Supplementary Information

Predicting global patterns of long-term climate change from short-term simulations using machine learning

L.A. Mansfield*[1,2], P.J. Nowack[1,3,4,5], M. Kasoar[1,3,6], R.G. Everitt[7], W.J. Collins[8], A. Voulgarakis[1,6,9]

Affiliations:

[1]Department of Physics, Imperial College London, South Kensington Campus, London, SW7 2BW, United Kingdom,

[2]School of Mathematics and Statistics, University of Reading, Whiteknights, Berkshire, RG6 6AX, United Kingdom.

[3]Grantham Institute, Imperial College London, South Kensington Campus, London, SW7 2AZ, United Kingdom.

[4]Data Science Institute, Imperial College London, South Kensington Campus, London, SW7 2AZ, United Kingdom.

[5]School of Environmental Sciences, University of East Anglia, Norwich, Norfolk, NR4 7TJ, United Kingdom.

[6]Leverhulme Centre for Wildfires, Environment and Society, Department of Physics, Imperial College London, South Kensington Campus, London, SW7 2BW, United Kingdom.

[7]Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom.

[8]Department of Meteorology, University of Reading, Whiteknights, Berkshire, RG6 6ET, United Kingdom.

[9]School of Environmental Engineering, Technical University of Crete, Chania Crete, 73100, Greece.

Corresponding author: Laura A Mansfield (laura.mansfield@pgr.reading.ac.uk)

Supplementary Notes

Alternative Approach to Pattern Scaling

To obtain a scenario prediction with pattern scaling, we require a scaler value to multiply by the reference temperature response, here the 2xCO2_PDRMIP scenario (see Methods). This scaler value estimates the ratio of global mean temperature response between the new and reference scenario. In climate prediction and impact studies, the long-term global mean temperature response is typically estimated from a computationally cheap model, such as an energy balance model[1,2], or by assuming linearity between effective radiative forcing and long-term temperature response[3]. The latter is calculated from a short GCM run with fixed sea-surface temperatures (see Methods). This choice is based on the well-established assumption of linearity between radiative forcing and long-term temperature response for well-mixed greenhouse gases[4].

We have also investigated an alternative approach to estimate the scaler value, using the ratio of short-term global mean temperature response between the new and reference scenario. Although this is not typically done in practice, we do this to directly compare our proposed machine learning methods with pattern scaling when using the same predictor variable. Under the assumption of linearity, for a sudden step-forcing we expect the short-term temperature response to be directly proportional to the ratio of long-term global mean temperature response[4], making this a valid choice for the scaler value.

Predicted Response Maps

The response maps predicted for each scenario are shown in Supplementary Fig. 1, where the top left plot shows the short-term GCM temperature response and all other maps show the predictions of long-term response: the true GCM prediction on the bottom of the first

column, the machine learning methods in the second column (Ridge regression and Gaussian process regression respectively) and the pattern scaling methods in the third column (using ERF and global mean short-term response as the scaler value respectively). Note that the pattern scaling reference patterns are from the 2xCO2_PDRMIP scenario so this response is exactly predicted with both pattern scaling methods.

There are increased regional variations in the prediction for Ridge and Gaussian process regression compared to the pattern scaling methods. Accounting for regional patterns in the short-term response leads to improved predictions in some regions in the short-lived pollutant perturbations, e.g. over Asia in the Gaussian process regression prediction of 10xBC_Asia, 10xSO4_Asia and No_SO2_China. Furthermore, highly warming and cooling regional short-term responses can cancel, giving a weak global mean radiative forcing (e.g. No_CO_Global and No_BC_NHML) or a weak global mean short-term temperature response (e.g. 0.005°C in No_BC_Global). This leads to weak pattern scaling predictions, (e.g. a global mean response of 0.01°C in No_BC_Global). This confirms that the spatial variability in the short-term response can be a valuable predictor of long-term response.

The short-lived pollutant scenarios (BC, SO_4, SO_2, CO) are predicted somewhat less skilfully with any method compared to the long-lived pollutants, primarily due to the weaker forcing in the idealised short-lived pollutant perturbations that were performed with the GCM. Additionally, amongst short-lived pollutant scenarios, those featuring the weakest forcings are also less accurately predicted, as is the case for No_BC_NHML, No_BC_Global and No_SO2_US.

Note that in some cases the machine learning methods incorrectly predict regional response patterns representative of the responses seen in the training data, such as the predicted cooling over Europe in the 10xBC Asia experiment. This occurrence would be less frequent with increased amount and variety of training data.

Spatial variability in response

Supplementary Fig. 3 shows the global temperature distribution for each scenario for all prediction methods (General circulation model (GCM), Ridge regression prediction, Gaussian Process regression (GPR) prediction, Pattern Scaling (PS) using effective radiative forcing (ERF) and Pattern Scaling using the short-term global mean surface temperature (T)). For short-lived pollutant scenarios (first row), pattern scaling methods consistently under-predict the variability in response over the globe. This is because the pattern scaling methods are restricted to the same distribution of temperature response, simply scaled based on the estimated mean response. This highlights that accurate regionally-specific predictions cannot be made with this approach.

For larger forcing scenarios (third row), all methods perform similarly well in terms of the distribution of temperature response. These are mostly long-lived greenhouse gas forcings, with a much more homogeneous response pattern. We suggest that for rapid predictions of the response due to greenhouse gas forcings, a pattern scaling approach is sufficient. However, for predicting response to short-lived forcings, we expect increased regional structure in the response and therefore one of the machine learning methods would be more appropriate.

Accuracy of Alternative Pattern Scaling Approach

Supplementary Fig. 4 shows the same prediction errors as Fig. 3 in the main text, with the additional results from the alternative pattern scaling approach that uses the global mean short-term temperature response as the scaler, labelled P(T). This method often performs better than the ERF approach. There are many possible reasons for this but the use of short-term temperature response rather than radiative forcing addresses, to a degree, the well-known issue regarding different pollutant types having different climate efficacies, i.e. different ratios between radiative forcing and temperature response[1,5,6].

In fact, Supplementary Fig. 4 shows that this approach to pattern scaling can outperform the machine learning methods in terms of regionally averaged response in some regions (e.g. Northern Africa, North America) and is comparable in several other regions. This indicates the short-term global mean response alone is a relatively effective predictor of the long-term response, even before accounting for regional anomalies in the short-term.

However, such a method is still strongly restricted in the pattern of response and cannot predict regional effects that are captured in the machine learning methods, as highlighted in Supplementary Figs. 1 and 2. Furthermore, pattern scaling does not present the opportunity to improve the performance with an increased dataset. Due to our relatively small dataset, we expect to see machine learning prediction errors improve with additional training data (Fig. 4) compared against the fixed median error from pattern scaling. This pattern scaling method is therefore ideal in the situation when there is limited training data but with increased datasets, we expect to obtain increasingly higher performance from using machine learning methods.

There are other possible choices to estimate the global mean temperature response, such as using an energy balance model[1,2]. However, due to the definition of pattern scaling, the spatial variability in response will always be tied to the magnitude of response, as highlighted in Supplementary Fig. 3.

Scenario prediction accuracy

The large spread in absolute prediction errors for all methods in Fig. 3 is specific to certain scenarios and regions. Firstly, some regions have a large magnitude of response in some scenarios, which leads to larger prediction error. Supplementary Fig. 5a shows the relationship between the mean response in a region and the associated prediction error for each method. The lines showing a relative error of 10%, 20%, 50%, 100% and 200% are also shown. Regions and scenarios with large magnitudes of response, greater than around 2°C tend to be predicted

with lower relative errors, falling below the 20% prediction error. For scenarios with weak magnitudes of response, we do not necessarily see small errors, with some of these prediction errors falling in the region where the relative error is greater than 200%.

Supplementary Fig. 5b shows the relative error for all scenarios over all regions compared to the global mean short-term response magnitude. The larger relative prediction errors tend to come from predictions of scenarios with weaker short-term responses, which will have lower signal-to-noise ratios. This is motivation for a training dataset with more strongly forced scenarios with greater signal-to-noise ratios in the short-term response.

Early indicators using Regression coefficients

There are a large number of regression coefficients; specifically, for each of the 27,840 outputs, there are 27,840 regression coefficients as described in *Methods*. However, most of these coefficients are close to zero and have little influence on the output[7]. For the long-term response regression at a single grid-cell, the coefficients of larger magnitudes highlight the regions in the short-term response that are the best predictors of this grid-cell response. Supplementary Fig. 6a and b shows the magnitude of the coefficients on a map for a selected example output grid cell over East Asia and Europe respectively, highlighted by the black star.

Following equation (3), it is the value of coefficient multiplied by the short-term temperature response (i.e. $\beta_j x_j$ for each grid-cell j) that contributes to the long-term response. Since x_j can take larger or smaller values depending on the region, we also show the results of the coefficient maps when x_j is scaled to be have zero mean and unit variance for all input grid-cells and find similar results (Supplementary Fig. 6d, e).

In Supplementary Fig. 6a, the dominant coefficients appear in regions close to the predicted grid cell, indicating a strong relationship between the short- and long-term responses

in the localized region over East Asia. In contrast, some regions draw more predictive power from remote regions. For example, Supplementary Fig. 6b shows coefficients for a prediction over Europe are strongly influenced by the short-term responses predominantly in sea ice regions over the Arctic. This feature is consistent amongst grid-cell predictions in the Europe region. As the short-term response in this Arctic region is highly variable (Supplementary Fig. 8c) and strongly responding (due to Arctic amplification, e.g. Supplementary Fig. 1a), this could contribute to the relatively poor prediction over Europe. This confirms a limitation of this approach, where predictions can be highly dependent on noise in the training data.

There are spatially similar features in the regression coefficient maps that appear regardless of prediction region, such as the larger coefficients over South Asia, Northern Africa and generally over continental regions. This is further highlighted when an average is taken across all 27,840 outputs to find the global mean regression coefficient map shown in Supplementary Fig. 6c and when we do the same for scaled input values in Supplementary Fig. 6f. Here we see familiar patterns associated with warming, such as increased magnitudes in the mid-latitude bands around the jet stream and at high latitudes[8–10]. Some of this can be partially explained as regions that typically respond more strongly which means signals indicating the sign and magnitude of the response can be picked up earlier. For example, ice and snow regions (e.g. Arctic) and high-altitude regions (e.g. Himalayas) are highlighted, both of which are known to warm more rapidly due to ice/snow albedo feedback[11] and faster upper tropospheric warming[12,13], respectively. However, when we account for the magnitude of typical response by scaling the input variables to the same magnitude everywhere in Supplementary Fig 6f, we see some patterns remain, suggesting they are robust early indicators of long-term response in the GCM. In particular, the mid-latitude jet stream regions are dominant, as well as the high latitude Arctic sea-ice regions.

Data constraints

As discussed in the main text, there are constraints due to the training data available that contributes to the poorer performance for data-driven methods in predictions over Europe. In particular, we suggest that this is due to:

a) Large variance in long-term responses over Europe across the training data (Supplementary Fig. 8a). This means prediction data is more likely to be further from any known training data points to constrain it. This is particularly problematic for Gaussian process regression since it relies on correlations between training and test data points.

b) Large internal variability in the long-term response over Europe and surrounding high latitude areas (Supplementary Fig. 8b). This makes this region inherently harder to predict with statistical or physical models, due to the weak signal-to-noise ratio relative to other regions.

c) Large internal variability in the short-term response over Europe and surrounding high latitude areas (Supplementary Fig. 8c). A model with strong dependence on the short-term response over Europe is therefore dependent on noisy, highly variable inputs.

Dimension Reduction

One of the key challenges is the high-dimensional nature of gridded data that is output from GCMs. We have explored statistical and physical approaches to dimension reduction on both the short-term and long-term temperature response. For the former, we use principal component analysis (PCA). We calculate these by collapsing the spatial data for each simulation into a single vector and performing a Singular Value Decomposition on this to

obtain the principal components. Since the number of components is limited by the number of simulations available, we use all components here (20)[7].

We also make predictions on key regions, informed by physical knowledge of coherent climate characteristics rather than statistical relationships. We use the regions shown in Fig. 3, along with additional regions over the oceans (divided into North Atlantic, South Atlantic, North Pacific, South Pacific, Indian Ocean, Southern Ocean and the Antarctic) to cover the full grid.

Supplementary Fig. 10 shows the absolute prediction errors in each region using these dimension reduction methods on the inputs (the short-term response), on the outputs (the long-term response) and on both the inputs and outputs. By using dimension reduction on the short-term responses, the problem becomes better constrained. However, we do not find improvements in the predictions with either approaches.

Alternative predictor variables

There are a range of short-term predictor variables that could be chosen as inputs to the regression for predicting the long-term temperature response. Supplementary Fig. 9 shows the prediction errors when using various predictor variables in the regression, where errors are calculated from the absolute error over the regions shown in Fig. 3. These predictor variables are surface air temperature, air temperature at 500 hPa, geopotential height at 500 hPa, effective radiative forcing (ERF; calculated as the difference between outgoing and incoming radiation in the response of a GCM with fixed surface temperatures[14]) and sea level air pressure. Both sea level pressure and ERF produce large absolute errors suggesting these are not suitable predictors for long-term surface temperature response patterns. However, air temperature and geopotential height at 500 hPa offer predictions to a similar degree of accuracy as the surface temperature response. This suggests there is similar information encoded in these variables and

their patterns. Still, surface air temperature appears to be the predictor variable with consistently lower prediction errors and is most interpretable for predicting the long-term surface temperature response.

Short-term response period

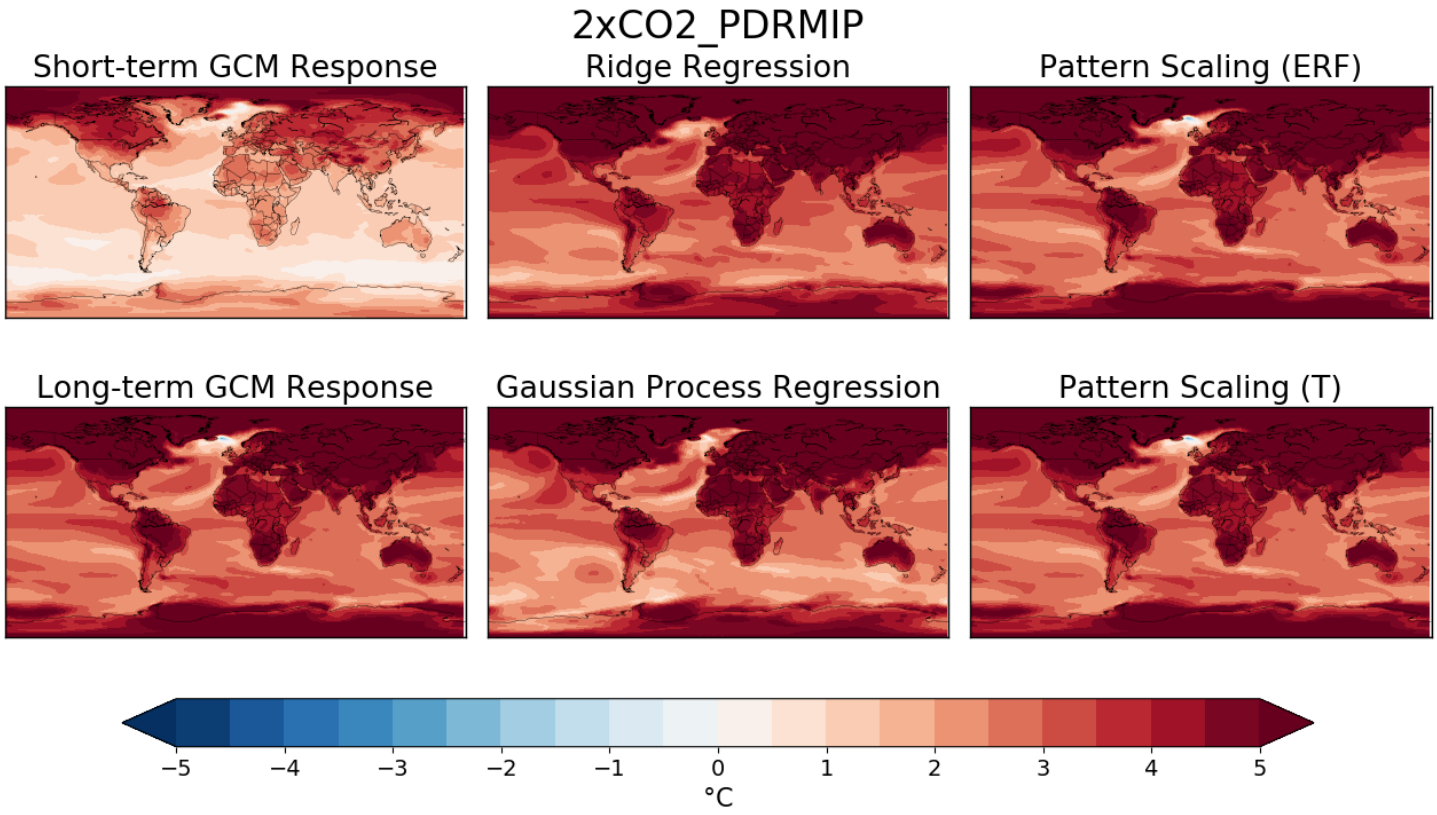
We define the short-term response to be the first 10 years of the GCM response, to allow the GCM some time to respond to the forcings and to take an average over sufficient years to remove natural variability[15]. However, we find that using shorter time periods already show promise. Supplementary Fig. 11 shows the absolute prediction errors in °C when the short-term response is defined as the first 5 years of the GCM response. The prediction errors for Ridge regression and Gaussian process regression are increased compared to Fig. 3, but in most regions are competitive when compared against both pattern scaling methods, but particularly the ERF approach. As before, it is expected that an increased training dataset will further reduce prediction error. This would make a strong enhancement to the speed of prediction of new unseen scenarios, as fewer years of the GCM are required.

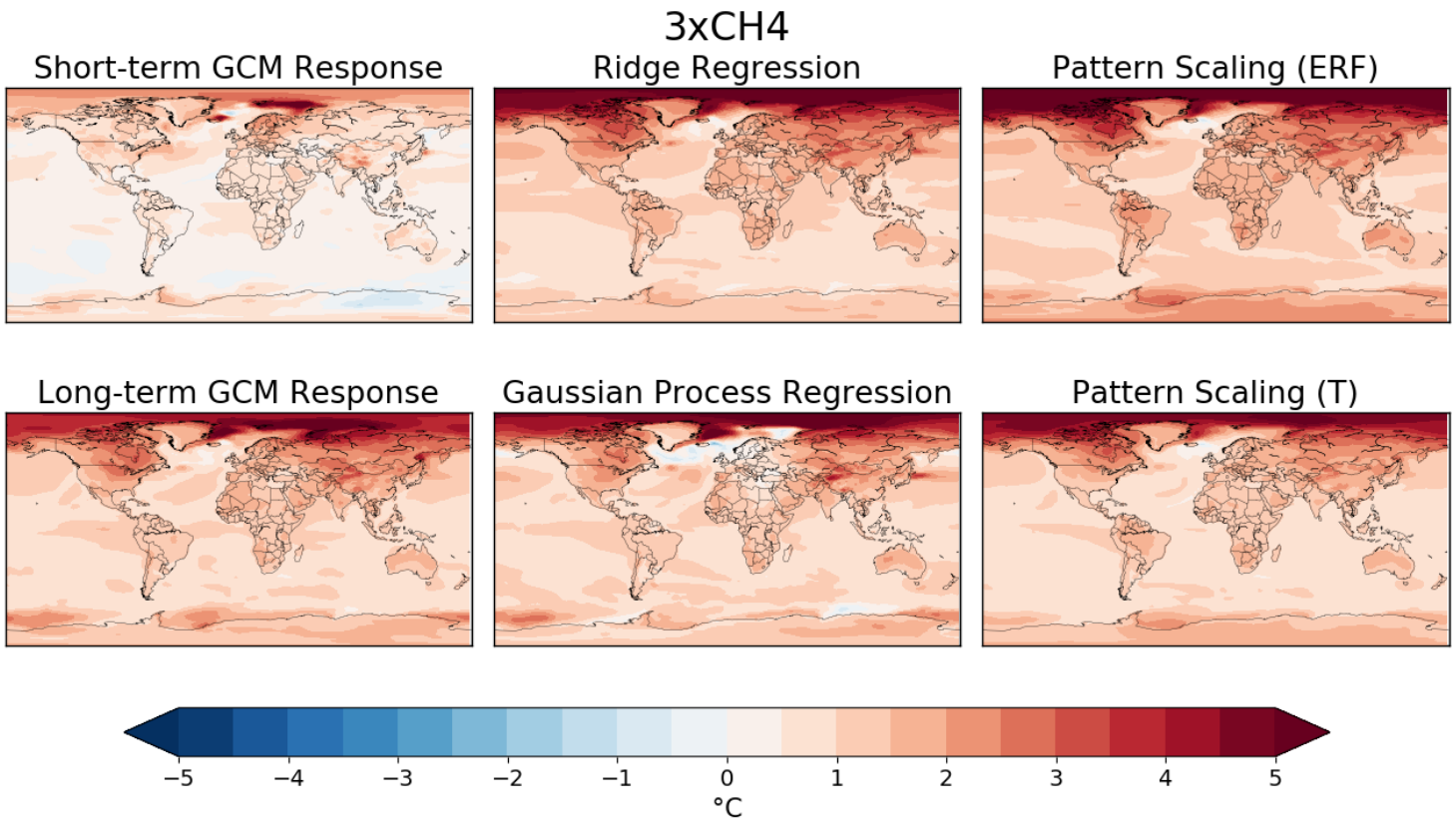
Supplementary Tables

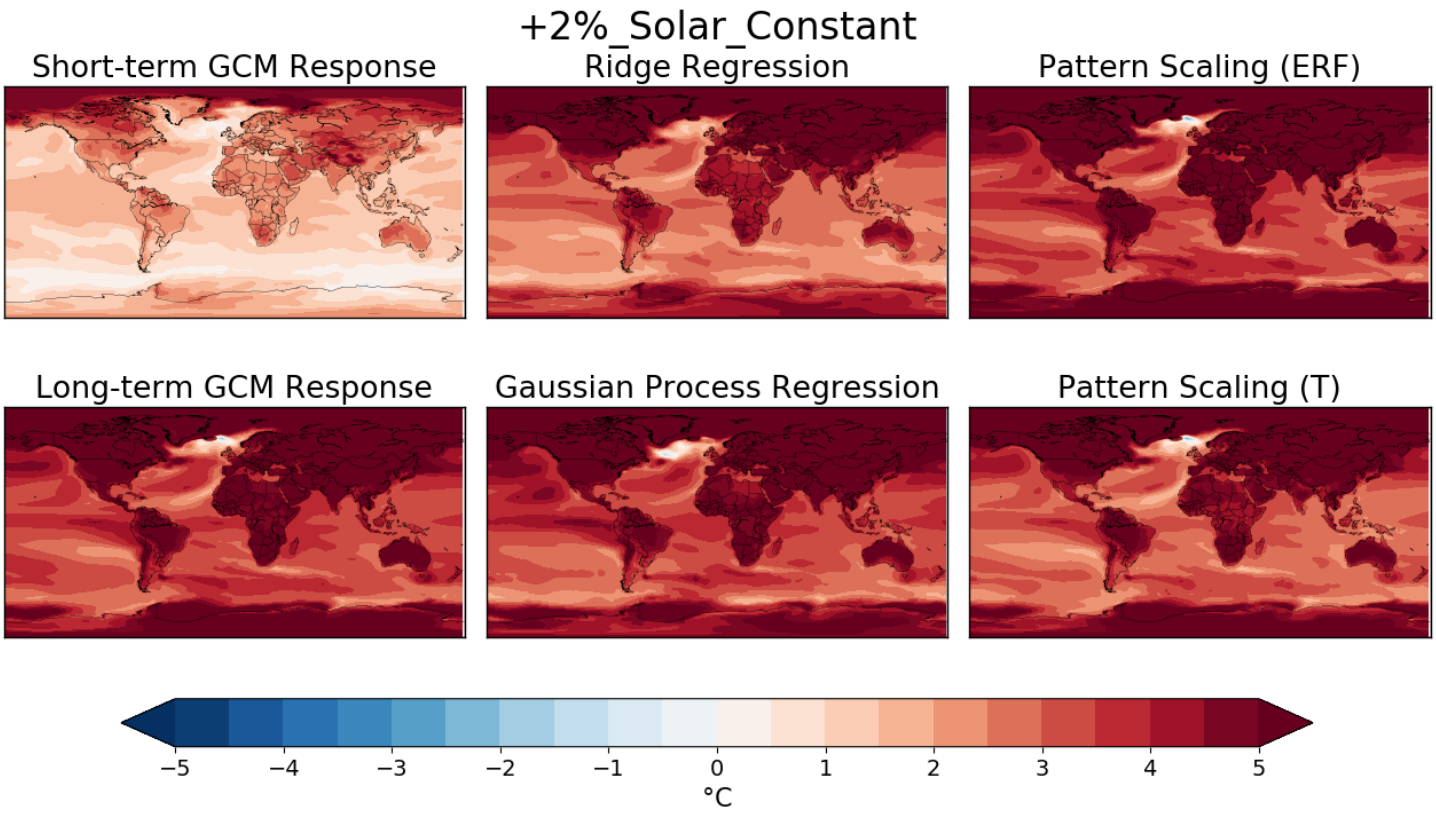
Supplementary Table 1: List of available simulations and their sources[8,16–23]**.** Note that for all long-lived pollutants (CO_2, CH_4, CFC-12) perturbations are global and for short-lived pollutants (SO_4, SO_2, BC, CO) are a mix of both global and regional perturbations, specified in the table.

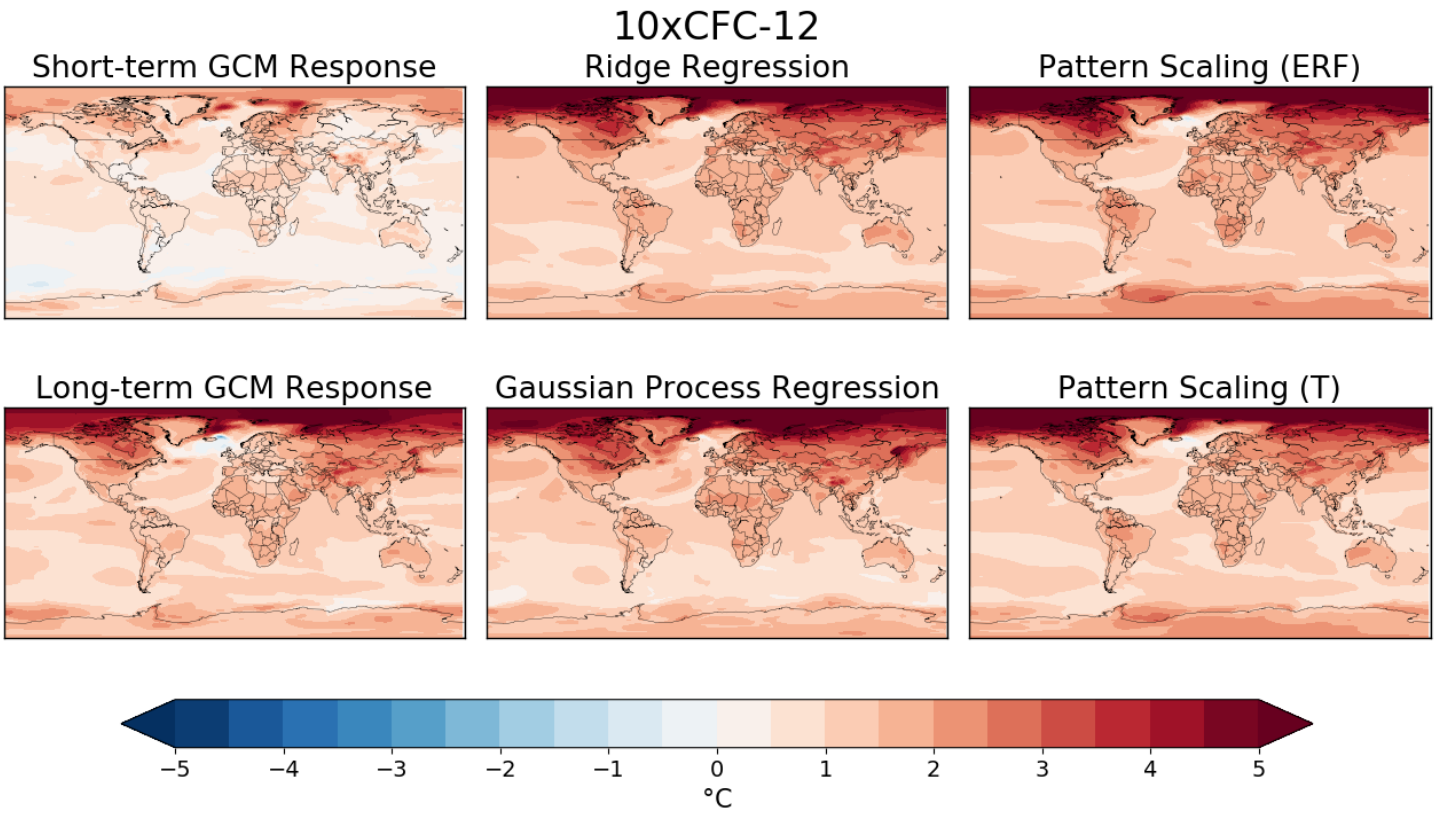
Source	Short name	Description
PDRMIP [17–19] (100 years of simulation)	2xCO_2_PDRMIP	Global doubling of CO_2 concentration (PDRMIP)
	3xCH_4	Global tripling of methane concentration
	10xCFC-12	10x increase in chloroflorocarbon-12 globally
	+2%_Solar_Constant	2% increase in solar forcing
	5xSO_4_Global	5x increase in sulfate aerosol concentration globally
	10xBC_Global	10x increase in black carbon concentration globally
	10xSO_4_Europe	10x increase in sulfate aerosol concentration over Europe
	10xSO_4_Asia	10x increase in sulfate aerosol concentration over Asia
	10xBC_Asia	10x increase in black carbon concentration over Asia
	SO_4_pre-industrial	Pre-industrial sulfate levels
ECLIPSE[20–22] (80 years of simulation)	2xCO_2_ECLIPSE	Global doubling of CO_2 concentration (ECLIPSE)
	20%_CH_4	20% reduction in methane emissions globally
	No_BC_Global	100% reduction in black carbon emissions globally
	No_SO_2_Global	100% reduction in sulfur dioxide emissions globally
	No_CO_Global	100% reduction in carbon monoxide emissions globally
Kasoar *et al*. (2018)[8,16,23] (200 years of simulation; used first 100)	No_SO_2_NHML	100% reduction in sulfur dioxide emissions over the northern hemisphere mid-latitudes
	No_SO_2_China	100% reduction in sulfur dioxide emissions over China
	No_SO_2_East_Asia	100% reduction in sulfur dioxide emissions over East Asia
	No_SO_2_Europe	100% reduction in sulfur dioxide emissions over Europe
	No_SO_2_US	100% reduction in sulfur dioxide emissions over the US
	No_BC_NHML	100% reduction in black carbon emissions over the northern hemisphere mid-latitudes

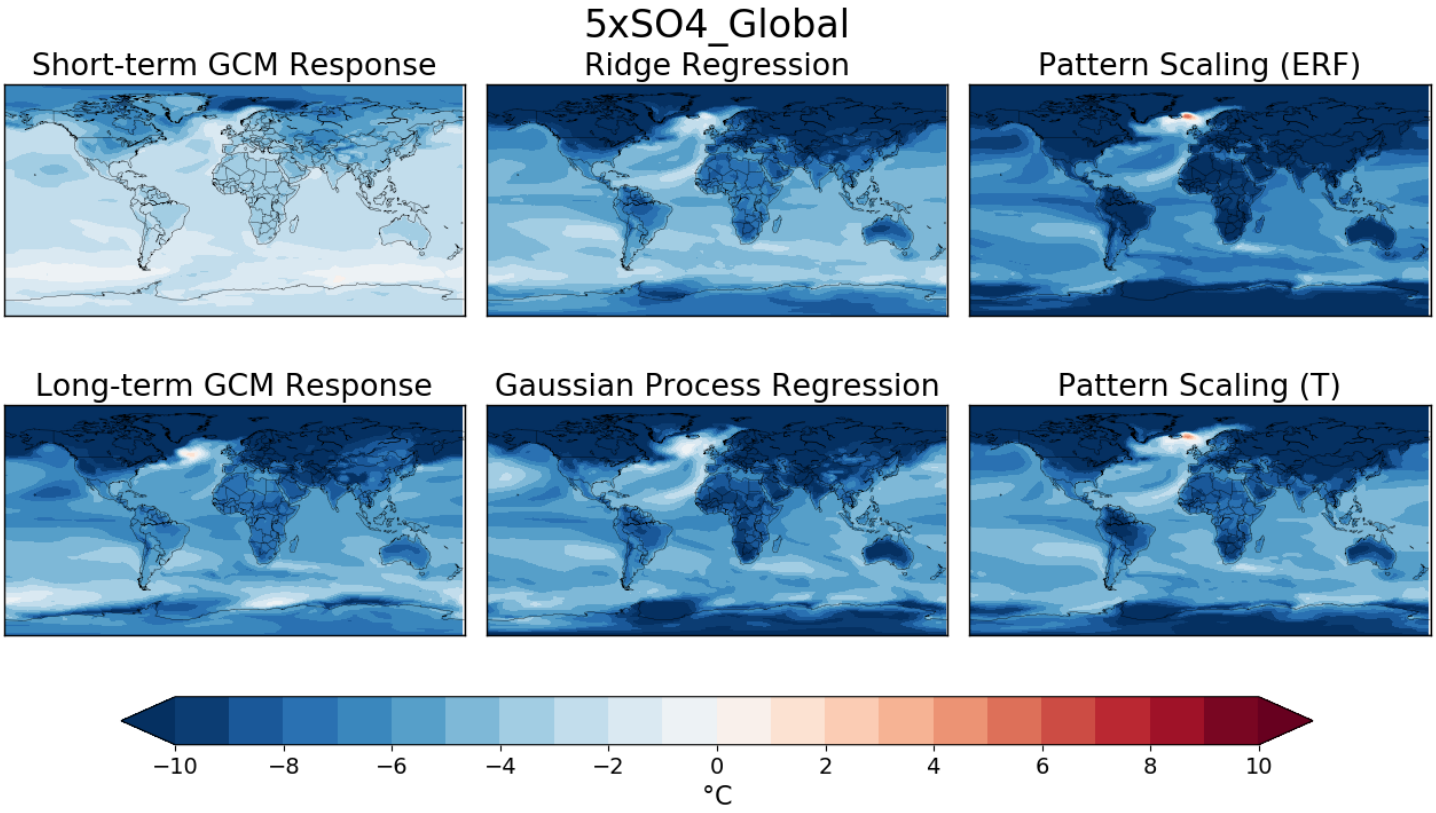
Supplementary Figures
(for multi-page figures note that the caption is placed below the last subfigure)

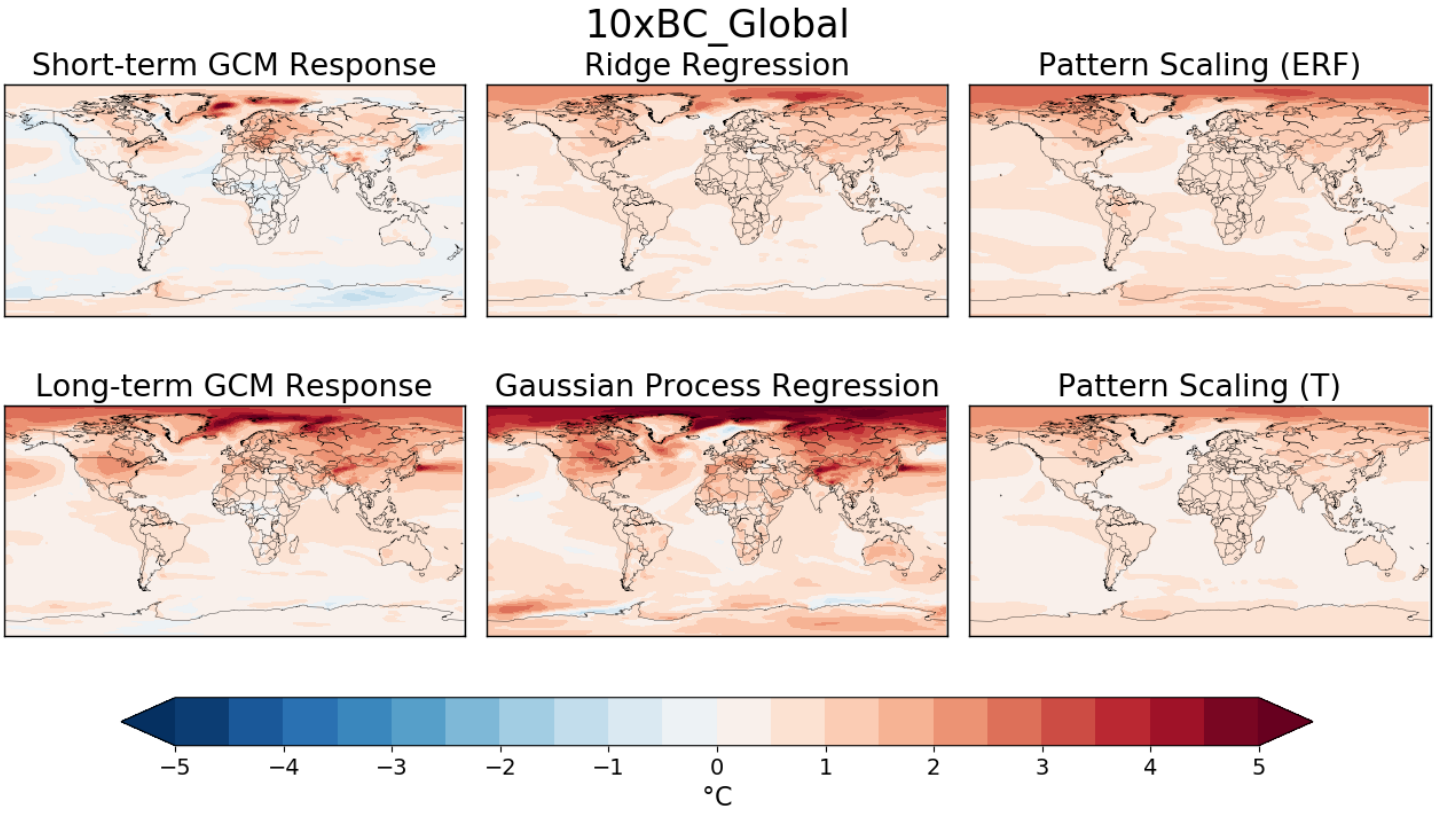

2xCO2_PDRMIP
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−5 −4 −3 −2 −1 0 1 2 3 4 5
°C


3xCH4
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−5 −4 −3 −2 −1 0 1 2 3 4 5
°C


+2%_Solar_Constant
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−5 −4 −3 −2 −1 0 1 2 3 4 5
°C

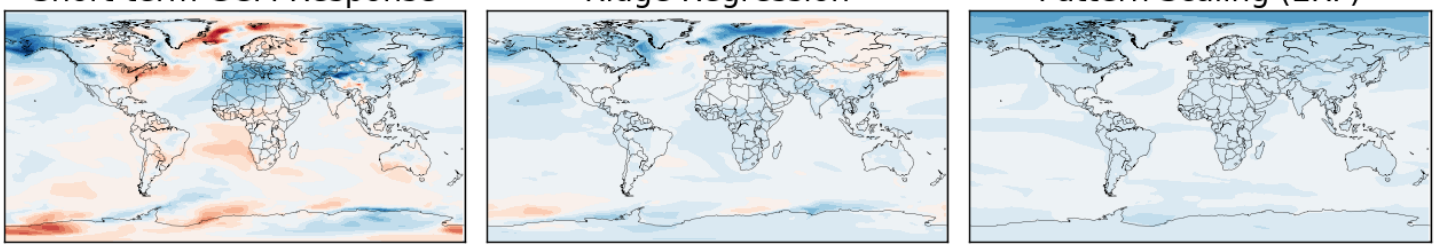

10xCFC-12
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−5 −4 −3 −2 −1 0 1 2 3 4 5
°C


5xSO4_Global
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−10
−8
−6
−4
−2
0
2
4
6
8
10
°C

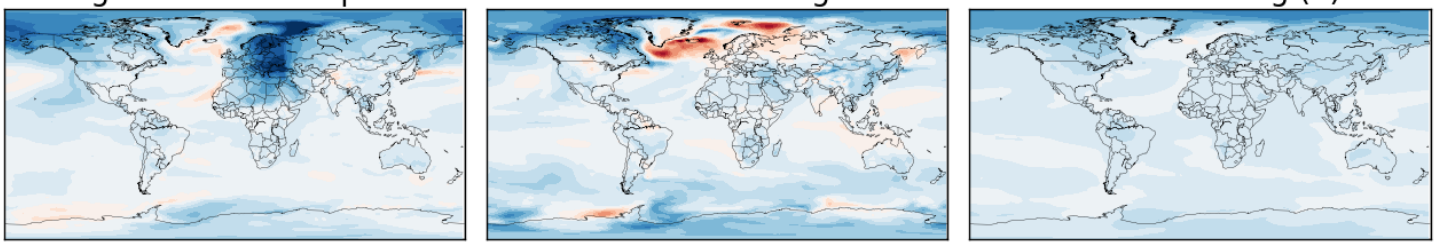

10xBC_Global
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−5
−4
−3
−2
−1
0
1
2
3
4
5
°C

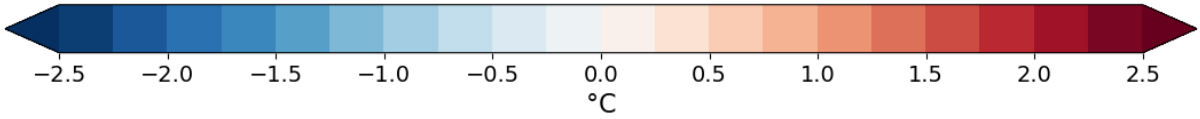
10xSO4_Europe

Short-term GCM Response | Ridge Regression | Pattern Scaling (ERF)



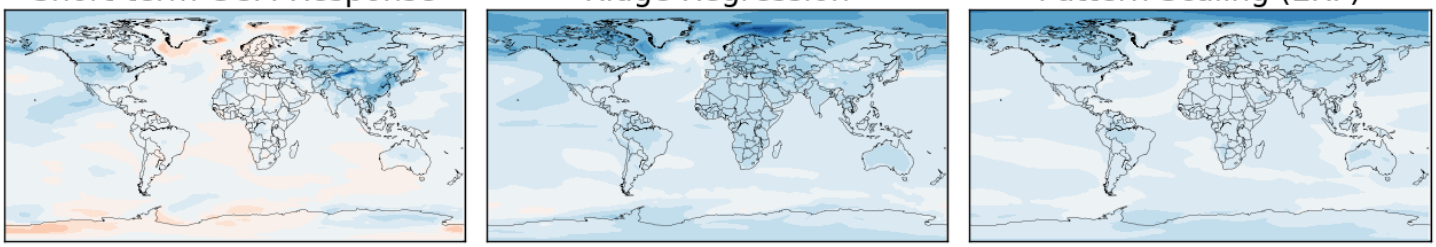
Long-term GCM Response | Gaussian Process Regression | Pattern Scaling (T)



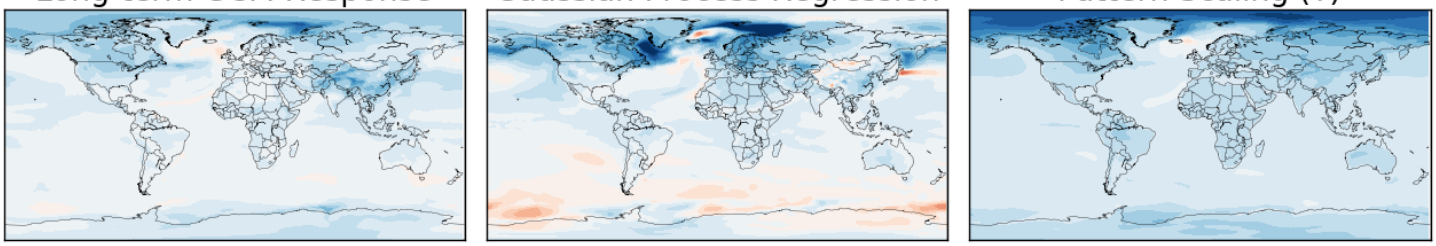


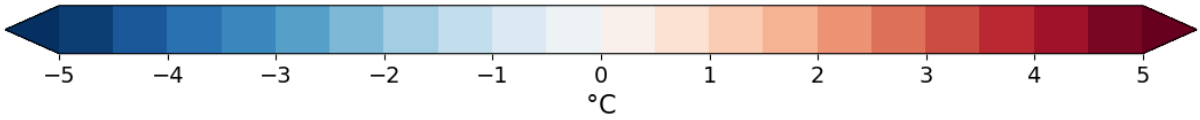
10xSO4_Asia

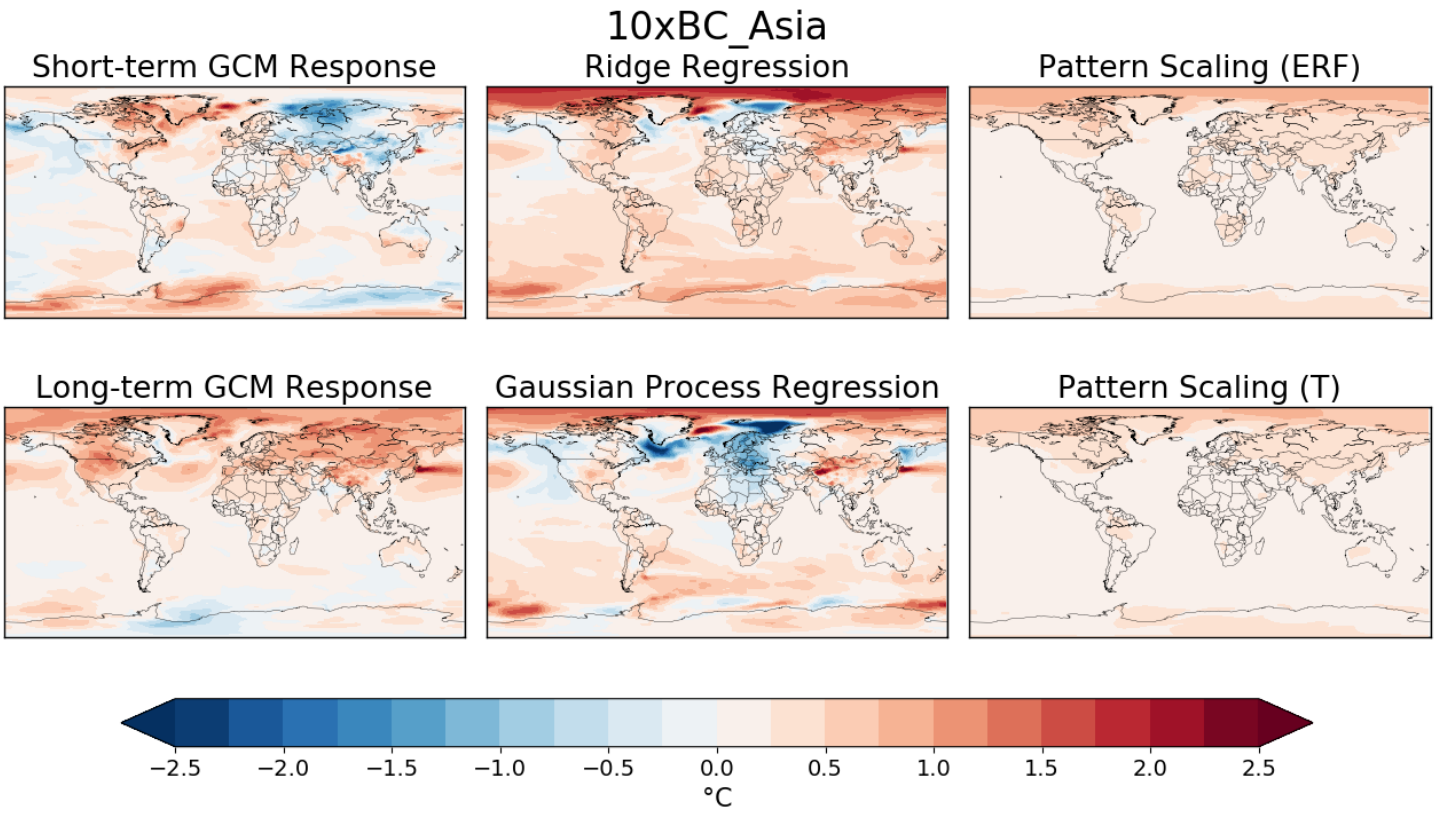
Short-term GCM Response | Ridge Regression | Pattern Scaling (ERF)

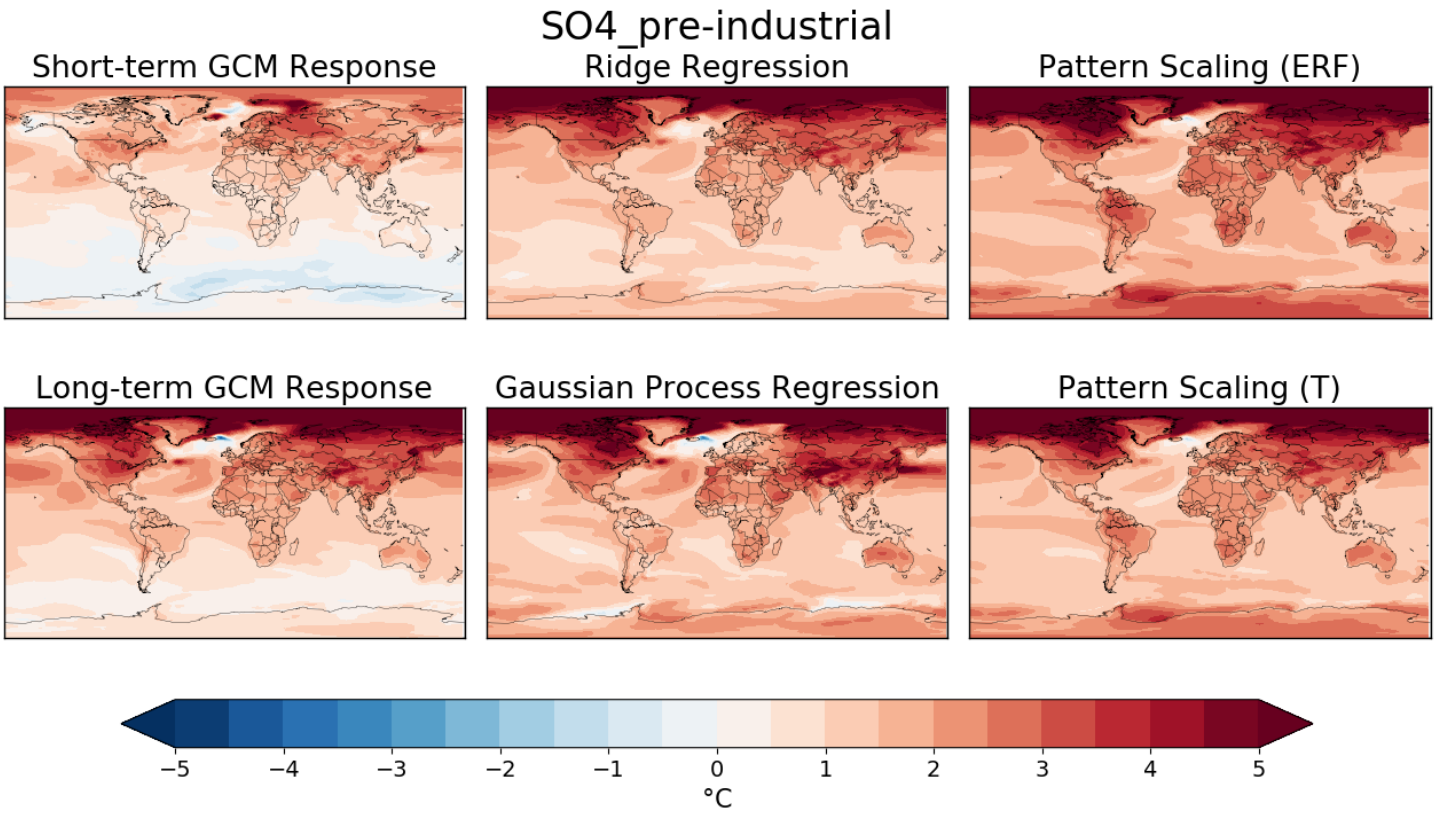


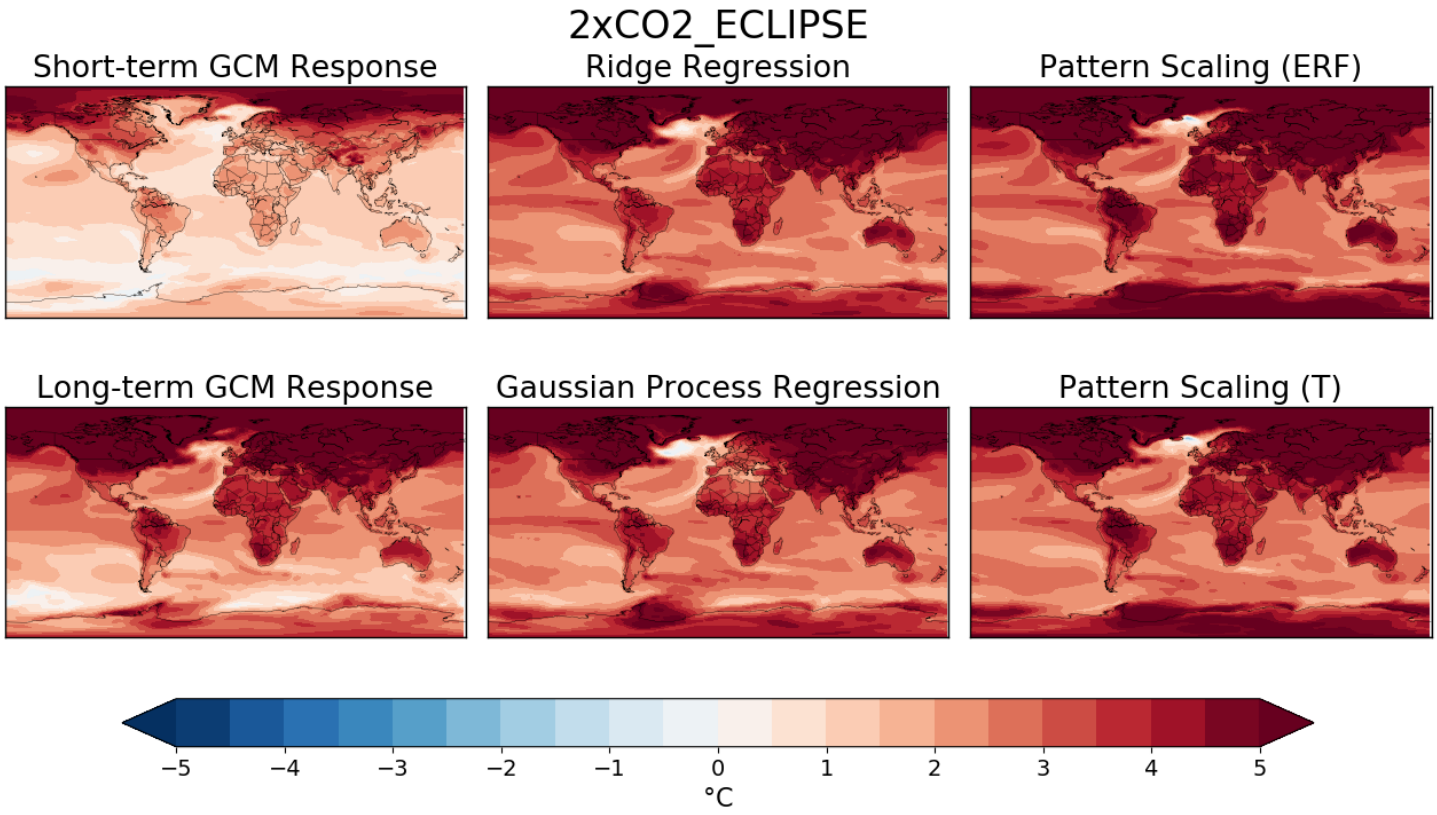
Long-term GCM Response | Gaussian Process Regression | Pattern Scaling (T)

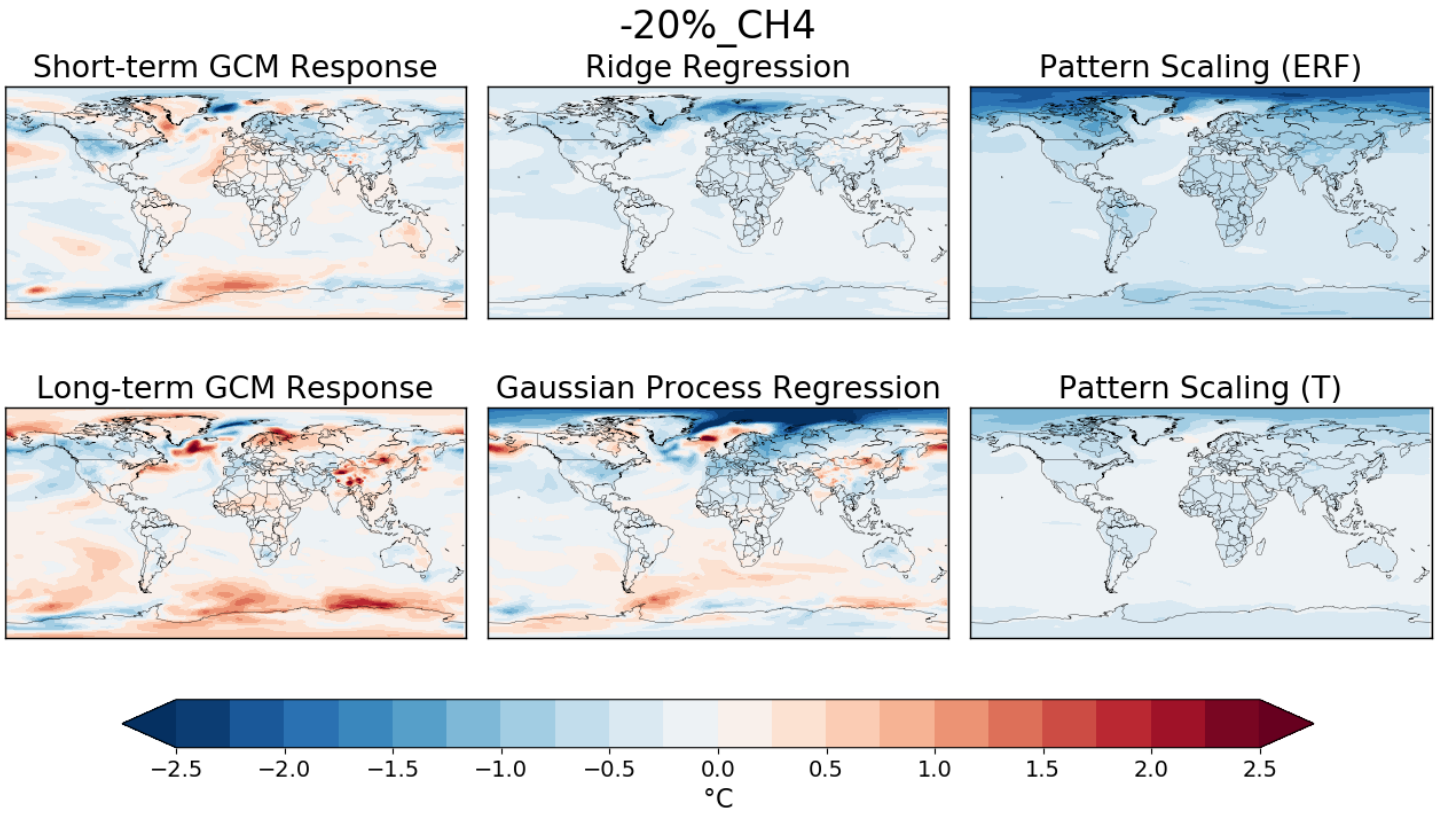


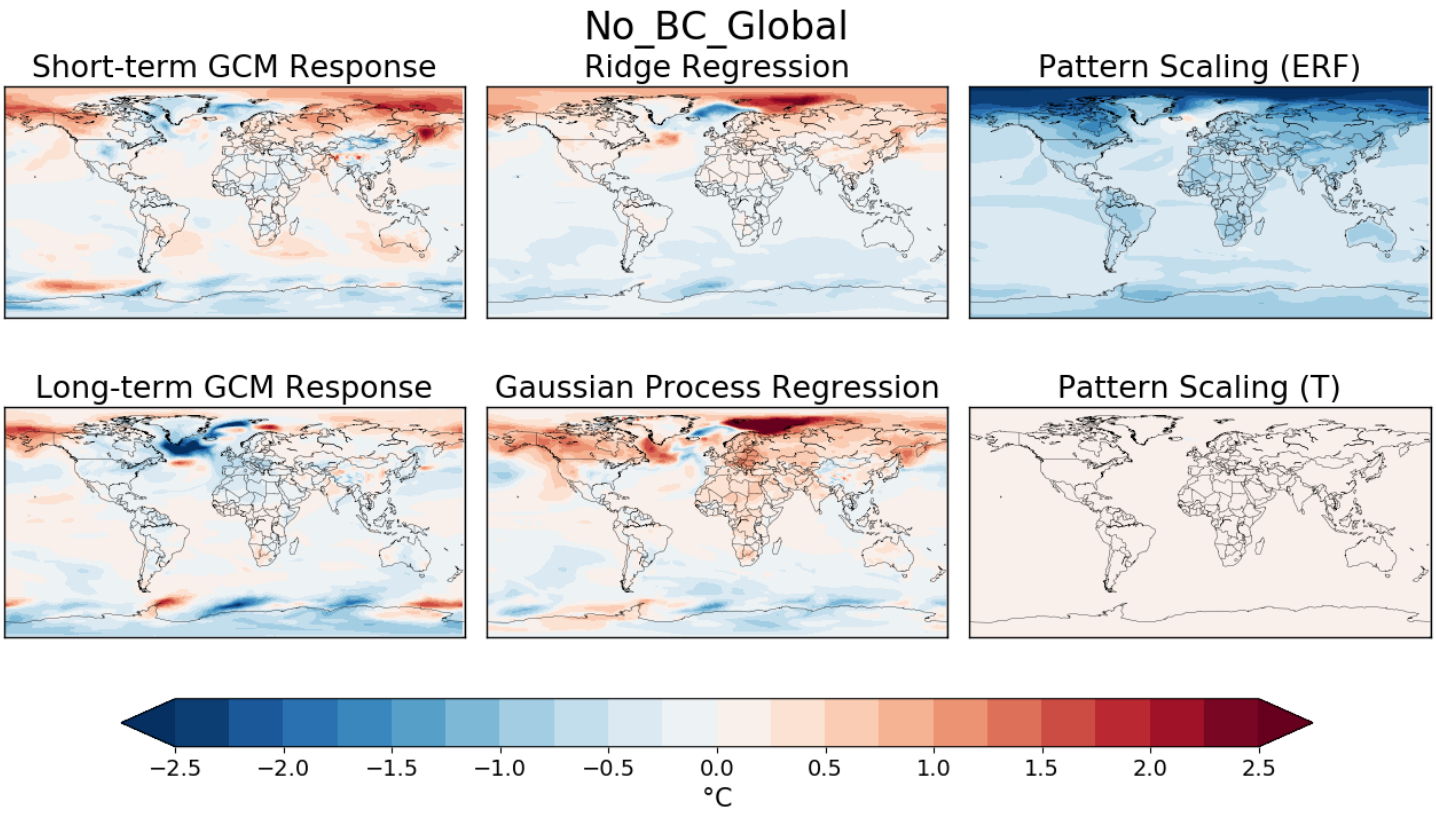


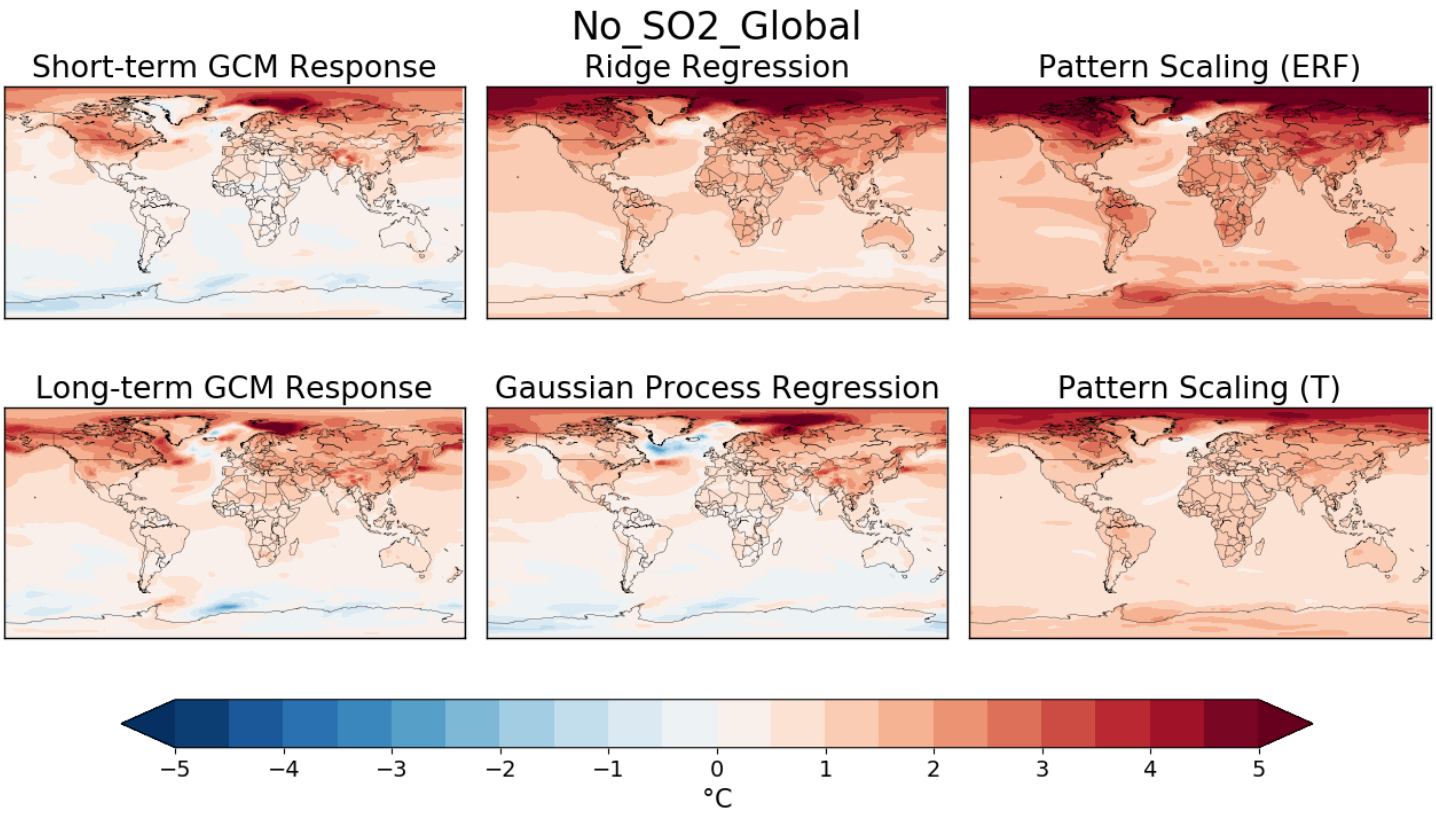

10xBC_Asia
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
°C

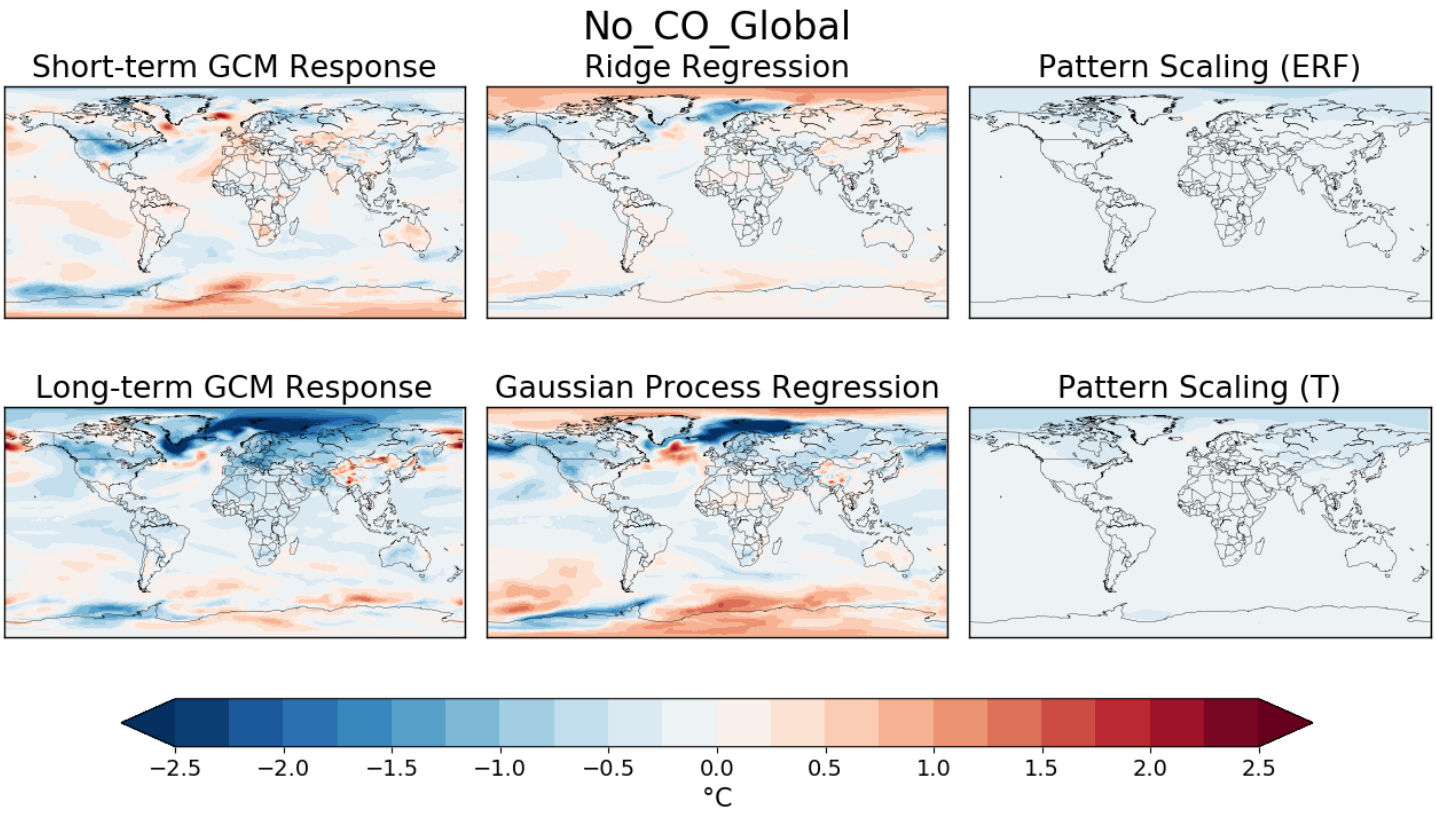

SO4_pre-industrial
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−5
−4
−3
−2
−1
0
1
2
3
4
5
°C

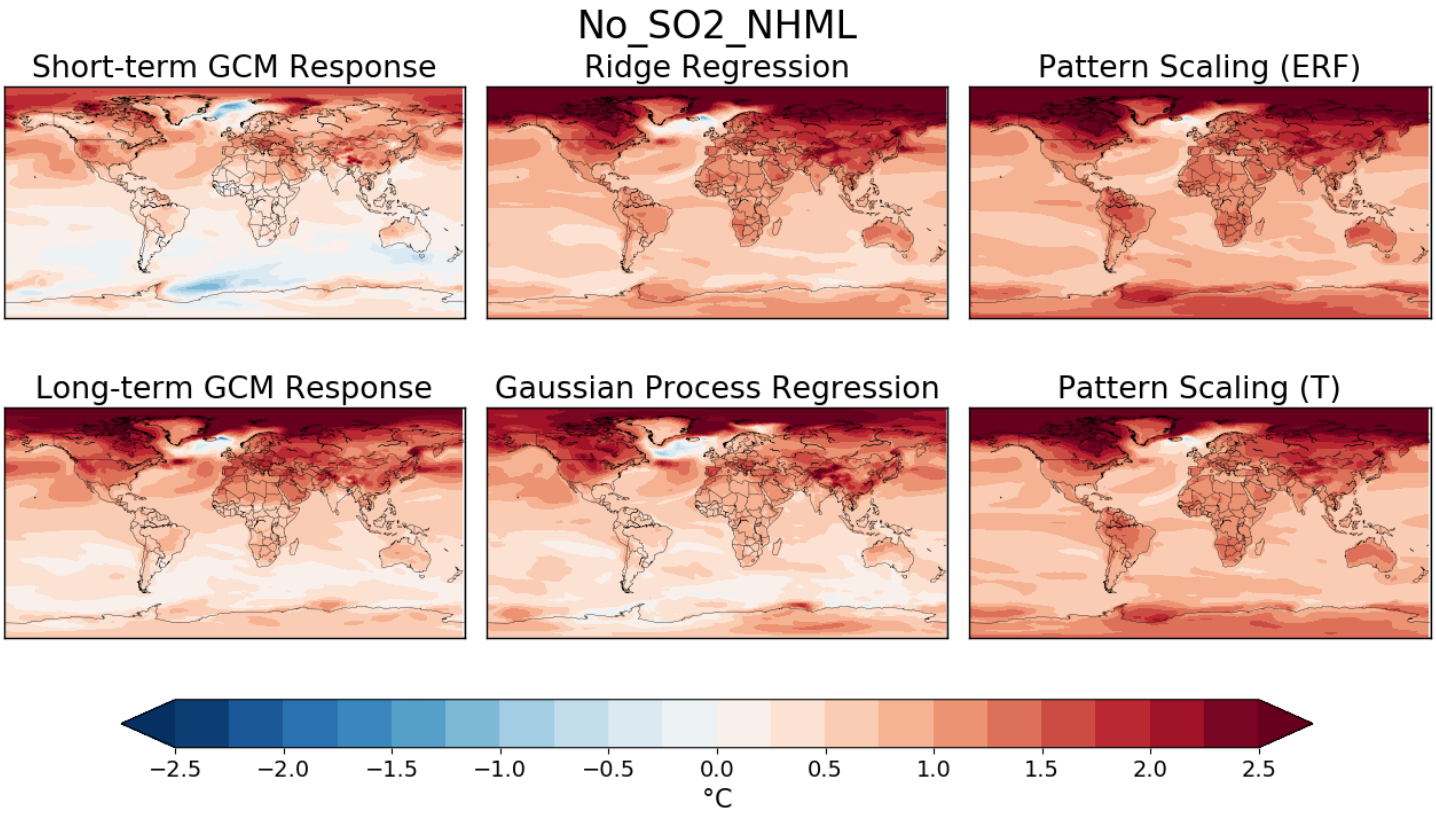

2xCO2_ECLIPSE
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−5
−4
−3
−2
−1
0
1
2
3
4
5
°C

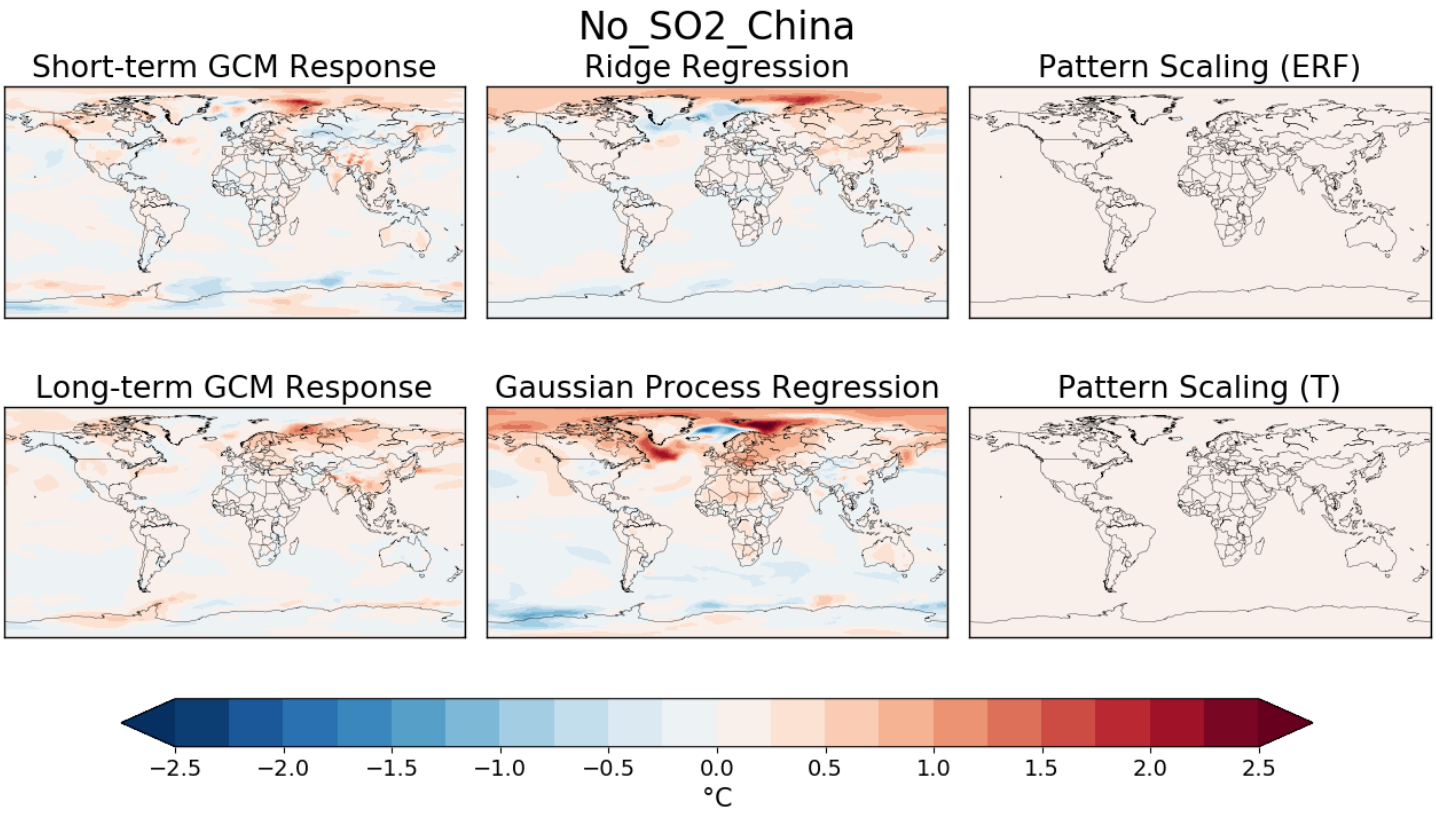

-20%_CH4
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
°C

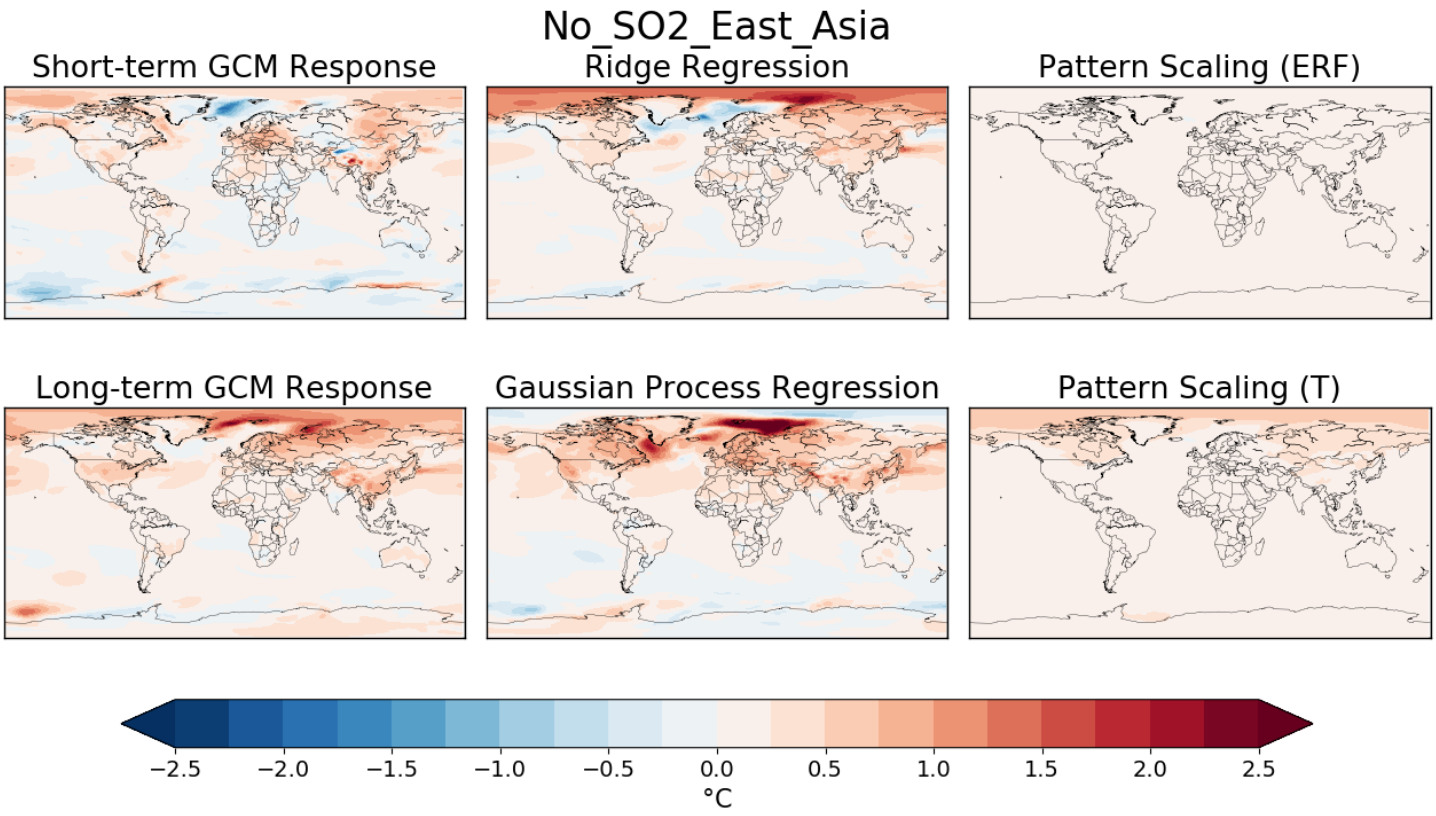

No_BC_Global
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
°C

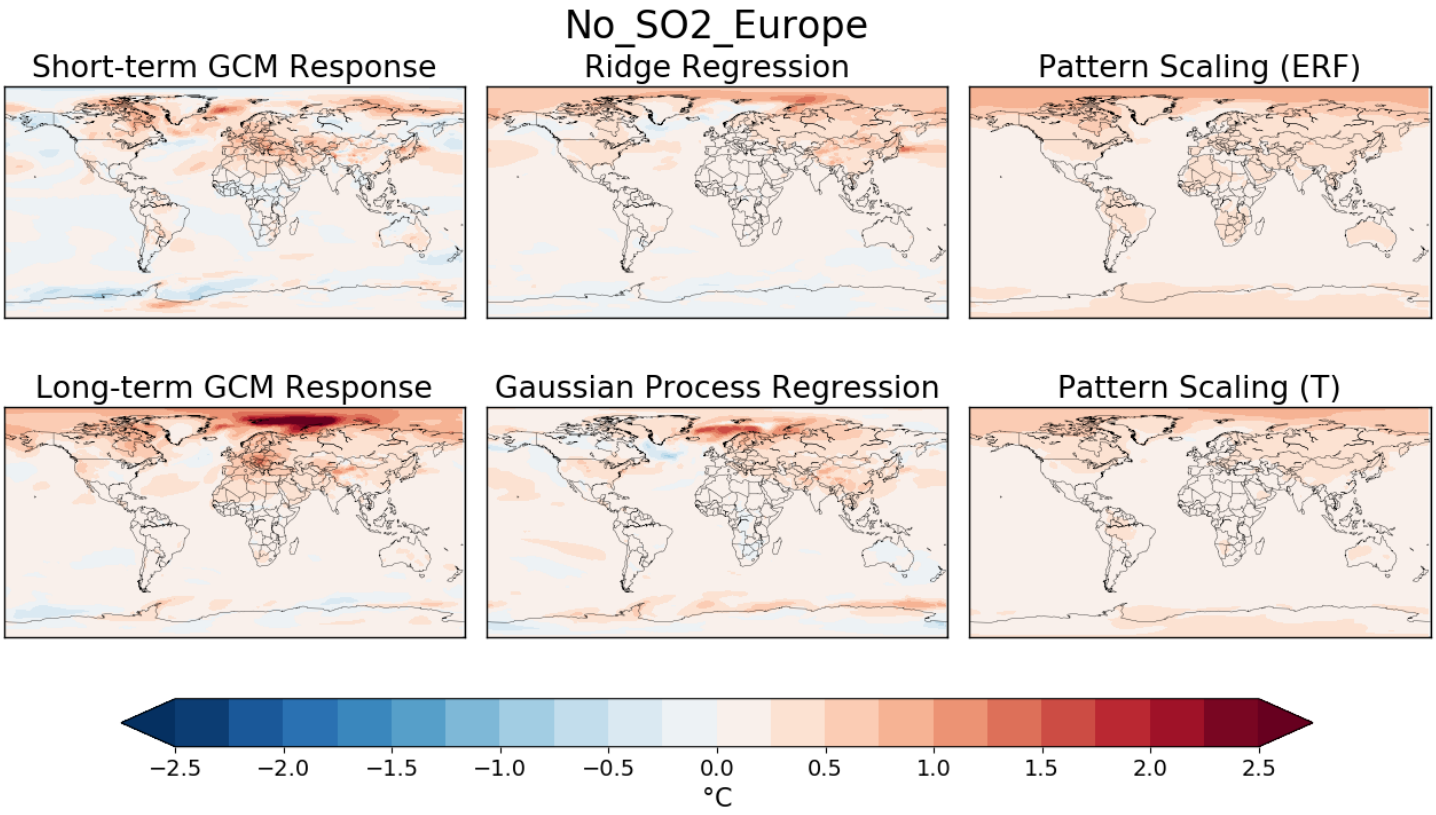

No_SO2_Global
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−5
−4
−3
−2
−1
0
1
2
3
4
5
°C

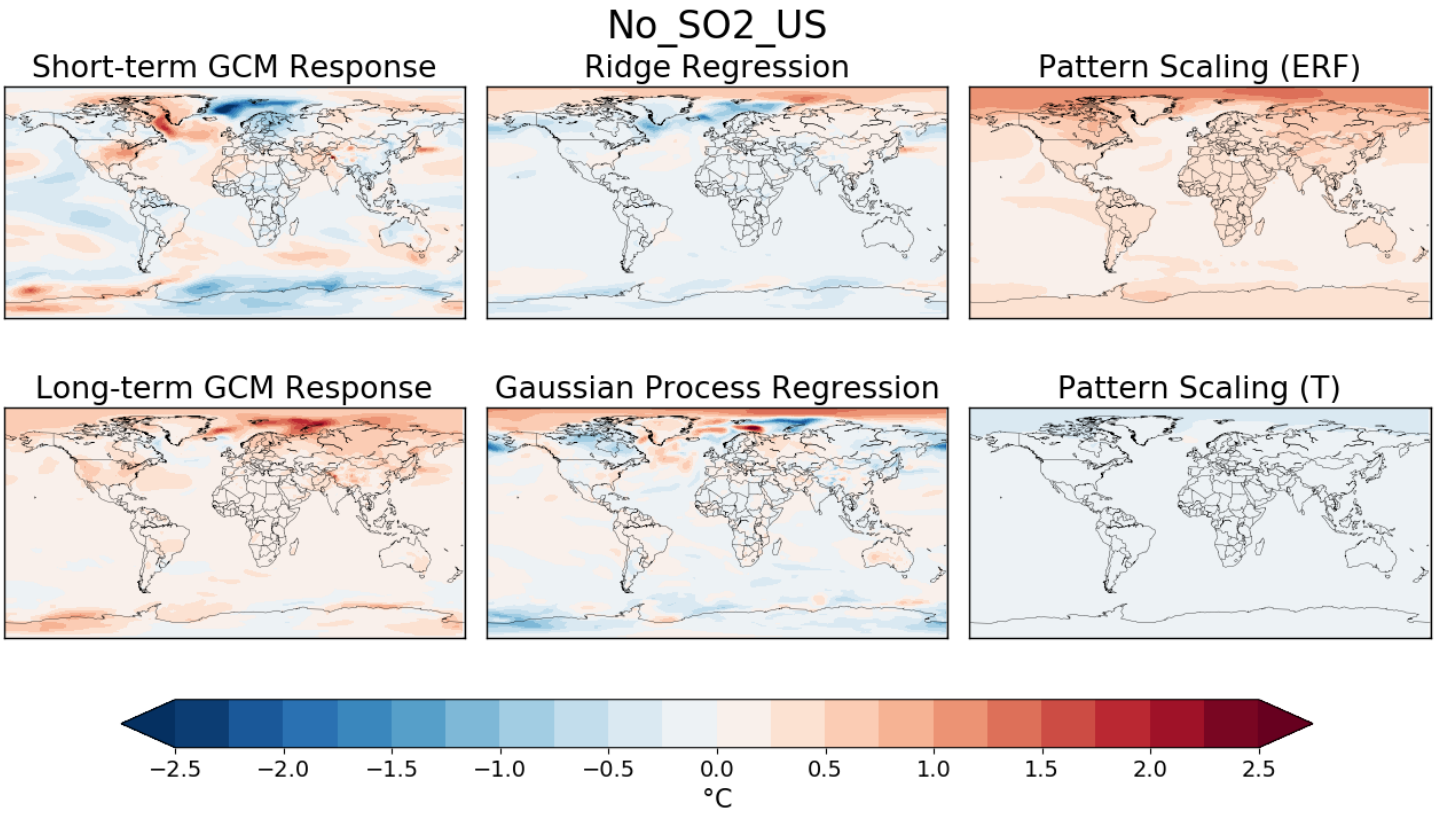

No_CO_Global
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
°C

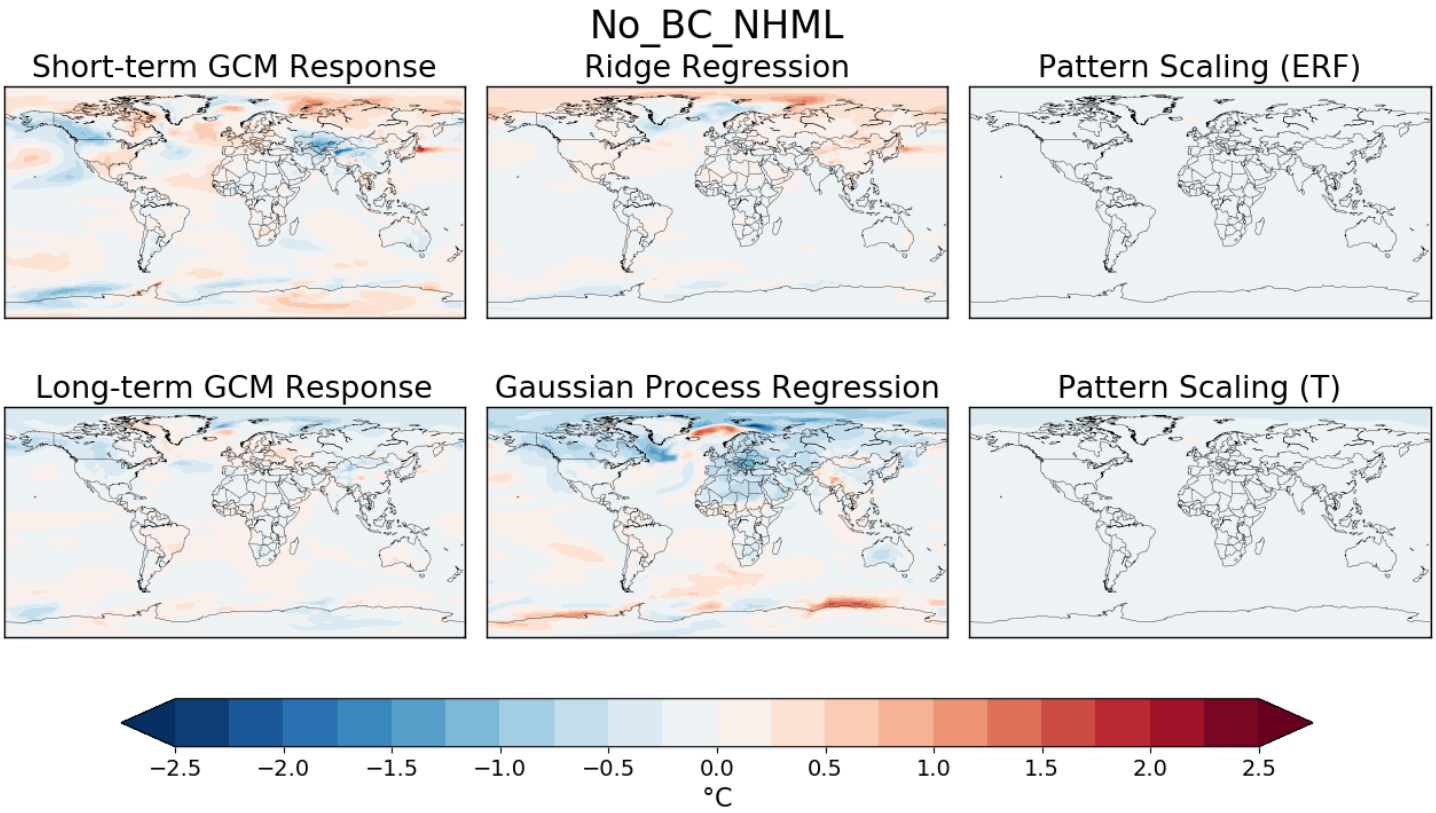

No_SO2_NHML
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
°C


No_SO2_China
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
°C

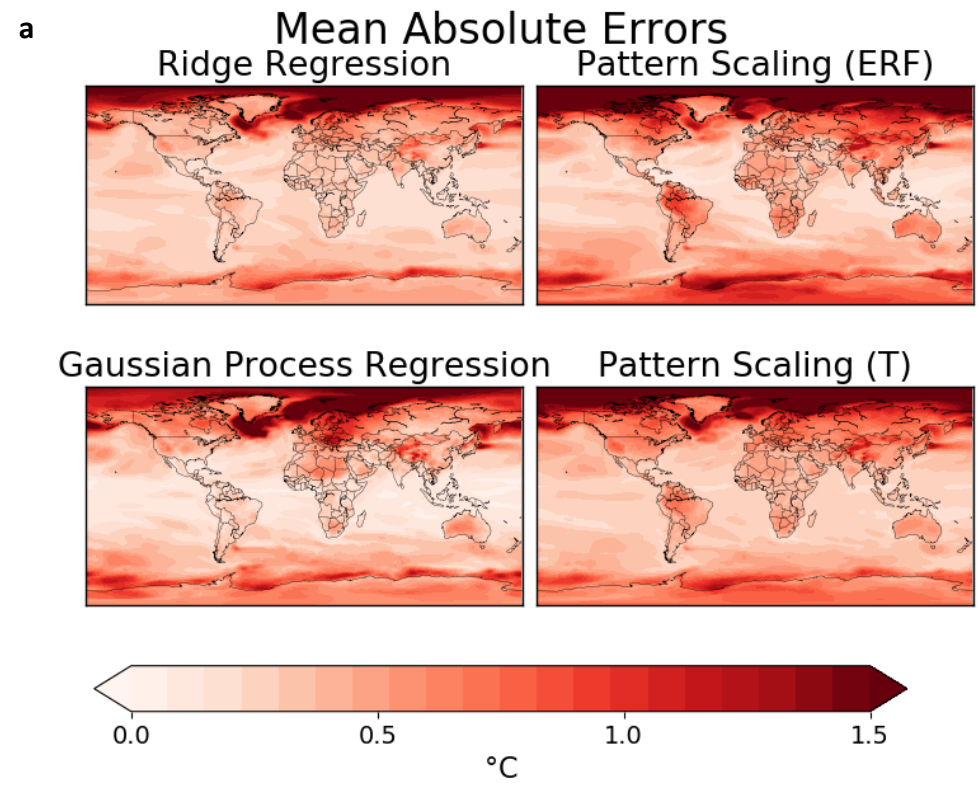

No_SO2_East_Asia
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
°C

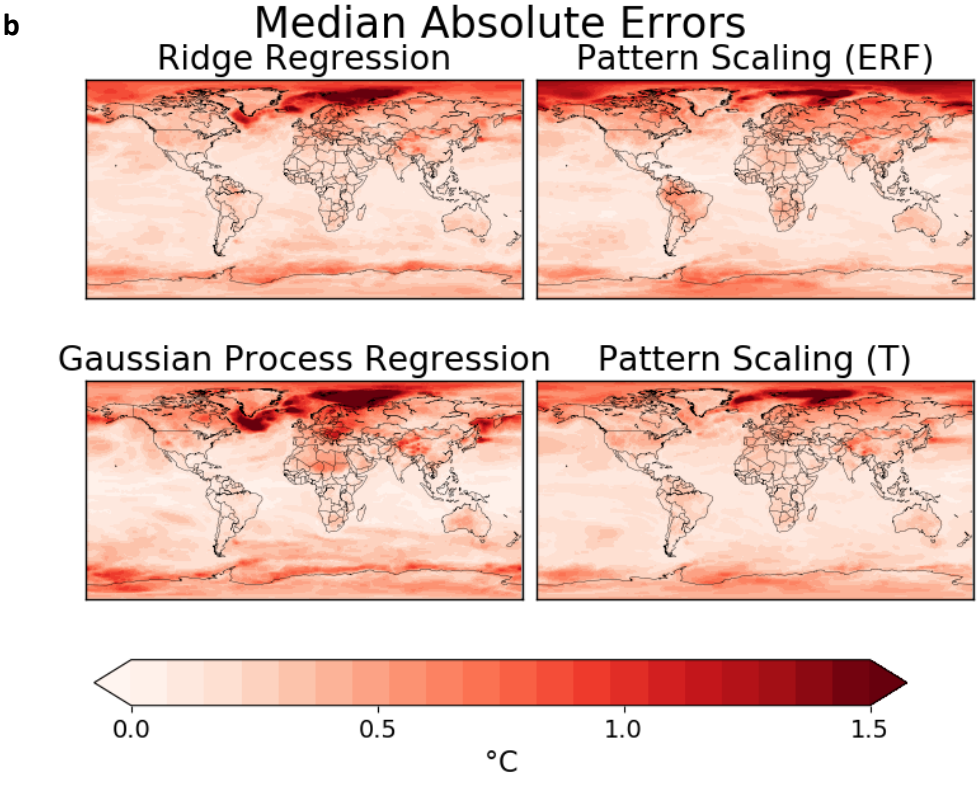

No_SO2_Europe
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
°C


No_SO2_US
Short-term GCM Response
Ridge Regression
Pattern Scaling (ERF)
Long-term GCM Response
Gaussian Process Regression
Pattern Scaling (T)
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
°C

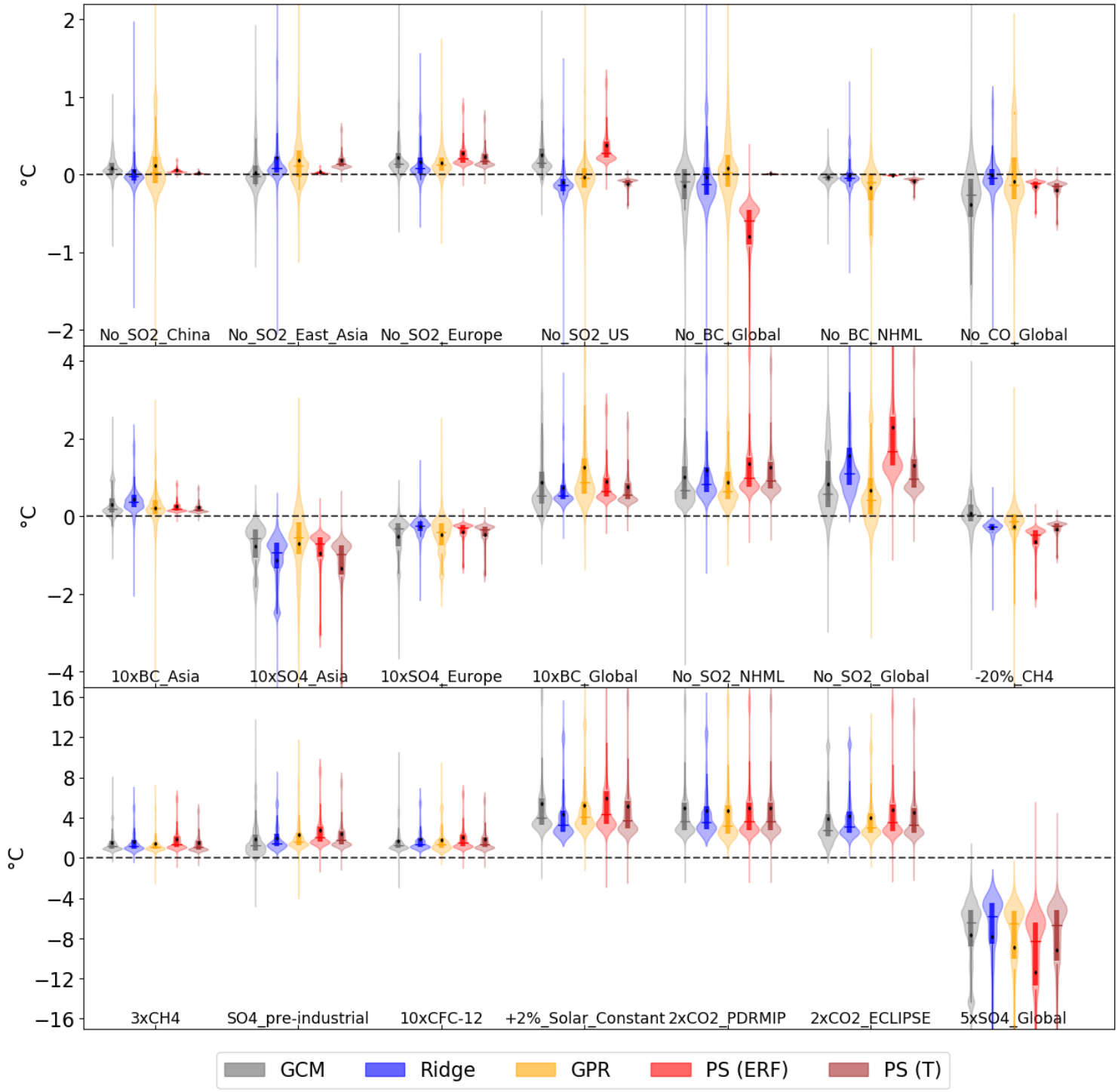


Supplementary Fig. 1: Predicted surface temperature response patterns for all scenarios in °C. First column: GCM responses in short-term (top) and long-term (bottom); Second column: the predicted long-term response using machine learning methods Ridge regression (top) and Gaussian Process Regression (bottom); Third column: the predicted long-term response using Pattern Scaling methods based on effective radiative forcing (ERF; top) and short-term global mean surface-temperature (T; bottom). Scenarios are described in Supplementary Table 1. Note the different colour bar scales for different scenarios.



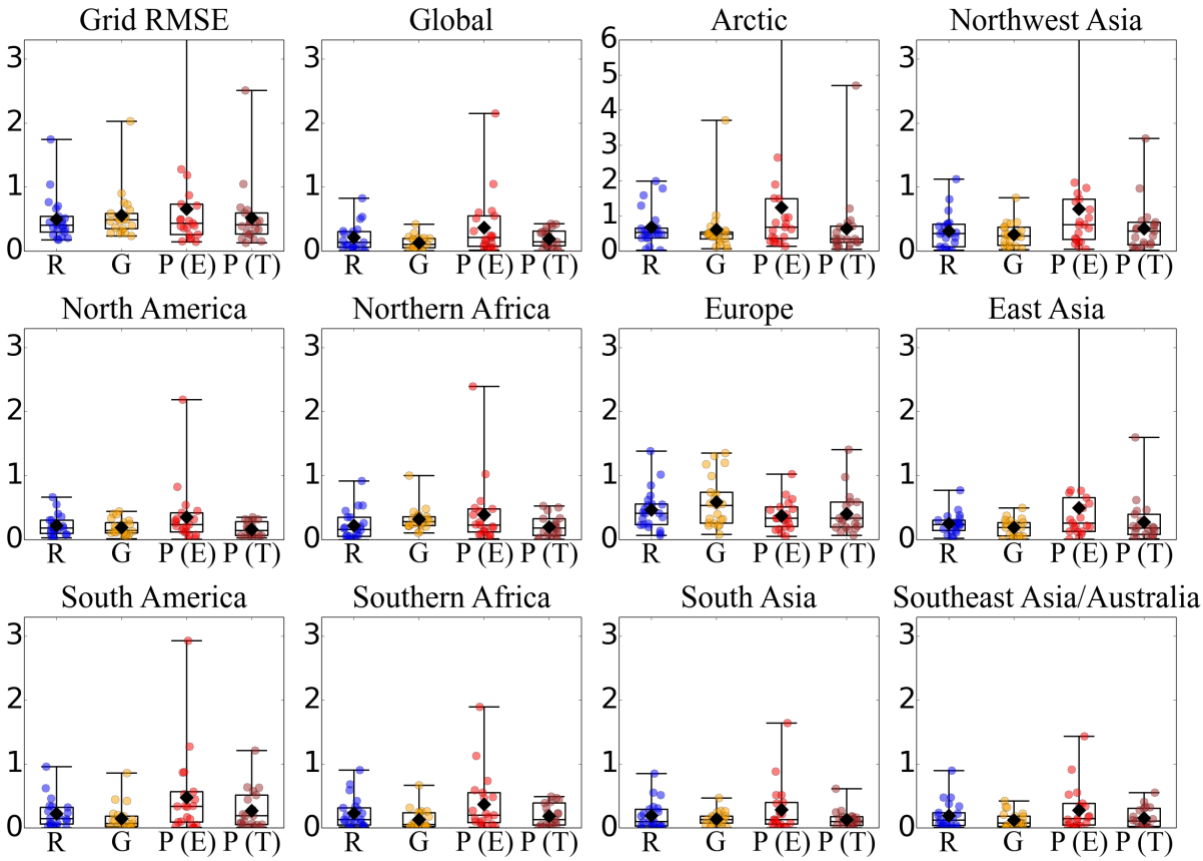


Supplementary Fig. 2: Absolute error maps for all methods, a mean, **b** median over all predicted scenarios



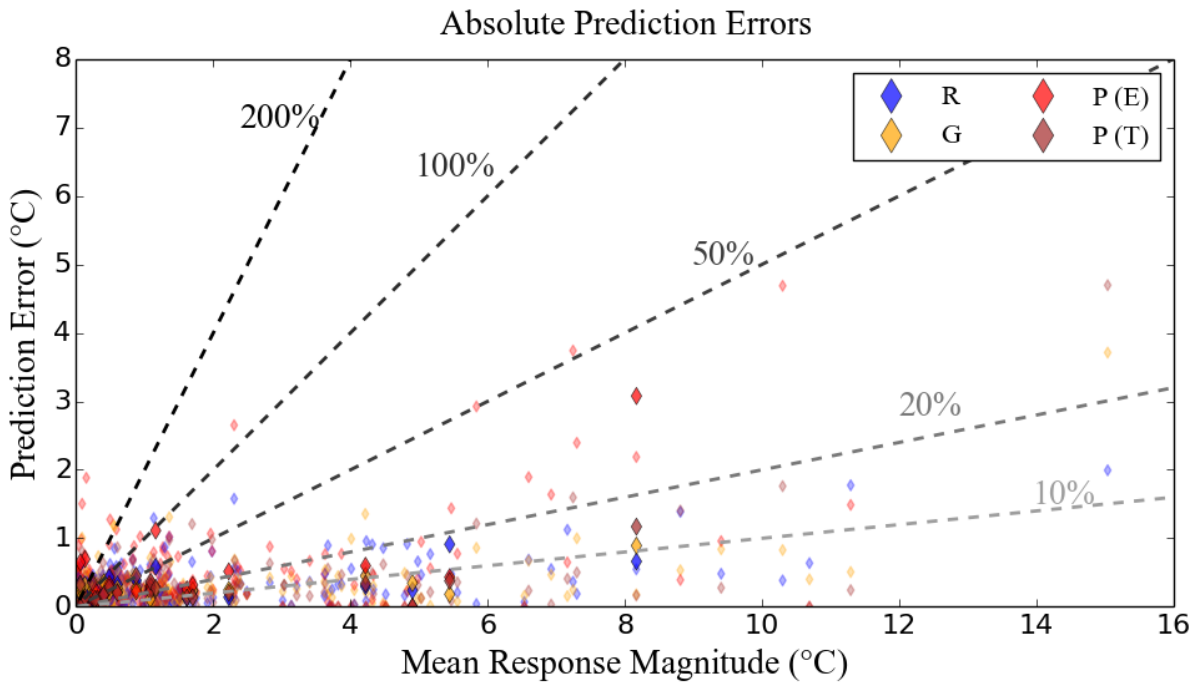
Supplementary Fig. 3: Spatial variability of long-term surface temperature response in °C for all prediction methods and for all scenarios. The distribution of predicted surface temperature responses constructed from all spatially weighted grid-points is shown along the vertical axis for each prediction method. From left to right the plots show the prediction from the general circulation model (GCM), Ridge regression prediction, Gaussian Process regression (GPR) prediction and Pattern Scaling (PS) using effective radiative forcing (ERF) and using the short-

term global mean surface temperature (T). The central vertical boxes indicate the interquartile range shown on a standard box plot, the horizontal line shows the median and the black point shows the mean. Note the different vertical scales for each row.

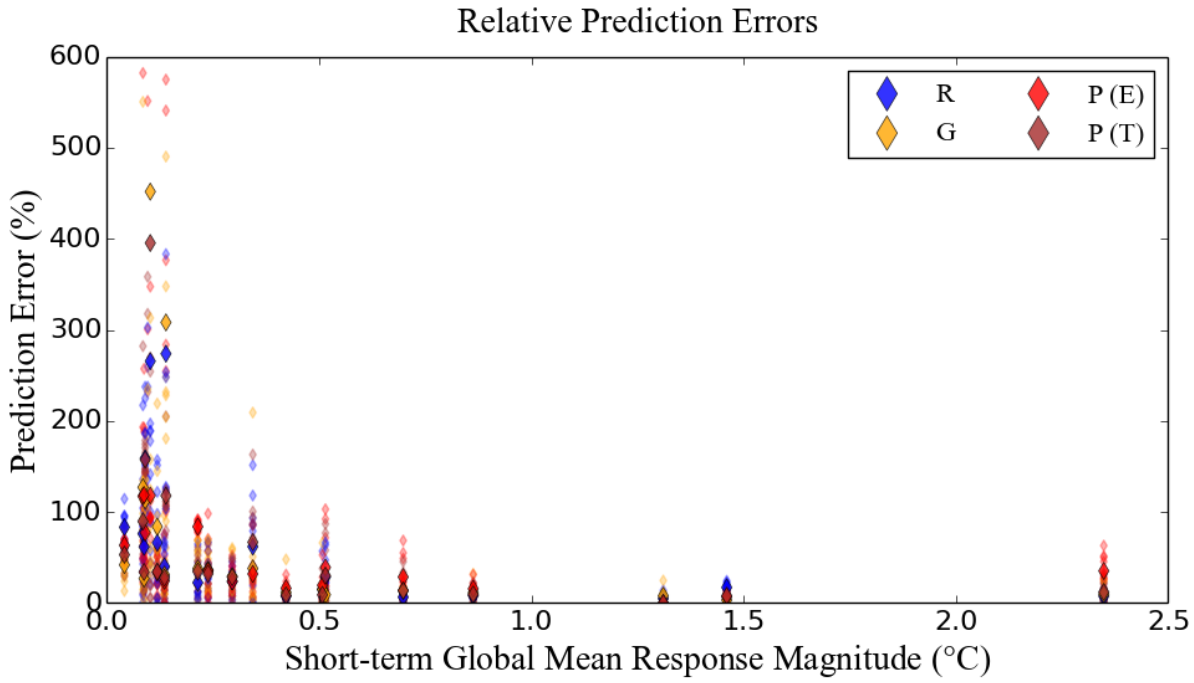


Supplementary Fig. 4: Absolute errors in °C for all scenarios in each selected region highlighted in Fig. 3 for long-term climate response prediction using four methods: R=Ridge regression, G=Gaussian Process regression, P(E)=pattern scaling using ERF as the scaler value P(T)=pattern scaling using global mean short-term temperature response as the scaler value.

a

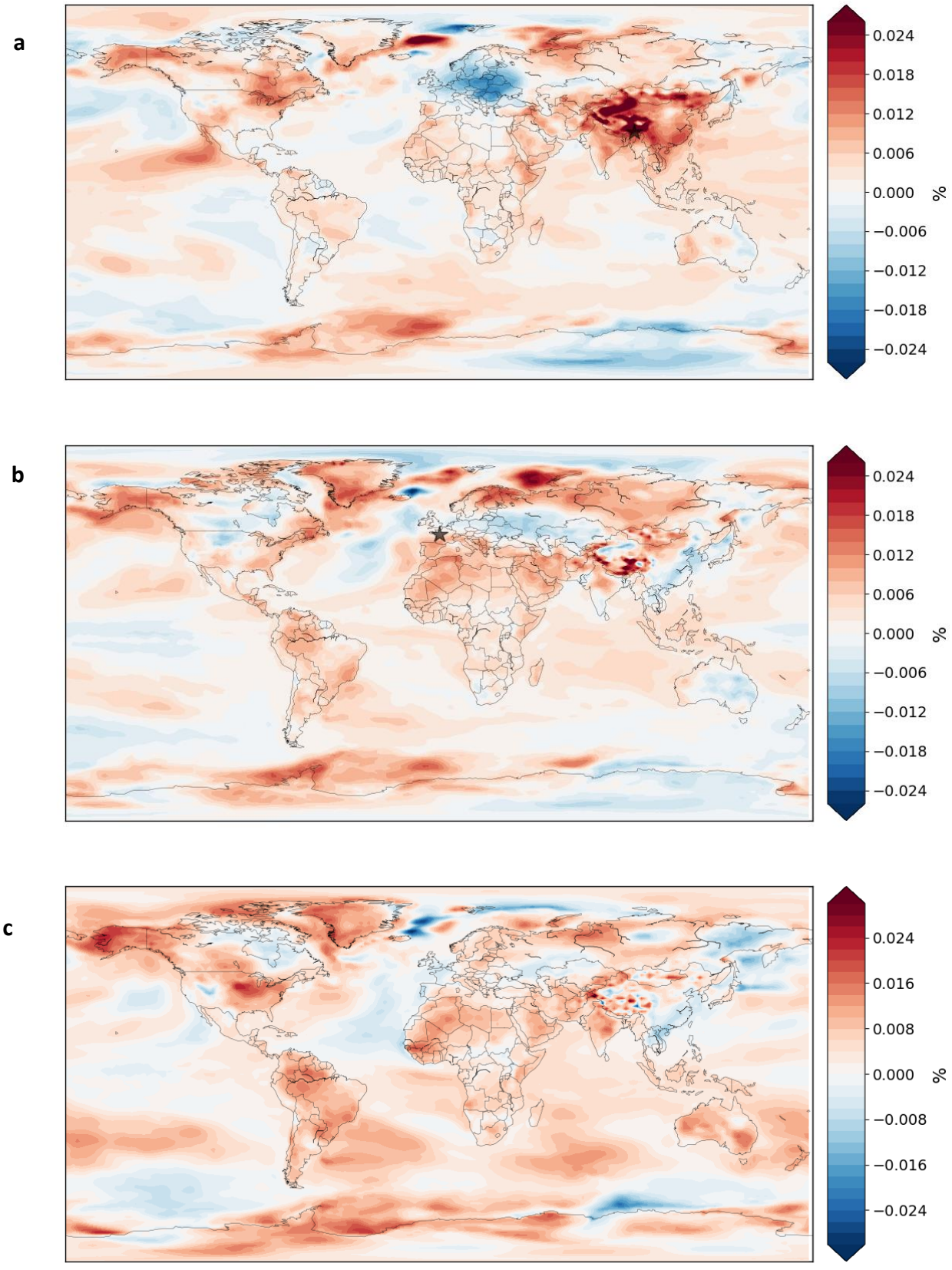


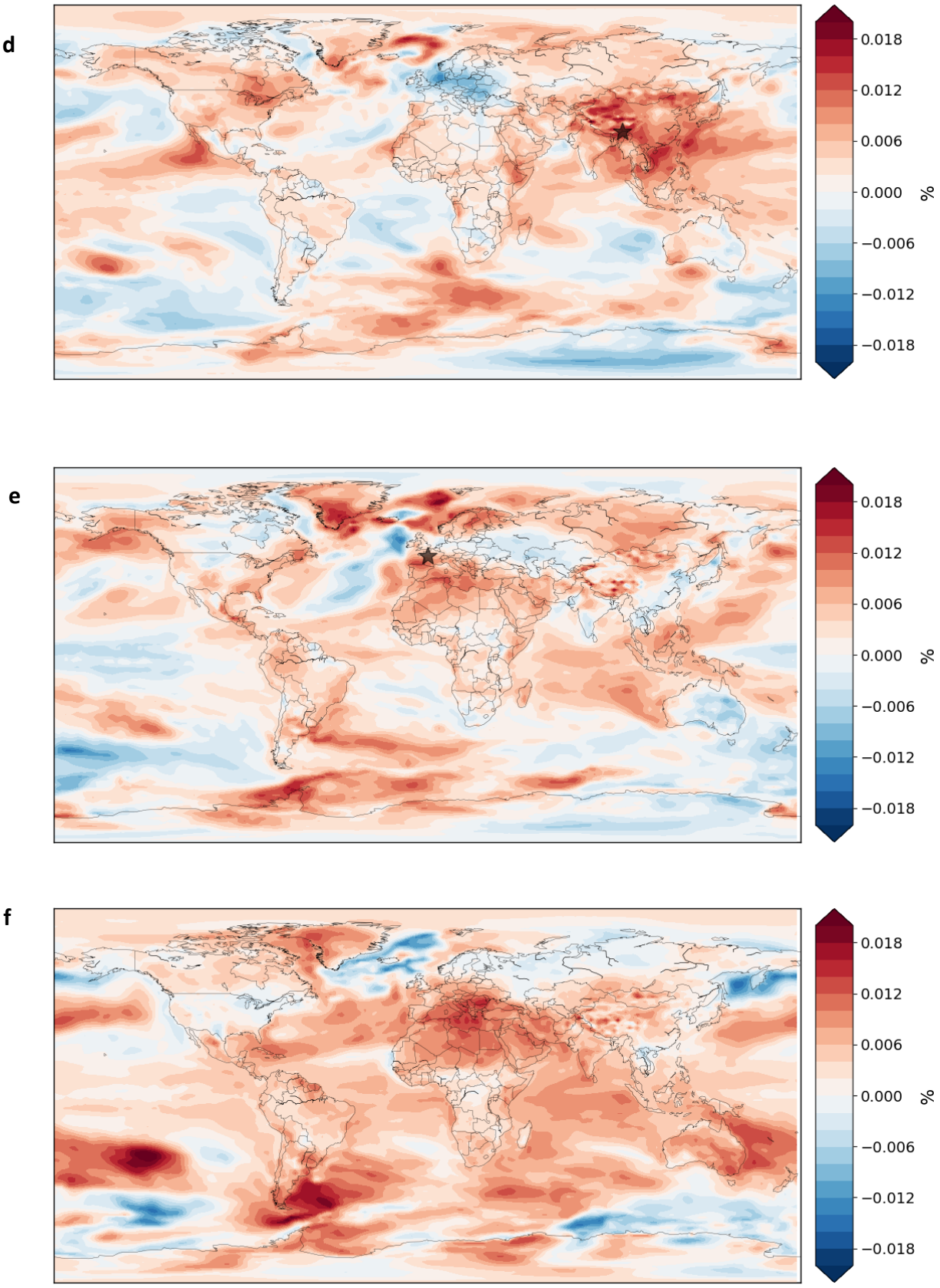
b



Supplementary Fig. 5: Absolute and relative prediction errors °C, a Absolute prediction error in °C compared against response magnitude for all predicted scenarios over key regions in Fig. 3. The faded points show regional predictions for each scenario with the bold points

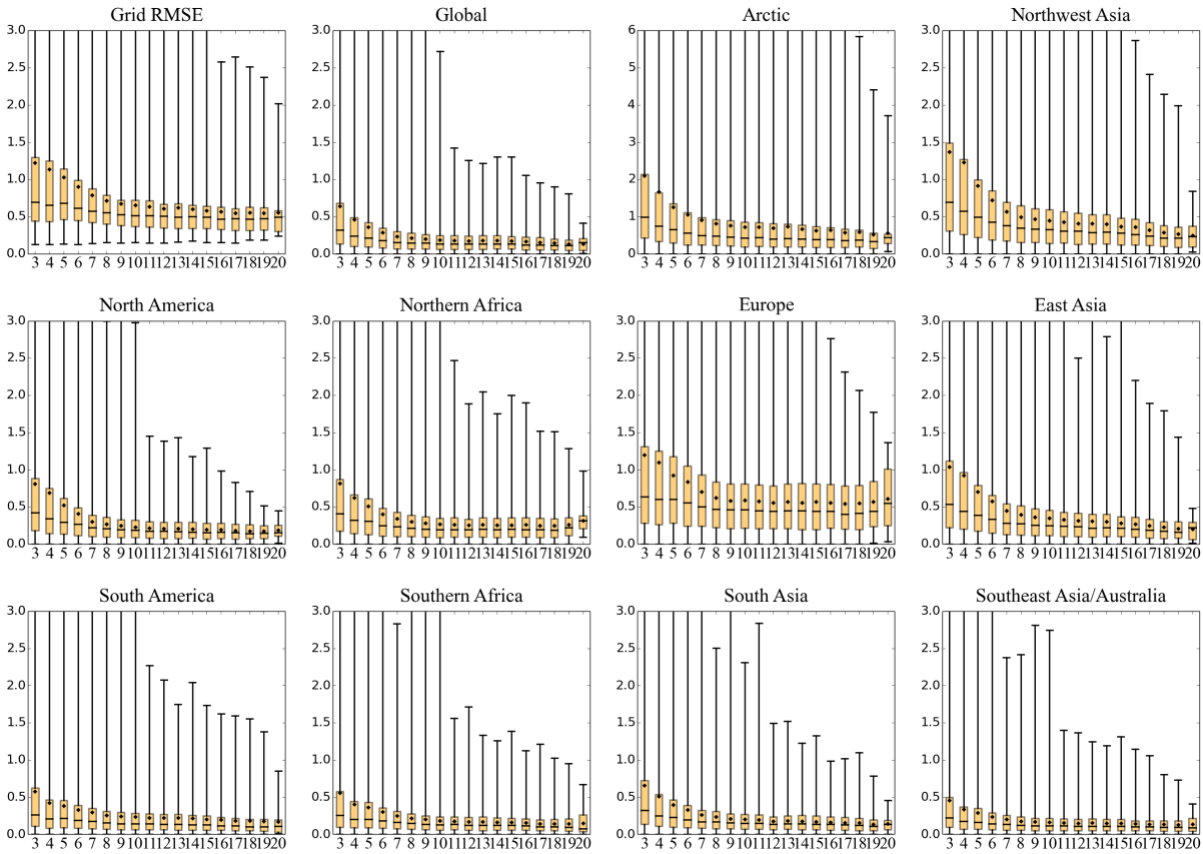
showing the average absolute prediction error for one scenario across all regions. The lines for relative errors of 10%, 20%, 50%, 100% and 200% are also shown. **b** Relative prediction errors as a percentage against short-term global mean response. The faded points show the regional prediction errors and the bold points show the average relative prediction error for each scenario across all regions. R=Ridge regression, G=Gaussian Process regression, P(E)=Pattern scaling using ERF as the scaler value P(T)=Pattern scaling using global mean short-term temperature response as the scaler value.


a
0.024
0.018
0.012
0.006
0.000
−0.006
−0.012
−0.018
−0.024
%
b
0.024
0.018
0.012
0.006
0.000
−0.006
−0.012
−0.018
−0.024
%
c
0.024
0.016
0.008
0.000
−0.008
−0.016
−0.024
%

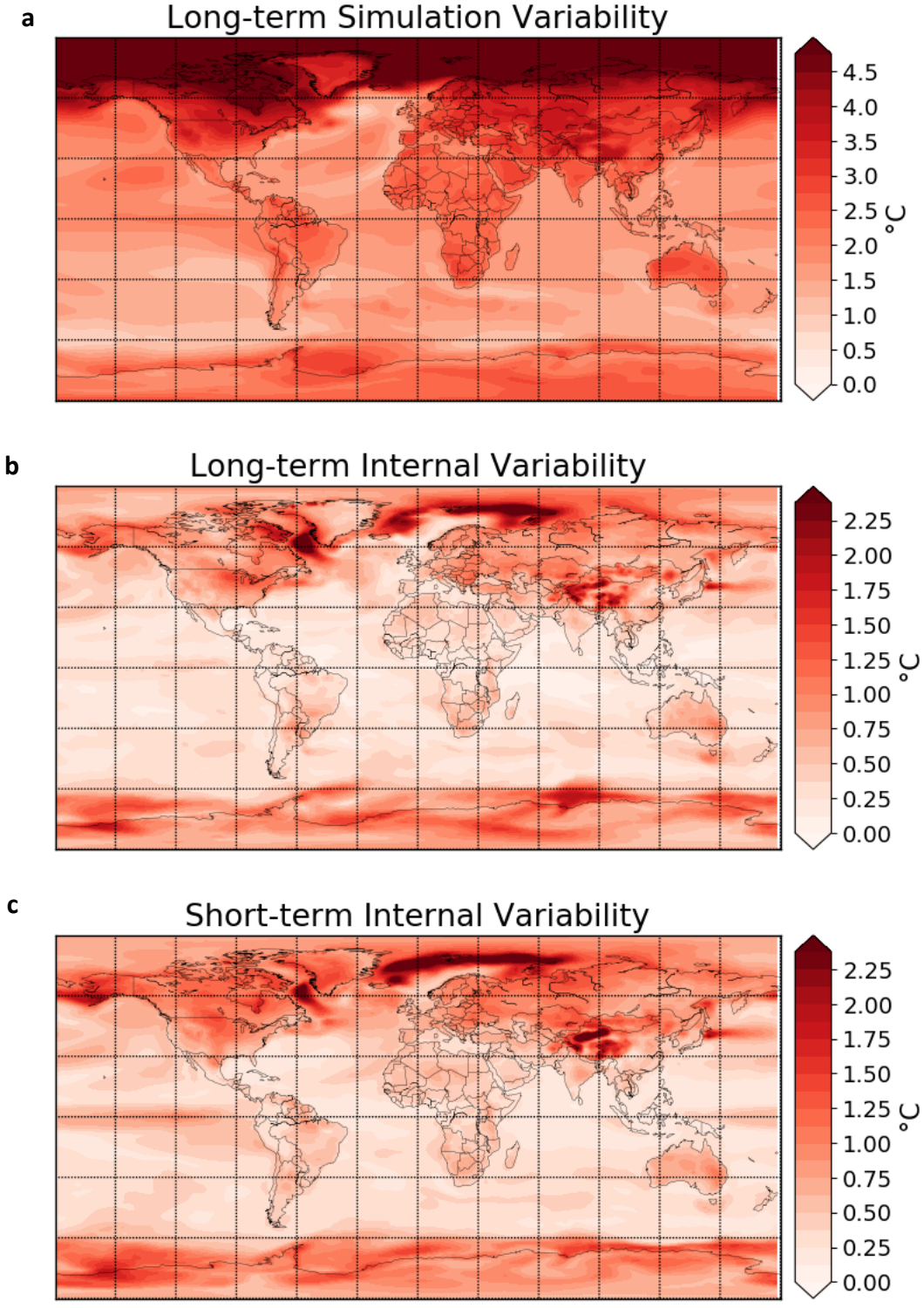


Supplementary Fig. 6: Ridge regression coefficients for a single selected grid-cell

regression at the point indicated by the star a over East Asia, **b** over Europe **c** the global mean Ridge regression coefficients calculated by taking a spatially weighted average over all output grid-cells and displayed as a percentage. **d e f** the same as **a b c** but calculated with inputs normalized independently for each predictor to remove any dependence on magnitude of response.

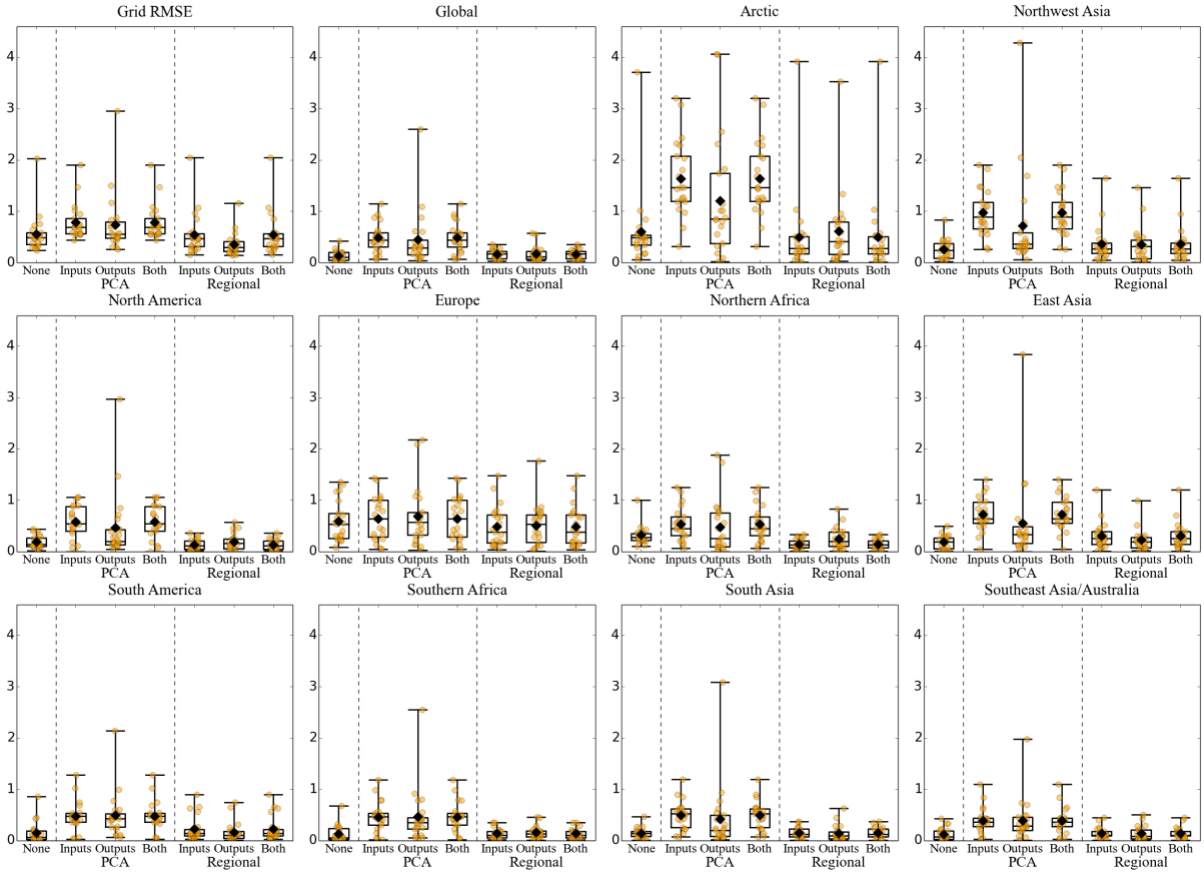


Supplementary Fig. 7: Boxplots showing the distribution of absolute prediction errors across all scenario predictions trained on an increasing number of simulations. For a fixed number of training data points, the process of training and predicting is repeated several times over different combinations of training data to obtain multiple prediction errors for each scenario. Boxes show the interquartile range, whiskers show the extrema, lines show the medians and black diamonds show the mean values (also plotted in Fig. 4). Note the different scale for the Arctic and Europe.

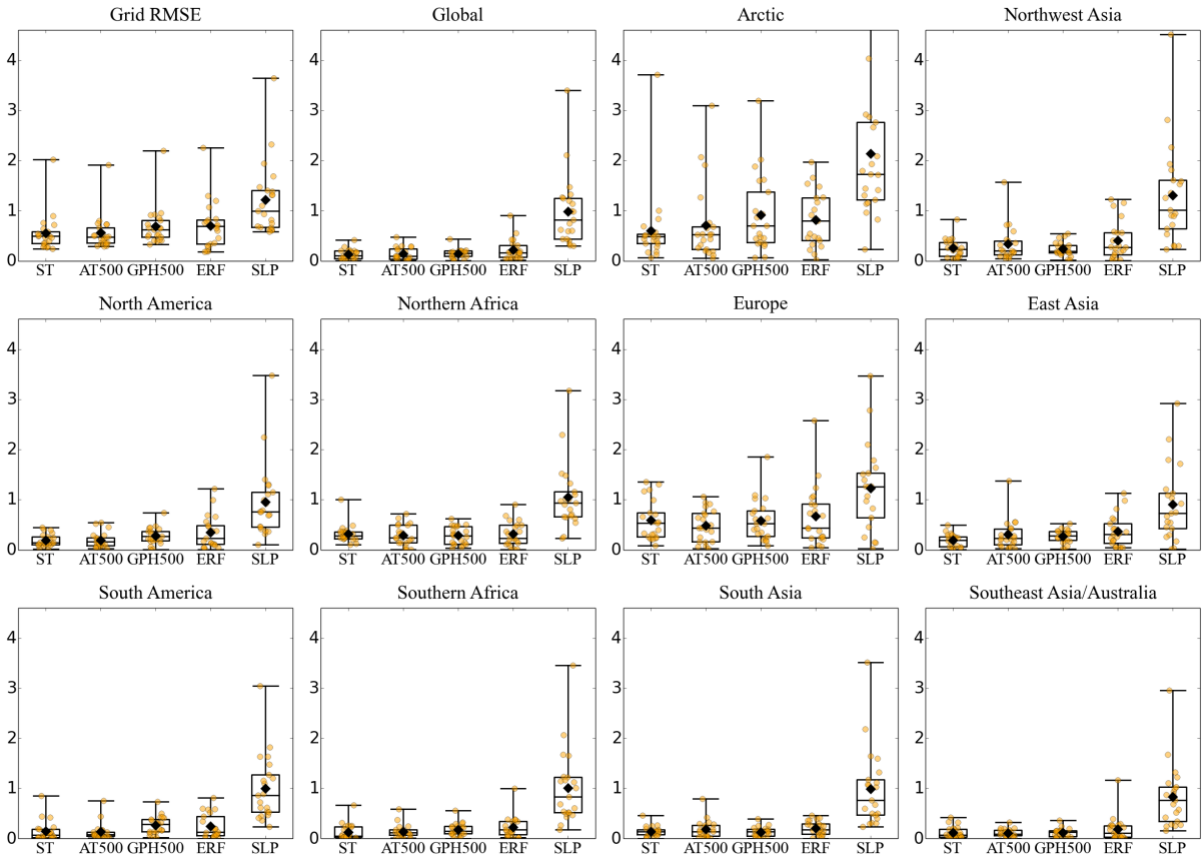


Supplementary Fig. 8: Maps of variability in data, a Simulation variability calculated as the standard deviation in long-term surface temperature response across all available data, **b** Internal variability in long-term temperature response calculated as the standard deviation at each grid point across the long-term response time series in the control run from Kasoar *et al.*

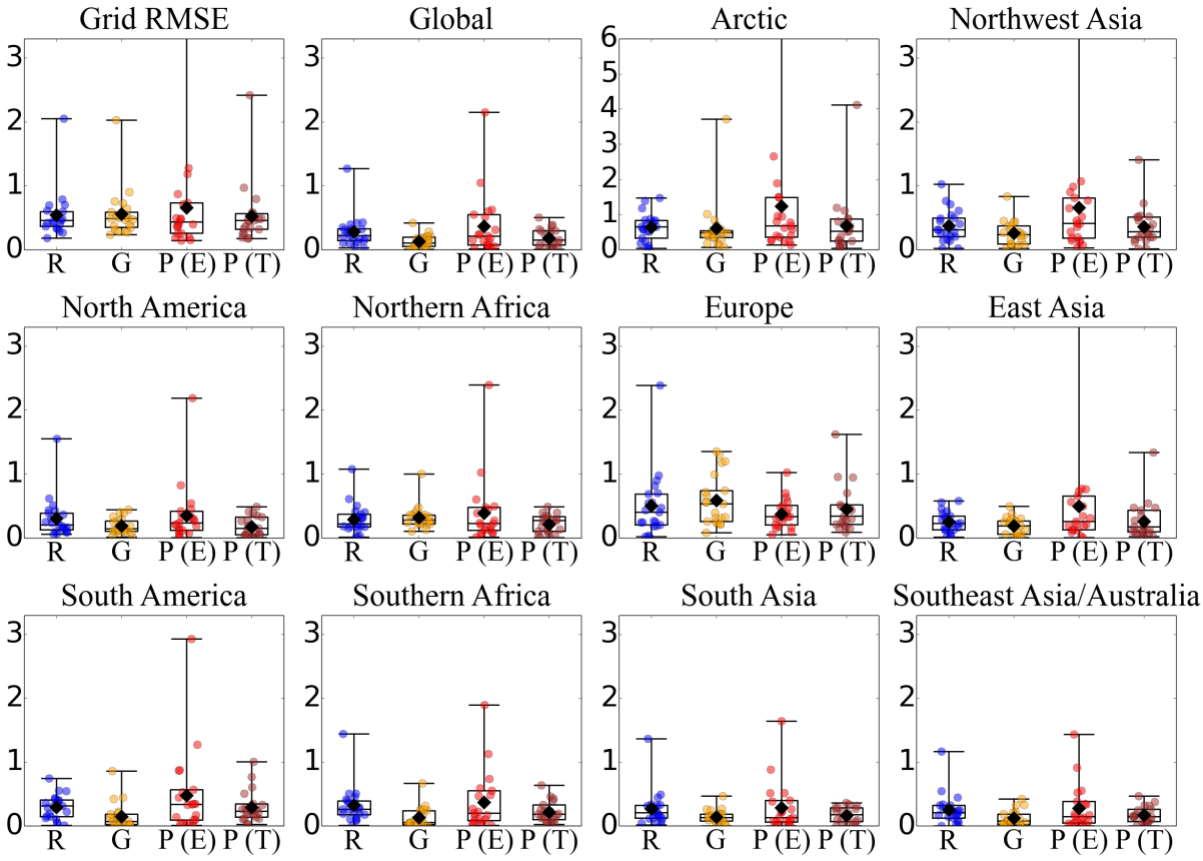
(2018) simulations, **c** Internal variability in short-term temperature response calculated as the standard deviation at each grid point across the short-term response time series in the control run from the Kasoar *et al.* (2018) simulations.



Supplementary Fig. 9: Absolute errors in °C for the key regions highlighted in Fig. 3 for Gaussian process regression where different approaches to dimension reduction are used. The first column shows no dimension reduction (the same as Fig. 3), the next three columns use principal component analysis (PCA), on the inputs, the outputs and both inputs and outputs respectively and the last two columns use regional dimension reduction on the inputs, the outputs and both inputs and outputs respectively.



Supplementary Fig. 10: **Absolute errors in °C for Gaussian process regression where different predictor variables are used as inputs** for all key regions highlighted in Fig. 3: ST=Surface Temperature, AT500=Air Temperature at 500hPa, GPH500=Geopotential Height at 500hPa, ERF=Effective Radiative Forcing, SLP=Sea Level Pressure. Note that when using SLP as a predictor for the response for the Arctic, two points exceed the axis maximum (5.0 and 5.6 °C).



Supplementary Fig. 11: Same as Supplementary Fig. 4 but using only 5 years as short-term response in the training and prediction process. Absolute error in °C for all scenarios in each selected region, for long-term climate response prediction using four methods: R=Ridge regression, G= Gaussian Process regression, P(E)=Pattern scaling using ERF as the scaler value P(T)=Pattern scaling using global mean short-term temperature response as the scaler value.

Supplementary References

1. Tebaldi, C. & Arblaster, J. M. Pattern scaling: Its strengths and limitations, and an update on the latest model simulations. *Clim. Change* **122**, 459–471 (2014).
2. Ishizaki, Y. *et al.* Temperature scaling pattern dependence on representative concentration pathway emission scenarios. *Clim. Change* **112**, 535–546 (2012).
3. Huntingford, C. & Cox, P. M. An analogue model to derive additional climate change scenarios from existing GCM simulations. *Clim. Dyn.* **16**, 575–586 (2000).
4. Gregory, J. M. *et al.* A new method for diagnosing radiative forcing and climate sensitivity. *Geophys. Res. Lett.* **31**, (2004).
5. Shindell, D. T., Voulgarakis, A., Faluvegi, G. & Milly, G. Precipitation response to regional radiative forcing. *Atmos. Chem. Phys.* **12**, 6969–6982 (2012).
6. Hansen, J. *et al.* Efficacy of climate forcings. *J. Geophys. Res. Atmos.* **110**, (2005).
7. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*. (Springer New York Inc., 2001).
8. Kasoar, M., Shawki, D. & Voulgarakis, A. Similar spatial patterns of global climate response to aerosols from different regions. *npj Clim. Atmos. Sci.* **1**, 12 (2018).
9. Meehl, G. A., Washington, W. M., Erickson III, D. J., Briegleb, B. P. & Jaumann, P. J. Climate change from increased CO2 and direct and indirect effects of sulfate aerosols. *Geophys. Res. Lett.* **23**, 3755–3758 (1996).
10. Pachauri, R. K. & Meyer, L. A. IPCC Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. *Intergov. Panel Clim. Chang.* (2014).
11. Hall, A., Cox, P., Huntingford, C. & Klein, S. Progressing emergent constraints on future climate change. *Nature Climate Change* (2019). doi:10.1038/s41558-019-0436-6

12. Nowack, P. J., Braesicke, P., Luke Abraham, N. & Pyle, J. A. On the role of ozone feedback in the ENSO amplitude response under global warming. *Geophys. Res. Lett.* (2017). doi:10.1002/2016GL072418

13. Fu, Q., Manabe, S. & Johanson, C. M. On the warming in the tropical upper troposphere: Models versus observations. *Geophys. Res. Lett.* (2011). doi:10.1029/2011GL048101

14. Anthropogenic and natural radiative forcing. in *Climate Change 2013 the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (2013). doi:10.1017/CBO9781107415324.018

15. Mitchell, T. D. Pattern Scaling: An Examination of the Accuracy of the Technique for Describing Future Climates. *Clim. Change* **60**, 217–242 (2003).

16. Kasoar, M. *et al.* Regional and global temperature response to anthropogenic SO2 emissions from China in three climate models. *Atmos. Chem. Phys.* (2016). doi:10.5194/acp-16-9785-2016

17. Myhre, G. *et al.* PDRMIP: A precipitation driver and response model intercomparison project-protocol and preliminary results. in *Bulletin of the American Meteorological Society* **98**, 1185–1198 (2017).

18. Liu, L. *et al.* A PDRMIP Multimodel study on the impacts of regional aerosol forcings on global and regional precipitation. *J. Clim.* (2018). doi:10.1175/JCLI-D-17-0439.1

19. Samset, B. H. *et al.* Fast and slow precipitation responses to individual climate forcers: A PDRMIP multimodel study. *Geophys. Res. Lett.* **43**, 2782–2791 (2016).

20. Stohl, A. *et al.* Evaluating the climate and air quality impacts of short-lived pollutants. *Atmos. Chem. Phys.* **15**, 10529–10566 (2015).

21. Aamaas, B., Berntsen, T. K., Fuglestvedt, J. S., Shine, K. P. & Collins, W. J. Regional

temperature change potentials for short-lived climate forcers basedon radiative forcing from multiple models. *Atmos. Chem. Phys.* **17**, 10795–10809 (2017).

22. Baker, L. H. *et al.* Climate responses to anthropogenic emissions of short-lived climate pollutants. *Atmos. Chem. Phys.* **15**, 8201–8216 (2015).

23. Shawki, D., Voulgarakis, A., Chakraborty, A., Kasoar, M. & Srinivasan, J. The South Asian Monsoon Response to Remote Aerosols: Global and Regional Mechanisms. *J. Geophys. Res. Atmos.* (2018). doi:10.1029/2018JD028623