ENVIRONMENTAL RESEARCH
CLIMATE

PURPOSE-LED PUBLISHING™ Part of the ENVIRONMENTAL RESEARCH SERIES

PAPER • OPEN ACCESS

Detecting changes in large-scale metrics of climate in short integrations of a global storm-resolving model of the atmosphere

To cite this article: Ilai Guendelman *et al* 2025 *Environ. Res.: Climate* **4** 025010

ENVIRONMENTAL RESEARCH
CLIMATE

PAPER

Detecting changes in large-scale metrics of climate in short integrations of a global storm-resolving model of the atmosphere

Ilai Guendelman[1,*], **Timothy M Merlis**[1], **Kai-Yuan Cheng**[1], **Lucas M Harris**[2], **Christopher S Bretherton**[3], **Maximilien Bolot**[1], **Linjiong Zhou**[1], **Alex Kaltenbaugh**[2], **Spencer K Clark**[2,3] and **Stephan Fueglistaler**[1]

[1] Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, United States of America
[2] Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, NJ, United States of America
[3] Allen Institute for Artificial Intelligence, Seattle, WA, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: ig1245@princeton.edu

Keywords: climate change, global storm resolving models, large-scale circulation

OPEN ACCESS

RECEIVED
11 November 2024

REVISED
2 May 2025

ACCEPTED FOR PUBLICATION
8 May 2025

PUBLISHED
27 May 2025

Abstract

Recent advances have allowed for integration of global storm resolving models (GSRMs) to a timescale of several years. These short simulations are sufficient for studying aggregated statistics of short-timescale and small spatial-scale phenomena; however, it is questionable what we can learn from these integrations about the large-scale climate response to perturbations. To address this question, we use the response of X-SHiELD (a GSRM) to uniform sea surface temperature warming and CO_2 increase in two-year integrations and compare it to similar CMIP experiments. Specifically, we assess the statistical meaning of having two years in one model outside the spread of another model or model ensemble. This is of particular interest because X-SHiELD shows a distinct response of the global-mean precipitation to uniform warming and the northern hemisphere jet shift response to isolated CO_2 increase. To estimate the probability of X-SHiELD's and the CMIP models having different means, we take the approach of Bayesian inference. We derive a posterior distribution for the differences in the mean between X-SHiELD and the CMIP models taking into account the X-SHiELD values for the global-mean precipitation response to uniform warming and the response of the norther hemisphere jet latitude to isolated CO_2 increase. We find that the most probable value for the difference between X-SHiELD and the CMIP mean is larger than one standard deviation, representing both internal variability and inter-model spread of the CMIP models. We also find that there is an important base-state dependence for some large-scale metrics that, when taken into account, can qualitatively change the interpretation of the results. We note that a year-to-year comparison is meaningful due to the use of prescribed sea-surface-temperature simulations.

1. Introduction

Traditional global climate models rely on parameterizations of small-scale phenomena. A major gap in traditional climate models is the representation of convection and moist processes (e.g. Stevens and Bony 2013, Bony *et al* 2015, Palmer and Stevens 2019, Balaji *et al* 2022). Through interaction between scales, different representations of convection can be a source of large-scale biases and spread in the models' response to perturbations.

One pathway to overcome the need for deep convection parameterizations is increasing model resolution to a point where deep convection is resolved, i.e. kilometer scale (km-scale) or global storm resolving models (GSRMs). Indeed, with recent computational advancements, GSRM simulations are becoming more abundant. For example, the DYnamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains intercomparison consists of 40 day integrations of different atmosphere only GSRMs (Stevens *et al*

2019). Following that, recent work has suggested a protocol for a longer, year-long integration for a control climate and a uniform 4 K increase in sea surface temperature (SST, Takasuka *et al* 2024). In addition recent studies have used longer integrations on the scale of months to several years in present-day and perturbed climates with different GSRMs (Cheng *et al* 2022, Bolot *et al* 2023, Bao *et al* 2024, Guendelman *et al* 2024, Rackow *et al* 2025, Merlis *et al* 2024a, 2024b). A main limitation for GSRM simulations is still computational time and cost. For example, a year-long integration of the eXperimental System for High-resolution prediction on Earth-to-Local Domains (X-SHiELD) model, an atmosphere-only GSRM requires ∼10M CPU hours. This computational cost is several orders of magnitude more expensive compared to a coarse-grid full earth system model and even more compared to a coarse atmosphere-only run. For comparison, the computational cost of GFDL-ESM4.1 is ∼10K CPU hours per simulated year (Dunne *et al* 2020) and for the atmosphere-only model, AM4.1, it is ∼1K CPU hours for a year (Zhao *et al* 2018). Other modeling centers have had a similar experience (Hohenegger *et al* 2023). Since a climate-timescale simulation (at least ∼30 years) of a GSRM would be extremely expensive with current computational limitations, it is useful to examine what we can learn from a GSRM simulation of only a few years. More specifically, can an integration of individual years, say one to three years, give indications on how different or similar the response of GSRMs is when compared to traditional Coupled Model Intercomparison Project Phase 6 (CMIP) models and other GSRMs?

Short integrations are sufficient when the focus is on short-timescale and small spatial-scales, where it is possible to aggregate their statistics and study their distributions. Alternatively, examining variables that quickly converge such as global means, is feasible with short integrations. On the one hand, focusing on aggregated statistics and global means is limiting because it is difficult to translate changes in a distribution or global mean to regional ones. On the other hand, regional changes pose a challenge because they are noisy, so short integrations are inappropriate for this purpose. Zonal-mean variables and large-scale climate metrics are a compromise, as they are more descriptive of the spatial dynamics than global means and are less variable in time compared to regional climate. However, large-scale metrics are still known to be influenced by interannual variability (e.g. Waugh *et al* 2018), that arises both from interannual variation in the SST and from internal atmospheric variability. It is, therefore, ambiguous whether year-long (or two or three year-long) integrations are adequate to study them. Nevertheless, one of the distinctive capabilities of GSRMs is that they simultaneously and consistently simulate large scales and convective scales, in contrast to traditional climate models, so it is natural to investigate how their large-scale responses differ from coarser global models.

In this study, we assess what can we learn about the response of different large-scale climate metrics from two-year simulations of a GSRM. Specifically, we compare the responses of X-SHiELD (a GSRM) to uniform SST warming and the isolated effect of increased CO_2 concentrations against responses from CMIP models. More specifically, we focus on the case where the two years are outside one standard deviation (σ) from the mean of another model or a multi-model ensemble. In section 2, we present the large-scale circulation metrics used to compare X-SHiELD and the CMIP models in this study. In section 3, we we compare the response in X-SHiELD to that of the CMIP models. In section 4, we use CMIP models to test the statistical meaning of having two years in one model outside the mean of another model or a model ensemble. In section 5, we show that the response has a base-state dependence for some of the metrics. That is, the response in a specific year depends on the control values of the metric. We show that taking this base-state dependence into account can qualitatively shift the comparison between X-SHiELD and CMIP models. We end by discussing the limitation of using two years and other potential ways to overcome the computational limitations of GSRMs in section 6.

2. Methods

The models used in this study are X-SHiELD, a GSRM, and traditional, coarse resolution (order 100 km), atmosphere-only (AMIP) models with prescribed SST from CMIP. We analyze models with a uniform 4 K warming in SST (amip-p4K in CMIP) and an increase in CO_2 (amip-4xCO_2 in CMIP) with fixed control SST. Note that the CO_2 increase in X-SHiELD is to 1270 ppmv that is ≈3.12 times the control but we rescale the response to a 4xCO_2 perturbation. We run X-SHiELD from 20 October 2019 to 10 January 2022 with the SST nudged to real-time European Centre for Medium-Range Weather Forecasts SST analyses with a 15 day relaxation timescale, and we analyze the output for the simulated 2020 and 2021. The choise of the period is in accordance to the protocol of the future year-long GSRM intercomparison (Takasuka *et al* 2024). X-SHiELD has a horizontal resolution of ≈3.25 km globally with 79 vertical layers. The deep convection parameterization is turned off and a shallow convection parameterization is used. More details on model configuration and model biases can be found in Cheng *et al* (2022) and Guendelman *et al* (2024). We compare X-SHiELD's response with a single member from each of the following CMIP models:

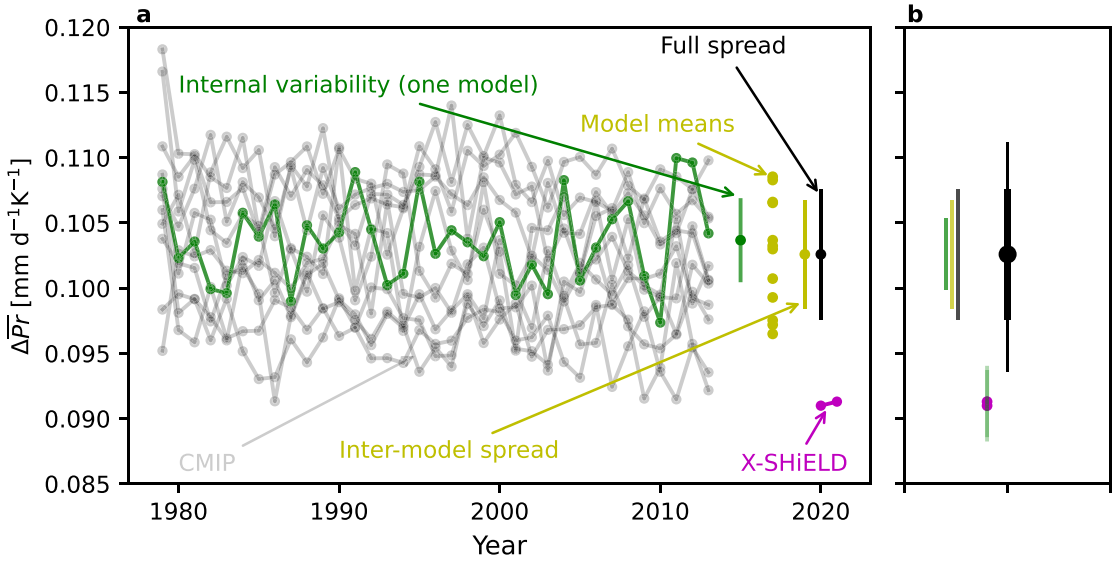

Figure 1. Comparison between X-SHiELD and CMIP global mean precipitation response to uniform warming and estimation of different sources of spread in the CMIP models. (a) Faded black curves show the time series of $\overline{Pr}$ response to uniform warming for the different models and the black vertical line represents one standard deviation across all years and all models. The green curve is the timeseries of an individual model and the green vertical line represents the internal variability of that model. Yellow points represent the mean response of the different CMIP models and the yellow vertical line represents one standard deviation of that quantity. Magenta points are for the two X-SHiELD years. (b) A summary of panel (a): in the center of the plot is a whisker plot of the response across all years and models in CMIP, where the dot is the mean, the thick line is one standard deviation, and the thin black line represents the range between the 2.5th and 97.5th percentiles. The black and the yellow lines to the left of the whisker plot correspond to the black and yellow vertical lines in panel (a). The green line represents the mean internal variability across the models.

BCC-CSM2-MR, CanESM5, CNRM-CM6-1, IPSL-CM6A-LR, E3SM-1-0, MIROC6, HadGEM3-GC31-LL, MRI-ESM2-0, GISS-E2-1-G, CESM2, NorESM2-LM, GFDL-CM4. The CMIP models are all integrated from 1979 to 2014.

We focus on the following metrics:

(i) $\overline{Pr}$ — Global-mean precipitation.
(ii) S_f — Subsidence fraction, that is, the fraction of the area with downward motion at 500 hPa in the tropics (30S–30N), calculated as the annual mean of the monthly mean subsidence fraction.
(iii) ϕ_{Pr-E} — The annual-mean width of the tropics, defined as the latitude range between $Pr - E = 0$ (beyond the subtropics).
(iv) ϕ_{U_e} — The latitude of the northern hemisphere (NH) eddy-driven jet, defined as the latitude of the maximum 850 hPa wind speed in the NH.

Intuitively, the global-mean precipitation and subsidence fraction are metrics that one would assume to be influenced by the fact that the deep convection is resolved and not parameterized. Conversely, the eddy-driven jet and the tropical width are expected to be mostly determined by large-scale environmental conditions such as the static stability and the meridional temperature gradients (e.g. Waugh *et al* 2018, Chemke and Polvani 2019). That said, these metrics can be indirectly influenced by the representations of convective processes (e.g. Garfinkel *et al* 2024).

Figure 1(a) highlights the difference between using a short, two-year integration in a single model (X-SHiELD, magenta) and analyzing more than 30 years across multiple models (CMIP, black), using the global-mean precipitation response to uniform warming as an example, which depicts the challenges of comparing X-SHiELD and CMIP responses. The main challenge is that while the spread in CMIP models can be estimated, we cannot estimate it for X-SHiELD. Moreover, the CMIP spread is comprised of two sources:

- **Internal variability**, which accounts for interannual and internal atmospheric variability, represented by the green curve and line in figure 1(a), using an individual model as an example. The mean value of the internal variability across the CMIP models is represented by the green line in figure 1(b).
- **Inter-model spread** reflects the differences in the mean response of different models, represented by the yellow points and a line in figure 1(a) (with the same representation in figure 1(b).

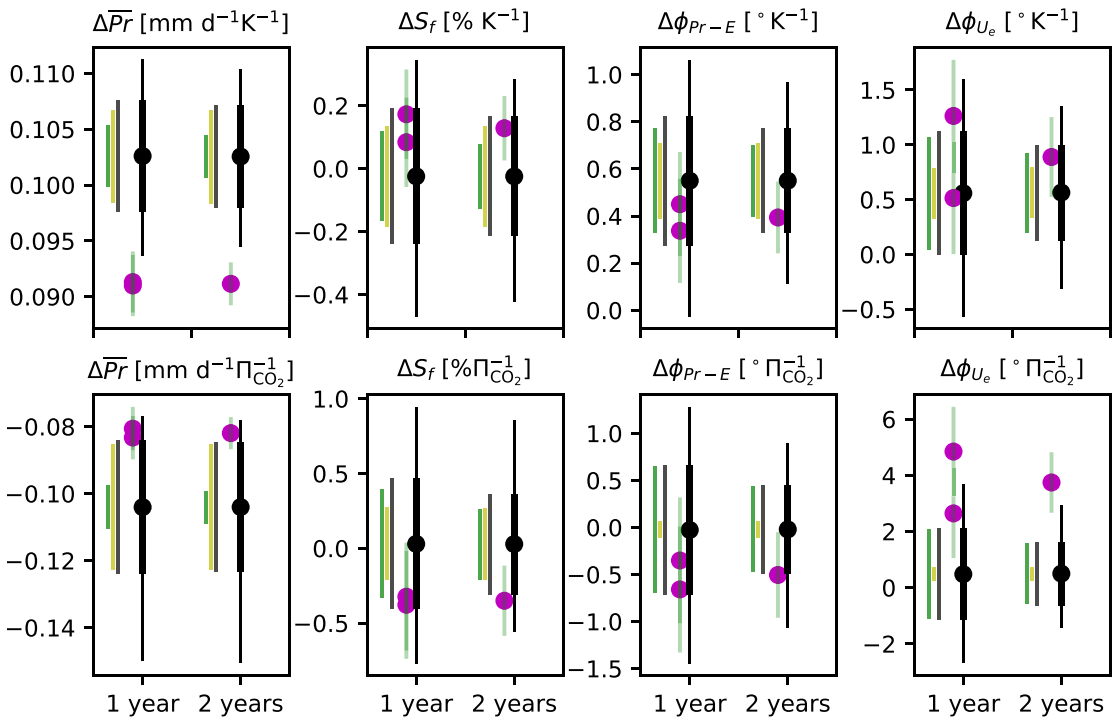

Figure 2. A comparison between the response of CMIP models and X-SHiELD to uniform warming (top row) and CO_2 increase (bottom row), normalized by the perturbation magnitude, i.e. divided by 4 K for the warming and by the appropriate logarithmic scaling with CO_2 denoted by Π_{CO_2}, that is log(4) and log(3.12) for CMIP and X-SHiELD, respectively. For each subplot the left side lines are for yearly data and the right side lines are for moving two-year mean. Black dots represent the CMIP multi-model mean, thick black lines show one standard deviation (σ) of all CMIP years, thin lines show all years between the 2.5 and 97.5 percentiles. The green line represents the mean standard deviation of individual years from individual CMIP models, a measure of internal variability, these lines are analogus to ones detailed in figure 1. The yellow line represents the standard deviation of the time mean CMIP models, that is the inter-model spread. The faint black line is the same as the thick black line, i.e. the standard deviation across models and years. Magenta points are for X-SHiELD's two years; faint magenta lines have the same magnitudes as the green lines.

In contrast, X-SHiELD is a single model with only two years of data, meaning that we lack information about its mean or internal variability. This poses a challenge when interpreting its response when compared to the CMIP ensemble. In figure 1(b), we introduce a whisker plot representation for the comparison between the models that summarizes figure 1(a). We will use this representation again in figure 2 for the response of the different metrics discussed above.

A second challenge with these simulations is the fact that X-SHiELD is integrated over a different period than the CMIP models (compare the x-axis position of the black and magenta dots in figure 1(a)). To address these challenges, we will take the following steps:

- We will use the CMIP models, where both the mean state and spread are known, to assess what we can learn about the differences between two model means when given only two years of data from one model.
- We will use values from the control simulation to correct for potential base-state dependencies, accounting for a potential bias resulting from the base-state, which can account for the different integration periods.

While differences in the El Niño state may account for some of the interannual variability in climate metrics (e.g. Seager *et al* 2003), a more direct comparison is to use the control values. This is physically consistent since we use AMIP simulations where the SSTs in the control and perturbation simulations are similar, allowing for a direct comparison.

3. Model comparison

Figure 2 shows the distribution of the response to uniform warming (top) and CO_2 increase (bottom) for both CMIP models (black and green lines and dots) and X-SHiELD (magenta dots). The whisker plots and vertical lines follow the same format as in figure 1(b), with the left side representing the spread among individual years and the right side representing the spread among rolling two-year means (including overlaps). For clarity, we repeat the plot description here. Black dots represent the CMIP multi-model mean, thick lines represent one standard deviation (σ) of all CMIP years, and thin lines represent all years between

the 2.5 and 97.5 percentiles. In this way, the spread represents both internal variability and inter-model spread. In each plot, the left side lines is for the spread of individual years and the right side lines for two-year mean, where for the CMIP models it is a rolling two-year mean.

To get a sense of the relative roles of internal variability and model spread, to the left of each whisker plot, we plot a green, yellow, and faded black line. The faded black line is the same as the thick black line in the whisker plot and is plotted to ease the comparison. The green line represents the mean standard deviation of individual models, that is, calculating the standard deviation of each model and taking the multi-model mean of this quantity. In other words, this line represents the mean internal variability across the different models. The yellow line represents the standard deviation of all model means, that is, calculating the time mean of each model and subsequently calculating the standard deviation. In other words, this line represents the inter-model spread of the means.

For the global-mean precipitation, the internal variability is significantly smaller than the inter-model spread. This is expected, as the global-mean precipitation is constrained by the energy balance (Allen and Ingram 2002), which converges quickly and has smaller internal variability compared to the other metrics. For the other metrics, the internal variability is comparable or larger than the inter-model spread of the means.

The internal variability decreases when considering the spread across two-year rolling means instead of individual years, and in some cases the two-year mean variability becomes comparable to the inter-model spread. This is not the case for the response of $\phi_{\mathrm{Pr}-E}$ and ϕ_{U_e} to increase in CO_2 and ϕ_{U_e} response to uniform warming, where the two-year mean internal variability is still larger.

X-SHiELD's response to uniform SST warming is mostly qualitatively similar to the response in the CMIP models. It consists of an increase in the global-mean precipitation, a small increase in the subsidence fraction, widening of the tropics, and a poleward shift of the eddy-driven NH jet. While for S_f, $\phi_{\mathrm{Pr}-E}$ and ϕ_{U_e} at least one of X-SHiELD's years falls within one σ of the CMIP multi-model mean, the global mean precipitation response to uniform warming in X-SHiELD is weaker, and both X-SHiELD years are below the 2.5 percentile of all CMIP years. Note that when considering the relative change ($\Delta\overline{Pr}/\overline{Pr}$), this difference is smaller, but X-SHiELD years are still well outside one σ from CMIP mean (Guendelman *et al* 2024).

X-SHiELD's response to CO_2 increase with perscribed SST consists of a decrease in global-mean precipitation, subsidence fraction, and the width of the tropics and has a poleward shift of the NH eddy-driven jet. The response of $\phi_{\mathrm{Pr}-E}$ in both X-SHiELD years is within one σ of the CMIP multi-model mean. The response of $\overline{Pr}$ and S_f in both X-SHiELD years are approximately one σ from the CMIP mean, and X-SHiELD's response of ϕ_{U_e} is outside one σ from the CMIP mean.

4. What can we learn from two years about the model mean?

The main caveat in this comparison is that due to computational limitations the X-SHiELD simulations are limited to two years. This raises the question about the implication of having the two X-SHiELD years outside one σ (or more) of the CMIP mean? Such as the one observed for the global mean precipitation ($\overline{Pr}$) response to warming and eddy-driven jet latitude (ϕ_{U_e}) response to CO_2 increase.

One possible way to assess the difference between the response of X-SHiELD and the response of the CMIP models is to assume that X-SHiELD has similar internal variability as the CMIP models. We can overlay the presumed variability on the different X-SHiELD points and examine their overlap with the CMIP spread. This is represented in figure 2 by the faded green lines. Based on this comparison one can claim that the precipitation response in X-SHiELD is significantly different. However, this interpretation is potentially misleading, as for the CMIP models, the black points in figure 2 represents the mean and the lines represent the spread, while for X-SHiELD the point represent only one data point (either one year or a two-year mean) so the spread is not necessarily centered around it. Additionally, figure 2 shows that focusing on the two-year mean can be also misleading. For example, taking the response of ϕ_{U_e} to increase in CO_2, if we would only look at the two-year mean data, we would be more inclined to state that X-SHiELD's response is significantly different from CMIP models. However, when examining the two years separately the picture becomes less clear as one of the years is closer to the mean of the CMIP models. If we assume that this year is a better representation of the X-SHiELD mean, we would come to a different conclusion in regards the comparison between the response in X-SHiELD and CMIP. This highlights that the analysis of separate years incorporates additional information that is lost when considering the mean; in particular, there is additional information about the model internal variability. This also points to the need for a more comprehensive statistical analysis, especially when we work with small sample. Note that a year-to-year comparison is more justified due to the use of fixed SST simulations, that is, the boundary conditions are the same and the perturbations are well defined. This is not the case for coupled models, as their SST evolve, making a year-to-year comparison difficult in a statistical sense, especially when considering metrics that are not global mean.

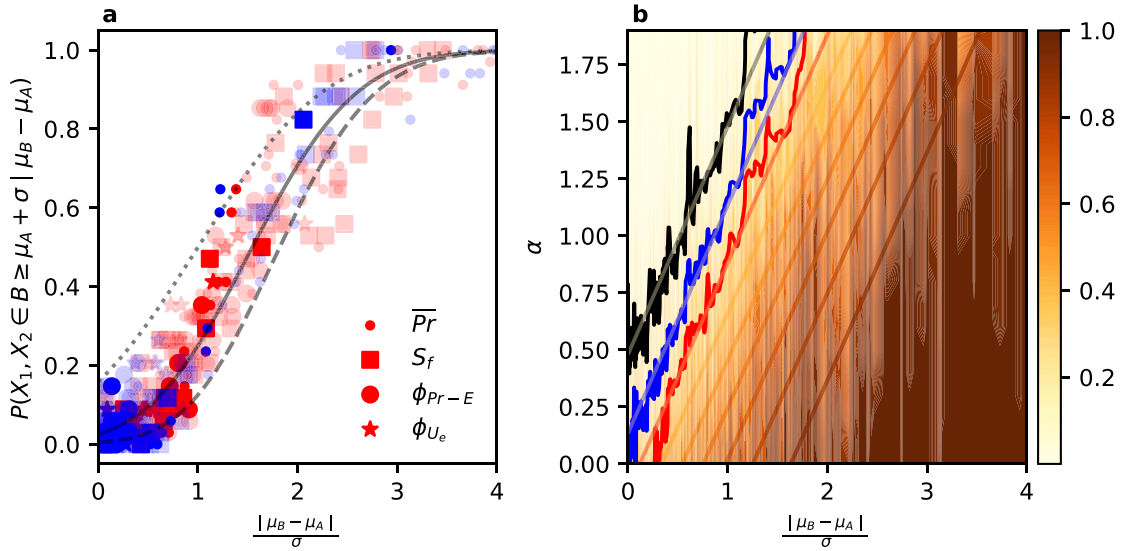

Figure 3. Bootstrapped estimates from the CMIP models to address the question of what are the chances for two years in one model (B) to be one σ (panel (a), or $\alpha\sigma$ in panel (b)) out of the mean of model A, μ_A as a function of the distance between the two model means $(\mu_A - \mu_B)/\sigma$? Panel (a) shows the chance for two years in B to be one σ out of μ_A for the different metrics, with symbols given in the legend. Solid symbols are individual models with respect to the remaining multi-model and transparent symbols are comparisons between different model pairs. Red colors are for the response to uniform SST increase and blue colors are for the response to CO_2 increase. The lines are predictions from normal distribution for one year (dotted), two years (solid), and three years (dashed) to be outside one σ (equation (3)). Panel (b) adds α, a parameter that represents the distance of two years in B from the mean of A in multiple of σ. Colors show the probability of two years in model B to be $\alpha\sigma$ away from the mean of A knowing $\mu_B - \mu_A$, in different pairs of the CMIP model ensemble for the different metrics and experiments. Black, blue, and red line represent the 0.1, 0.2, and 0.3 probabilities, respectively (after smoothing). The faint lines are predictions from normal distribution.

Since some of the metrics appear to be outliers in X-SHiELD when compared to CMIP models, we ask: What can we infer about how different the mean value of X-SHiELD is from the CMIP mean, knowing that two X-SHiELD years are outliers? This is a conditional probability question and a Bayesian approach is appropriate here. For the sake of clarity, we start with a generalized formulation, and following that we will apply that to the comparison between X-SHiELD and the CMIP models.

Assume that we have two distributions $A \sim \mathcal{N}(\mu_A, \sigma^2)$ and $B \sim \mathcal{N}(\mu_B, \sigma^2)$, in the context of this study, A can be the response of a metric in CMIP models and B is X-SHiELD's response of the same metric. For simplicity we assume that they distribute normally, however, this can be generalized to other distributions. We also assume that the distributions have the same standard deviation, when applying this to CMIP and X-SHiELD this is actually a conservative assumption as will be made clear later. We know μ_A and we have 2 points (in the context of this study two years), $X_1, X_2 \in B$ that we measured $X_1, X_2 \geqslant \mu_A + \sigma$. Formally, the question we posed can be expressed as $p(\mu_B - \mu_A \mid X_1, X_2 \geqslant \mu_A + \sigma)$. For simplicity, we focus on this case; the symmetrical case $p(\mu_B - \mu_A \mid X_1, X_2 \leqslant \mu_A - \sigma)$ is treated analogously. Using Bayes theorem we can write:

$$p(\mu_B - \mu_A \mid X_1, X_2 \geqslant \mu_A + \sigma) = \frac{p(X_1, X_2 \geqslant \mu_A + \sigma \mid \mu_B - \mu_A)\, p(\mu_B - \mu_A)}{p(X_1, X_2 \geqslant \mu_A + \sigma)}. \tag{1}$$

Note that we will use capital P for the cumulative probability function (cdf) and lower-case p for the probability density function (pdf). We can easily estimate $P(X_1, X_2 \geqslant \mu_A + \sigma \mid \mu_B - \mu_A \leqslant \sigma)$, that is the cdf for two points in model B to be $\geqslant \mu_A + \sigma$, knowing $\mu_B - \mu_A$, using the CMIP models by bootstrapping (resampling with replacement) different 2 sequential years of one model (B) and calculating the rate of occurrence of these two years being one standard deviation outside of the rest of the models (solid symbols in figure 3) or another individual model (faded symbols) mean. Figure 3 shows $P(X_1, X_2 \in B \geqslant \mu_A + \sigma \mid \mu_B - \mu_A)$, for orientation it is easy to first focus on the two extremes. Starting with the case where $\mu_B - \mu_A \approx 0$, in this case we expect $P(X_1, X_2 \in B \geqslant \mu_A + \sigma \mid \mu_B - \mu_A)$ to be low. On the other extreme, that is, $\mu_B - \mu_A \approx 4\sigma$ we expect all years in B to be more than σ away from μ_A, that is, $P(X_1, X_2 \in B \geqslant \mu_A + \sigma \mid \mu_B - \mu_A) \approx 1$. We can also calculate the expected probability for two normal distributions and can generalize it for n points,

$$\begin{aligned} P(\forall k \in \{1, \ldots, n\} \mid X_k \in B \geqslant \mu_A + \sigma \mid \mu_B - \mu_A) &= (1 - P(X \leqslant \mu_A + \sigma))^n \\ &= \left(1 - P\left(\frac{X - \mu_B}{\sigma} \leqslant \frac{\mu_A - \mu_B + \sigma}{\sigma}\right)\right)^n, \end{aligned} \tag{2}$$

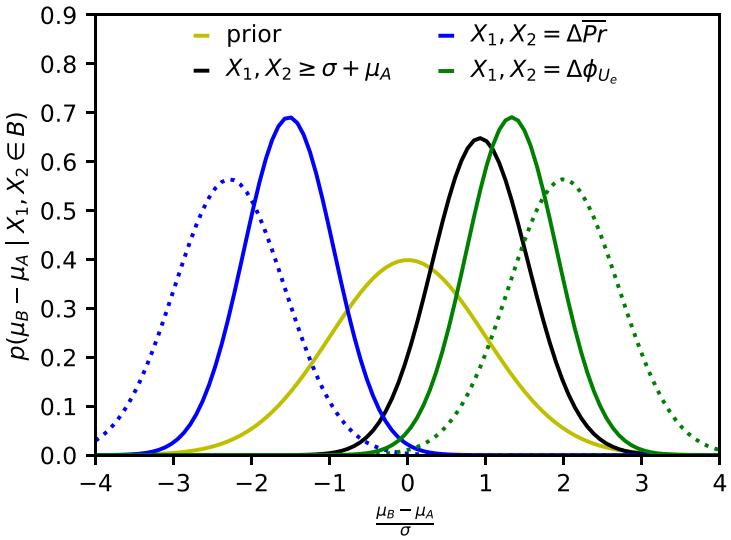

Figure 4. Posterior density probability functions of $\mu_B - \mu_A$ (normalized by σ) knowing $X_1, X_2 \in B$. This figure shows the prior (yellow) and posterior distributions for X_1, X_2 larger than σ (black), taken using the X-SHiELD and CMIP values for $\overline{Pr}$ (blue) and ϕ_{U_e} (green). The dotted lines are for a distribution of $\sim\mathcal{N}\left(\frac{X1+X2}{2} - \mu_{\text{CMIP}}, \frac{\sigma^2}{2}\right)$, where μ_{CMIP} is the the CMIP multi-model mean.

because X_k is in B and B is a normal distribution, this means that $\frac{X-\mu_B}{\sigma}$ is distributed in standard normal distribution and this gives:

$$P(\forall k \in \{1,\ldots,n\} \mid X_k \in B \geqslant \mu_A + \sigma \mid \mu_B - \mu_A) = \left\{\frac{1}{2}\left[1 - \text{erf}\left(\frac{\mu_A - \mu_B + \sigma}{\sqrt{2}\sigma}\right)\right]\right\}^n, \tag{3}$$

where erf is the error function. Figure 3(a) shows equation (3) for $n = 1, 2, 3$ (gray dotted, solid and dashed lines, respectively). Additionally, we can generalize the question and ask what is the probability of two points in B to be $\alpha\sigma + \mu_A$, which is plotted in figure 3(b) (using only pairwise comparisons). In both cases we see that the CMIP data follows the one expected from a normal distribution (compare the points in panel (a) and the noisy lines with the predicted lines in panel (b) of figure 3) motivating further exploration and the use of normal distributions for simplicity. Note that the case of $\alpha\sigma$ also accounts for the case where both distributions have different standard deviations. Moreover, in the case of comparing a single model with a model ensemble, the standard deviation of the single model will be, generally speaking, less than the multi-model spread, as it accounts for only the internal variability, that is $\alpha \geqslant 1$ meaning that the assumption of both distributions having the same σ is a conservative one.

We can use equation (1) to get an estimate for how the pdf for $\mu_B - \mu_A$ looks like, knowing that $X_1, X_2 \geqslant \mu_A + \sigma$. To do that, we first need to assume a prior for the distribution $\mu_B - \mu_A$, note that this is a prior on μ_B given that we know μ_A. We choose a prior of $(\mu_B - \mu_a) \sim \mathcal{N}(0, \sigma^2)$, that is, in the CMIP and X-SHiELD context, this prior states that X-SHiELD's mean is within the CMIP model spread. We can now use the information that we have about distribution B to update our prior. The yellow curve in figure 4 is the prior distribution $(p(\mu_B - \mu_A))$ and the black curve is the posterior pdf $p(\mu_B - \mu_A \mid X_1, X_2 \geqslant \mu_A + \sigma)$. We can see that the posterior pdf shifts to the right and its variance decreases, assigning higher probabilities for larger values for $\mu_B - \mu_A$.

Up to this point, we looked at $p(\mu_B - \mu_A | X_1, X_2 \geqslant \mu_A + \sigma)$. However, this is a specific case and we can generalize it to $p(\mu_B - \mu_A | X_1, X_2)$, that is, what is the probability for $\mu_B - \mu_A$ to have a specific value, given that we know the values of X_1 and X_2 in distribution B. This way we can apply this to the comparison between CMIP and X-SHiELD. We perform this analysis to get a posterior distribution of the difference between X-SHiELD and CMIP means for the response of $\overline{Pr}$ to uniform warming and ϕ_{U_e} response to CO_2 increase, knowing the two year values (blue and green curves in figure 4 respectively). We can see that applying the values of the two years shifts the posterior distribution towards the mean of the two year values with respect to the CMIP mean. However, it is not centered around the two-year mean values, that is if we compare the solid (that is the posterior pdf) and dotted lines in figure 4, with the dotted lines representing a normal distribution centered on $\frac{X_1+X_2}{2} - \mu_{\text{CMIP}}$ with a standard deviation of $\frac{\sigma}{\sqrt{2}}$. This is a result of the assumed prior and highlights that this analysis is a more conservative estimation on the difference between the means compared to assuming that the mean of B is distributed around the mean of the two individual

points. It is possible to perform a similar analysis using the two-year mean. In this case, the posterior distribution would differ slightly from the one presented here, as averaging over two years reduces information about the year-to-year variability in the model.

This framework can be modified in the future to account for more years to produce different posterior distributions. Additionally, this can also be generalized to account for differences in the standard deviation between a multi-model spread and individual model (X-SHiELD in our case). We avoid assuming different standard deviations, given that the assumption that the standard deviation of the multi-model ensemble and the individual model are the same is a more conservative one and as such serves as an lower limit for the differences in the means.

5. Base-state dependence of the response

The comparison presented in figure 2 considers the changes in response to uniform warming and isolated CO_2 increase. However, given that the X-SHiELD and the CMIP simulations are conducted over different periods and given the potential effects of model and internal variability, it is possible that the response to the perturbation will have some dependence on the base state and if we account for it the comparison between X-SHiELD and CMIP models can differ qualitatively. Indeed, when examining the response, it is evident that for some of the metrics, the response depends on the values in the control simulation (that is, some have a base-state dependence, figure 5). Take for example, the response of the width of the tropics ($\phi_{\mathrm{Pr}-E}$) to uniform warming (third panel from the top in the left column of figure 5), we see that years with wider tropics in the control simulations expand to a lesser extent in response to uniform warming than years with narrower tropics.

Figure 6 shows the effect of the base state dependence on the comparison between CMIP and X-SHiELD. The black lines in figure 6 represent the absolute changes (Δ), are the same as in figure 2, the blue lines represent the changes after accounting for the state dependence ($\overline{\Delta}$), i.e. removing the linear fit from all dots, including the X-SHiELD dots. The underlying assumption of removing the same linear fit from the X-SHiELD points is that the state dependence found for CMIP models will be the same for X-SHiELD (and other GSRMs). There are three qualitatively different ways the result can be affected when accounting for this state dependence. The first is that accounting for the state dependence results in no significant change. For example, the response of the global mean precipitation ($\overline{Pr}$) and the subsidence fraction (S_{f}) to uniform warming, where there is no or weak base-state dependence; or the NH jet shift ($\phi_{U_{\mathrm{e}}}$) response to CO_2 increase (figure 6) where there is a moderate base-state dependence. The second way that the base-state dependence can influence the comparison is shifting the response in X-SHiELD from being close to the CMIP mean to being outside or very close to $\sim\sigma$ away from the mean. For example, the response to uniform warming of the NH jet shift ($\phi_{U_{\mathrm{e}}}$) and the response of the tropics width ($\phi_{\mathrm{Pr}-E}$, figure 6). The third way that the base state dependence can influence the comparison is by shifting the response in X-SHiELD from being $\sim\sigma$ away from the CMIP mean to being close to the mean. For example the response to CO_2 increase of the subsidence fraction (S_{f}) and tropics width ($\phi_{\mathrm{Pr}-E}$, figure 6). The correlation between the response and control years has a physical meaning in our case because we are using a model that is forced by SST. That is, in both the control and perturbation, the SST in equivalent years is the same up to a known prescribed perturbation, that is a uniform warming or a CO_2 increase. This will not be the case in a coupled model, where the SST evolves freely, and thus a year-to-year comparison is less meaningful.

6. Discussion

The computational cost limits the ability to run global climate model simulations with storm-resolving resolution. Given these limitations, the purpose of this study is to examine the question of what can we learn from a short integration of two years, with a focus on zonal mean variables and large-scale climate metrics. While two years (or even less) can be sufficient to study aggregated statistics of short- and small-scale events, or alternatively, study metrics that converge and have small interannual variability, this is not the case when considering large-scale climate metrics. An example of this are metrics that describe the large-scale circulation, such as the tropics width and jet position that are known to have strong interannual variability (e.g. Staten *et al* 2018, Waugh *et al* 2018).

To address this question, we compare results from X-SHiELD, an atmosphere-only GSRM with CMIP models. We show that in some of the large-scale metrics X-SHiELD years are outside a measure of the spread of the CMIP models that account for both internal variability and inter-model spread. This result leads to the question of, given that two years in one model are outside the spread of another model, what can we say about the relationship between the means of two models. We show that by using Bayesian inference together with reasonable assumptions, one can estimate a lower limit to the answer to this question. More specifically,

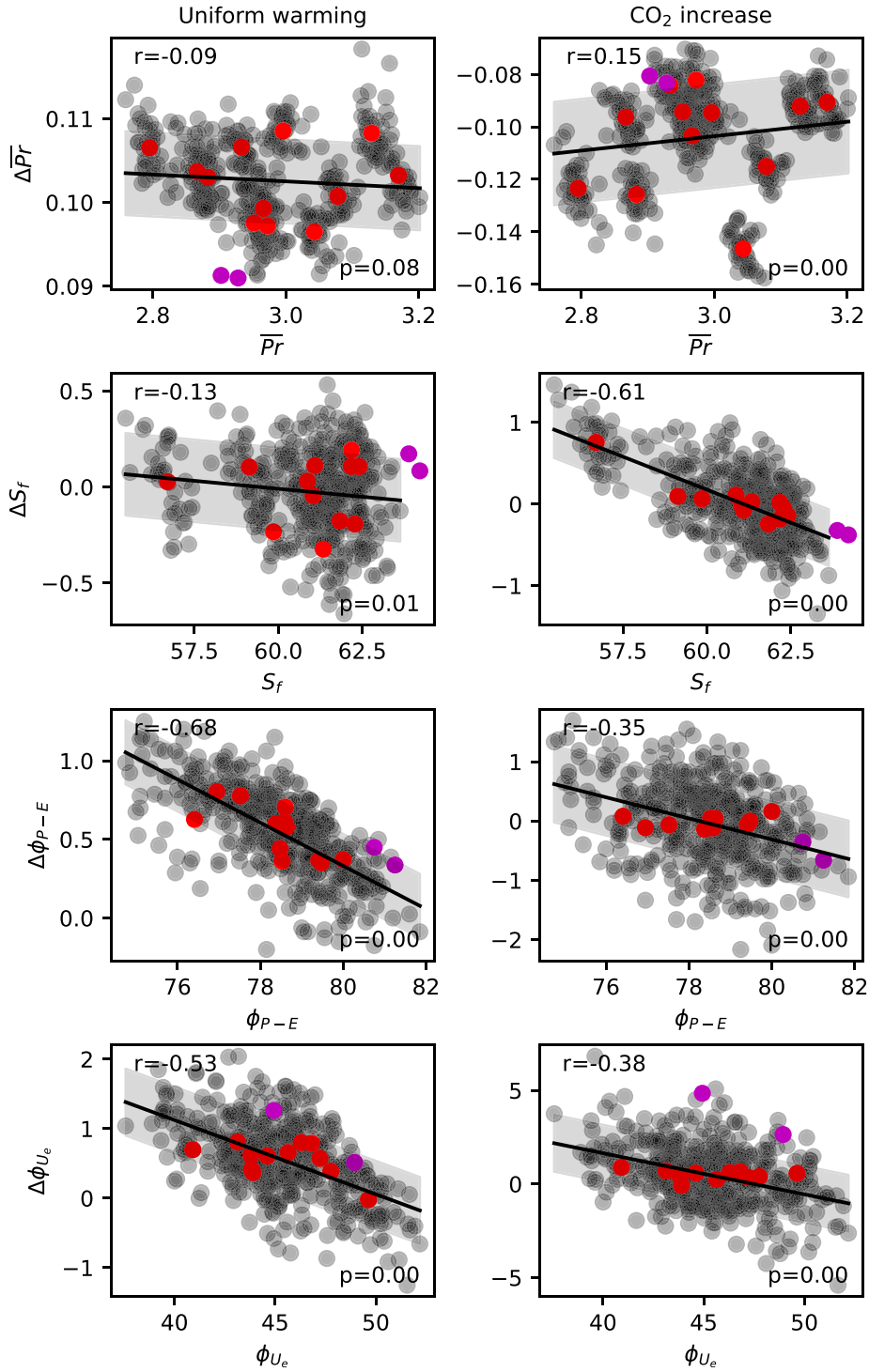

Figure 5. Scatter plots of the response for the different years in the models as a function of the control values. Gray dots are the different years in the CMIP models. Red dots represent the different models mean. Line is a linear fit for the gray dots, and shading is one σ of the fit. Magenta dots are the values of X-SHiELD years.

we estimate the probability for how different the mean response of global-mean precipitation to uniform warming and latitude shift of the NH jet response to CO_2 in X-SHiELD is compared to the CMIP multi-model mean given the values in the two year simulation (figure 5). We show that in both cases the center of the posterior pdf for the difference in means, $\mu_B - \mu_A$ shifts more than a standard deviation compared to the prior distribution (blue and green curves in figure 4).

We note that our result depends on different assumptions, such as the standard deviation of distribution B (X-SHiELD in our case) and the prior for $\mu_B - \mu_A$. That said, across the entire process we used conservative assumptions regarding the assessment of how different X-SHiELD is from the CMIP model. Given the small sample nature of this problem, this posterior distribution is a conservative estimate, and the true value of

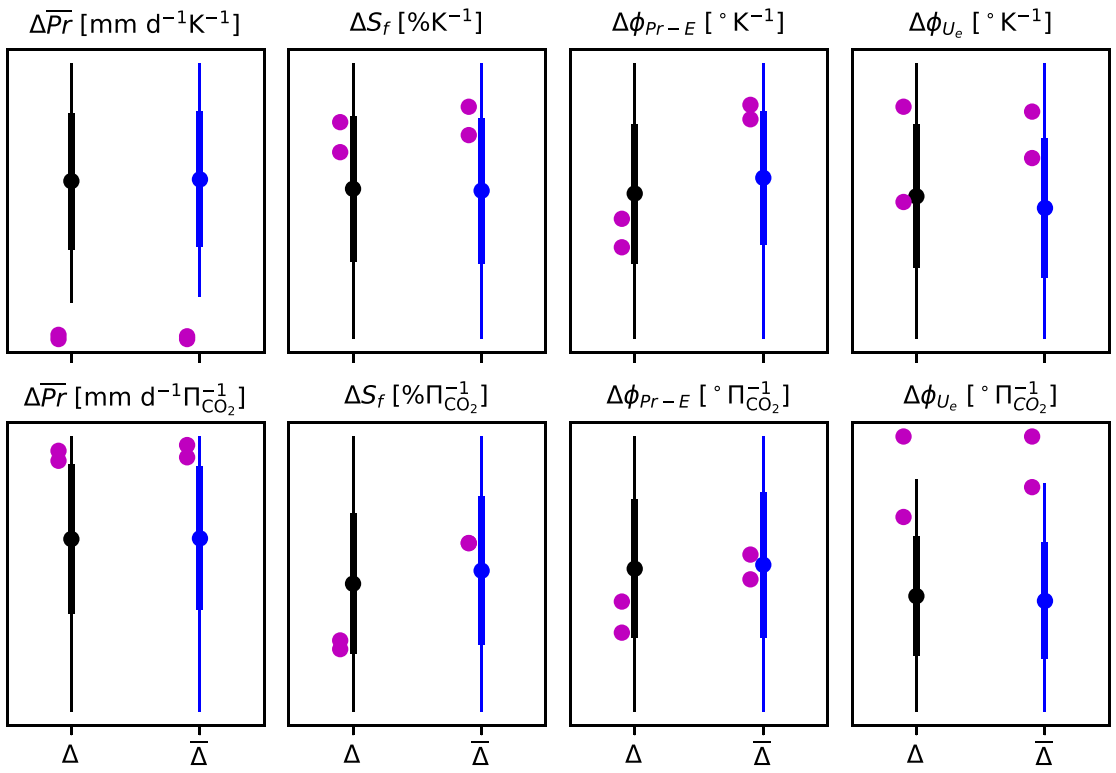

Figure 6. Similar to figure 2, showing a comparison between the response of CMIP models and X-SHiELD to uniform warming (upper row) and CO_2 increase (bottom row). Black colors represent the absolute response (Δ), and blue colors represent the response after accounting for the state dependence ($\overline{\Delta}$). The black and blue dots represent the CMIP multi-model mean. The thick line is one standard deviation (σ) of all CMIP years, thin line are all years between the 2.5th and 97.5th percentiles. Magenta points are for X-SHiELD's two years.

X-SHiELD's mean might be different than the most probable ones presented in the derived pdf. However, given the computational limitations, we find that this analysis is sufficient to examine the possibility of a GSRM being an outlier, especially if we consider the current stage of GSRM modeling as a *beta-testing* stage. Conducting a comprehensive analysis for different climate metrics, over a range of GSRMs will provide a better indication which of the different metrics exhibit an inherently different response when resolution is increased to the point of resolved deep convection. This can also be a basis for the case (or a lack of) climate-scale simulations using GSRMs. It is important to note that a configuration of an atmosphere-only model forced by or nudged-to SST is crucial for this type of comparison. Namely, this configuration enables a more statistically meaningful year-to-year comparison, something that is lacking in coupled models, where the SST freely evolves and the interannual variability de-correlates between the perturbed and control integrations. This is even more evident when considering the correlation between the response and the control values: there is a base-state dependence of the response in some of the analyzed metrics. These correlations are meaningful due to the shared control SST in the different simulations. This base-state dependence can account for the differences between the integration period of X-SHiELD and CMIP.

An additional limitation is the multiple comparison problem, that is, by selecting a large number of parameters, the chance of detecting a case in which X-SHiELD outside the spread of CMIP models simply by chance increases. This point highlights the benefit of analyzing the individual years, as if both years in X-SHiELD show a similar response and are outside of the CMIP spread, this will indicate that the difference is not spurious but rather a more significant one. However, this should be considered when interpreting the results of this and future studies.

The two years in X-SHiELD show some differences when they are compared with the CMIP multi-model ensembles, in particular, for the response of $\overline{Pr}$ to uniform warming and the response of ϕ_{U_e} to increase in CO_2 also when accounting for the base-state dependence (figures 2 and 6). In both cases, the calculated posterior probability distribution suggests that the most probable value of the mean response of X-SHiELD is more than one σ away from the CMIP mean (figure 4). These differences between X-SHiELD and CMIP models may be attributed to the role of X-SHiELD's higher resolution and ability to resolve some aspects of deep convection. For example, the difference in the response of the global-mean precipitation seems to arise

from a different response in the mid-latitude land (Guendelman *et al* 2024), which highlights the role of land-atmosphere interaction and the importance of resolving it (see also Lee and Hohenegger 2024). A potential explanation for the difference in the jet shift response to CO_2 increase in X-SHiELD could be differences in diabatic and latent heating (Tamarin and Kaspi 2017, Garfinkel *et al* 2024). However, a detailed analysis is needed that is beyond the scope of this study.

This analysis aims at overcoming the computational limitations that comes with convection-permitting simulations with realistic conditions, that is, realistic boundary conditions and seasonally varying SST. Another possibility to overcome the computational cost is to reduce the realism of the simulation. An example of this would be an aquaplanet with no seasonal cycle with a GSRM configuration (Miura *et al* 2005, Tomita *et al* 2005, Narenpitak *et al* 2017, O'Gorman *et al* 2021, Lin *et al* 2023, Clark *et al* 2024). On the one hand, this will allow for short runs to be statistically meaningful; on the other hand, the reduction in realism will limit our ability to come to conclusions about the real world. For example, in a previous study, we found that the main source of difference in precipitation response to warming between X-SHiELD and CMIP models is over land (Guendelman *et al* 2024). Knowing that, we can hypothesize that there will not be a similar significant difference in an aquaplanet configuration. As this is the case, a combination of idealized and realistic simulations is needed to test the added value of resolving deep convection to examine its effect on the large-scale climate.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.15114620. The source code of X-SHiELD is available on https://github.com/NOAA-GFDL/SHiELD_build.

ORCID iDs

Ilai Guendelman https://orcid.org/0000-0002-6873-0320
Timothy M Merlis https://orcid.org/0000-0002-5593-8210
Kai-Yuan Cheng https://orcid.org/0000-0002-4246-7659
Lucas M Harris https://orcid.org/0000-0001-6072-8624
Christopher S Bretherton https://orcid.org/0000-0002-6712-8856
Maximilien Bolot https://orcid.org/0000-0002-2171-5924
Linjiong Zhou https://orcid.org/0000-0002-5772-6203
Spencer K Clark https://orcid.org/0000-0001-5595-7895
Stephan Fueglistaler https://orcid.org/0000-0002-0419-440X

References

Allen M R and Ingram W J 2002 *Nature* **419** 224–32
Balaji V, Couvreux F, Deshayes J, Gautrais J, Hourdin F and Rio C 2022 *Proc. Natl Acad. Sci. USA* **119** e2202075119
Bao J, Stevens B, Kluft L and Muller C 2024 *Sci. Adv.* **10** eadj6801
Bolot M, Harris L M, Cheng K Y, Merlis T M, Blossey P N, Bretherton C S, Clark S K, Kaltenbaugh A, Zhou L and Fueglistaler S 2023 *npj Clim. Atmos. Sci.* **6** 209
Bony S *et al* 2015 *Nat. Geosci.* **8** 261–8
Chemke R and Polvani L M 2019 *J. Clim.* **32** 859–75
Cheng K, Harris L, Bretherton C, Merlis T M, Bolot M, Zhou L, Kaltenbaugh A, Clark S and Fueglistaler S 2022 *Geophys. Res. Lett.* **49** e2022GL099796
Clark J P, Lin P and Hill S A 2024 *J. Adv. Model. Earth Syst.* **16** e2023MS003968
Dunne J P *et al* 2020 *J. Adv. Model. Earth Syst.* **12** e2019MS002015
Garfinkel C I, Keller B, Lachmy O, White I, Gerber E P, Jucker M and Adam O 2024 *J. Clim.* **37** 2541–64
Guendelman I, Merlis T M, Cheng K, Harris L M, Bretherton C S, Bolot M, Zhou L, Kaltenbaugh A, Clark S K and Fueglistaler S 2024 *Geophys. Res. Lett.* **51** e2023GL107008
Hohenegger C *et al* 2023 *Geosci. Model Dev.* **16** 779–811
Lee J and Hohenegger C 2024 *Proc. Natl Acad. Sci. USA* **121** e2314265121
Lin P, Ming Y and Robinson T 2023 *J. Adv. Model. Earth Syst.* **15** e2022MS003300
Merlis T M *et al* 2024a *Sci. Adv.* **10** eadn5217
Merlis T M *et al* 2024b *Geophys. Res. Lett.* **51** e2024GL111549
Miura H, Tomita H, Nasuno T, Iga S-I, Satoh M and Matsuno T 2005 *Geophys. Res. Lett.* **32** L19717
Narenpitak P, Bretherton C S and Khairoutdinov M F 2017 *J. Adv. Model. Earth Syst.* **9** 1069–90
O'Gorman P A, Li Z, Boos W R and Yuval J 2021 *Phil. Trans. R. Soc. A* **379** 20190543
Palmer T and Stevens B 2019 *Proc. Natl Acad. Sci. USA* **116** 24390–5
Rackow T *et al* 2025 *GMD* **18** 33–69
Seager R, Harnik N, Kushnir Y, Robinson W and Miller J 2003 *J. Clim.* **16** 2960–78

Staten P W, Lu J, Grise K M, Davis S M and Birner T 2018 *Nat. Clim. Change* **8** 768–75
Stevens B *et al* 2019 *Prog. Earth Planet. Sci.* **6** 61
Stevens B and Bony S 2013 *Science* **340** 1053–4
Takasuka D, Satoh M, Miyakawa T, Kodama C, Klocke D, Stevens B, Vidale P L and Terai C R 2024 *Prog. Earth Planet. Sci.* **11** 66
Tamarin T and Kaspi Y 2017 *J. Atmos. Sci.* **74** 553–72
Tomita H, Miura H, Iga S, Nasuno T and Satoh M 2005 *Geophys. Res. Lett.* **32** L08805
Waugh D W *et al* 2018 *J. Clim.* **31** 7565–81
Zhao M *et al* 2018 *J. Adv. Model. Earth Syst.* **10** 735–69