Reviewer comments are in *black italics.* Author responses are in orange, with new additions to the manuscript in **bold orange.**

Reviewer 1

There is a real need for emulators that can capture the spatial patterns of the response to GHGs and short-lived climate forcers, and there is work in this manuscript that moves us towards that goal. However, I am afraid I need to recommend rejection of the manuscript in its current form. The simulations used to train the emulator are not fit for the often-implied purpose, and there is confusing messaging throughout the manuscript about the suitability of the emulator for approximating the climate responses seen in the SSPs. This is an emulator of the fast temperature response to sudden changes in forcing, which we do not expect to capture the slow climate responses seen in the SSPs, yet there are parallels drawn with the SSPs and implications made about direct policy-relevance throughout the manuscript. There are some acknowledgements that this emulator is in fact only a step towards this, and a rewrite of the manuscript that presents it as the proof of concept that it is would be more appropriate.

We thank the referee for their time spent reviewing our manuscript and for the thoughtful comments. We agree with the reviewer that there is a need for climate emulators that capture the spatial response patterns and would like to stress that our study is intended as a proof of concept that aims to bring us a step closer to this, rather than directly addressing this need. We did not intend the results to be directly applied to policy, so we greatly appreciate reading the reviewer's interpretation of our submitted manuscript. We believe we can address this through revisions following the referee's suggestions to frame the study more clearly as proof of concept.

To summarise, in our revisions:

- We change the title to specify that this is a *short-term* temperature response emulator only. This is a key step towards full emulation, but it would require more expensive simulations to do this.
- We state the purpose of the emulator more clearly in the introduction and highlight that it is solely for the fast response.
- We remove the application of the emulator to the SSPs (Section 5) and put this in the Supplementary part. The purpose of this section was to give an indication of how the emulator could be used, but at this stage we cannot draw strong conclusions because as the reviewer points out, our approach cannot be directly compared to transient SSP scenarios. So as not to distract the reader, we have moved this to the Supplementary.
- Throughout the paper, we now frame the emulator as proof of concept.

- In the conclusions, we discuss the additional work needed to create a valuable policy tool.

We outline our revisions in more detail in our other point-by-point responses below.

For policy implications, the focus on fast responses is a key limitation of the study, and it needs to be clearer in the title and conclusions that this is an emulator of fast temperature responses only. This is very different to what many readers will understand from the 'climate response' currently in the title (for many, this will be decadal- to centennial-scale responses), and may well be markedly different from the long-term response in many regions. Since the regional aerosol perturbations used in the short simulations produced to train the emulator are based on regional perturbations also used in long equilibrium experiments with similar scalings, it would be helpful to see a comparison of the fast and slow responses to regional aerosol perturbations to help the reader to better understand the limitations of trying to apply an emulator of the fast response to real-world or CMIP-style coupled-transient scenarios. The patterns and magnitude are different, as can be seen in existing literature, so the claims throughout the manuscript of 'policy-relevance' and the emulator as a 'basis for rapid climate change projection' must be toned down, and must appear alongside appropriate caveats.

We agree this is a limitation of the study and this should be made clearer. We have implemented the reviewers suggestions. New additions are in bold.

New title:

> "Using a Gaussian Process Emulator to approximate **short-term** climate response patterns to greenhouse gas and aerosol forcings."

In the abstract, we removed the sentence directly linking the emulator to policy relevance, we add a sentence about how it paves the way for future policy-relevant emulators and use more conservative language:

> ~~Finally, we demonstrate how this type of emulator could be used in policy-relevant studies to predict fast adjustments of regional climate to changes in atmospheric emissions for a given scenario.~~ ~~This establishes a basis for rapid climate change projection, without the need for computationally expensive climate model simulations, and increases the number of climate change scenarios that can be explored simultaneously.~~ **Our work demonstrates the potential for rapid climate change projections subject to various climate forcing agents, focusing on the short-term response. This paves the way for the**

future development of more policy-relevant emulators, that can target decadal and multi-decadal timescales.

In the introduction, we state the purpose of the emulator more clearly:

Lines 64-85:

> In this paper, we present an emulator based on the HadGEM3 GCM**, that predicts the short-term** global surface temperature response to a wide range of forcing types, including regional aerosol emissions and global greenhouse gas concentration changes. **This is the** fast adjustment of the surface temperature in the first 5-years following the perturbation. This fast adjustment emulator predicts the response to an abrupt change in emissions, such as the simulations run in idealized forcing-response studies (e.g., Myhre et al., 2017; Richardson et al., 2019; Stohl et al., 2015; Samset et al., 2018; Smith et al., 2018). The emulator is designed to predict the entire spatial pattern of the global climate change response, rather than just the global mean, which could prove useful in future policy and impact studies. **Our approach** lays the foundations for the development of long-term climate response emulators **which would be valuable in informing policy. Furthermore, we propose that this type of emulator could be coupled to other emulators designed to predict across different timescales. For instance, it could be combined with the emulator developed in Mansfield et al. (2020), which predicts the long-term temperature response given the short-term temperature response. Additionally, while such a fast-response emulator is not directly relevant for long-term policy design, it can have real-world use for predicting regional and global temperature responses to short-term forcings that are triggered by episodic pollution sources such as wildfires or sudden anthropogenic pollution decreases (e.g. as during the COVID-19 pandemic).”**

Within this new framing, in the conclusions, we also highlight how future studies can learn from our design choices when developing future emulators. We add the new paragraph:

Line 732-750:

> **“The overarching goal of this study was to identify a framework for policy-relevant emulators. We made several choices when designing the training and test dataset using GCM simulations, from which future studies can learn. For instance, we chose a maximin Latin hypercube design to sample uniformly across the distribution of each**

parameter. We extended the parameter space for the aerosol forcings beyond their feasible ranges to account for the weaker aerosol signals compared to the greenhouse gas forcings. Still, we found that many simulations are dominated by the response to greenhouse gas forcings. Future studies may wish to compare the emulator performance under different ratios of extended parameters to realistic parameters. Alternative designs, such as selecting some simulations with restricted greenhouse gas perturbations, which lead to stronger aerosol signals could potentially improve performance in this range. Furthermore, Bayesian optimization techniques could be used to directly identify new parameter choices for additional GCM simulations that would reduce emulator error or uncertainty (Shahriari et al., 2016)."

The emulator derives most of its variance in the temperature response from variance in the CO2 forcing, which is to be expected from the design of the training experiments, and is not representative of the real-world responses. In the real world, we have seen strong temperature responses to aerosol, which are not captured by this emulator. And we do not expect them to be. Most of the aerosol perturbations considered in this work are fast responses to SO2 changes. The fast temperature response to SO2 is a fraction of the magnitude of the slow response that we see in long coupled simulations and the real world, and it has a markedly different spatial pattern, see e.g. Figure 2: https://journals.ametsoc.org/view/journals/clim/31/11/jcli-d-17-0439.1.xml The design of the training experiments mean that this emulator will always underestimate the importance of aerosol relative to CO2 for any real-world applications, and that it will also fail to capture the pattern of the real-world aerosol response, as this differs markedly to the pattern of the fast temperature response. As the training dataset uses a series of 5yr coupled simulations, the emulator is also not showing us a clean version of the fast response to the forcings considered, as it is conflating them with responses across multiple timescales and internal variability (although an attempt was made to address the latter by having an ensemble of short simulations for each case).

We would like to highlight that the emulator does not aim to capture real-world aerosol response as these are idealized experiments where we perform abrupt perturbations. The new framing of the paper highlights this (see responses above). We remind the reader of this when we introduce Figure 1.

Line 345:

"Note that these perturbations are not intended to be realistic perturbations over a five-year period, but rather they estimate the short-term response to abrupt perturbations, which at this stage would be used for scientific experimentation, with the view to develop the ideas further in the future for more direct policy relevance."

Putting aside the caveats of focusing on the fast response to SO2 perturbations, and of using 5yr simulations to do this, the choice of SO2 perturbation regions, and the magnitude of the SO2 perturbations themselves, which are scaled to be large enough to generate a clear signal useful for emulator training also seem sensible. Exploration of BB OC/BC is more limited, using only a tropics-wide training perturbation, and is likely to be missing key regional patterns in the response as a result. The choice of a Gaussian Process approach seems sensible, and the emulator does appear to represent the GCM responses well. However, the design of the training dataset mean that it is not an appropriate tool for emulating temperature responses in the SSPs.

We add a comment to address the issue that the tropics-wide perturbation of BB OC/BC will likely miss key regional patterns:

Line 212:

> "**We note that by using tropics-wide perturbations, the emulator may not capture highly regionalised forcing-response patterns. However,** we do not consider OC/BC on smaller regional scales because previous studies have found that large, unrealistic OC/BC forcings are required to obtain climate responses that exceed the internal climate noise (e.g., Myhre et al., 2017; Baker et al., 2015; Stohl et al., 2015)."

We also remove the emulation of temperature response for SSPs in the main text and have moved it to the Supplementary Material and framed it as an example for how this type of emulator could be used once trained on a dataset designed for this task.

Specific comments follow.

Title: Study is only trying to emulate the fast temperature response. This must be specified in the title in place of 'climate'.

We have edited the title to read as follows:

> 'Using a Gaussian Process Emulator to approximate **short-term** climate response patterns to greenhouse gas and aerosol forcings.'

Introduction: Would be helpful to have a sentence in the introduction explaining the policy-relevance of the fast temperature response if that framing is retained. How often does the real-world/slow response look like the fast response? For SO2 in particular, there are large differences that should be addressed. It would be better to bring out the narrative that this emulator is an interim step, which provides proof of concept, then to try to force a narrative of policy-relevance.

In the introduction paragraph, we have re-framed the purpose of the emulator as the reviewer suggests. Rather than linking it directly to policy, we discuss how it forms an interim step towards long-term climate response emulators. The updated third paragraph in introduction with changes addressing this point in bold:

Lines 66-87:

> "In this paper, we present an emulator based on the HadGEM3 GCM, that predicts the short-term global surface temperature response to a wide range of forcing types, including regional aerosol emissions and global greenhouse gas concentration changes. This is the fast adjustment of the surface temperature in the first 5-years following the perturbation. This fast adjustment emulator predicts the response to an abrupt change in emissions, such as the simulations run in idealized forcing-response studies (e.g., Myhre et al., 2017; Richardson et al., 2019; Stohl et al., 2015; Samset et al., 2018; Smith et al., 2018). The emulator is designed to predict the entire spatial pattern of the global climate change response, rather than just the global mean, which could prove useful in future policy and impact studies. **Our approach** lays the foundations for the development of long-term climate response emulators **which would be valuable in informing policy. Furthermore, we propose that this type of emulator could be coupled to other emulators designed to predict across different timescales. For instance, it could be combined with the emulator developed in Mansfield et al. (2020), which predicts the long-term temperature response given the short-term temperature response. Additionally, while such a fast-response emulator is not relevant for long-term policy design, it can have real-world use for predicting regional and global temperature responses to short-term forcings that are triggered by episodic pollution sources such as wildfires or sudden anthropogenic pollution decreases (e.g. as during the COVID-19 pandemic).**"

L32: Since the perturbation regions have been used in previous studies (listed on L147), can you include a figure that compares the spatial pattern of the fast and slow responses to your regional aerosol perturbations, and also a figure comparing your fast responses to those calculated in a traditional fixed SST framework?

We thank the reviewer for the suggestion. While we like the idea to include a comparison of fast and slow responses, we do not want to distract the reader from the primary focus of emulation here. Instead, we have chosen to add a comment in our conclusions about how the fast and slow responses differ, and discuss how the fast-response emulator could be coupled to a slow-response emulator:

Lines 576-580:

> **"Furthermore, the emulator is also a step towards more policy-relevant emulators that would require long-term or even transient climate responses. This could be achieved through coupling to other emulators that predict the long-term climate response given the short-term climate response (e.g., Mansfield et al., 2020). Based on past studies, we would expect to see stronger and more spatially homogeneous responses to forcings over longer timescales, as the slow responses associated with ocean interactions and climate feedbacks dominate (Kasoar et al., 2018; Lewinschal et al., 2019; Liu et al., 2018; Mansfield et al., 2020)."**

L59: For Gaussian process representation of regional temperature responses, see also: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2025JH000741

This is a useful citation. We include it in Line 63:

> "Most past emulation studies estimate the *global mean* response to forcings over a timeseries (Castruccio et al., 2014; Meinshausen et al., 2011; Smith, Forster, et al., 2018), with some extending this to include spatial patterns (Bao et al., 2016; Beusch et al., 2022; **Dewey et al., 2025**; Nath et al., 2022)."

L74: As you say on L70, emulating the fast response lays the foundations for emulating the long-term climate response, but emulating the fast response itself has rather limited policy application, whether calculated from fixed-SST simulations or the from the first few years following a step perturbation in a fully coupled simulation.

On its own, its policy application is limited but we propose it could be combined with other emulators to capture long-term or transient response. We also suggest emulators for short-term response could have uses in predicting response to sudden short-term forcings, such as wildfires or sudden pollution decreases. Line 74:

> "**Furthermore, we propose that this type of emulator could be coupled** to other emulators designed to predict across different timescales. **For instance, it could be combined with the emulator developed in Mansfield et al. (2020), which predicts the long-term temperature response given the short-term temperature response. Additionally, while such a fast-response emulator is not relevant for long-term policy design, it can have real-world use for predicting regional and global temperature**

responses to short-term forcings that are triggered by episodic pollution sources such as wildfires or sudden anthropogenic pollution decreases (e.g. as during the COVID-19 pandemic).”

L161: You've taken care to perturb SO2 in multiple regions, but then only apply one tropics-wide perturbation for BB OC/BC. Because of the nature of the response to tropical absorbing aerosol, the response to tropical BB is likely to be strongly dependent on the longitude of the perturbation e.g.: https://www.nature.com/articles/s41558-022-01415-4 It would be interesting and worthwhile to see if there are distinct responses in an Asia-Pacific and America-Atlantic-Africa perturbation. There is an argument made here that unrealistically large BB OC/BC forcings will be required for small-scale forcing, but, due to the physical mechanisms involved in the response, you might actually find larger responses when splitting the tropical band in two in this way. Westervelt et al. (2020) perturb South American and African biomass burning emissions in 3 GCMs, by similar amounts to those in this study, and find significant regional and global temperature responses in all models (https://acp.copernicus.org/articles/20/3009/2020/). Besides, in the paragraph beginning on L184, you describe using unrealistically large SO2 perturbations to ensure that you see a signal suitable for training the emulator. Why is it acceptable to do this for SO2, but not for BB OC/BC?

The reason to use a tropics-wide perturbation was primarily for scientific and computational simplicity. We added this perturbation as an additional component, as the biomass burning aerosols are generally less well explored than the sulfate aerosol perturbations. Including regional perturbations would substantially increase the number of inputs and require more expensive GCM simulations, and we did not feel that this was merited when designing our study. However, this could be a useful future step in further developments of this or other emulators.

Also, please note that we do still double the range of the BB OC/BC perturbations to increase the strength of the perturbations for the emulator experiments, as we do for SO2 (e.g., see Supplementary Fig. 2).

We add a comment that we might miss regional patterns due to this perturbation.

Line 212:

“**We note that by using tropics-wide perturbations, the emulator may not capture highly regionalised forcing-response patterns (Westervelt et al., 2020; Williams et al.,**

> **2022). However,** we do not consider OC/BC on smaller regional scales because previous studies have found that large, unrealistic OC/BC forcings are required to obtain climate responses that exceed the internal climate noise (e.g., Myhre et al., 2017; Baker et al., 2015; Stohl et al., 2015)."

L268: 'Most test simulations are reasonably well predicted by the emulator and show patterns of warming or cooling generally in the correct spatial location, similar to Figure 1' . Looking at this figure, I see that the emulator does not correctly capture the sign of the response over some very populous regions (Europe, the Arabian peninsula) and has absolute errors in excess of 1 standard deviation over China, North America, the Amazon basin. By eye, there is clearly a reasonable pattern correlation between the GCM and emulator responses, but some of the regional differences are large. How does one decide when the emulator is good enough? The example in Figure 1 has an incorrect sign and a large absolute error for Europe, and appears to be suggesting a warming for a 5x European SO2 experiment, which seems unphysical to me, despite the moderately high CO2 in the experiment.

We are glad that the reviewer agrees that most test simulations are reasonably well predicted by the emulator and show patterns of warming or cooling generally in the correct spatial location. We are aware that the emulator does not provide a perfect match, but we emphasise that it is a proof concept exercise at this stage. We therefore use conservative language when describing emulator performance, e.g., 'reasonably well predicted'.

We add a comment on this.

Line 355:

> The absolute differences are generally larger at high latitudes, particularly around the Arctic, Greenland and Northern Europe, **for instance, in Figure 1a where the emulator does not capture the cooling over Europe due to the strong SO2 Europe perturbation.**

We agree it is difficult to decide when the emulator is 'good enough'. We believe that the global mean response scatter plots in Figure 2 (and the R^2 metric used in them) are a suitable validation method at this stage. In that figure, we see the emulator is generally correct over the broader regions and the uncertainty bands mostly cover the truth. The uncertainty bands show 1 σ, so we expect that around 30% of the time, the truth falls outside of the predicted uncertainty bands, which aligns with our results here. For example, see Line 389-390, (not changed): "Based on all 18 test simulations, 76.1% of grid points are predicted to within 1 σ of

the true response and 95.0% are predicted to within 2 σ, which is roughly in line with expectations, assuming a Gaussian distribution of responses."

L277: If your points are more likely to fall outside the accuracy threshold in the Northern Hemisphere, does that mean the emulator is particularly struggling with the response to SO2?

We are unsure if this applies to the Northern hemisphere in general, however, we take the reviewer's point that the emulator under-predicts the cooling response to very strong SO2 for this case:

"The absolute differences are generally larger at high latitudes, particularly around the Arctic, Greenland and Northern Europe, **for instance, in Figure 1a where the emulator does not capture the cooling over Europe due to the strong SO2 Europe perturbation.**"

Figure 2: It would be helpful to see these regions on a map. Are they the same as the emission regions? Since one of the more useful features of this emulator is its ability to capture the spatial pattern of the response, it would be interesting and useful to see a similar evaluation of this that assesses the emulated pattern. Even just a pattern correlation between the true and predicted responses would be helpful. I'm certainly more curious about how well the pattern of the response compares to the training data. This figure implies that it is well captured since there is a strong linear relationship in all the regions shown, but it would be nice to have an objective metric. The relationships shown in Figure 2 are stronger than I expected based on Figure 1 and the Supplementary figures.

We thank the reviewer for this suggestion and will include map comparisons of the mean true and predicted responses in the revised manuscript.

L370: Typo on this line – unnecessary ')'

We have removed the ')'.

L389: Doesn't Westervelt et al. 2020 include South America and Africa cases? Since you point the reader to studies with an extratropical focus, it would be helpful also to highlight those that do include regional tropical perturbations, e.g.
https://acp.copernicus.org/articles/23/3575/2023/ ,
https://acp.copernicus.org/articles/20/3009/2020/

We have added these suggested citations:

"This is also one of the first studies to explore the response to regional tropical perturbations (**e.g., Wells et al., 2022; Westervelt et al., 2020)**, …"

Figures 4 and 5: I like what you are trying to do with these figures, and with figure 5 especially, but I am not quite sure how to interpret them. Your maximum CO2 case is a doubling of present day CO2, so essentially you have a 5-yr simulation where you have jumped from the present day to 2100 in SSP5-8.5, where we have a more rapid increase in GHG emissions than current policy. Minimum CO2 is a pre-industrial value, so we are essentially thinking about the variance in the temperature between a pre-industrial climate and something 4-6C warmer (I can't remember where HadGEM sits in this range, but I assume it is towards the upper end). This is then compared to regional aerosol perturbations of various scalings. Both figures show us, essentially, CO2 contributing to most of the variance in most regions. Aerosol contributes to local variance in the emission region, but is essentially swamped by the CO2-driven variance. In the SSPs, most of the aerosol emission uncertainty is seen before 2050, with the pathways largely converging after that, so we would, in the real world, expect most of the aerosol-driven contributions to variance to be seen before 2050, with the CO2-driven variance in the response across different pathways to become larger with time. This is a somewhat long-winded way of saying that, I expect CO2 to contribute most of the variance in most of the regions in this experiment design because the range of CO2 concentrations considered is very large, but I think this is actively unhelpful if you are presenting the emulator as a policy-relevant tool, because we do not expect to see CO2 emissions and aerosol emissions changing across the ranges tested on the same kind of timescales in the real world. In reality, aerosols are likely more important for variance in the response in the near future, while CO2 will become steadily more important out to 2100. I'm also surprised that, even in the regions with very large aerosol emission ranges applied, that they are only contributing to around 5% of the variance. Aerosol has offset around a third of GHG-driven warming over the industrial era, with some large temperature responses over the main emission regions that are also on the order of a third of the GHG-driven response. The CO2 range in these simulations is to then double again on top of this historical change, so if aerosol offset around a third of the warming from 282-400ppm, we might expect it to offset around one fifteenth of the warming from 282-834ppm, or around 5%. But the aerosol ranges being considered are between 0 and 2 to 7x the present day values, so why don't we expect something of the order 10-30% globally, and around 30% regionally? Why is the aerosol contribution so small in these figures? The ranges considered all go from 0 to some regional scaling, so it is not a case of the aerosol effect buffering as we go to higher scalings, as the aerosol reduction, which should produce a larger temperature response for a given emission change, is also considered. Is this simply a reflection of the small magnitude of the fast response

to SO2 compared to the magnitude of the slow response? Or a reflection of the emulator not capturing the magnitude of the regional responses to the aerosol? Or are the colourbars in Figure 5 saturating, so that those dark blue patches are actually showing us something more like 10-20%?

The aerosol contribution is small compared to the CO2 response because we are considering large CO2 perturbations compared to the amount of aerosol perturbations.

We agree that the large perturbations of CO2 sound unrealistic in the short-term time frame, but we selected this to cover all plausible scenarios, keeping in mind that we also have short-to-long-term response relationships (Mansfield et al., 2020) that can help us translate the short-term responses to long-term responses in future work. In the revised version of the paper, we have re-framed the manuscript to show the purpose of the emulator is the first step in a series of emulators, where the subsequent step would be the linking of the framework presented here with the long-term response.

In our revisions, we remind the reader that this is not a likely five-year scenario (as we are modelling the short-term response to an abrupt jump in emissions) when presenting Figure 1, and as the reviewer mentions, this amount of CO2 would not be emitted in such a short time-frame. However, if linked to another emulator that could allow us to infer the long-term response, these magnitudes of CO2 perturbations would be necessary.

Line 345:

> **"Note that these perturbations are not intended to be realistic perturbations over a five-year period, but rather they estimate the short-term response to abrupt perturbations, which at this stage would be used for scientific experimentation, with the view to develop the ideas further in the future for more direct policy relevance."**

As to why the aerosol contribution appears so small, we expect they are small as a percentage of total change when varying across the full range of input values and because CO2 has such a wide range of inputs. But the magnitude of SO2 response is not small in terms of temperature change (e.g., Supplementary Figure S6 shows we are seeing regional temperature around 0.1-0.5K). In some regions, yes, the colourbars are saturating, with some regional values reaching 0.2. We agree that non-linear colour bars significantly improve this figure to highlight this.

Line 508:

> "Firstly, the variance due to the CO_2 concentration is responsible for the majority of the variance in the emulator (~90%) across all regions, which is not surprising given the wide input ranges covered by the emulator based on the high uncertainty of future CO_2 emissions. **The aerosol contributions appear small relative to these large perturbations, and we expect this reflects the weaker range of aerosol perturbations in comparison to the large CO_2 perturbations**."

We think that this does partially mask some of the aerosol signal. Future studies interested in understanding the response to aerosols only may wish to develop emulators without the strong CO2 range, as we comment on in the conclusions:

Lines 725-733:

> **"We made several choices when designing the training and test dataset using GCM simulations, from which future studies can learn. For instance, we chose a maximin Latin hypercube design to sample uniformly across the distribution of each parameter. We extended the parameter space for the aerosol forcings beyond their feasible ranges to account for the weaker aerosol signals compared to the greenhouse gas forcings. Still, we found that many simulations are dominated by the response to greenhouse gas forcings. Future studies may wish to compare the emulator performance under different ratios of extended parameters to realistic parameters. Alternative designs, such as selecting some simulations with restricted greenhouse gas perturbations, which lead to stronger aerosol signals could potentially improve performance in this range. Furthermore, Bayesian optimization techniques could be used to directly identify new parameter choices for additional GCM simulations that would reduce emulator error or uncertainty (Shahriari et al., 2016)"**

Figure 5: Regardless of your thoughts on my above comments, I think you need to use nonlinear colour scales in this figure so that you're not getting so much saturation in those blue colours.

In the revisions, we use non-linear colour bars, as suggested.

Figure 6: I appreciate the note at L465 that the 'emulator does not predict the transient response to the SSP scenarios between present day and 2050, but rather it estimates the short-term climate response to an abrupt jump in emissions'. However, this comes soon after the text at L487: ' we now explore how the emulator could be used to estimate the response to mid-century (2050) conditions from the Shared Socioeconomic Pathways (SSPs), which project socio-economic changes and their associated emissions into the future', which the emulator cannot do. As Figure 6 is presented, it looks like you are using the emulator to predict the transient

response in the SSPs. The titles of the panels are 'predicted response for SSP1 and SSP3'. This is not what they show, nor can it be what they show. The figure caption needs to clearly state that this is a fast temperature response to an instantaneous time slice of emissions representative of 2050 emissions, and the panel titles also need to be clear that this is a fast response. If the emulator can be used to estimate the responses seen in SSP1-1.9 and SSP3-7.0 then we really need to see the current maps in Figure 6 alongside actual SSP1-1.9 and SSP3-7.0 projections for a period centred on 2050 from HadGEM3-GC3.1. For a policy-relevant emulator of the spatial temperature response to various forcings, surely this is the most exciting test, and it's not shown. However, I am not convinced that the emulator can be used for this, in which case this figure is, at best, misleading without clear caveats.

We agree that this is misleading as the emulator prediction cannot easily be compared to SSP simulations. We have removed the SSP demonstration from the main paper and transferred it to the Supplementary, where we also show that the global mean of the CMIP projections are, as expected, higher than the fast-adjustment emulator predictions. We have decided to present this as an example of how an emulator may be used for policy, although currently it is an incomplete and intermediate step when it comes to this purpose. We comment on this in the conclusions:

Line 561-565:

> **"This type of emulator holds value for both scientific experimentation and for policy-relevant studies. For training, we used idealised simulations that isolate the effect of perturbations (i.e., abrupt changes), which makes the emulator well suited for scientific experimentation to rapidly identify the short-term response patterns of GHG and aerosols. Furthermore, the emulator is also a step towards more policy-relevant emulators that would require long-term or even transient climate responses. This could be achieved through coupling to other emulators that predict the long-term climate response given the short-term climate response (e.g., Mansfield et al., 2020). Based on past studies, we would expect to see stronger and more spatially homogeneous responses to forcings over longer timescales, as the slow responses associated with ocean interactions and climate feedbacks dominate (Kasoar et al., 2018; Lewinschal et al., 2019; Liu et al., 2018; Mansfield et al., 2020). Supplementary Fig. S7 demonstrates an example of how this emulator could provide an intermediate step in predicting climate change under SSP scenarios by estimating the fast response. The global mean prediction shows good agreement with the impulse-response model, FAIR, but our approach additionally predicts the full global patterns. As these are only fast responses, they underestimate the full response of the CMIP multi-model means for the SSP scenarios, but the next step would be to map the fast response to the more**

policy-relevant transient climate responses for policy-related applications (Mansfield et al., 2020). ”

L472 and L514: are there any policy-related studies that would consider the fast response in isolation?

We agree that this is less directly related to policy but instead, focus on the scientific results that show the emulator can be used for estimating sensitivity indices. We have removed the section on SSPs, which the reviewer refers to on Lines 472 and 514. Instead, we comment on it in the conclusions as preliminary step, see our response above.

Also, as mentioned earlier, there is an application for considering episodic emissions perturbations, such as wildfires or sudden emissions reductions during the COVID-19 pandemic.

Line 85: **Additionally, while such a fast-response emulator is not directly relevant for long-term policy design, it can have real-world value for predicting regional and global temperature responses to short-term forcings that are triggered by episodic pollution sources such as wildfires or sudden anthropogenic pollution decreases (e.g. as during the COVID-19 pandemic).”**

‘These projections demonstrate how the emulator can be used to predict the spatial response of surface temperature to a specific scenario featuring emission changes from different pollutants at a fraction of the cost of the complete GCM.’ I strongly disagree with this statement. The response of surface temperature in specific scenarios is a slow response to changes in forcing, and I do not expect it to be well captured by an emulator of 5-year responses to a sudden change in forcing. An emulator that could be used to explore a wider range of scenarios than we can afford with GCMs would be a valuable tool, but this emulator is a step towards that, not the finished article.

We agree and have removed the section on SSPs from the main text. We now show it in the Supplementary Fig. 7, which also shows the global mean CMIP transient response is significantly larger than the fast response emulator.