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.

Reviewer 2

Review: Using a Gaussian Process Emulator to approximate the climate response patterns to greenhouse gas and aerosol forcings (egusphere-2025-6046)

1 Overview

This manuscript presents a statistical model in the form of a Gaussian Process (GP) emulator to approximate the surface temperature response as a function of a collection of climate forcings including global greenhouse gas concentrations and regional aerosol emissions. The GP model is built on strong statistical theory and methodology. It yields substantial computational efficiency gains compared with using a Global Climate Model (GCM) directly. The model is tested and applied to the HadGEM3 GCM (for which a good description of the setup is provided) where it is employed to perform a sensitivity analysis, as well as demonstrating how it may be used for policy-assessments under rapid climate change projections. The results are consistent with those obtained in other studies such as the sixth Coupled Model Intercomparison Project (CMIP6), and using reduced complexity models. Furthermore, given the model description and the provided Python code, it would be possible to apply the GP emulator to other GCMs.

This review focuses on the modelling aspect of the manuscript. The research is within the scope of Geoscientific Model Development (GMD) with the application illustrating its use to answer scientific questions of interest to EGU. In addition, the presented methodology constitutes a novel contribution to statistical modelling of climate models by aiming to model spatial patterns rather than the global mean response with respect to time, as well as the capability to incorporate a larger number of climate forcings as input parameters.

The manuscript is logically structured in its presentation of the GCM, input parameters, GP emulator, model training data, and its application and evaluation. Justification is provided throughout for the modelling choices and assumptions explaining potential limitations and how these are addressed. The title clearly describes the content of the manuscript. It is also well written, uses good quality and necessary figures to illustrate the results, contains appropriate references to existing literature, and the abstract provides a good summary.

In summary, this manuscript presents a new form of statistical model as an approximation to output of a GCM which provides an efficient means of performing important uncertainty quantification tasks such as sensitivity analyses and policy assessments. I recommend publication subject to a few minor revisions.

We would like to thank the referee for their time spent reading our manuscript and providing helpful comments which will improve the quality of the paper. We are glad that the reviewer has a positive view of our work and recommends publication of the manuscript following minor revisions. We address the referee's comments as we outline below.

2 Main Points and Suggestions
Below I outline several suggestions to enhance the presentation of the model and clarity of the manuscript.
2.1 Univariate versus Multivariate Emulation
The aim of the GP emulator presented in this manuscript is to predict the entire spatial pattern of the global climate change response. In section 2.2 it is outlined that the Earth is split into N grid cells with independent GP emulators constructed for each grid cell. This is a common approach which has been successfully employed in other climate modelling studies, for example, Johnson et al. (2020) and Lee et al. (2012). However, there is a risk that the emulator mean predictions and their uncertainty for neighbouring grid cells are spatially inconsistent. It is observed in figure 1 and in Supplementary figure S4 that for a variety of input settings that this is not the case in this application, although it is plausible that using a different GCM, training data, or input parameters, that such issues may occur.
Have the authors considered multivariate GP emulators? There exists wide ranging methodology with many having been employed in applications to climate modelling. Some examples include: separable GPs in Conti et al. (2010) and Higdon et al. (2008); outer-product GP emulator in Rougier (2008) and Rougier et al. (2009); dimension reduction of the spatial field, such as in Salter et al. (2022, 2019) and Wilkinson (2010); or parallel partial GP emulators in Gu et al. (2016, 2019). These incorporate extra spatial structure information, thus ensuring spatial consistency, and with the potential for smaller uncertainties through the sharing of information and patterns between grid cells. Moreover, in the spatial statistics literature the Kriging methodology is well established and amounts to GP emulation of spatial fields (see for example Cressie (1993)) with specific covariance functions, such as the C4-Wendland class of correlation functions, designed to incorporate spatial inputs on spheres (Gneiting (2013)). It is acknowledged that these methods are generally result in more complex emulators to formulate and can be more computationally expensive to train, although this is often negligible compared with the cost of running the GCM. There is a trade-off to be achieved with the emulator accuracy. For this manuscript, what is the justification for N independent emulators versus a multivariate approach?

We did consider multivariate GP emulators, however, we found them to be computationally prohibitive to train due to the large number of grid cells, and we found that the independent emulators performed well for our task. We did also explore dimension reduction such as PCA to induce appropriate correlations on the spatial field and found this to give similar results to those presented in our manuscript.

We add a comment on this in the methods:

Line 230:

> "Here, we use independent emulators for each grid-cell, **however, we also found similar results with GP emulators applied to the top principal components, a dimension reduction technique that also enforces spatial correlations between grid-cells (e.g., Salter et al., 2019; Salter & Williamson, 2022; Wilkinson, 2010)**."

We appreciate the suggestion that multivariate GPs may be able to share spatial information, thereby reducing uncertainties and we also comment on this as an extension in the conclusions:

Line 729-731:

> "**Another approach to reducing uncertainties could be to develop multivariate GP emulators which account for the spatial structure and may be able to share information between neighbouring grid cells (e.g., Conti & O'Hagan, 2010; Gu et al., 2019; Gu & Berger, 2016; Higdon et al., 2008; Rougier, 2009; Rougier et al., 2009).**"

In the application to HadGEM3, how many grid cells are there and hence what is N? This will aid the reader in understanding whether this methodology would be applicable to their model where for larger values of N when using a higher resolution model this adds to the computational expense of fitting, testing, and evaluating the emulator.

Updated to include number of grid cells:
Line 292: "Here, we use independent emulators **for each of the $27,648$ grid cells**"

2.2 GP Emulator Structure
Section 2.5 describes the GP emulator structure. In line 226 it is stated that "the prior mean function is often assumed to be zero", for which it is implicit that this is the assumption made throughout this manuscript. An alternative approach is to build more global structure into the mean function, such as using low order polynomials of the input parameters, thus leaving the covariance kernel to handle the smaller scale

local variability, such as in Craig et al. (1996), Edwards et al. (2019), and Vernon et al. (2010). If the underlying mean behaviour is constant with respect to the inputs, then emulators formulated in such a way are still capable of identifying them. Have the authors considered a structured mean function? Would this enable building more accurate emulators? The choice of variables to incorporate in a structured mean function may also be driven by the results shown in the sensitivity analysis where the global CO2 concentration forcing parameter has a large sensitivity index and thus may be informative as a term in the mean function.

In Line 226, we explicitly say 'as done here' to clarify that this assumption is made: "The prior mean function is often assumed to be zero so that all choices are determined by the covariance function, as done here."
We have not extensively explored building global structure into the mean function, although we have used a non-stationary linear covariance function with ARD to determine the slope independently for each input. This should be equivalent to a linear mean function as the reviewer describes and we comment on this.
Line 419:

> "**This captures the first-order linear relationships between emission perturbations and temperature response, while the remaining terms aim to capture the local variability (Craig et al., 1996; Edwards et al., 2019; Vernon et al., 2010).**"

2.3 Covariance Kernel
The construction of the covariance kernel consists of the sum of a linear kernel, a squared exponential kernel, and a white noise process (also known to as a nugget term). However, it is unclear the exact structure this refers to, and the associated hyperparameters. Examples of ambiguities include:
1. Whether it is a homogenous or an inhomogeneous linear kernel,
2. The exponent form in the squared exponential kernel such as using either a single common correlation length (also known as length scale) hyperparameter, separate and potentially distinct correlation lengths, or encompassing normalisation of parameter vector differences via an input variance matrix (the fully anisotropic form of the squared exponential covariance function),
3. Whether a separate variance hyperparameter is used for each additive component, or if a shared variance hyperparameter is employed, potentially with an additional weighting hyperparameter between the two components.
This could be resolved by including the mathematical formula for K(x, x'), highlighting each of the hyperparameters which are estimated via Maximum Likelihood Estimation.

In this formulation, does σGCM correspond to the white noise variance hyperparameter? How does σGP relate to the variance hyperparameter of the non-white noise part of the kernel? Also, is there some intuition for why this additive combination works for the surface temperature response being modelled, for example, due to the different properties of the kernels? Please clarify these details. Furthermore, it is initially unclear that σGP and σtotal are mean uncertainties over the 18 test scenarios. It would be helpful to clarify this and remind the reader that the GP emulator's uncertainty depends on the new input settings at which it is evaluated, and that σGP and σtotal are summary statistics for these. Investigations shown in Supplementary figure S3 illustrate similar results irrespective of which of a variety of covariance kernels are used in the emulator formulation. Why did the authors opt for the more complex kernel described, rather than for one with fewer hyperparameters, such as MatÅLern-5/2, which is commonly employed across climate modelling? Also, what is the intuition for why this form of covariance kernel works? Can certain modes of variability be associated to each component of the covariance function?

We thank the reviewer for pointing this out and have clarified the full covariance structure in Section 3.3, see below. This also includes some descriptions of the meaning of the hyperparameters, and that the white noise kernel used is either fixed based on the GCM values or is a learnable hyperparameter.

Line 270:

The kernel defines the similarity of two inputs $\boldsymbol{x}$ and $\boldsymbol{x}'$ and how this propagates through to the similarity of the outputs $\boldsymbol{f_{GP}(x)}$ **and** $\boldsymbol{f_{GP}(x')}$**. We use**

$$K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{d=1}^{9} \sigma_{Ld}^2 \, x_d \, x_d' + \sigma_{SE}^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{9} \frac{(x_d - x_d')^2}{l_d^2}\right) + \sigma_W^2$$

This describes an additive kernel consisting of three components. The first term represents a non-stationary linear kernel which learns independent hyperparameters, σ_{Ld}^2 , for each input parameter, d (known as Automatic Relevance Determination, ARD), representing the rate of change of the output for each input parameter. This captures the first-order linear relationships between emission perturbations and the temperature response, while the remaining terms aim to capture the local variability (Craig et al., 1996; Edwards et al., 2019; Vernon et al., 2010). The second term is a squared exponential kernel which also uses ARD to learn independent lengthscales l_d for each input parameter, d. These lengthscales describe how rapidly the output changes with respect to varying input parameters. The hyperparameter σ_{SE}^2 is a scale

that describes the total amount of variation in the output (Rasmussen & Williams, 2006). The last term is a white noise kernel which represents the internal variability of the GCM, assumed to be constant, σ_W^2, for all predictions. We explored both fixing σ_W^2 at the internal variability derived from the GCM (described further below and used throughout the paper unless stated otherwise), and allowing σ_W^2 to be learned as an additional hyperparameter (described in Section 3.3). We chose this kernel function because of its interpretability, however, we found the emulator to be robust to a range of kernel choices (Supplementary Fig. S3).

Following the reviewer's comments, we also comment on which terms are dependent on *x* on Line 315 which discusses the internal variability component of the noise:

Line 315:

"We expect this to be present for any prediction and therefore, we include this as a fixed noise term in our emulator, **by setting $\sigma_W^2 = {\sigma_{GCM}}^2$ for the white noise covariance function of the GP emulator.** When the emulator predicts the temperature response at new unseen input values, there is also predictive uncertainty provided by the GP emulator, which we label σ_{GP}. **Note that this comes from the other covariance functions and is dependent on the inputs, x.** The total uncertainty arising from both the GCM internal variability and the GP emulator prediction is denoted σ_{total} and these are related via ${\sigma_{total}}^2 = {\sigma_{GCM}}^2 + {\sigma_{GP}}^2$, discussed later in Section 3.3."

In section 3.3 and in Supplementary figure 5 it is discussed how similar uncertainty estimates for σtotal and σGPlearn are obtained where the latter does not use the fixed estimate of σGCM. Please add information on whether the two emulator structures are otherwise identical. In particular, does the second emulator possess a white noise process for which the variance hyperparameter is estimated simultaneously with all other emulator hyperparameters? This is important for a fair comparison and interpretation of the results. In addition, reporting and contrasting the hyperparameter estimates may divulge why σtotal and σGPlearn are similar, as well as provide an explanation for the main sources of the GP emulators predictive uncertainty.

We do use the same emulator structure with a white noise process and our edits to the kernel structure described above should address this comment and clarify that the emulator structure remains the same. In section 3.3, we add:

Line 405: "For comparison, we built an alternative version of the GP emulator**, where rather than fixing the internal variability based on the GCM internal variability, $\sigma_w^2 = \sigma_{GCM}^2$, we treat the white noise covariance σ_w^2, as an additional hyperparameter to be optimized."**

A related minor point is in the learning of σGCM. In lines 242 – 250 this is estimated based on 14 control runs, 8 of which are obtained by splitting a 40-year control run into 8 non-overlapping segments. It is rightly acknowledged that the sample size is small and that the autocorrelation is likely to have an effect. A potential consideration would be to adjust the estimate of σGCM to account for the autocorrelation in the variance estimate calculation.

We appreciate this point, but we do not pursue this here for simplicity. We have checked that autocorrelation between the 8 simulations is relatively weak, falling between -0.3 and 0.3 for most grid cells for the lagged 1 to 4 autocorrelations functions We also find similar patterns if we compute σ_{GCM} over the 6 independent control runs only.

2.4 Training Design
In section 2.3, lines 184 – 193, the design construction mechanism is described. An interesting feature is the split where 75% (60 simulations) are sampled within the parameter ranges given in table 1, whilst the remaining 25% (20 simulations) are sampled from extended parameter ranges with the 7 regional aerosol forcing parameters upper range doubled. The justification seems reasonable based on improving signal-to-noise ratio for enhanced emulator predictive fit. Linking to section 2.4, where it is stated that a maximin Latin hypercube sampler is implemented, it is unclear how this weighted sampling from a nested parameter space is achieved. Please provide more details as to the sampling algorithm. For example:

• Are two separate maximin Latin hypercube designs over each space constructed (potentially with rejection of points from the larger space)?

We thank the reviewer for highlighting that this was unclear in the manuscript. We do not use two separate Latin hypercube designs. The "75%-25%" split was not clearly defined, as we do not sample 60 simulations from within the parameter space and 20 in the extended range. Rather, we mean that the distribution for the parameters ensures that 75% of the samples are within the defined range, and the remaining 25% are spread further across the parameter space. This means that the points are more spread out for the larger forcings. We clarify in the text:

Line 255-262:

> **"To generate the samples with the distribution described above, we first sample 80 points using a maximin LHS for values between 0 and 1 for all inputs. For the GHGs, we use a linear transform to scale the input range so that all values fall within the defined bounds. For the short-lived pollutants, we use a linear transform to scale the first 75% of samples to lie within the ranges defined in Table 1. We then scale the remaining 25% of samples to fall between 1 and 2X the maximum values in Table 1. This provides us with samples that are further separated for the larger perturbations values, but more closely sampled in the region we are most interested in."**

• Is there a joint optimisation (maximisation of inter-design point distance) across the two components of the design to ensure that points outside of the standard parameter range are not too close to a design point within the range?

No, we use a standard LHS design for simplicity. We clarify:

Line 261: "We also perform an additional 18 simulations for testing, which are sampled separately from the ranges in Table 1 using **a standard LHS design**"

Further questions raised by the design approach are:
1. What is the reason for using a 75-25 split? The extended parameter space possesses a volume 27 = 128 times that of the standard parameter space, yet only a quarter of the design points are sampled from this reason.

We hope that our new description of the sampling approach clarifies this. We do not sample 20 training points from the extended region only (which would mean that all 20 samples would have unrealistically large parameter perturbations), but instead we use a scaling to ensure that 25% of samples come from the extended region *for each input parameter*. This means it is fairly likely that each training point has one large parameter perturbation.

2. Why is the range doubled? This is particularly important if the aim is to have an accurate emulator over the standard parameter space, and given how much larger this extended parameter space is. Moreover, if it is reasonable to believe that aerosol forcings may be large, why is the input parameter range not extended to account for this?

The range is doubled to increase the signal to noise ratio in some of the perturbations and to help the emulator learn the signals. Even with these larger perturbations, we see relatively weak signals compared to the GHG forcing perturbations.

Line 225-230 (not changed):

> "The fact that we apply regional perturbations to the short-lived pollutants inevitably leads to small changes in abundances and weak forcings. This can give weaker grid-scale climate response signals that are difficult to separate from the internal noise (Kasoar et al., 2018; Shawki et al., 2018). For this reason, we extend the input space beyond the ranges described in Table 1 for these forcers, in order to create an emulator that gives less uncertain results. We expect that including simulations with stronger perturbations, which should have larger signal-to-noise ratios, could be beneficial for constraining the emulator prediction to these perturbations. We therefore extend the input space for the short-lived pollutants only so that 75% of simulations fall within the input range indicated in **Error! Reference source not found.** and the remaining 25% lie outside this but within a range that is double this (see Supplementary Fig. S2)."

3. How is fitting emulators using the wider parameter range reflected in estimates of correlation length parameters? If design points in the extended parameter space are a long distance with respect to the correlation lengths from the nearest point at which the emulator is evaluated then they will provide little extra information.

We have looked at the emulator parameters in more detail. The RBF correlation lengths vary significantly between grid cells. There are some regions where the correlation lengths are short suggesting that the additional points do add extra information. We also expect that the wider parameter range help other kernel parameters, including the linear parameters which are estimated to have larger values over the emission region for the aerosol perturbations.

4. What is the additional cost of running the 20 simulations outside the standard parameter ranges, and is this justified by the improvements in emulator predictive uncertainty? In an ideal scenario, although not practical here due to the cost of obtaining further simulations from HadGEM3, a fair comparison would be the emulators obtained using the presented design with an emulator fitted to an 80-point maximin Latin hypercube design over the standard parameter space only.

The reviewer makes an interesting point but we have not explored the additional costs involved or the gain in emulator accuracy. This would be an interesting topic for future work, which we now comment on this in the conclusions:

Line 733-759:

> **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).”**

3 Technical Corrections and Suggestions

• Lines 97-114 – This paragraph describes how the surface temperature ‘response’ is defined. It would be clearer to have this as a mathematical formula as the key quantity of interested to be emulated.

Lines 111-139:

> **We estimate the control surface temperature, by averaging over all years and all control simulations $\overline{\boldsymbol{T}}^{\boldsymbol{CTRL}}$.** All perturbation runs are initialized from the **same state and run for 5 years.** The well-mixed GHGs are perturbed via their global concentrations, while for the short-lived pollutants we use scaling factors for emissions over broad regions, described further in Section 2.2. **We define these as inputs to the GCM, denoted $\boldsymbol{x^p}$ for perturbation $\boldsymbol{p}$. We estimate the perturbation surface temperatures by averaging over the 5 years and we denote this as $\boldsymbol{f_{GCM}(x^p)}$.** We define the **short-term** surface temperature ‘response’ as
>
> $$\boldsymbol{y^p} = f_{GCM}(\boldsymbol{x^p}) - \overline{\boldsymbol{T}}^{CTRL}$$
>
> The surface temperature response is defined at every grid cell of the GCM, giving a complete spatial map, **i.e., $\boldsymbol{y} = (\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_N})$ for all $\boldsymbol{N = 27,648}$ grid cells.**

• Line 112 – It is written that yGP ≈ y. This may be ambiguous and require a more

mathematically precise definition. For example, is it that the GP mean is a close approximation of y, or does this also encapsulate that the uncertainty is "small"?

We have removed this and improved our description of the GP (this directly follows on from paragraph pasted above):

Lines 140 – 145:

> **2.2 Gaussian process emulator**
>
> **Our perturbation simulations create $P = 80$ input-output pairs which we call our training data $D = (x, y)$.** We use a GP emulator to estimate the surface temperature response to a **new** emissions scenario, x^*. **The GP provides a probability distribution for each x^*, from which we can sample, i.e.,**
>
> $$y_{GP} \sim GP(x^*|D)$$
>
> **This means we can also estimate the mean μ_{GP} and uncertainty σ^2_{GP} for each x^*.**

• Lines 211-212 – This is more commonly referred to as a maximin Latin hypercube design.

We add "maximin"

• Line 228 – The kernel K(x, x') = Cov[f(x), f(x')], thus G(x) and G(x') should be replaced f(x) and f(x') respectively.

Thank you for spotting this typo, we have corrected this.

• Line 243 – This is the first reference to a 40-year control run. Should this be introduced in section 2.1 where the six 5-year control runs are first described?

Yes, thank you, we added to Section 2.1 Global Climate Model Configuration:

Line 107-111:

> "**The last 40-years of this simulation is treated as a control run** and this was then used as the starting point for simulations that serve as training data for the emulator.

Before computing any perturbation runs, **an additional** six 5-year-long control simulations were performed starting from this equilibrium state**, giving a total of 70-years of control simulations.**"

• *Line 360 – Point 1 – For the purpose of reproducibility, please provide more information on how the inputs are sampled including their variance (or standard deviation), and whether they are jointly or independently sampled.*

Line 449 "we sample all other inputs from **independent** normal distributions, **centered at present day levels with standard deviations set to ¼× the input range presented in Table 1**."

• *Line 360 – Point 2 – Ranges are defined for each of the input parameters. Are these limits enforced here?*

These limits are not strictly enforced here.

• *Line 511 – Add a citation to the heteroscedastic GP emulation paper Binois et al. (2018), and variance emulation, such as in Andrianakis et al. (2017).*

We have added these citations.

References

Andrianakis, Ioannis, Ian R. Vernon, Nicky McCreesh, Trevelyan J. McKinley, Jeremy E. Oakley, Rebecca N. Nsubuga, Michael Goldstein, and Richard G. White (2017). "History matching of a complex epidemiological model of human immunodeficiency virus transmission by using variance emulation". In: Journal of the Royal Statistical Society: Series C (Applied Statistics) 66.4, pp. 717–740. issn: 0035-9254. doi: 10.1111/rssc.12198.
5
Binois, MickaÅNel, Robert B. Gramacy, and Mike Ludkovski (2018). "Practical Heteroscedastic Gaussian Process Modeling for Large Simulation Experiments". In: Journal of Computational and Graphical Statistics 27.4, pp. 808–821. doi: 10.1080/10618600.2018 . 1458625. url: https : / / www . tandfonline . com / doi / full / 10 . 1080 / 10618600.2018.1458625.
Conti, Stefano and Anthony O'Hagan (2010). "Bayesian emulation of complex multioutput and dynamic computer models". In: Journal of Statistical Planning and Inference

140 (3), pp. 640–651. doi: 0.1016/j.jspi.2009.08.006.
Craig, Peter S., Michael Goldstein, Allan H. Seheult, and James A. Smith (1996). “Bayes linear strategies for matching hydrocarbon reservoir history”. In: Bayesian Statistics 5. Ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. Proceedings of the Valencia International Meeting 5. Clarendon Press, pp. 69–95. isbn: 978-0198523567. doi: 10.1007/978-1-4612-2290-3_2.
Cressie, Noel A.C. (1993). Statistics for Spatial Data. Revised. Wiley Series in Probability and Statistics. New York, NY: John Wiley & Sons. isbn: 978-0-471-00255-0. doi: 10.1002/9781119115151.
Edwards, Tamsin L., Mark A. Brandon, Gael Durand, Neil R. Edwards, Nicholas R. Golledge, Philip B. Holden, Osabel J. Nias, Antony J. Payne, Catherine Ritz, and Andreas Wernecke (2019). “Revisiting Antarctic ice loss due to marine ice-cliff instability”. In: Nature 566, pp. 58–64. doi: 10.1038/s41586-019-0901-4.
Gneiting, Tilmann (2013). “Strictly and non-strictly positive definite functions on spheres”. In: Bernoulli 19.4, pp. 1327–1349. doi: 10.3150/12-BEJSP06.
Gu, Mengyang and James O. Berger (2016). “Parallel Partial Gaussian Process Emulation for Computer Models with Massive Output”. In: The Annals of Applied Statistics 10.3, pp. 1317–1347. issn: 1932-6157. doi: 10.1214/16-AOAS934.
Gu, Mengyang, Jesus Palomo, and James O. Berger (2019). “RobustGaSP: Robust Gaussian Stochastic Process Emulation in R”. In: The R Journal 11.1, pp. 112–136. doi: 10.32614/RJ-2019-011.
Higdon, Dave, James Gattiker, Biran Williams, and Maria Rightley (2008). “Computer Model Calibration Using High-Dimensional Output”. In: Journal of the American Statistical Association 103.482, pp. 570–583. doi: 10.1198/016214507000000888.
Johnson, J. S. et al. (2020). “Robust observational constraint of uncertain aerosol processes and emissions in a climate model and the effect on aerosol radiative forcing”. In: Atmospheric Chemistry and Physics 20.15, pp. 9491–9524. doi: 10.5194/acp-20-9491-2020. url: https://acp.copernicus.org/articles/20/9491/2020/.
Lee, L. A., K. S. Carslaw, K. J. Pringle, and G. W. Mann (2012). “Mapping the uncertainty in global CCN using emulation”. In: Atmospheric Chemistry and Physics 12.20, pp. 9739–9751. doi: 10. 5194/acp- 12 - 9739- 2012. url: https: //acp. copernicus.org/articles/12/9739/2012/.
Rougier, Jonathan (2008). “Efficient Emulators for Multivariate Deterministic Functions”. In: Journal of Computational and Graphical Statistics 17.4, pp. 827–843. doi: 10.1198/106186008X384032.

Rougier, Jonathan, Serge Guillas, Astrid Maute, and Arthur D. Richmond (2009). “Expert Knowledge and Multivariate Emulation: The Thermosphere–Ionosphere Electrodynamics

General Circulation Model (TIE-GCM)". In: Technometrics 51.4, pp. 414–424. doi: 10.1198/TECH.2009.07123.
Salter, James M. and Daniel B. Williamson (2022). "Efficient Calibration for High-Dimensional Computer Model Output Using Basis Methods". In: International Journal for Uncertainty Quantification 12.6, pp. 47–69. issn: 2152-5099. doi: 10.1615/Int.J.UncertaintyQuantification.2022039747.
Salter, James M., Daniel B. Williamson, John Scinocca, and Viatcheslav Kharin (2019). "Uncertainty Quantification for Computer Models With Spatial Output Using Calibration-Optimal Bases". In: Journal of the American Statistical Association 114.528, pp. 1800–1814. doi: 10.1080/01621459.2018.1514306.
Vernon, Ian, Michael Goldstein, and Richard G. Bower (2010). "Galaxy Formation: a Bayesian Uncertainty Analysis". In: Bayesian Analysis 5.4, pp. 619–670.
Wilkinson, Richard D. (2010). "Bayesian Calibration of Expensive Multivariate Computer Experiments". In: Large-Scale Inverse Problems and Quantification of Uncertainty. Ed. by Lorenz Biegler, George Biros, Omar Ghattas, Matthias Heinkenschloss, David Keyes, Bani Mallick, Youssef Marzouk, Luis Tenorio, Bart van BloemenWaanders, and Karen Willcox. Wiley. Chap. 10, pp. 195–215. isbn: 9780470697436. doi: 10.1002/9780470685853.ch10.