We thank the reviewers for their careful reading of the manuscript and their constructive comments and suggestions. We have revised the manuscript accordingly and provide below a detailed point-by-point response to all comments. In each case, we also indicate the corresponding changes made in the revised manuscript to address the reviewers' concerns.

Reviewer #1

Comment #1:

The key limitation of the approach (an assumed mapping between global mean temperature and impacts) - is acknowledged by the authors as limiting applibability to questions of path depedence, such as deep overshoot where warming patterns may significantly differ to a non-mitigation scenario. The extended work on overshoot impacts by many of the coauthors exemplifies this concern. Though this limitation is acknowledged - a useful extension would be to apply the technique to a scenario like SSP534over and gauge which variables are impacted by path dependency, and to what degree. Similarly, it would be useful to operationally test with SSP370 as a test of applicability to strongly differing aerosol emissions pathways. The authors could also consider what the optimal training set should be - e.g. is it an asset to keep SSP585 in the training set, if that scenario exhibits high warming rates which are unlike policy-relevant scenarios and it might actually reduce performance?

We thank the reviewer for this feedback. We agree that, while path dependence and aerosol effects are explicitly acknowledged as limitations of the method, their quantitative impact on emulator performance warrants further analysis.

To address this, we extended the leave-one-out validation framework described in Section 3 by conducting additional experiments using SSP1-2.6 and SSP3-7.0. The figures displaying the results of the additional experiments are shown in Appendix C of the paper. We additionally adjusted Section 3.2 to discuss and describe the additional figures. In addition to smaller edits generalizing the text in Section 3.2 to account for the extra validation experiments, we describe and discuss the additional figures in detail at the end of subsection 3.2.1:

"To investigate the effect of potential sources of bias on RIME-X, we repeat the validation framework using \textit{SSP1-2.6} and \textit{SSP3-7.0} as test scenarios, while training the emulator on the remaining scenarios. \textit{SSP1-2.6}, with its slight overshoot and prolonged stabilization period, provides a useful test case for assessing biases arising from path dependence, including changes in circulation patterns, land–atmosphere feedbacks, SST-driven pattern effects, and monsoon dynamics. In contrast, \textit{SSP3-7.0} allows us to evaluate the sensitivity of the emulator to aerosol forcing pathways that differ from those represented in the training scenarios. The resulting plots are provided in the Supplementary Material as Figures~\ref{fig:anmaemap_ssp126}, \ref{fig:anmaemap_ssp370}, \ref{fig:QQplots_ssp126} and \ref{fig:QQplots_ssp370}. The results show that, for \textit{SSP1-2.6}, biases increase, particularly for precipitation-related indicators. Minimum errors remain approximately unchanged for temperature and increase slightly for precipitation (+0.4 pp). Median errors remain approximately unchanged for temperature and increase substantially for precipitation (+1.3 pp). Maximum errors increase slightly for temperature (+0.7 pp) and more then double for precipitation (increases of up to +10 pp). The Q–Q plots (Figure~\ref{fig:QQplots_ssp126}) indicate that the increase in maximum precipitation errors is associated with a progressive shift of the simulated distribution away from the emulated distribution over time, suggesting the presence of path-

dependent effects. This behavior is consistent with the expected limitations and could, for example, be explained by continued ocean warming and associated circulation adjustments influencing regional precipitation patterns even after global mean temperatures stabilize. For \textit{SSP3-7.0}, we likewise observe a moderate increase in biases, particularly in regions strongly affected by aerosol forcing changes, including parts of Africa and Asia. Minimum errors decrease slightly for temperature (−0.1 pp) and increase slightly for precipitation (+0.3 pp). Median errors increase moderately for both temperature (+0.7 pp) and precipitation (+0.9 pp). Maximum errors increase more noticeably for temperature (+2 pp) and nearly double for precipitation (up to +7 pp). The error map plots (Figure~\ref{fig:anmaemap_ssp370}) indicate that these larger deviations are concentrated primarily in regions with strong aerosol-driven circulation responses, consistent with the expected limitations of the framework when emulating a scenario with substantially different aerosol forcing than the training scenarios. Overall, these results indicate that both path dependence and aerosol forcing differences can introduce systematic deviations in the emulated distributions. However, the resulting biases remain moderate in most regions and are consistent with the discussed limitations."

We chose not to include SSP5-3.4-OS in this validation, as the number of available CMIP6 simulations for this scenario is substantially smaller than for the other scenarios. This would introduce additional sampling uncertainty in the ground-truth distributions and limit the comparability of the error metrics to validation experiments with other scenarios.

Regarding the question on the optimal training set, we agree that this is an important consideration. In general, the training dataset should be as large and diverse as possible, covering a wide range of scenario characteristics. In particular, including high-warming scenarios is beneficial, as it ensures that models with lower transient climate responses also contribute information at higher GMT levels.

To clarify this point, we have expanded the discussion of training data limitations in the manuscript (Section 3.1) and added the following:

"As a result, higher-emission scenarios generally sample a broader span of GMT levels, which counteracts—at least to some extent—the reduction in available samples for any specific high-GMT bin. **The inclusion of high-warming scenarios such as \textit{SSP5-8.5} in the training data also provides additional coverage at high-GMT levels.**"

We have also clarified this point in the conclusions:

"Moreover, as fewer ESMs reach high warming levels, the robustness of emulations decreases with warming level due to reduced training data**, making the inclusion of high warming scenarios like \textit{SSP5-8.5} in the training data particularly useful for a wide applicability of the method.**"

Comment #2:

Out of sample validation of tempature/precip emulations are convincing and well presented. But while the model is demonstrated for its use in impact emulation (extremes, crop yields etc), and these results are highlighted in the abstract, validation is missing for these quantities. Some assessment of the out-of-sample performance of impact-relevant metrics would be very useful to demonstrate the model is viable for real-world applications. This is especially true given the

model use in the Climate Impact Explorer - where end-of-chain results are the focus, but they are not properly validated. For these end-user applications, an assessment of reliability as a function of impact would be invaluable. For the purposes of a development-model description paper, this is all understandable - but the title/abstract should be slightly adjusted to create a clearer impression of what is validated and what is not.

We thank the reviewer for this valuable feedback. We agree that validation of impact-relevant indicators is important, particularly in the context of end-user applications such as the Climate Impact Explorer.

Our primary validation focuses on CMIP6 temperature and precipitation, as these datasets provide sufficiently large ensembles to derive robust ground-truth distributions from the simulations (each simulation only provides one sample per timestep, so to retrieve a relatively robust time-dependent distribution for a particular scenario, we require a high number of simulations for that particular scenario) and to meaningfully quantify errors between simulated and emulated distributions. For many impact indicators (e.g., from ISIMIP), such validation is more challenging due to the limited number of available simulations, which can lead to substantial sampling uncertainty in the estimated ground-truth distributions.

Nevertheless, to address this point, we have added an additional evaluation of impact-relevant and extreme indicators included in the Climate Impact Explorer. Specifically, we compare time series of selected quantiles (median, 5th, and 95th percentiles) emulated by RIME-X and shown in the climate impact explorer against the corresponding simulated time series. To avoid cherry-picking from a large set of possible indicators, we restrict this analysis to the indicators already presented in the showcase section of the manuscript. The resulting figures can now be found in Appendix D of the revised manuscript. In addition to a reference to the additional validation in Appendix D in the Showcase section we added this description and discussion of the the additional validation in Appendix D:

“We provide an additional evaluation of the RIME-X emulator for the indicators presented in the Section~\ref{showcase}. The evaluation compares the empirical median and 90\% confidence interval derived directly from the ISIMIP ensemble simulations for the SSP3-7.0 scenario with the corresponding emulated time series generated by RIME-X in five year steps from 2020 to 2090, based on all available ISIMIP simulations as training data and the GMT timeseries retrieved from the SSP3-7.0 ISIMIP simulations. We use the SSP3-7.0 scenario, because it represents the closest available primary ISIMIP scenario to the CAT Current Policies scenario used in the showcase in terms of its approximately linear warming trajectory, although with a stronger warming signal and different regional aerosol distribution. In addition, SSP3-7.0 provides complete coverage across all 14 ESMs and broad availability of agriculture-model simulations in case of the the maize yield indicator. The evaluation Figure~\ref{fig:ISIMIP1} to Figure~\ref{fig:ISIMIP5} present results for 15 regions selected to represent the full range of emulator performance. For each of the 5th, 50th, and 95th percentile time series, we compute the average normalized mean absolute error (ANMAE; see Section~\ref{validationresultsm1}) across all timesteps and select the regions nearest to the minimum, 5th percentile, median, 95th percentile, and maximum of the resulting error distribution. Figure~\ref{fig:ISIMIP_countries} shows the ANMAE for each country and indicator. Some degree of discrepancy is expected because the empirical quantiles are estimated from only 14 ESM simulations, whereas the emulated quantiles are derived from all available simulations within the warming levels corresponding to the SSP3-7.0 trajectory for a given year. In particular, for extreme precipitation

and for cooling degree days in Greenland (GRL), sampling uncertainty appears to account for a substantial share of the error. This is reflected in the relatively noisy empirical time series (Figure~\ref{fig:ISIMIP4} and the Greenland panel in Figure~\ref{fig:ISIMIP3}). Despite this sampling uncertainty, the evaluation shows that RIME-X reproduces the main characteristics of the evolving distributions across indicators and regions. The ANMAE remains below 10\% in most cases and is often substantially lower (Figure~\ref{fig:ISIMIP_countries}). The largest deviations occur for the 95th percentile of the maize yield indicator in the highest-error regions (Figure~\ref{fig:ISIMIP2}). These discrepancies can largely be attributed to the impact model DSSAT-Pythia, which does not provide values in SSP3-7.0 but exhibits comparatively strong yield responses under SSP5-8.5 in the high-error regions Pakistan (PAK), Liechtenstein (LIE), and Slovenia (SVN). As a result, these high values enter the RIME-X-derived distributions but are absent from the corresponding SSP3-7.0 reference ensemble, leading to inflated errors in the 95th percentile in these cases. Overall, maize yield change and extreme precipitation exhibit the highest error levels across indicators (Figure~\ref{fig:ISIMIP_countries}). This is consistent with the known limitations of the RIME-X approach discussed in Section~\ref{theoretical_limitations}. In particular, extreme precipitation is, in addition to being sensitive to sampling uncertainty, strongly influenced by circulation changes, sea-surface-temperature-driven pattern effects, and monsoon dynamics, which may not be fully captured by a GMT-conditioned emulator. Similarly, maize yield changes depend not only on GMT but also on CO$_2$ concentration effects, which may introduce additional bias despite the strong correlation between GMT and CO$_2$ concentration in the scenarios considered."

Finally, following the reviewers suggestion, we have clarified in the abstract that the primary validation focuses on temperature and precipitation from CMIP6, while results for impact indicators are demonstrated and qualitatively evaluated.

"By jointly quantifying global and regional sources of uncertainty from the start, RIME-X enables systematic exploration of the full space of plausible regional climate impact trajectories under different emissions scenarios. The method is **conceptually** applicable to **all** regional indicators whose distributions are predominantly determined by warming level and provides a computationally efficient framework for uncertainty-aware regional indicator emulation. **We evaluate the method using out-of-sample validation on temperature and precipitation simulations from the Coupled Model Intercomparison Project 6 (CMIP-6) and demonstrate its applicability to a range of impact-relevant and extreme event indicators derived from the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP).** We provide an open-source Python implementation of RIME-X, including preprocessing workflows for data from ISIMIP and support for user-defined indicators."

Comment #3:

Code is generally well written - seperation of preprocessing and emulation, well structured config files and dependencies. However - the lack of unit testing is a significant concern for robustness of future development - and I would highly recommend the authors to implement a formal test suite. More documentation & tutorials would also improve the usability of the codebase for new users.

We thank the reviewer for this helpful feedback. We agree that a more comprehensive testing framework and improved documentation are important for the robustness and usability of the codebase.

We are currently extending the repository to include a formal unit and regression testing suite, as well as additional documentation and user-oriented tutorials. These improvements, along with further functionality, like enabling changing socioeconomic conditions throughout the emulations, applicability to more datasets, and preprocessing of user-defined region masks from shapefiles, will be incorporated into the public repository over the course of this year.

Comment #4:

The code also seems to slighly differ from the paper in the mapping of GMT distributions to impact distributions. The paper describes a Monte Carlo sampling of temperatures from the SCM ensemble distribution, for each sample drawing corresponding impact values from the conditional distribution. The code instead converts the GMT ensemble into quantiles, then uses linear interpolation between pre-computed GMT bins to obtain impact values. This is a sensible approach, but slightly conceptually different from the paper's description and relies on the validity of linear interpolation between GMT bins. This should be discussed.

We thank the reviewer for this comment. The apparent difference arises because there are two versions of the code implementing RIME-X.

The version referred to here (https://github.com/iiasa/rimeX/blob/main/rimeX/emulator.py) is the first explicit implementation of the method, which deterministically calculates the quantiles of the marginal distribution.

However, this version proved difficult to parallelize efficiently on a full grid, which is required for the Climate Impact Explorer. For this reason, we developed a second implementation in (https://github.com/iiasa/rimeX/blob/main/rimeX/preproc/quantilemaps.py). In this version, the function make_quantile_map_array is a preprocessing step that uses regional averages and warming level information from the MIP to create “quantile maps” representing the conditional distributions given warming levels, as described in the Implementation section of the paper. The emulation is then performed by make_quantilemap_prediction, which applies the sampling procedure using the SCM distribution together with the precomputed quantile map. We intend to retain both implementations of the method, as the explicit version proved valuable for potential future use cases of the emulator that require a sampling strategy from the RIME-X distribution. However, to avoid confusion about the implementation, in the repository extension outlined in response to Comment #3 we will provide more comprehensive documentation and tutorials to better explain the codebase and improve users’ ability to effectively apply it.

Apart from small sampling differences, which quickly become negligible (we tested with up to 10,000 samples and typically found <1% difference between outputs) both implementations are output-equivalent. The advantage of the quantile-map version is that it can be applied gridpoint-wise in parallel, enabling full spatial emulation and producing distribution quantiles for each grid point, as shown in Figure 5 in the manuscript.

Comment #5:

Compound events - RIME-X is demonstrated to produce marginal distributions for temperature and precip conditional on global temperature. A useful extension would be to think about joint conditional distributions of multiple variables - which cannot be trivially inferred from the single distributions. This should be discussed as a limitation (as in a caveat not to trust the implicit joint distributions from the current model), and a possible future extension could be to do this formally.

We thank the reviewer for this suggestion. This is indeed a very relevant point, as not capturing spatial and temporal dependencies between distributions is a limitation to the applicability of the method, although it does not affect the accuracy of the marginal emulations. We have therefore added this explicitly as a limitation in the Conclusion, including the following sentence:

"This epistemic uncertainty is, however, partly mitigated through model weighting and GMT-based constraints. **Additionally, the method does not capture temporal, spatial, or inter-variable correlations between the emulated distributions of multiple variables, regions, or time steps, even when they originate from the same underlying data source, meaning that dependencies across regions, variables, and time steps are not preserved between several emulated distributions."**

At the same time, we note that there is a practical way to emulate compound indicators within the current RIME-X framework. Specifically, compound metrics can first be computed from the underlying climate model data, thereby fully capturing spatial, temporal, and cross-variable dependencies, and can then be emulated as a single derived indicator using RIME-X. For example, wet-bulb temperature can be calculated from the raw simulation output of temperature and humidity and subsequently emulated as a compound variable.

Finally, we agree that a more general treatment of joint distributions across multiple variables, regions, or time steps would be highly valuable and methodologically feasible. We have therefore added this as a direction for future work in the Conclusion:

"Future work could address some of these limitations—for instance, by conditioning indicator distributions on changing socioeconomic variables like GDP or population, improving robustness at high warming levels using training data with prescribed GMT shifts like, e.g., from the upcoming Tipping Points Modelling Intercomparison Project \citep[TIPMIP;][]{winkelmann2025tipping}, testing additional global predictor indicators such as the likelihood of being before or after peak GMT**, or extending the framework to produce multivariate output that captures spatial, temporal, or cross-indicator correlations particularly relevant for the analysis of compound events."**

Comment #6:

The 21-year running mean is deeply embedded as part of the pre-processing step, but this step smooths over annual variability and infrequent features, and potentially cutting off an important part of the impact space associated with even mildly extreme years. The implication of this should be discussed more - how can the framework inform about once per decade events which might in fact be the dominant impacts on infrastructure - and this could be an avenue the authors explore in future versions.

We thank the reviewer for this comment. The 21-year running mean is only used as a definition for global mean temperature (GMT) in the framework. For individual regional indicators, the temporal

aggregation is configurable via the running_mean_window parameter in the configuration files, which is set to 21-years by default as this choice was made for the Climate Impact Explorer.

However, this is not a structural limitation of the method. Users can choose different aggregation windows, and in the implementation, it is even possible to compute distributions for individual months. In principle, there is no constraint preventing even finer temporal resolutions, such as daily values, depending on the underlying input data and the intended application. This is not included in the implementation, though.

At the same time, there is a high level of structural flexibility for the indicator definition within the RIME-X framework. In principle, also 1-in-20 year extreme events can be identified and mapped, using a running window approach. However, such approaches would require large training data sets, as there is a clear trade-off here with stochastic natural variability. In practice, the available training data limits what's possible, and a running mean approach has proven to be numerically robust. This is a structural limit of a mapping instead of a generative emulator approach.

Comment #7:

It would be useful to provide guidance for end-users on where the model is currently reliable, where it's merely informative and where it's just hypothetical. The provided validation indicates that the model is relatively robust at coarse-scale: temperature/precip projections under non-overshoot scenarios. Downstream Impacts, and high-frequency outputs are perhaps less robust - and it would be useful for end-users to flag these as areas in development, as well as scenarios where the underlying assumptions would be expected to fail (large aerosol signals, SRM, large overshoots, impacts dominated by land use assumptions). Laying these out (especially in the impact explorer) would strengthen the model framework in terms of providing areas of operational applicability.

We thank the reviewer for this helpful suggestion. We agree that providing clearer guidance on the domain of applicability is important for end users. However, we consider a strict separation into "robust", "informative", and "hypothetical" usage categories not to be robustly definable in a general and transferable way, as applicability depends on the specific indicator, region, scenario, aggregation scale, and used case.

Instead, we prefer to give an extensive list of possible features of indicators, regions, and scenarios that could introduce biases or corrupt the emulations. We have strengthened the manuscript by further expanding the discussion of limitations and scenarios where the underlying assumptions are expected to break down. Some of these limitations were already included, and we have now added the additional cases highlighted in the comment. Specifically, we have expanded the text as follows (bold additions):

"Though supported and applied in the literature for many indicators \citep{herger2015improved,schleussner2016differential,hausfather2022climate}, this assumption can introduce biases, particularly for overshoot or stabilization scenarios, and in regions **or indicators** affected by pattern effects, high regional aerosol concentrations, **changing circulation patterns, land–atmosphere feedbacks, SST-driven pattern effects, monsoon dynamics**, or changes in regional land use \citep{schleussner2024overconfidence,pfleiderer2024limited,wells2023understanding,giani202

4origin}. **These limitations also restrict the applicability of our approach to idealized model experiments deploying solar radiation modification.**"

In addition, we have strengthened the conclusions by providing a consolidated list of limitations and applicability constraints:

"However, some limitations can affect RIME-X's accuracy. It assumes that the regional indicator depends primarily on GMT, excluding influences like aerosols, **shifts in socioeconomic conditions, land use, or circulation patterns (including SST-driven pattern effects, monsoon dynamics, and land-atmosphere feedbacks)**. This limits applicability **to regions or** indicators where this assumption approximately holds. It also does not account for the path to reaching GMT levels, limiting applicability in scenarios **or indicators** where time-lag effects play a major role for the assessment of regional indicators, such as overshoot scenarios. **In addition, the framework may be less robust in scenarios where regional aerosol forcing differs strongly from the training scenarios**. **Our approach is also not validated for idealized model experiments deploying solar radiation modification.**"

Reviewer #2

Comments related to the Introduction

Comment 1:*L60+ the rest of the intro is really model description; I get that you want to preview the method here but this feels like overkill. Readers scanning through won't come to the intro for this information.*

We thank the reviewer for this comment. We understand the concern, but we prefer to retain this level of methodological preview in the introduction, as it helps readers quickly understand the core idea and positioning of the framework before entering the Methods section. To address the issue of clarity and scannability, we have added a short orienting sentence at the beginning of the methodological preview in the introduction of the revised manuscript to make explicit that this part serves as a brief overview rather than the full methodological presentation:

"To overcome these challenges, we present the Rapid Impact Model Emulator Extended (RIME-X) — a novel, top-down probabilistic emulator designed as a probabilistic extension to the deterministic Rapid Impact Model Emulator (RIME) \citep{byers2025fast}. **Both approaches are briefly outlined here, while a detailed description of RIME-X is provided in Section~\ref{methodologysection}.**"

Comment 2:

L80 I think "assumption-light" may be a bit much if you don't include some of the assumptions this does rely on – e.g. that GWL timeslices are more appropriate than simple pattern scaling. You go into the assumptions later so maybe worth referring to that section here, or move this to that section.

We thank the reviewer for this suggestion. We removed "assumption-light" from the sentence in the revised manuscript.

Comment 3:

I think it'd be worth noting more clearly other couplings of SCMs with regional emulators which are relevant to your work, e.g. the MESMER-MAGICC coupling https://gmd.copernicus.org/articles/15/2085/2022/ and STITCHES https://esd.copernicus.org/articles/13/1557/2022/ (which as I understand it is designed to be driven with any GMST; you cite this but not really in this context).

We thank the reviewer for this helpful suggestion. We agree that it is important to more clearly acknowledge existing approaches that couple SCMs with regional emulators, such as the MESMER–MAGICC framework, and to better place our work in this context.

In the revised manuscript, we will explicitly reference those approaches and clarify the role of SCM–emulator coupling in enabling bottom-up uncertainty analysis. In particular, we will revise the sentence in lines 53–55 as follows:

Original:

“These methods support a more holistic bottom-up exploration of scenario uncertainty, global climate response uncertainty, model uncertainty, and natural variability for a set of regional indicators by repeatedly emulating those indicators for multiple scenarios and GMT pathways using emulator instances calibrated on different ESMs.”

Revised:

“Coupling these methods with SCMs enables a more holistic, bottom-up exploration of uncertainties by deriving a probabilistic set of GMT pathways across scenarios from SCMs and repeatedly emulating regional indicators along these pathways using emulator instances calibrated on different ESMs \citep[e.g.][]{Beusch_Nicholls_2022, tebaldi2022stitches, Schwaab_al._2024}.”

Comments related to the Methodology

Comment 1:

Figure 2 – on panel b at the bottom, are dots the best way to show this? why not a distribution? If it must be dots, would suggest adding some noise to the y value so the density can be more easily seen; it doesn't convey much info currently I think. Also, the cmap used is quite narrow; it's hard to see much difference across the decades. Would consider using one with more contrast if possible. Panel a legend should be "ensemble" singular I think, and you don't need to have "reached in" in each one, in fact I don't think you need any of the decadal labels in either plot as the reader can see from the dots in a what they correspond to.

We thank the reviewer for this helpful suggestion. In the revised version of the manuscript, we improved the figure by using a colormap with higher contrast to better distinguish the decadal differences. We will also improve the labels, as suggested.

The distributions shown in panel (b) are directly derived from the points shown in the same panel; therefore, a distribution-based visualization is already available. The dots are intentionally kept to provide a visual link between panel (a) and panel (b), allowing readers to trace how the ensemble samples map into the resulting distributions.

Comment 2:

L135 in the equation, why is it $T_i \leq GMT < T_i + \Delta T$ and not $T_i - \Delta T/2 \leq GMT < T_i + \Delta T/2$ ie why is the GMT of the bin not in the middle of the bin? Maybe as delT is small this isn't important (though this can be changed by the user I understand, so it could be) but seems in theory to be the less obvious choice.

We thank the reviewer for this question. This is primarily an implementation choice driven by how warming level bins are defined in the configuration. In the current setup, users specify a minimum and maximum warming level together with a fixed warming level step size.

With this definition, bins are constructed directly from the specified bounds and step (e.g. min = 0, max = 5.0, step = 0.1), which yields bin edges aligned with these values (i.e. 0.0, 0.1, ..., 5.0). If instead a centred bin definition of the form were used, this would either shift the effective bin centres to values such as 0.05, ..., 4.95, which is less intuitive given the user-defined

configuration, or it would require including values outside the specified minimum and maximum warming levels.

For consistency with the configuration interface and interpretability of user-defined warming level grids, we therefore use the left-inclusive bin definition .

Comment 3:

L152 "Note that we only have the ability to collect several values of IM in bins BIM(GMT) and extract the distribution because we discretized the GMT variable." Is nearly a repeat of L139 "Note that we can only define probabilities for GMT levels like we do here because we discretized the GMT variable. "- I think this could be rephrased or just kept once, and notedagain in the discussion.

We thank the reviewer for pointing this out. In the revised manuscript, we have consolidated the two sentences into a single statement at the end of Section 2.2.3 to avoid repetition and improve clarity.

Comment 4:

Fig 3 are the data points on the left below the lowest of the cbar? They seem lighter – which I guess could be the case as 1C GMT in Germany might be before 2005-15. I'm not clear on why the number of points in the distribution is decreasing as temperatures increase – what determines this? Fewer input model runs?

We thank the reviewer for this observation. We checked this, and the data points on the left correspond to a GMT level of 0.9°C and should therefore match the minimum of the colorbar.

Regarding the decreasing number of samples at higher GMT levels, this is indeed due to the reduced number of available simulated data points in higher warming ranges. This effect arises from the fact that fewer model simulations reach higher GMT levels. We have already discussed this limitation in the limitations and discussion sections of the manuscript.

Comment 5:

Fig 6 don't think you need "value" on x axes. By the time we're down to province scales (bottom graphs), we run into issues with the data, right? Since you use ESM data, the province won't lie cleanly over gridcells, and may be much smaller than a gridcell? Just worth noting.

We thank the reviewer for this comment. We agree that the label "value" on the x-axis is unnecessary and will remove it.

Regarding the spatial scale, the province shown in Fig. 6 is Guangdong, which is relatively large (only slightly smaller than Cambodia and slightly larger than Uruguay). At this scale, the mismatch between administrative boundaries and ESM grid cells is limited, although not entirely negligible.

That said, we fully agree that this can become an issue for smaller provinces or regions, as well as for some countries with complex geometries or sizes comparable to or smaller than individual grid cells. Thus, we will add the following sentence to the list of limitations in the Conclusions of the

revised manuscript:

“Finally, the applicability of the method at finer spatial scales is limited by the resolution of the underlying ESM input data, as coarse model grids may not adequately resolve smaller regions or complex boundaries.”

Comment 6:

Fig 5 unsure on the x labels – it’s the change in TXx in deg C right? So units should be something like “TXx cf 2020 (deg C)”

We thank the reviewer for the comment. Yes, the x-axis represents the change in TXx relative to the year 2020. We have revised the label accordingly to “Change in TXx relative to 2020 (°C)” to make both the variable and units explicit.

Comment 7:

When you use ISIMIP data, it’s bias-corrected, right? How might this affect things compared to raw CMIP output?

We thank the reviewer for this question. The climatic ISIMIP data used in this study are bias-corrected while the ISIMIP data stemming from the impact models is not bias-corrected. As systematic biases present in, for example, raw CMIP6 output are reduced through this process, we expect that using bias-corrected data also reduces systematic biases in the emulated distributions, thereby improving their accuracy and applicability.

Comment 8:

Table A1 – the Experiments column isn’t clear to me – this is the total members across the scenarios listed? It’d be good to see this broken down by scenario for clarity, which you could do in the scenario column.

We thank the reviewer for pointing this out. The number reported in the “Experiments” column corresponds to the maximum number of ensemble members available for a single scenario, rather than the total across all listed scenarios. As this number can differ between scenarios, we agree that the current presentation may be unclear.

In the revised manuscript, we therefore edited Table A1 to improve clarity by adding the number of ensemble members for each scenario directly in the “Scenario” column (in brackets). This will make the distribution of experiments across scenarios more transparent. We also noticed that the Table didn’t list the full amount of models we used for our validation experiments, so we added all additional models we used to the Table, which forced us to split Table A1 into Tables A1 and A2.

Comment 9:

Table A2 you also show TXx in section 2.4 – shouldn’t this be in the table too? Or clarify it’s just Fig 6.

We thank the reviewer for pointing this out. He is absolutely right, we included TXx in Table A2 (Table B1 in the revised manuscript) in the revised version of the manuscript to ensure consistency with Section 2.4.

Comments related to the Discussion and Validation

Comment 1:

L255 this is a great bit of discussion; but I think it's worth noting this increased uncertainty only covers the training scenarios – so if someone looked at a scenario outside this range – eg extreme NETs or regional aerosols (even eg SRM), this uncertainty wouldn't cover the effects because they're not accounted for mechanistically. This isn't a weakness of your model (if anything it shows its robustness) but this in-sample vs out-of-sample issue is key and I think you should bring this into the discussion. Similarly, one could argue that an increased uncertainty may is a limitation, if the variations driving that uncertain could in theory be taken into account (again, just worth noting).

We thank the reviewer for this constructive addition. We agree that the distinction between in-sample and out-of-sample scenarios is important and should be made explicit in the discussion, we will thus add the following to L255 in the revised manuscript:

"However, when all scenarios and years **of a training dataset** are included in the calibration, any loss of predictability due to time lag or related effects will manifest as increased uncertainty (i.e., wider error bars) rather than as an overconfident estimate**, as long as the emulated scenario is not a strongly out-of-sample case with respect to the training data, for example, when emulating scenarios involving high-overshoot pathways using a training dataset that does not include such processes."**

We agree with the reviewer that increased uncertainty can, in itself, be interpreted as a limitation—particularly in cases where the underlying sources of variability could in principle be explicitly represented rather than reflected as broader distributions. At the same time, the sentence at L255 is formulated in an informative manner, aiming to describe how such limitations manifest within the emulator (i.e., as increased uncertainty rather than overconfident estimates), rather than to evaluate them normatively. For this reason, we retain the original phrasing in the manuscript, while clarifying the scope of its validity as discussed above.

Comment 2:

L263 "This effect is partly mitigated by" I don't think this is true – it's a separate issue. I'm not sure what the argument here is to mean a wider GMT range from the SCMs counteracts the loss of samples – but am open to being convinced.

We thank the reviewer for this comment. The logic is that a wider GMT range means that more GMT bins are considered when constructing the final marginal distribution. As a result, even though individual conditional distributions are likely estimated from fewer samples per bin at higher warming level, the total number of samples contributing to the marginal distribution does not decrease as strongly as the sampling density within each GMT bin, because more GMT bins are included due to the wider GMT range. We therefore think that the larger GMT range at higher

emissions levels partly mitigates the loss of samples in each individual bin at higher warming levels.

Comment 3:

L281 worth being clear you don't use 1500 runs (right?) – maybe say how many you use here. Also Table A1 mentions SSP119 which you don't list here; but you don't mention SSP460 in the table – so maybe these are mixed up? In any case needs clarifying.

We thank the reviewer for this comment. You are correct that this needs clarification, and also that there was a mix-up in the scenario naming in Table A1. We will correct this in the revised manuscript by replacing SSP1-1.9 with SSP4-6.0 and ensuring consistency across the text and tables.

Regarding the number of simulations: we do not use 1500 fixed runs, but instead include all available ensemble members from the models listed in the table. This results in a total of 1858 simulations for all scenarios combined and the historical period for *tas*, and 1850 simulations for *pr*. Of these, SSP2-4.5 contributes 337 simulations for *tas* and 336 for *pr*. For the calibration of the emulator, we use all simulations from models and ensemble members that also provide SSP2-4.5, excluding only those scenarios or ensemble members without corresponding SSP2-4.5 data. This yields approximately 1500 simulations (slightly fewer in practice due to these exclusions).

Comment 4:

This evaluation section is nice, but you only look at annual mean T and precip. I'd expect T to be pretty good, though precip does surprisingly well – although the importance of getting extremes right means the errors at those ends are more important than the other percentiles, and I think this should be noted. It's nice to see this at the country level too. But it's a shame to only look at these 2 variables as you make a point of how flexible RIME-X is in terms of temporal scale and the range of impact variables you can look at. I think it'd be very useful to have some analysis of a less standard variable – you looked at maize yield earlier so maybe some kind of ISIMIP outputlike that could be good – and/or a shorter timescale (you had some plots of TXx earlier but not here). This would greatly enhance the argument here I think.

We thank the reviewer for this valuable and constructive comment. We agree that extending the evaluation beyond mean temperature and precipitation strenghens the argument, especially given the flexibility of RIME-X in terms of variables and timescales.

Our primary quantitative validation is based on CMIP6 temperature and precipitation, as these variables provide sufficiently large ensembles to construct robust ground truth distributions and to meaningfully assess distributional errors. For many ISIMIP impact indicators, ensemble sizes are substantially smaller, which can introduce considerable sampling uncertainty in the estimated ground truth distributions and complicate a comparable quantitative evaluation.

To address your suggestion, we have now considered how we can do an additional validation that includes all variables in the showcase section of the paper, including the maize yield change variable as an example of an impact variable and TXx and extreme daily rainfall variables as examples of extreme variables on smaller timescales. We also added this extra validation to the

revised manuscript as Appendix D. For the full discussion and added text refering to this extra validation in the manuscript we refer to the response to the second comment by the first reviewer.

Additionally, we have clarified in the abstract that the core quantitative validation is based on temperature and precipitation from CMIP6, while additional results for impact and extreme indicators are presented as demonstrative applications of the framework.

"By jointly quantifying global and regional sources of uncertainty from the start, RIME-X enables systematic exploration of the full space of plausible regional climate impact trajectories under different emissions scenarios. The method is **conceptually** applicable to regional indicators whose distributions are predominantly determined by warming level and provides a computationally efficient framework for uncertainty-aware regional indicator emulation. **We evaluate the method using out-of-sample validation on temperature and precipitation simulations from the Coupled Model Intercomparison Project 6 (CMIP-6) and demonstrate its applicability to a range of impact and extreme event indicators derived from the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP).** We provide an open-source Python implementation of RIME-X, including preprocessing workflows for data from ISIMIP and support for user-defined indicators."

Finally, we added a hint at the importance of the high and low percentiles in the description of Figure 8 in the revised manuscript:

"The expected deviation is quantified in the shaded area. **While agreement across all percentiles is desirable, deviations in the tails are particularly important, as accurate representation of extremes is critical for many impact assessments**."

List of textual and structural revisions we implemented in the revised manuscript due to comments from this reviewer:

At the end of this response, we provide a consolidated list of the requested textual and structural revisions that we implemented in the revised manuscript, and do not require additional clarification:

L14 would say "applicable to [those] regional indicators" to ensure clarity that this wouldn't apply to all indicators.

L33 this sentence on internal var feels a bit out of place but I can see you're relating it to the model uncertainty in MIPs; maybe use "additionally" or something to link to the broader paragraph.

L41 this intro is all one paragraph – it needs breaking up really. This could be a good spot for a paragraph break.

L42 would say "enables [the] running [of]"

L59 another paragraph break maybe?

L137 should note it's index i in the equation as you use j just before (and the equation should be numbered I think which would help here). Also I can't see N defined (number of bins?).

L139 “levels like we” -> “levels as we” I think.

L146 “more on them” a bit casual I think – maybe just “see…”?

Fig 4 “which distribution” doesn’t read right. “the Climate Action” don’t need “the”? And “point hints at” implies ambiguity – just say “denotes” or similar. Don’t need to repeat “year” in the legend.

L195 “equally spaced 101 quantile levels” should be “101 equally spaced quantile levels” right? And worth clarifying this gives the integer percentiles 0-100?

Figure 6 is introduced before Figure 5 currently.

L208 “formats: In” “formats. In” ?

L212 I think the sentence order needs switching. “We highlight the median and the 90% confidence interval. ” seems to refer to the left hand side currently but this is on the right. Would intro the left hand side first to make it easier in any case.

L228 “section 3.1[,] which”

L242 this Leach 2021 paper doesn’t seem to look at calibration specifically – probably best to refer to the FaIR calibration paper here https://gmd.copernicus.org/articles/17/8569/2024/

L248 new paragraph please!

L251 “depend mainly on the level of GMT”– entirely depends, right?

L277 not sure why CMIP6 is italicised now; I don’t think it needs to be but just be consistent.

L285 ISIMIP italicised too.

Reviewer #3

First Comment

Model's validation is discussed only for temperature and precipitation (figure 7 and 8). How's model performance about other indicators? How about providing a table summarizing the model performance for other indicators?

We thank the reviewer for this suggestion. Our main validation focuses on temperature and precipitation because those are popular impact-relevant indicators readily available in the CMIP6 ensemble, which provides a sufficiently large sample to derive robust ground truth distributions to reliably quantify emulator errors, while many impact indicators (e.g,. from ISIMIP) have ensemble sizes much smaller, which introduces substantial sampling uncertainty and makes a comprehensive quantitative benchmark across all indicators less meaningful.

Nevertheless, we agree that extending the validation beyond temperature and precipitation is valuable. Therefore, we expanded the manuscript by including additional validation exercises for selected impact-relevant and extreme indicators already featured in the showcase section in Appendix D. We refer to the answer to the second comment by the first reviewer for the added text and a discussion of it.

Second Comment:

The entire framework relies on the assumption that regional climate impact indicator distributions are predominantly determined by global mean temperatures. While this assumption is reasonable for some thermodynamic variables and some regions, it is problematic for tropical regions and for many regional indicators which are strongly influenced by SST gradients, circulation patterns, and land-atmospheric feedback. Mostly precipitation, extreme rainfall, wind, compound and multivariate extremes. So, the domain where the RIME-X framework can be applied needs to be clearly defined. Also, for the different indicators what is the fraction of variance that is explained by the global mean temperature can be mentioned. Limitations for hydroclimate and extremes, especially in monsoon regions, need to be discussed.

We thank the reviewer for this important comment. We agree that the assumption that regional indicators are primarily determined by GMT introduces limitations, particularly for indicators and regions strongly influenced by SST gradients, circulation changes, land–atmosphere feedbacks, and monsoon. To clearly define the domain of applicability of RIME-X, we explicitly list scenarios, regions, and indicators where these assumptions may break down in Section 3.1 and reiterate them in the Conclusions.

We have strengthened the manuscript by expanding the discussion of these limitations. Specifically, we have expanded the text in section 3.1 and the conclusion as follows (bold additions):

Though supported and applied in the literature for many indicators \citep{herger2015improved,schleussner2016differential,hausfather2022climate}, this assumption can introduce biases, particularly for overshoot or stabilization scenarios, and in regions **or indicators** affected by pattern effects, high regional aerosol concentrations, **changing circulation patterns, land–atmosphere feedbacks, SST-driven pattern effects, monsoon dynamics**, or changes in regional land use \citep{schleussner2024overconfidence,pfleiderer2024limited,wells2023understanding,giani2024origin}. **These limitations also restrict the applicability of our approach to idealized model experiments deploying solar radiation modification.**"

In addition, we have strengthened the conclusions by providing a consolidated list of limitations and applicability constraints:

“However, some limitations can affect RIME-X's accuracy. It assumes that the regional indicator depends primarily on GMT, excluding influences like aerosols, **shifts in socioeconomic conditions, land use, or circulation patterns (including SST-driven pattern effects, monsoon dynamics, and land-atmosphere feedbacks)**. This limits applicability **to regions, or** indicators where this assumption approximately holds. It also does not account for the path to reaching GMT levels, limiting applicability in scenarios **or indicators** where time-lag effects play a major role for the assessment of regional indicators, such as overshoot scenarios. **In addition, the framework may be less robust in scenarios where regional aerosol forcing differs strongly from the training scenarios**. **Our approach is also not validated for idealized model experiments deploying solar radiation modification.**”

In addition, to evaluate how RIME-X performes in some of these cases, we extended the leave-one-out validation framework described in Section 3 by conducting additional experiments using SSP1-2.6 and SSP3-7.0. Those results are additionally included and discussed in the revised paper. For the exact text and discussion of this addition we refer to our answer to the first comment by the first reviewer.

Third Comment

Secondly, the method assumes that the regional responses depend on warming levels but not on the pathway taken to reach the warming level. The manuscript needs to explore the conditions of overshoot, heterogenous anthropogenic aerosol forcing, rapid reductions of aerosols.

We thank the reviewer for highlighting this important point. We refer to our response to Comment 2, where we already addressed pathway dependence and aerosoles in more detail. In particular, we extend the leave-one-out validation framework introduced in Section 3 by repeating the analysis using SSP1-2.6 and SSP3-7.0 as test scenarios. This setup explicitly introduces differing forcing pathways, including cases with overshoot and heterogeneous anthropogenic aerosol forcing.

Overall, these results indicate that both path dependence and aerosol forcing can introduce systematic deviations, but that these remain moderate in most regions and are consistent with the limitations discussed in the manuscript.

Fourth Comment

The validation is done based on the CMIP6 SSP 245 scenario. The pathways for the scenarios in CMIP7 are different (https://doi.org/10.5194/egusphere-2024-3765, 2025). What is the expected deviation in results, if any, for this?

We thank the reviewer for this forward-looking question. We agree that differences between CMIP6 and CMIP7 scenario ensembles may influence the behavior of RIME-X, as the method depends on the characteristics of the training data.

The main differences we anticipate are related to (i) the broader representation of overshoot pathways in CMIP7 and (ii) generally lower peak warming levels in the CMIP7 high-emission

scenario compared to SSP5-8.5 in CMIP6. If CMIP7 scenarios are used as training data, the inclusion of more overshoot pathways would likely increase the spread of the emulated distributions, as a wider range of pathway-dependent responses is sampled. At the same time, the reduced availability of very high warming levels would limit the emulator's applicability at the upper end of the temperature range, as fewer training samples would be available for very high warming levels.

Overall, we therefore expect that a CMIP7-based training set could lead to broader uncertainty estimates in some regions and indicators due to increased pathway diversity, while potentially narrowing the range of high warming scenarios to which the emulator can be robustly applied.

Fifth Comment

The manuscript claims applicability to extremes. However, the methodology may not adequately capture the heavy tails, and compound extremes. It can smooth or underestimate tail risks, which are often the most relevant for impact assessments. Authors need to clarify whether the method is suitable for compound or cascading extremes

We thank the reviewer for this important comment and agree that these limitations should be clearly stated. We will clarify the point about heavy tails in the revised manuscript by discussing this limitation in the conclusion. Specifically, we will add the following text:

"In addition, the framework may be less robust in scenarios where regional aerosol forcing differs strongly from the training scenarios. Moreover, as fewer ESMs reach high warming levels, the robustness of emulations decreases with warming level due to reduced training data**, making the inclusion of high warming scenarios such as SSP5-8.5 in the training data particularly important for broad applicability. A further limitation arises from the fact that the emulator is not generative: the emulated distributions are constrained by the events present in the training simulations and may therefore miss low-probability extremes that are physically plausible but not sampled. While this limitation can be mitigated by increasing the size and diversity of the training dataset, this is not always feasible. It also implies that biases or corrupted simulations in the training data can propagate into the emulator outputs.** Emulations are additionally constrained by epistemic uncertainty due to the limited number of models available to quantify structural model uncertainty within the training dataset."

In addition, we clarify the limitations regarding compound and cascading extremes by adding the following sentence to the conclusion:

"This epistemic uncertainty is, however, partly mitigated through model weighting and GMT-based constraints. **Additionally, the method does not capture temporal, spatial, or inter-variable correlations between the emulated distributions of multiple variables, regions, or time steps, even when they originate from the same underlying data source, meaning that dependencies across regions, variables, and time steps are not preserved between several emulated distributions."**

At the same time, we note that there is a practical way to emulate compound indicators within the current RIME-X framework. Specifically, compound metrics can first be computed from the underlying climate model data, thereby fully capturing spatial, temporal, and cross-variable dependencies, and can then be emulated as a single derived indicator using RIME-X. For example,

wet-bulb temperature can be calculated from the raw simulation output and subsequently emulated as a compound variable.

Sixth Comment

The use of 21-year running average removes important components of variability and limits applicability to interannual variability and event-scale extremes. Authors need to discuss implications of this for risk assessments.

We thank the reviewer for this important comment. The 21-year running mean is only used as a definition for global mean temperature (GMT) in the framework. For individual regional indicators, the temporal aggregation is configurable via the running_mean_window parameter in the configuration files, which is set to 21-years by default as this choice was made for the Climate Impact Explorer.

However, this is not a structural limitation of the method. Users can choose different aggregation windows, and in the implementation, it is even possible to compute distributions for individual months. In principle, there is no constraint preventing even finer temporal resolutions, such as daily values, depending on the underlying input data and the intended application.

At the same time, there is a high level of structural flexibility for the indicator definition within the RIME-X framework. In principle, also 1-in-20 year extreme events can be identified and mapped, using a running window approach. However, such approaches would require large training data sets, as there is a clear trade-off here with stochastic natural variability. In practice, the available training data limits what's possible, and a running mean approach has proven to be numerically robust. This is a structural limit of a mapping instead of a generative emulator approach.

Minor comments:

Abstract: Model's performance (may be for temperature and precipitation) should be added quantitatively.

We thank the reveiwer for this suggestion. We agree that the scope of the validation should be more clearly reflected in the abstract. In the revised manuscript, we therefore extend the abstract to explicitly mention the validation approach and its application to both climate variables and impact indicators:

"By jointly quantifying global and regional sources of uncertainty from the start, RIME-X enables systematic exploration of the full space of plausible regional climate impact trajectories under different emissions scenarios. The method is **conceptually** applicable to **all** regional indicators whose distributions are predominantly determined by warming level and provides a computationally efficient framework for uncertainty-aware regional indicator emulation. **We evaluate the method using out-of-sample validation on temperature and precipitation simulations from the Coupled Model Intercomparison Project 6 (CMIP-6) and demonstrate its applicability to a range of impact and extreme event indicators derived from the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP).** We provide an open-source Python implementation of RIME-X, including preprocessing workflows for data from ISIMIP and support for user-defined indicators."

We do not include quantitative performance metrics in the abstract, as their interpretation requires the formal definitions introduced in the methods section, and including them without context may, in our view, reduce clarity rather than improve it.

L183: n=10000? Or 10.000 (fractional)?

Thanks, it is n = 10000, and we will clarify that in the revised manuscript.

List of textual and structural revisions that will be implemented in the revised manuscript

At the end of this response, we provide a consolidated list of the requested textual and structural revisions that will be implemented in the revised manuscript, and do not require additional clarification:

L100-105: Mention what "P_s,y" and "p" denote to avoid any confusion for the reader.

L150: "... tas ..." --> "... tas (temperature at surface) ..."

L175: "... IM, we ..."

L193: "... files, we call..."

Additional changes:

In addition to the revisions requested by the reviewers, we identified a minor error in the calculation of the global mean temperature in the CMIP6 simulations. After correcting this issue, we reproduced Figures 7 and 8 in Section 3.2. This resulted in minor changes to the figures, the error values, and the ranking of countries based on these error values, which required corresponding updates to the text in the manuscript.