Response to reviewers

Thank you to the reviewer for your detailed review of our manuscript. We have responded to each comment in full and outlined the changes we will make to the manuscript to address your comments in this document. Our responses are in black font in response to review comments in blue, and where we quote new text, this is in italic.

Reviewer general comments

Mathison et al. describe a new climate emulator that is coupled with a land-based impact model, which together form PRIME. This paper describes a significant advance in the use of FaIR, a commonly used 1D climate emulator because it allows users to simulate regional climate using patterns from many different climate models at relatively low computational cost. Furthermore, it allows for rapid impact assessment with its direct integration with an impact model.
I do not have major methodological concerns, but I have several questions and suggestions that I would like the authors to address before recommending publication. The more general ones are written here and several smaller line-by-line suggestions are listed below.

Thanks for your comments, we respond to each comment in turn here and in the revised manuscript.

The underlying assumption of matching an ensemble of FaIR output with an ensemble of climate model patterns is that there is no relationship between the two. In other words, the patterns in warm models do not look significantly different from the patterns in cold models. In some FaIR papers, specific climate models are emulated by using tuned parameters (e.g. Leach et al. 2021, Figures 3 and 4), but here the full range of models of global mean T is then matched with a full range of patterns. To address this, I could imagine a supplemental figure that, for example, plots the pattern correlation between each model and the multi-model mean, against the global mean temperature in a future year (e.g. 2050 or 2100). There may be other ways to do this, but I think a small additional analysis to demonstrate that the pattern is not a strong function of sensitivity would be useful to demonstrate whether this is a limitation or a non-issue. I would be surprised if this was a major problem, but I think it would be worth characterizing for potential future users.

Thank you for this insightful comment. We note that the literature suggests a link between the strength of feedbacks and the pattern of warming, e.g. Andrews and Webb (2018), Ringer et al. (2014). However, it is the difference in atmospheric physics that dominates the spread in climate model climate sensitivity (Dong et al. 2020) and is what FaIR is aiming to emulate. From this perspective, we expect the pattern effect to be a second order uncertainty. Furthermore, when considering our overall aim of sampling uncertain climate impacts the evidence from CMIP6 is that the a ‘hot-model’ is not necessarily correlated with a positive bias in the simulated impact (Swaminathan et al., 2024). Swaminathan demonstrates quantitative evidence that for a range of impacts metrics (such as flood, drought and fire weather) there is at most only weak correlation between these impacts and climate sensitivity. In other words there is no pattern dependence of being a “hot” or a “cold” model. In fact Swaminathan et al., (2024) goes further and suggests that to try to assume a link from the patterns to the global temperature can actually be misleading and lead to omission of some important impact-relevant information. The challenge then is to appropriately constrain the patterns without artificially reducing the impact spread. This is an area we will

certainly take forward in future assessments and revisions. In summary, our approach is to sample the combined uncertainty in climate and pattern effects noting that we may be oversampling the spread.

In response to this comment we have therefore added the following text to the manuscript:

"Our approach combines the full range of FaIR temperature responses with the full range of CMIP ESM patterns. We note a pattern effect relating warming to climate sensitivity (Andrews and Webb (2018), Ringer et al. (2014)) has been shown in the literature. However, assessments of simulated impacts in the CMIP6 ensemble sampling a wide range of impacts metrics from multiple regions found little or no correlation with climate sensitivity for most regions and climate drivers (Swaminathan et al., 2024), which contributes to justifying the approach to treat these independently. Other studies have found changes to circulation patterns and dynamical regimes more important for climate patterns than global scale thermodynamical response (Ribes et al., 2021 and 2022, Palmer et al 2023). To maximise our sampling of uncertainty we therefore take the pragmatic decision to co-vary all patterns with sampled temperature pathway."

As for other users, I'm wondering whether the pattern scaling component is a stand-alone model. The schematic (Figure 1) shows patterns that include the ocean, whereas the rest of the figures are limited to the land. I could imagine many applications in which a user would want an ensemble of patterns-scaled climate (including the ocean) but is not interested in the land-based impacts. It would be helpful if you could address that kind of hypothetical use case in the text.

Thank you for this comment, although the patterns generated include the ocean we chose to focus our analysis on land because JULES is a land surface model and only outputs data on land. Our focus for this framework has been mainly to look at Earth System and climate impacts on land as output from the JULES model when run as an impacts model. However, part of the benefit of running a framework like this is being able to output at various points along the pipeline, so we have added this possibility by including the following to the text:

"We generate global patterns that include land and ocean but in this analysis, we focus on the patterns over land for running JULES and considering land impacts. However, it would be possible to use the patterns over the ocean and exclude the land for other downstream applications."

Lastly, I think the text about the land model needs to acknowledge that atmospheric carbon (and other variables) have no feedback with the changes in vegetation. Although I still see use in this application, I think this deserves more serious consideration in the text, especially as comparisons are made to CMIP6, where some models do simulate the terrestrial carbon cycle (line 398). This seems like a significant difference between PRIME and some GCMs.

My comment on line 164 about adding anomalies from a 1850-1889 baseline to 1901-1930 climatology may also require some changes if indeed I'm understanding the methodology correctly.

We agree this could be an important feedback, we now discuss our aspiration to add it to a future development of PRIME. For now the structure of the framework does not allow this to operate. The comparison with CMIP6 outputs though is still a clean (like-for-like) one because we draw on simulations which are "concentration-driven" and so, even for carbon-cycle ESMs, also do not allow this feedback

onto atmospheric CO_2 to operate. We have now acknowledged this with the following text in the discussion:

"However, this simple one-way coupling between components, while a benefit in terms of flexibility, could also be deemed a limitation because changes in emissions from the land are not allowed to feedback on the scenario.This is a desirable capability that we plan to build into PRIME but in its current form, the structure of the framework does not allow this to operate. The comparison with CMIP6 outputs shown here draws on simulations which are "concentration-driven" therefore this feedback onto atmospheric CO_2 is not included even for carbon-cycle ESMs, which makes the analysis shown here, comparing CMIP6 simulations and PRIME, a clean comparison."

We address the comment on line 164 in the specific comments section below.

Specific Comments:

Line 4: I think "global picture" is a confusing choice of words. Also there are other climate emulators that have a spatial element, so I'm not really sure what this sentence is trying to say.

Here we are saying that some simple climate models often achieve their improved runtime efficiency through reductions in spatial detail, for example providing Global estimates of common climate metrics such as Mean Surface Temperature, CO_2 concentration and Effective radiative forcing. We have modified the text to say:

"Simple climate models are extremely efficient although some only provide global estimates of climate metrics such as mean surface temperature, CO_2 concentration and Effective Radiative forcing."

Line 5: "general information" is vague

Agreed we have removed this and this sentence now reads:

"*Within the Intergovernmental Panel on Climate Change (IPCC) framework, understanding of the regional impacts of scenarios that include the most recent science is needed to allow targeted policy decisions to be made quickly.*"

~60: the overview of ongoing efforts is great. I think the machine learning models like Climatebench probably deserve a mention for completeness, since they have similar objectives

We agree and have now mentioned Climatebench in the manuscript saying:
"the ClimateBench v1.0 (WatsonParris et al., 2022) benchmarks machine learning emulators that predict annual mean global distributions of temperature, diurnal temperature range and precipitation"

66-68: I found this sentence confusing. "this type of input"? Not sure what input is being referenced

We have made this clearer by redrafting this text to say:

"We use pattern-scaled climate variables instead of ESM output to drive our impacts model, because this approach offers a useful opportunity to more quickly derive impacts information from new scenarios. However, this does not imply that pattern-scaled climate variables should replace ESMs or ISIMIP bias-corrrected data but could provide a steer on which scenarios would be most useful for ESMs to run or which ones to bias-correct for use in more specialist impacts models."

98-99: Choices affect the way patterns are selected? Do you just mean users choose the patterns? Also, please define Rose suite.

Yes the user can choose which patterns to include from which ESMs. We have clarified the text to say:

"*PRIME is a flexible framework, with ensemble members and patterns selected by the user and therefore dependent on their chosen application.*"

Apologies for not defining what a Rose suite is in the manuscript, this is amended in the revised manuscript to say:

"we are developing software to simplify running the PRIME framework using the choices presented here using Rose and Cylc (Oliver et al., 2018 and Cylc documentation 2024) - a group of utilities and specifications which provide a common way to manage the development and running of scientific application suites in both research and production environments. Rose and Cylc are used to ensure a consistent framework for managing and running meteorological and climate models, they are therefore ideally suited to this application."

135: As stated repeatedly, part of the appeal of emulators is the low computational cost. However, the sampling done here is fairly sparse (Figure 3) and the reason given here is to "make the ensemble size manageable". I think it would be worth dedicating a bit of text to explain this discrepancy.

Thank you -this is a fair comment (no pun intended). The PRIME system is indeed intended to be fast in the sense that it is orders of magnitude faster than an ESM. The temporal and spatial detail of the land model means that PRIME is slower than FAIR itself as a global emulator, and the volume of data produced is substantial. As such PRIME sits between the two classes with the UKESM model achieving circa 2 years per day and FAIR running at approximately 11 years per second. PRIME runs approximately 20000 years per day using a moderate compute resource and applying limits to the number of runs at any one time.

Hence we still need to be mindful of run length and data volume. For uses where we expect substantial exploitation of results – e.g. if PRIME was to be used as a fast-response tool for new scenarios in AR7, then a much larger ensemble could still be fairly easily produced and data made available.

An additional appeal of emulators - including PRIME - is that they can run from global mean forcing, whereas to run an ESM requires many weeks of human effort to set up spatially-resolved forcing data - some of which is not available for novel scenarios from IAMs.

We have added the following text to Section 3 to explain this choice*:*

"Figure 3 shows the selection of ensemble members from the full FaIR distribution of 2237 members; these 9 percentiles (0, 1, 5, 25, 50, 75, 95, 99 and 100\%) are chosen to explore the full range of global

temperature sensitivity, but make the data more manageable because it increases considerably when combined with the CMIP6 patterns (see Table S1 for a full list of those used) and run through JULES."

136: "these are selected using one scenario so that scenarios can be compared against each other" is confusing

The temperature percentiles will be different for each scenario, so this means 1% will not be the same ensemble member for SSP5-8.5 as for example SSP1-2.6 or SSP5-3.4-OS. Unfortunately, this means if we choose different ensemble members for the 1% for each of the three scenarios it will be difficult to compare between scenarios. Selecting a single scenario for defining the ensemble members for each percentile makes it much easier to compare across scenarios. This is partly why we found it is useful to look at the joint temperature and CO_2 distribution for each scenario to convince ourselves this approach would work. We are actively considering our approach to ensemble selection across the temperature and CO_2 distributions for future versions of PRIME.

We have added the following text to better explain the use of a single scenario for selection of ensemble members.

"We use a single scenario to define the ensemble member per percentile because each scenario will have different ensemble members for each percentile. For example, the 50th percentile ensemble member for SSP5-8.5 would not be the same ensemble member as the 50th percentile for SSP5-3.4-OS. We choose the same ensemble members for all scenarios to make the comparison between scenarios easier."

147-148: Do the linear regressions include an intercept? If so, how is it treated?
Please also specify one realization of each model (i.e. not an ensemble and not all available)

Thank you for this comment, the linear regression intercepts are zero and the realization used for each CMIP6 ESM is in Table 1 of the supplementary information. We have added to the text referencing this table, to say that it includes which realization is used. The text has been updated to reflect this comment at the start of Section 2.2 (approximately line 165) in this latest revised version.

152: I think there needs to be a summary of this conversion from monthly to 3-hourly data. Not necessarily here, but somewhere.

The summary of the conversion from monthly to 3-hourly data is provided later in this section from line 190 onwards. We agree that this sentence would fit better later in this paragraph, where the weather generator is described, so we have now moved it to this location.

159: "this is the method currently used in PRIME" - it's not clear what is being referred to

This refers to the pattern scaling approach, to make this clearer the following text it now says:

"Within PRIME we use patterns for all input variables required to run the JULES land-surface model. JULES tends to be less sensitive to some of the input variables that do not typically scale as well with temperature, such as wind speed, pressure and longwave downwelling radiation, so we can include them without introducing erroneous output changes (see Section 3.2). It should be noted that we generate

global patterns that include land and ocean but in this analysis, we focus on the patterns over land for running JULES and considering land impacts. However, it would be possible also to use the patterns over the ocean for relevant downstream applications."

161: I thought the methods were described quite clearly up until paragraph, where it starts to get confusing. It's somewhat unclear what belongs to IMOGEN and what does not. I also don't understand why the diurnal cycle calculation could cause numerical instabilities (166) when there are no feedbacks. And what part of this is 3 hourly and where is the monthly temperature from the pattern scaling coming into play? Is there only one diurnal cycle for each month?

We have rewritten this section, making the steps from climate patterns to weather data to drive JULES clearer. It is the lack of the diurnal cycle in JULES that could cause instabilities. The text now says:

"The spatial distribution of the monthly mean meteorology for each month of the transient simulation is reconstructed from the climate patterns multiplied by the global mean temperature change (see Section 2.1) superimposed on an observed monthly climatology. This is done by IMOGEN. In this version of PRIME, the observed monthly climatology was constructed from the daily meteorological data provided by the GSWP3-W5E5 dataset from the ISIMIP3a project (Frieler et al., 2023) for the period 1901--1930. This was regridded to a resolution of N48 with a 3.75° longitude grid size and a 2.5° latitude grid size.

In addition, the weather generator in IMOGEN (huntingford et al.,2010) is used to downscale the weather data from the monthly to hourly timestep, which is the temporal resolution used to drive JULES. This method is described in detail in (mathison et al.,2022). One limitation of this method is the lack of variability in the driving humidity, temperature and radiation at both the sub-daily and daily resolution. In the next version of PRIME we will develop the temporal downscaling meteorology so that it coherently includes the effects of, for example, clouds on the diurnal cycle of the weather data."

164: Why are 1850-1889 anomalies added to 1901-1930 climatology? If it's observationally limited (i.e. no data prior to 1900), then it seems like the anomalies should also be relative to 1901-1930? Maybe you could at least demonstrate that this is a reasonable approximation, but it seems inconsistent. Also why did you decide to add it to an observational climatology instead of a modeled one?

The period 1901-1930 was a compromise between the pre-industrial (very limited observational data) and present day (lots of readily available observational data). We wanted an observational climatology where there was very little climate change, but where we had some (albeit quite limited) observational data. FaIR had a maximum global mean temperature change of 0.01 K by 1930 and less for the mean of the period 1901 to 1930 - this falls well within natural variability. We used the observed climate rather than a modelled climate so we could get the most realistic present day conditions and minimize differences between the model output and observed land surface.

Equation 3: I think that 3 should be an exponent - please check the equation

This equation has been updated.

194: typo - mode, not model

No this is correct and should say model

213: I don't think you mean "where ESMs have not been run" because the simulations exist

We do mean this because ultimately we want to use this framework to run scenarios that have not yet been run using ESMs, but in this paper we are showing that the framework is fit for purpose by running it using scenarios that are known and have previously been run using ESMs. We clarify this by modifying it to say:

"In this section, we evaluate PRIME. In this context, that means that we show that the framework is 'fit for purpose' by testing it on scenarios where ESM simulations already exist. However, ultimately we want to use PRIME to produce land simulations for scenarios where ESMs have not been run "

Figure 2 and S1. Caption should specify where the data are from. It says in Fig S1, but not in Fig 2. Could you also mention in the text again why you're subsampling? It hasn't been mentioned in several pages

I think the reviewer means Figure 3 and S1 as these are very similar plots showing the joint distribution from FaIR for CO_2 and temperature for the 3 scenarios covered with one scenario (SSP1-2.6) shown in Figure 3 and the other two scenarios (SSP5-3.4-OS and SSP5-8.5) shown in Figure S1. We have updated the caption of Figure 3 and Figure S1, so the details are now the same for these two figures.

"*Figure 3. Joint frequency distribution from the FaIR simulations of Temperature (TAS) and CO_2 concentration in 2100 for SSP1-2.6 emissions and the sub-selected percentiles (blue crosses) used to drive the JULES impacts model. Shades of green denote the density of points with individual histograms above and to the right of the main panel. 10% confidence intervals are shown by the contours.*"

We also now restate that the FaIR distribution consists of 2237 ensemble members, so we sub-sample from this distribution to make the amount of data more manageable. For each FaIR ensemble member selected, we use 34 patterns from CMIP6 and we run JULES for each of these models. To run JULES we generate 3-hourly data for each of the 8 input variables. Currently we select 9 ensemble members using the temperature distribution, so for each scenario we have over 300 runs of JULES with all the output that generates.

"Figure 3 shows the selection of ensemble members from the full FaIR distribution of 2237 members, these 9 percentiles (0, 1, 5, 25, 50, 75, 95, 99 and 100%) are chosen to explore the full range of global temperature sensitivity, but make the data more manageable because it increases considerably when combined with the CMIP6 patterns (see Table S1 for a full list of those used) and run through JULES."

Figure 4: I don't think the differences should be on the same scale as the anomalies. For example, 4c just shows that most differences are less than 0.6 degrees, but grouping 0.01 to 0.6 together does not make sense when almost all values are less than 0.6. Plus the colors are faint (for the same reason) so they are hard to see. The figure captions (for 4 and accompanying SI figures) should also specify "anomalies" for them "predicted ensemble means"

Thank you for this comment, we agree the colours are faint in the difference plot and there is a balance to be found between being able to see the detail but also trying to show that these differences are small. As mentioned in the caption, the colour bar magnitude for the differences in this revision is the same as that

for the anomalies, in order to show that the prediction error is small compared to the change induced by the scenario. However we do accept that this does make the colours very faint and it is difficult to see the differences. We experimented with this, but found that those differences for temperature are faint no matter what range we use in the colour bar. In any case, we have adjusted the colour bars for all the variables, this includes Figure 4 and the associated supplementary figures (Figure S2-S9) to try and make the differences clearer but also not overstate these, because in most cases these differences are quite small.

The figure captions of Figure 4 and S2 to S9 have also been updated to say “Evaluation of the pattern predicted ensemble mean anomalies…”

Figure 5: I think adding a ratio column (MAE / IQR) would better make the point.

We have added an MAE/IQR column to Figure 5 and the equivalent plots in the Supplementary information.

S15: All colors in all of the related figure are shown from light to dark, except pressure in S15. Although it’s not wrong, I would recommend reversing the colors for clarity

We have modified the plots showing the IQR/MAE for Pressure (i.e. plots equivalent to Figure 5) to go from light to dark. The modified plots are in the supplementary and are S11 and S15.

307: This sentence and several sentences in this paragraph are a little wordy and could be edited for clarity. For example, the reference to Burton et al does not seem to fit well in the context, and “actual values” on line 315 is vague.

Agree that the reference to Burton et al is a little separate to where we discuss the variables that are important for JULES, so this has been removed in the revised version and also added CMIP6 to the sentence highlighted, so it now reads “CMIP6 values”, this occurs at line 346 in the latest revision. We have also edited this paragraph to make it “less wordy”.

345: What does it mean for the pattern scaling technique to be well understood by the literature?

We agree that the meaning of this sentence is unclear so have removed this reference to the literature as it is enough to say that the method is well understood.

366: I’m a little confused here - you would not expect the values to be identical but “hope for a similar spread”. I agree it seems to work impressively well, but there is some disconnect between saying there are several differences (e.g. only using one land model) and yet expecting similar spread. The rest of the paper is about emulating spread when using many models, so why would you expect the spread here to be similar when only using one model? A bit of clarifying text would be helpful.

Yes, you are correct. It is true that we would not expect the spread to be exactly the same (in fact if we did then this would imply that different land models add no value!). Although here we have an additional

source of spread by the sampling of FAIR sensitivity ranges for all climate patterns. What is important is that the use of a single land model here does not overly restrict the output and negate the benefits of being able to sample climate sensitivity and climate patterns fully. We have revised the text to clarify this:

"It is important to check that the use of a single land model here does not overly restrict the output and negate the benefits of being able to sample climate sensitivity and climate patterns fully, so while we would not expect PRIME values to be identical to CMIP values for these JULES outputs, we check that the use of a single land model does not result in too narrow a range of outcomes."

Figure 9: The caption says "median and uncertainty ranges" but I don't see that in the figure. I am confused about what the lines are - can you clarify in the text and caption? Are they for different pattern scaling from each of the different climate models?

Thank you for this comment, this was an oversight. Each of the lines in Figure 9 represent output from each PRIME ensemble member. In an earlier version of this plot, we only showed the median and uncertainty ranges but during the peer-review process we have revised this to show each ensemble member. We have revised the caption to reflect this.

"Figure 9. Maps of net ecosystem production (top) and tree fraction (bottom) with timeseries showing PRIME output for each ensemble member for each study region: Amazon, Siberia, India and the USA (labelled) for SSP1-2.6 between 1850–2100."

469: You mention that local land-atmosphere feedbacks are not included, but that seems secondary to me given the fact that the land model produces carbon-cycle relevant parameters like GPP that do not feed back into FaIR. As mentioned above, I think that qualification should be added.

Please see the response to the query in the section - Reviewer general comments, where we explain that we are using simulations which are "concentration-driven" which means that even for carbon-cycle ESMs, these models also do not allow this feedback onto atmospheric CO_2 to operate. This means that the comparison made between PRIME and CMIP6 is still a fair one. However, the authors accept that the one-way coupling of the PRIME framework means that the land -surface cannot then have an effect on the emissions driving FaIR and have added text to the manuscript in the discussion at the end of Section 5.0 (around line 484 in the latest revised version) to explain that this is a future ambition but is not possible in PRIME in its current form.