in sector definitions, these comparisons have sparked limited interest so far.

Comment: *Fifth, the discussion on tool limitations lacks depth. A more comprehensive examination of potential biases, uncertainties and challenges in applying piamValidation across diverse IAMs would strengthen the manuscript. This reviewer suggests providing clearer guidelines for identifying and mitigating these issues, ensuring that users can navigate the tool's constraints effectively. Addressing these limitations in greater detail would enhance transparency and reinforce the reliability of validation outcomes.*

1.5 Response: Thank you for pointing this out. We restructure the section on limitation by providing a clearer distinction into the following three aspects:

1. Limitations indirectly related to the piamValidation tool,

2. tools limitation,
3. limitations regarding data management.

This restructuring will be similar to:

The piamValidation tool is subject to several limitations that fall into three broad areas. First, many aspects are indirectly related to the tool itself and instead depend on user and community choices. For instance, the identification of meaningful validation cases, the selection of appropriate reference data, and the definition of reasonable thresholds all substantially influence the outcome of the validation exercise. Although these challenges are not technical limitations of the piamValidation tool, they substantially influence the quality, consistency, and acceptance of the validation results. Second, certain limitations arise from the design of the tool. In particular, caution is required when thresholds are defined in terms of relative deviations and the validation values approach zero. Under such conditions, even very small absolute deviations can manifest as disproportionately large relative differences, complicating the interpretation of results. In these cases, the use of absolute deviations is preferable. Finally, the tool is constrained by challenges in data management. Harmonizing scenario data with reference sources often requires substantial effort to ensure consistency in units, definitions, sectoral coverage, and technological detail. This integration process can be time-consuming.

Comment: *Finally, additional specific comments are provided in the annotated manuscript file.*

1.6 Response: Thank you for the detailed language review. The revised manuscript will be reformulated accordingly where appropriate.

Comment: *This reviewer offers an overall endorsement and recommends acceptance, contingent on minor revisions to address the areas for improvement outlined above. These revisions would further enhance the paper's contribution to strengthening the credibility of IAMs.*