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

Mansfield, L. A., Nowack, P. J., Kasoar, M., Everitt, R. G., Collins, W. J. ORCID: https://orcid.org/0000-0002-7419-0850 and Voulgarakis, A. (2020) Predicting global patterns of long-term climate change from short-term simulations using machine learning. npj Climate and Atmospheric Science, 3 (1). 44. ISSN 2397-3722 doi: 10.1038/s41612-020-00148-5 Available at https://centaur.reading.ac.uk/93016/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1038/s41612-020-00148-5

Publisher: Nature Publishing Group

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

ARTICLE   OPEN

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] and A. Voulgarakis [1,6,9]

Understanding and estimating regional climate change under different anthropogenic emission scenarios is pivotal for informing societal adaptation and mitigation measures. However, the high computational complexity of state-of-the-art climate models remains a central bottleneck in this endeavour. Here we introduce a machine learning approach, which utilises a unique dataset of existing climate model simulations to learn relationships between short-term and long-term temperature responses to different climate forcing scenarios. This approach not only has the potential to accelerate climate change projections by reducing the costs of scenario computations, but also helps uncover early indicators of modelled long-term climate responses, which is of relevance to climate change detection, predictability, and attribution. Our results highlight challenges and opportunities for data-driven climate modelling, especially concerning the incorporation of even larger model datasets in the future. We therefore encourage extensive data sharing among research institutes to build ever more powerful climate response emulators, and thus to enable faster climate change projections.

INTRODUCTION

To achieve long-term climate change mitigation and adaptation goals, such as limiting global warming to 1.5 or 2 °C, there must be a global effort to decide and act upon effective but realistic emission pathways[1]. This requires an understanding of the consequences of such pathways, which are often diverse and involve changes in multiple climate forcers[1–3]. In particular, different emission scenarios of, for example, greenhouse gases and aerosols are responsible for diverse changes in regional climate, which are not always well captured by a metric such as global temperature-change potential[4–9]. Exploring more detailed relationships between emissions and multiregional climate responses still requires the application of Global Climate Models (GCMs) that allow the behaviour of the climate to be simulated under various conditions (e.g. different atmospheric greenhouse gas and aerosol concentrations or emissions fields)[10–12] on decadal to multi-centennial timescales (e.g. refs. [5,13–16]). However, modelling climate at increasingly high spatial resolutions has significantly increased the computational complexity of GCMs[2], a tendency that has been accelerated by the incorporation and enhancement of a number of new Earth system model components and processes[17–20]. This high computational cost has driven us to investigate how machine learning methods can help accelerate estimates of global and regional climate change under different climate forcing scenarios.

Our work is further motivated by studies that have suggested links between characteristic short-term and long-term response patterns to different climate forcing agents[5,21,22]. Here, we seek a fast 'surrogate model'[23] to find a mapping from short-term to long-term response patterns within a given GCM (Fig. 1). Once learned, this surrogate model can be used to rapidly predict other

outputs (long-term responses) given new unseen inputs (short-term responses i.e. the results of easier to perform short-term simulations). While data science methods are increasingly used within climate science (e.g. refs. [24–30]), no study has attempted the application we present here, i.e. to predict the magnitude and patterns of long-term climate response to a wide range of global and regional forcing scenarios.

Building surrogate climate models

To train our learning algorithms, we take advantage of a unique set of GCM simulations performed in recent years using the Hadley Centre Global Environment Model 3 (HadGEM3). In these, step-wise perturbations were applied to various forcing agents to explore characteristic short- and long-term climate responses to them[5,7,8,14,16,31–34]. The set of simulations includes global perturbations of long-lived greenhouse gases such as carbon dioxide (CO_2) and methane (CH_4), as well as global and local perturbations to key short-lived pollutants such as sulfate (SO_4) and black carbon (BC) particles, amongst others (Supplementary Table 1). A key difference between these two types of perturbations is that long-lived forcers are homogeneously distributed in the atmosphere so that the region of emission is effectively inconsequential for the global temperature response pattern. In contrast, the response pattern does depend on the region of emission for short-lived forcers.

The evolution of the GCM's global mean temperature response to some example forcing scenarios is highlighted in Fig. 1a. All scenarios show an initial sudden response in the first few years, which we label the 'short-term response'. The global mean temperature then converges towards a new (approximately) equilibrated steady state, which we label the 'long-term response'.

[1]Department of Physics, Imperial College London, South Kensington Campus, London SW7 2BW, UK. [2]School of Mathematics and Statistics, University of Reading, Whiteknights, Berkshire RG6 6AX, UK. [3]Grantham Institute, Imperial College London, South Kensington Campus, London SW7 2AZ, UK. [4]Climatic Research Unit, Data Science Institute, Imperial College London, South Kensington Campus, London SW7 2AZ, UK. [5]School of Environmental Sciences, University of East Anglia, Norwich, Norfolk NR4 7TJ, UK. [6]Leverhulme Centre for Wildfires, Environment and Society, Department of Physics, Imperial College London, South Kensington Campus, London SW7 2BW, UK. [7]Department of Statistics, University of Warwick, Coventry CV4 7AL, UK. [8]Department of Meteorology, University of Reading, Whiteknights, Berkshire RG6 6ET, UK. [9]School of Environmental Engineering, Technical University of Crete, Chania, Crete 73100, Greece. ✉email: laura.mansfield@pgr.reading.ac.uk
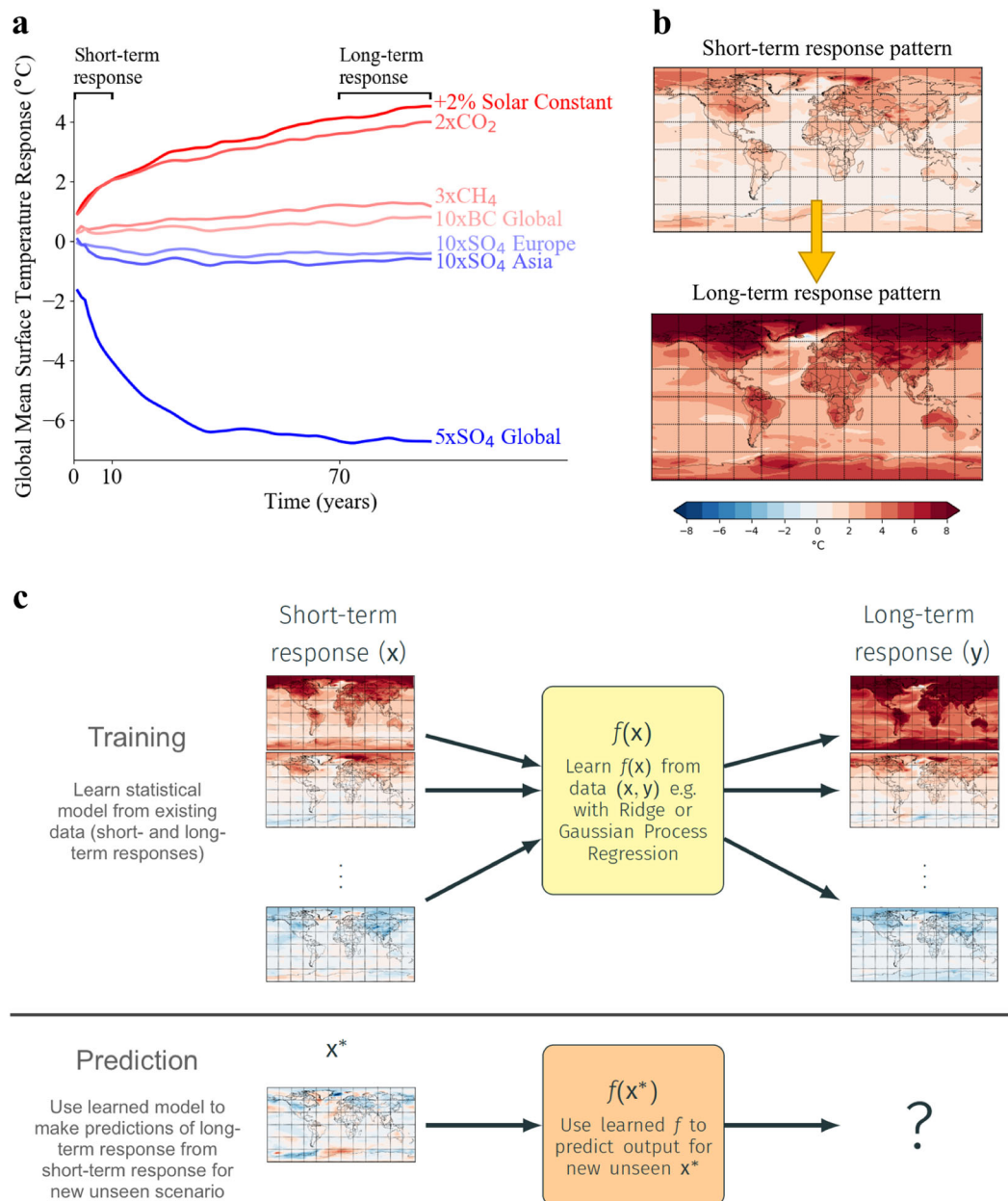
Fig. 1 Data-driven approach to learning relationships between short-term and long-term climate response patterns. a Global mean surface temperature response of a GCM (HadGEM3) to selected global and regional sudden step perturbations, e.g. to changes in long-lived greenhouse gases (CO_2, CH_4), the solar constant and short-lived aerosols (SO_4, BC). **b** Example of the short-term and long-term surface temperature response patterns for $2xCO_2$ scenario, defined as an average over the first 10 years and years 70–100, respectively. **c** Process diagram highlighting the training and prediction stages. In the training stage, a regression function is learned for pairs of short-term and long-term response maps, where the data are obtained from existing HadGEM3 simulations. In the prediction stage, the long-term response for a new unseen scenario is predicted by applying the already learned function to the short-term response to this new scenario, which is cheaper to obtain (here only 10 climate model years).

We are interested in not just the global mean response but, more importantly, in the global response patterns, such as the example shown in Fig. 1b for the $2xCO_2$ scenario.

In essence, GCMs map the initial state of the climate system and its boundary conditions, such as emission fields, to a state of the climate at a later time, using complicated functions representing the model physics, chemistry, and biology[17]. Our statistical model approximates the behaviour of the full GCM for a specific target climate variable of interest; here we choose surface temperature at each GCM grid cell, a central variable of interest in climate science and impact studies. This model is trained on simulations from the full global climate model (supervised learning[35]), in order

to predict the long-term surface temperature response of the GCM from the short-term temperature responses to perturbations (Fig. 1c). Then we can make effectively instantaneous predictions using results from new short-term simulations as input so that repeated long GCM runs can be avoided. Based on the available GCM data, we define the 'long-term' as the quasi-equilibrium response after removing the initial transient response (first 70 years) and averaging over the remaining years of the simulations, similarly to previous studies (see Methods)[5,14,36]. We define 'short-term' as the response over the first 10 years of each simulation.

The task is to learn the function $f(x)$ that maps these short-term responses (x) to the long-term responses (y) ('TRAINING' in Fig. 1c).

We use an independent regression model of the long-term response for each grid cell. Each one depends on the short-term response at all grid cells, so that predictions are not only based on local information but can also draw predictive capability from any changes in surface temperature worldwide. We present Ridge regression[37] and Gaussian Process Regression (GPR)[38] with a linear kernel (see Methods) as approaches for constructing this mapping. Then, the learned regression functions can be used to predict the long-term response for new, unseen inputs ($\mathbf{x}^*$), ('PREDICTION' in Fig. 1c). We choose Ridge regression and GPR, because these two methods handle well the limited sample size (number of simulations available) for training, which also limits how effectively the number of free parameters for other approaches such as deep learning, including convolutional neural networks, could be constrained. Future data collaborations, discussed below, could make the adaptation of our methodology to incorporate deep leaning an option. For the learning process, we use all but one of the available simulations at a time for training and cross-validation. The trained model is then used to make a temperature response prediction for the simulation that was left out each time. Finally, we assess the prediction skill of our machine learning models by comparing the predicted response maps $f(\mathbf{x}^*)$ to the results of the complex GCM simulations. This is repeated so that each simulation is predicted once based on the information learned from all other independent simulations (Methods).

RESULTS AND DISCUSSION

Overall method performance

We evaluate the performance of the two different machine learning methods (Ridge, GPR) by benchmarking them against a traditional pattern scaling approach[36,39], often used for estimating future patterns of climate change[40–42]. The latter relies on multiplying the long-term response pattern for the $2xCO_2$ scenario by the relative magnitude of global mean response for each individual climate forcer. This is approximated as the ratio of global mean effective radiative forcing (ERF) between the forcer and the $2xCO_2$ scenario (Methods)[36]. Alternative approaches are discussed in Methods and Supplementary.

We compare the predictions of long-term regional surface temperature changes with those produced by the complex GCM. From analysis at a grid-cell level, both Ridge regression and GPR capture some broad features that pattern scaling is also known to predict effectively, such as enhanced warming over the Northern Hemisphere, particularly over land, and Arctic amplification[43] (Supplementary Figs. 1 and 2). However, the key advantage of both machine learning methods is that they capture regional patterns and diversity in the response not predicted by pattern scaling. In particular, aerosol forcing scenarios show highly specific regional imprints on surface temperature due to the spatial heterogeneity of the emissions and their short lifetimes[4,7,33]. It is the ability to learn these patterns that gives data-driven methods the edge over any pattern scaling method for such predictions. The example in Fig. 2 shows the distribution of predicted temperature responses over all individual grid boxes for one short-lived and one long-lived forcing scenario. For the long-lived forcings all three types of model predictions produce a similar distribution of surface temperature responses to the GCM. However, for short-lived forcing scenarios, the range and variability of responses is highly underestimated in the case of pattern scaling. This is consistent across short-lived forcing scenario predictions (Supplementary Fig. 3) and exists because pattern scaling is constrained to the same pattern, regardless of the scaling factor used to estimate the global mean response (Methods, Supplementary Fig. 4).

In the following, we quantify how well the two machine learning models and pattern scaling perform on different spatial scales. At the grid-scale level, we calculate the Root Mean
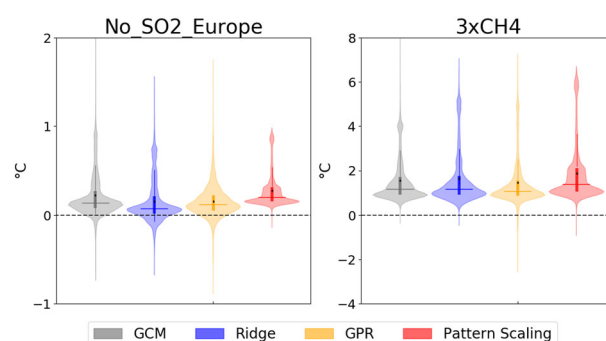

Fig. 2 Distribution of predicted grid-point scale surface temperature responses in °C for all methods for one short-lived forcing, No_SO2_Europe, and one long-lived forcing, 3xCH4. The central vertical boxes indicate the interquartile range shown on a standard box plot, the horizontal line shows the median and the black point shows the mean. The horizontal width shows the distribution of temperature values overall grid points, i.e. the wider regions highlight that more grid points have this value of predicted temperature response. Note the different vertical scales.

Squared Error (RMSE) by comparing the prediction and GCM response at every grid point (Methods). We highlight that grid-scale error metrics need to be interpreted with care because they can present misleading results, particularly for higher resolution models. For example, they penalize patterns that—as broad features—are predicted correctly but displaced marginally on the spatial grid[44]. This issue is necessarily more prevalent for the machine learning approaches where smaller scale patterns are more frequently predicted, while pattern scaling predicts more consistently smooth, cautious patterns with reduced spatial variability (Supplementary Fig. 1). This consideration is a key reason why predictions for larger scale domains are often selected in impact studies[11,12]. We therefore also compare the absolute errors in global mean temperature and in regional mean temperature over ten broad regions (Fig. 3); four of which are the main emission regions (North America, Europe, South Asia, and East Asia) and the remaining cover primarily land areas where responses affect the majority of the world's population. The boxplots in Fig. 3 show how these errors are distributed overall predicted scenarios for each regression method.

Both Ridge and GPR generally outperform the pattern scaling approach, but we find that, in most cases, it is GPR errors that are lowest. Note that scenario-specific pattern scaling errors are necessarily dependent on the approach chosen to scale the global CO_2-response pattern (Methods, Supplementary Fig. 4), but all pattern scaling approaches share their fundamental limitation in predicting spatial variability (Fig. 2). The large spread in absolute errors in Fig. 3 is due to the large spread in response magnitude for the different scenarios. Specifically, the large errors (e.g. 1–2 °C for the machine learning models and >3 °C for pattern scaling) come mostly from regions/scenarios with a large magnitude of response, which expectedly tend to be for strong forcings (e.g. strong solar or greenhouse gas forcings), but these errors can be small relative to the overall magnitude of scenario response. In contrast, small absolute errors can be large relative to the magnitude of response (Supplementary Fig. 5), making prediction more challenging for weakly forced scenarios. This is also consistent with the finding that regional aerosol perturbations, with typically weaker forcings, are more difficult to predict compared to long-lived pollutant perturbations (Fig. 2).

Learning early indicators

As well as advancing our predictability skills, the machine learning methods inform us about regions that experience the earliest indicators of long-term climate change in the GCM. By assessing the structure of learned Ridge regression coefficients,
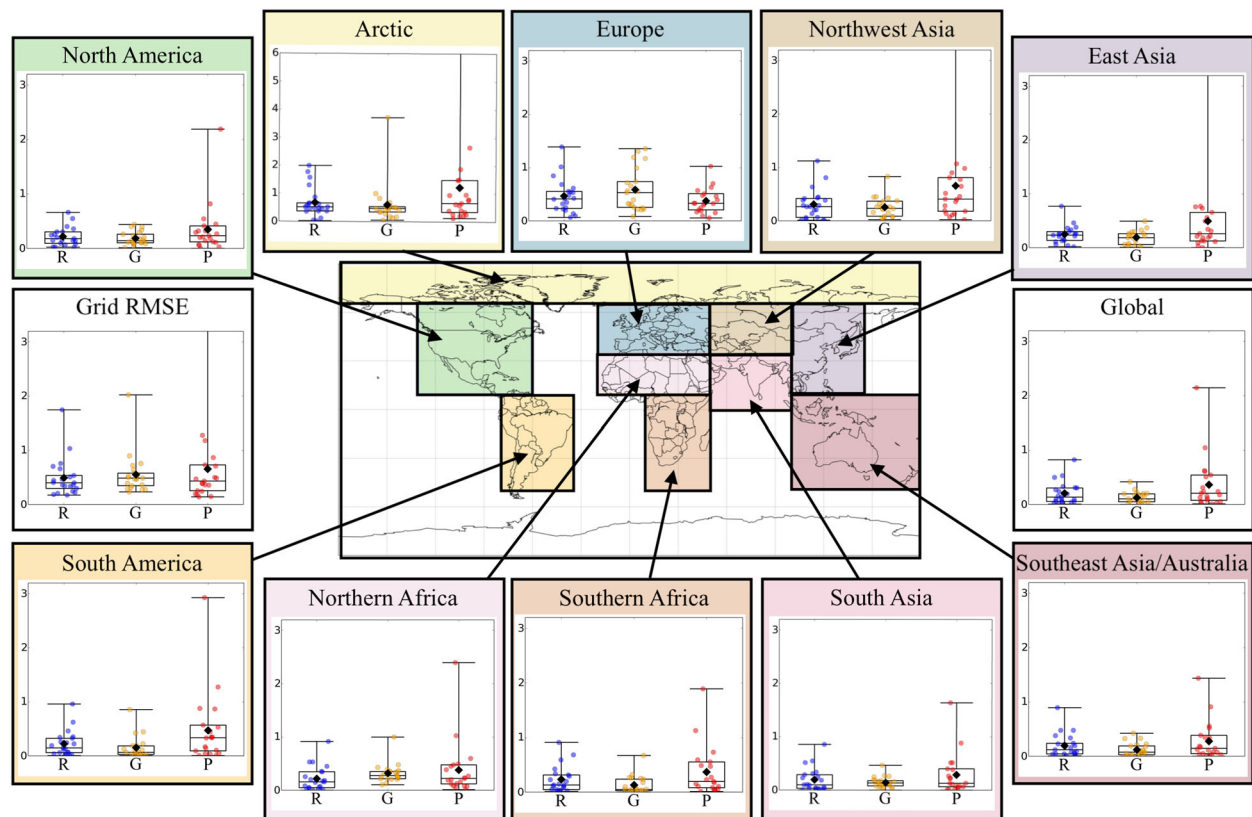
Fig. 3 Prediction skill comparison for entire globe and ten major world regions. RMSE at grid-cell level and global/regional absolute errors in °C for all scenarios, calculated by averaging the predicted response over each region and taking the difference between the GCM output and the prediction using three methods: R = Ridge regression, G = Gaussian Process Regression, and P = Pattern scaling. Boxplots show the distribution of errors across scenario predictions. Boxes show the interquartile range, whiskers show the extrema, lines show the medians and black diamonds show the mean. The dots indicate the errors for each individual scenario. Note the different scale for the Arctic and that points exceed the scale in Arctic (9.5), Northwest Asia (4.7), East Asia (3.7) and the Grid RMSE (3.8).

we find patterns in the short-term response that consistently indicate the long-term temperature response (Supplementary Fig. 6). In some regions (e.g. East Asia) the dominant coefficients appear in regions close to the predicted grid cell, whereas in other regions (e.g. Europe) predictions are strongly influenced by the short-term responses over relatively remote areas, such as sea-ice regions over the Arctic. This highlights the fact that climate model response predictability varies strongly depending on the region of interest, and often involves interactions with regions very far from the region of interest as well as from the emission region.

We also examine which areas are overall the most influential for long-term predictability, by averaging magnitude of coefficients across all grid cells to find a global mean coefficient map (Supplementary Fig. 6c, f). This coefficient map mimics warming patterns seen in previous studies (enhanced at high latitudes, over land and over the subtropics)[14] but also shows amplified coefficient weights in sea-ice regions, high-altitude regions, primary emission regions and mid-latitude jet stream regions. Arctic and high-altitude regions are known to warm more rapidly due to ice and snow albedo feedbacks[45] and faster upper tropospheric warming[11,46] respectively. These regions exhibit accelerated warming in the simulation compared to their surroundings, making them robust harbingers of long-term change within the model. We highlight the implications for future studies that attempt to interpret already observed warming patterns from a climate change perspective.

Data constraints and future directions

We identify more extensive training data (additional simulations and forcing scenarios) as key to further improving the skill of our machine learning methods. In Fig. 4 it is demonstrated that as the number of data training samples increases, the mean prediction accuracy significantly increases and becomes more consistent. We therefore expect significant potential for further improvements in predictions with even more training data. More simulations would better constrain parameters of the statistical models and improve the chances that a predicted scenario contains features previously seen by the statistical model (e.g. refs. [38,47], Methods).

Since obtaining training data from the GCM is expensive, sensible choices can also be made about how to increase the dataset by choosing which new scenarios will benefit the accuracy of the method the most, e.g. to address some complex regional aspects of the responses to short-lived pollutants. We recommend increasing the dataset to include more short-lived pollutant scenarios, noting that those with large forcings may reduce the noise in the training data so as to better constrain learned relationships (e.g. Supplementary Fig. 5). Some regions stand out as particularly challenging for our machine learning approaches, with Europe being a prominent example (Supplementary Fig. 2). This is partly due to large variations in the long-term response across the training data over Europe relative to other regions, which means predictions are less well constrained and would benefit more from increased training data. Additionally, the variability in the GCM-predicted temperature time series is generally larger over Europe compared to other regions in both
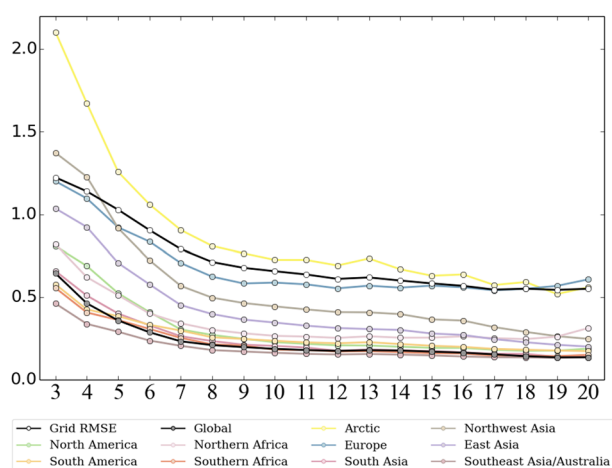
Fig. 4 Prediction skill for Gaussian Process Regression trained on an increasing number of simulations. Mean of absolute errors in °C across all predicted scenarios against number of training simulations, with each line representing a different region (Fig. 3). RMSE at the grid-scale level is also shown in black with white dots. 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. Full boxplots showing the distribution of errors across scenario predictions given these different combinations of training simulations can be found in Supplementary Fig. 7.

the control and perturbation simulations (Supplementary Fig. 8). This gives rise to a weaker signal-to-noise ratio for both short- and long-term responses in this region, increasing the difficulty of learning meaningful predictive relationships. It is also noteworthy that Ridge regression predictions for Europe depend strongly on remote parts of the Arctic where the short-term response is stronger but also highly variable (Supplementary Figs. 5 and 6). This points to the issue that internal variability can introduce noise to the inputs and outputs of the regression. This is partially addressed with multidecadal averages in the definitions of the short- and long-term responses, under the limitation that we have only a single realization of each simulation available. If, in future work, we have available an ensemble of simulations for each perturbation, an average over these would more effectively separate the internal variability from the response. The use of several diverse simulations in the training dataset also allows the noise in the inputs and outputs to be treated as random noise in the regression, which would be even better determined with increased training data.

A key challenge of working with the climate model information here is its high dimensionality (27,840 grid cells) given the small scenario sample size of 21 simulations. We note that we tried sensible approaches to dimension reduction for decreasing the number of points in both inputs and outputs, including physical dimension reduction by regional averaging, and statistical dimension reduction with principal component analysis (PCA)[47]. However, the resulting regressions generated larger prediction errors (Supplementary Fig. 9). Furthermore, we explored the use of different variables as the short-term predictors, such as air temperature at 500 hPa, geopotential height at 500 hPa (as an indicator of the large-scale dynamical responses), radiative forcing or sea level pressure. Surface temperature consistently outperforms other predictors, although a similar degree of accuracy is achieved with 500 hPa air temperature and geopotential height, suggesting the information encoded by these is similar (Supplementary Fig. 10). Throughout, we have selected the first 10 years of the GCM simulations as the inputs to our regression, but we find promising results for even shorter periods, e.g. the first 5 years

(Supplementary Fig. 11). Finally, we also tested other linear (e.g. LASSO[47]) and nonlinear (e.g. Random Forest) methods for the same learning task. However, these provided weaker results so that we focused our discussion on Ridge and GPR here. We have explored the use of these methods in the context of predicting temperature responses; however, we leave open the topic of predicting other variables such as precipitation, which we expect to be more challenging due to its spatial and temporal variability[48,49], but for which pattern scaling approaches are well-known to perform particularly poorly[36,41,43,50].

We also wish to highlight another long-term perspective in which the framework presented here could be useful. 'Emulators' that approximate model output given specific inputs, are a popular tool of choice for prediction, sensitivity analysis, uncertainty quantification and calibration and have great potential for climate prediction and impact studies[23,51–59]. However, long-term, spatially resolved climate prediction for diverse forcings has not yet been addressed due to the cost of training such emulators. A major implication of the approach presented here is that it can catalyse designing long-term climate emulators, by using a combination of the short-term/long-term relationships presented here and trained emulators of the short-term climate response to different forcings (i.e. multilevel emulation[52,59]). Training an emulator that predicts the spatial patterns of long-term response to a range of forcings would be an extremely challenging task, as it would require tens of simulations, all of them multidecadal in length, in order to train the emulator. Our method drastically accelerates this process by reducing the length of such simulations to be of the order of 5–10 years, with subsequent use of the relationships presented here for translating short-term responses to long-term responses.

Our study made use of existing simulations from a single global climate model. However, it opens the door for similar approaches to be taken with datasets from other individual climate models. The same GCMs are typically run by several different research centres across the world so that additional simulation data should be an effort of (inter)national collaboration. We therefore encourage widespread data sharing to test the limits of our approach as an important part of future research efforts in this direction. We hope that our work will catalyse developments for coordinated efforts in which carefully selected perturbation experiments will be performed in a multi-model framework. Increased availability of training datasets through model inter-comparison exercises, along with increasing access to powerful computing hardware can only help with this endeavour, leading to further advances in climate model emulation.

METHODS

Available simulations

To learn the regression models, we use data from long-term simulations from the Hadley Centre Global Environment Model 3 (HadGEM3) HadGEM3, a climate model developed by the UK Met Office[17]. HadGEM3 is a GCM for the atmosphere, land[18], ocean[19], and sea-ice[20]. In the configuration used here, the horizontal resolution is 1.875° by 1.25°, giving grid boxes ~140 km wide in the mid-latitudes[17]. The simulations were run in previous academic studies and model intercomparison projects, namely the Precipitation Driver and Response Model Intercomparison Project (PDRMIP)[16,31,32], Evaluating the Climate and Air Quality Impacts of Short-lived pollutants (ECLIPSE)[7,8,33] and Kasoar et al. (2018)[5,14,34]. There are 21 such simulations for a range of forcings, including long-lived green-house gas perturbations (e.g. carbon dioxide (CO_2), methane (CH_4), CFC-12), short-lived pollutant perturbations (e.g. sulfur dioxide emissions (SO_2, the precursor to sulfate aerosol (SO_4)), black carbon (BC), organic carbon (OC)) and a solar forcing perturbation. For the short-term pollutants, regional perturbations exist, to account for the influence of emission region to the response[4,60].

The long-lived greenhouse gas (CO_2, CH_4, CFC-12) simulations were performed by altering the atmospheric mixing ratios. The short-lived

pollutant experiments were performed by abruptly scaling present-day emission fields in simulations performed by ECLIPSE[7,8,33] and Kasoar et al. (2018)[5,14,34] or by scaling multi-model mean concentration fields in PDRMIP[16,31,32]. The solar forcing experiment was performed by changing the solar irradiance constant[31]. The GCM is run until it converges towards a new climate state, to reach an approximate equilibrium (70–100 years). The response is calculated by differencing this with its corresponding control simulation (independent control simulations were run for each project[5,7,8,14,16,31–34]). For the long-term response, we discard the transient response and average from year 70–100 for PDRMIP and Kasoar et al. (2018) to smooth out internal variability over the 30-year period[36]. For the 5 ECLIPSE simulations, we average from year 70 to year 80, since this is the full temporal extent of ECLIPSE simulations. For the short-term response, we average over the first 10 years of the simulation to reduce the influence of natural variability of the GCM[36].

The experiments from PDRMIP consist of simulations with a doubling of CO_2 concentration, tripling of CH_4 concentration, a 10× increase in CFC-12 concentration, a 2% increase in total solar irradiance, 5× increase in sulfate concentrations (SO_4), a 10× increase in black carbon (BC) concentrations, a 10× increase in SO_4 concentrations over Europe only, a 10× increase in SO_4 concentrations over Asia only, and a reduction to preindustrial SO_4 concentrations[16,31]. From ECLIPSE project simulations, we use a 20% reduction in CH_4 emissions, a doubling in CO_2 concentration, a 100% reduction in BC emissions, 100% reduction in SO_2 emissions, and a 100% reduction in carbon monoxide (CO) emissions[7,8,33]. The simulations performed by Kasoar et al. (2018) consist of a 100% reduction in SO_2 over the Northern Hemisphere mid-latitudes (NHML), a 100% reduction in BC over the NHML, a 100% reduction in SO_2 over China only, a 100% reduction in SO_2 over East Asia, a 100% reduction in SO_2 over Europe and a 100% reduction in SO_2 over US[5,14,34]. Additional simulations had also been performed by the groups, but we only consider simulations where the global mean response exceeds natural variability, calculated as the standard deviation among the control simulations. This is because we want to limit the noise in the small dataset we have. Scenarios that we did not use for this reason were the global removals of organic carbon, volatile organic compounds and nitrogen oxides (ECLIPSE[7,8,33]) and the removal of SO_2 over India (Kasoar et al. (2018)[5,14,34]).

Regression methods

We construct the mapping between short-term temperature response (x) and long-term temperature response (y) described in Fig. 1b using Ridge regression[37] and Gaussian Process Regression (GPR)[38]. These were found to be strongest from a range of machine learning methods tested, including Random Forest and Lasso.

Ridge regression

Given output variable y and input variable x, linear regression uses the mapping

$$y = \beta_0 + \sum_j \beta_j x_j \qquad (1)$$

where there are p predictors, indexed by $j = 1, \cdots, p$. The parameters to fit are the intercept, β_0, and the coefficients, β_j, associated with each predictor x_j. The method of least squares is used to fit the parameters by minimising the sum of the residual squared error for the training data pairs (x_i, y_i) for grid points $i = 1, \cdots, N$:

$$\sum_i \left[y_i - \left(\beta_0 + \sum_j \beta_j x_{ij} \right) \right]^2 \qquad (2)$$

When the number of samples exactly equals the number of parameters, $N = p + 1$, this can be minimised to give a unique solution. When $N > p + 1$ the parameters are overdetermined and this is an optimisation problem in β_j. In contrast, when $N < p + 1$, there are more free parameters, β_j, than there are observed data points to constrain them[47]. There are many possible values of β_j that satisfy (2) equal to zero, making this an underdetermined problem. Our problem falls under this regime since we have many predictors (one for each grid point, i.e. $p = 27,840$) but few training simulations ($N = 20$). This is why we introduce a regularisation constraint which penalises large values of β_j. Thus, we minimise[47,61]:

$$\sum_i \left\{ \left[y_i - \left(\beta_0 + \sum_j \beta_j x_{ij} \right) \right]^2 + \lambda \sum_j |\beta_j|^2 \right\} \qquad (3)$$

The last term shrinks many of the β_j coefficients close to zero, so that the remaining large coefficients can be viewed as stronger predictors of y. This introduces a bias but lowers the variance[5]. The regularisation parameter λ controls the amount of shrinkage and is chosen through cross-validation, described below. Once β_0 and β_j have been learned, we can use (1) to make predictions. We carried out the regression with and without inputs x normalised to zero mean and unit variance with very little difference in results. We use Python package scikit-learn to implement Ridge regression and cross-validation[62].

Cross-validation

Cross-validation is used here to estimate the best value of λ for prediction based on the available training data. First, we split the training dataset (of size N) into a chosen number of subsets of size N_{CV}. We use three subsets so N_{CV} is around 6–7. Then, we iterate through a list of possible values of λ, and for each one, the following steps are taken.

(1) Set λ from list.

    (a) Set aside one of the smaller datasets as the validation data (size N_{CV}).

    (b) Train the regression model with the remaining data ($N - N_{CV}$) by minimising (3).

    (c) Use the inputs of the validation dataset on the trained model to make predictions on the outputs using (1) and call this $\mathbf{y}_*$.

    (d) Compare these predictions with the true outputs of the validation dataset using an error metric such as root-mean-squared error (RMSE), accounting for all grid cells $i = 1, \ldots, p$ and weighting by the grid-cell area, w_i,

$$RMSE_{cv,\lambda} = \left(\sum_i^p w_i |y_i^{*2} - y_i^2| \right)^{1/2} \qquad (4)$$

    (e) Repeat steps a-d for other subsets of validation data (we use 3 in total).

(2) Calculate the cross-validation score as the mean RMSE for this value of λ for all three subsets.

$$RMSE_\lambda = \sum_{cv}^3 RMSE_{cv,\lambda} \qquad (5)$$

This process is repeated for all values of λ in the list. The value of λ that produces the lowest $RMSE_\lambda$ is selected as the parameter for use in the final stage of training of the model, where all training data is used.

Gaussian Process Regression

Rather than learning the parameters β_0 and β_j, Gaussian Process Regression is a non-parametric approach, where we seek a distribution over possible functions that fit the data. This is done from a Bayesian perspective, where we define a prior distribution over the possible functions. Then after observing the data, we use Bayes' theorem to obtain a posterior distribution over possible functions. The prior distribution is a Gaussian process,

$$y \sim GP(\mu_0(x), C_0(x, x')) \qquad (6)$$

where μ_0 is the prior mean function, which we assume to be linear with slope β, $\mu_0(x) = \beta x$, and $C_0(x, x')$ is the prior covariance function, which describes the covariance between two points, x and x'[38]. We choose the following squared exponential covariance function,

$$C(x, x') = \sigma^2 \exp\left(-\frac{|x - x'|^2}{2l^2} \right) \qquad (7)$$

where σ^2 and l are the output variance and lengthscale, respectively, which reflect the sensitivity of the outputs to changes in inputs[38].

The prior Gaussian process is combined with the data using Bayes' Theorem to obtain a posterior distribution over functions. This is another Gaussian process, with an updated mean function, $\mu^*(x)$, and covariance function, $C^*(x, x')$,

$$y \sim GP(\mu^*(x), C^*(x, x')) \qquad (8)$$

The details can be found in relevant textbooks[38]. Predictions of the output can then be made at unseen values of x, where the Gaussian process provides both an expected value and the variance around this value. Since

the prediction is effectively built on correlations between the new inputs and the training data inputs, this variance will be lower for predictions at values of x that are closer to values already seen in training data. We follow these steps with the framework provided by *GPy* in Python. The values of β, σ^2, and l are learned through optimisation (the L-BGFS optimiser) in *GPy*[63].

Pattern scaling

We benchmark our machine learning models against pattern scaling, a traditional method for obtaining spatial response patterns to forcings without running a full GCM[36,39]. It has been widely used for conducting regional climate change projections[40–42] in impact studies[64] and to extend simplified models to predict spatial outputs[58,65]. Pattern scaling requires one previous GCM run to obtain the long-term response of the variable of interest for a reference scenario. Typically, a strong greenhouse gas perturbation, such as a doubling of CO_2 is used as this reference response pattern on the longitude-latitude grid, $V_{ref}(lat, lon)$. We use the 2xCO_2 scenario from PDRMIP (since more than half of the simulations are from PDRMIP we expect this to be a more valid reference pattern than the 2xCO_2 ECLIPSE scenario)[16,31,32]. Then, the variable of interest is estimated at each grid point for a new scenario, $V^*(lat, lon)$ by multiplying the reference pattern by scaler value s, i.e.

$$V^*(lat, lon) = s \times V_{ref}(lat, lon) \tag{9}$$

The scaler value s is the ratio of long-term global mean temperature response between the prediction and reference scenario. This can be derived from either a simplified climate model, such as a global energy balance model[43,66]; a statistical model[58]; or a mathematical relationship, such as the assumed linear relationship between long-term temperature response and effective radiative forcing (ERF)[64,67]. We take the latter approach due to the availability of variables required to calculate ERF for the relevant perturbations studied here.

ERF is defined as the energy imbalance between the surface and the top of the atmosphere in a GCM run in which the atmosphere is allowed to respond, while sea-surface temperatures are kept fixed (i.e. no ocean coupling)[1,5,8,33]. These simulations were run for 5 years in previous studies[5,7,8,14,16,31–34] and therefore we average over the first 5 years of the simulations to reduce noise in the estimate of global mean ERFs.

Pattern scaling is generally considered as a fair approximation[36,43,66] but it assumes that the magnitude of the response scales linearly with the amount of radiative forcing, which is not necessarily true, particularly for climate forcings of a different type to the reference scenario[36]. Furthermore, it cannot necessarily predict the highly inhomogeneous effects of certain types of climate forcings such as from aerosol emissions.

There are alternative approaches for obtaining a sensible scaler value s such as using the ratio of short-term temperature response between the predicted and reference scenarios (see Supplementary Fig. 4). We note that such a method can sometimes achieve a higher performance in predicting the mean response in some regions than our machine learning approach. However, it suffers the same limitations as the method presented here, in that the spatial variability in the response is not captured, particularly for short-lived pollutants (Supplementary Fig. 3). This limitation will be true regardless of the choice of scaler value, since the spatial variability is fixed based on the reference pattern.

Prediction errors

We predict long-term climate response, y^* for each scenario following the three methods described above. We calculate the Root Mean Squared Error (RMSE) at the grid-cell level with

$$RMSE = \left(\sum_i^p w_i |y_i^{*2} - y_i^2| \right)^{1/2} \tag{10}$$

where subscript $i = 1, \ldots, p$ indexes the grid cell and w_i is the normalised weight of grid cell i. We note that measuring errors at these scales can introduce unintended biases in the evaluation of our methods. For example, even small spatial offsets in climate response patterns can lead to large, nonphysical quantitative errors[44]. We also show the absolute error in mean response over ten world regions that cover a broader spatial scale (Fig. 3). These are the four main emission regions; North America, Europe, South Asia and East Asia, as defined in the Hemispheric Transport of Air Pollution experiments[68]; and six remaining regions; the Arctic, Northwest Asia, Northern Africa, Southern Africa, South America and Australia. These cover the land regions where climate responses are of

interest due to societal relevance. Here we defined the prediction error as the absolute difference between the predicted response in each region, y_r^*, and the response from the complex GCM in the same region, y_r:

$$E_{abs} = \left| y_r^* - y_r \right| \tag{11}$$

where subscript r indicates the mean response overall grid boxes in that region, weighted by the grid box area. We also calculate the absolute error for the global mean response in the same way. These RMSE, regional and global error metrics are presented in Fig. 3 for all prediction methods.

REFERENCES

1. *IPCC Climate Change 2014: Synthesis Report* (eds. Pachauri, R. K. & Meyer, L. A.) (Cambridge Univ. Press, 2015).
2. Collins, M. et al. Quantifying future climate change. *Nat. Clim. Change* **2**, 403 EP (2012).
3. Rogelj, J., Mccollum, D. L., O'Neill, B. C. & Riahi, K. 2020 emissions levels required to limit warming to below 2 °C. *Nat. Clim. Change* **3**, 405–412 (2013).
4. Shindell, D. & Faluvegi, G. Climate response to regional radiative forcing during the twentieth century. *Nat. Geosci.* **2**, 294 EP (2009).
5. 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).
6. Shine, K. P., Fuglestvedt, J. S., Hailemariam, K. & Stuber, N. Alternatives to the Global Warming Potential for comparing climate impacts of emissions of greenhouse gases. *Clim. Change* **68**, 281–302, (2005).
7. Baker, L. H. et al. Climate responses to anthropogenic emissions of short-lived climate pollutants. *Atmos. Chem. Phys.* **15**, 8201–8216 (2015).
8. Aamaas, B., Berntsen, T. K., Fuglestvedt, J. S., Shine, K. P. & Collins, W. J. Regional temperature change potentials for short-lived climate forcers based on radiative forcing from multiple models. *Atmos. Chem. Phys.* **17**, 10795–10809 (2017).
9. Collins, W. J. et al. Global and regional temperature-change potentials for near-term climate forcers. *Atmos. Chem. Phys.* **13**, 2471–2485 (2013).
10. Bitz, C. M. & Polvani, L. M. Antarctic climate response to stratospheric ozone depletion in a fine resolution ocean climate model. *Geophys. Res. Lett.* **39**, L20705 (2012).
11. 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.* **44**, 3858–3866 (2017).
12. Hartmann, D. L., Blossey, P. N. & Dygert, B. D. Convection and climate: what have we learned from simple models and simplified settings? *Curr. Clim. Chang. Rep.* **5**, 196–206 (2019).
13. Persad, G. G. & Caldeira, K. Divergent global-scale temperature effects from identical aerosols emitted in different regions. *Nat. Commun.* **9**, 3289 (2018).
14. 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.* **123**, 11585–11601 (2018).
15. Conley, A. J. et al. Multimodel surface temperature responses to removal of U.S. sulfur dioxide emissions. *J. Geophys. Res. Atmos.* **123**, 2773–2796 (2018).
16. Liu, L. et al. A PDRMIP Multimodel study on the impacts of regional aerosol forcings on global and regional precipitation. *J. Clim.* **31**, 4429–4447 (2018).
17. Williams, K. D. et al. The Met Office Global Coupled Model 3.0 and 3.1 (GC3.0 and GC3.1) Configurations. *J. Adv. Model. Earth Syst.* **10**, 357–380 (2018).
18. Walters, D. et al. The met Office Unified Model Global Atmosphere 7.0/7.1 and JULES Global Land 7.0 configurations. *Geosci. Model Dev.* **12**, 1909–1963 (2019).

19. Storkey, D. et al. UK Global Ocean GO6 and GO7: a traceable hierarchy of model resolutions. *Geosci. Model Dev.* **11**, 3187–3213 (2018).

20. Ridley, J. K. et al. The sea ice model component of HadGEM3-GC3.1. *Geosci. Model Dev.* **11**, 713–723 (2018).

21. Ceppi, P., Zappa, G., Shepherd, T. G. & Gregory, J. M. Fast and slow components of the extratropical atmospheric circulation response to CO_2 forcing. *J. Clim.* **31**, 1091–1105 (2017).

22. Persad, G. G., Ming, Y., Shen, Z. & Ramaswamy, V. Spatially similar surface energy flux perturbations due to greenhouse gases and aerosols. *Nat. Commun.* **9**, 3247 (2018).

23. Ryan, E., Wild, O., Voulgarakis, A. & Lee, L. Fast sensitivity analysis methods for computationally expensive models with multi-dimensional output. *Geosci. Model Dev.* **11**, 3131–3146 (2018).

24. Bracco, A., Falasca, F., Nenes, A., Fountalis, I. & Dovrolis, C. Advancing climate science with knowledge-discovery through data mining. *npj Clim. Atmos. Sci.* **1**, 20174 (2018).

25. Kretschmer, M., Runge, J. & Coumou, D. Early prediction of extreme stratospheric polar vortex states based on causal precursors. *Geophys. Res. Lett.* **44**, 8592–8600 (2017).

26. Nowack, P. et al. Using machine learning to build temperature-based ozone parameterizations for climate sensitivity simulations. *Environ. Res. Lett.* **13**, 104016 (2018).

27. Sippel, S. et al. Uncovering the forced climate response from a single ensemble member using statistical learning. *J. Clim.* **32**, 5677–5699 (2019).

28. Runge, J. et al. Inferring causation from time series in Earth system sciences. *Nat. Commun.* **10**, 2553 (2019).

29. Knüsel, B. et al. Applying big data beyond small problems in climate research. *Nat. Clim. Change* **9**, 196–202 (2019).

30. Nowack, P., Runge, J., Eyring, V. & Haigh, J. D. Causal networks for climate model evaluation and constrained projections. *Nat. Commun.* **11**, 1415 (2020).

31. Myhre, G. et al. PDRMIP: a precipitation driver and response model intercomparison project-protocol and preliminary results. *Bull. Am. Meteorol. Soc.* **98**, 1185–1198 (2017).

32. 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).

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

34. Kasoar, M. et al. Regional and global temperature response to anthropogenic SO_2 emissions from China in three climate models. *Atmos. Chem. Phys.* **16**, 9785–9804 (2016).

35. Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag, 2006).

36. Mitchell, T. D. Pattern scaling: an examination of the accuracy of the technique for describing future climates. *Clim. Change* **60**, 217–242 (2003).

37. Hoerl, A. E. & Kennard, R. W. Ridge regression: applications to nonorthogonal problems. *Technometrics* **12**, 69–82 (1970).

38. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning.* Cambridge MA: MIT Press (2006).

39. Santer, B. D., Wigley, T. M. L., Schlesinger, M. E. & Mitchell, J. F. B. *Developing Climate Scenarios from Equilibrium GCM Results* (1990) Max Planck Institut für Meteorologie, Report 47, Hamburg.

40. Hulme, M., Raper, S. C. B. & Wigley, T. M. L. An integrated framework to address climate change (ESCAPE) and further developments of the global and regional climate modules (MAGICC). *Energy Policy* **23**, 347–355 (1995).

41. Murphy, J. M. et al. A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **365**, 1993–2028 (2007).

42. Watterson, I. G. Calculation of probability density functions for temperature and precipitation change under global warming. *J. Geophys. Res. Atmos* **113**, D12106 (2008).

43. 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).

44. Rougier, J. Ensemble averaging and mean squared error. *J. Clim.* **29**, 8865–8870 (2016).

45. Hall, A., Cox, P., Huntingford, C. & Klein, S. Progressing emergent constraints on future climate change. *Nat. Clim. Change* **9**, 269–278 (2019)

46. Fu, Q., Manabe, S. & Johanson, C. M. On the warming in the tropical upper troposphere: models versus observations. *Geophys. Res. Lett.* **38**, L15704 (2011).

47. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer New York Inc., 2001).

48. Pendergrass, A. G., Knutti, R., Lehner, F., Deser, C. & Sanderson, B. M. Precipitation variability increases in a warmer climate. *Sci. Rep.* **7**, 17966 (2017).

49. Pendergrass, A. G. & Knutti, R. The uneven nature of daily precipitation and its change. *Geophys. Res. Lett.* **45**, 11980–11988 (2018).

50. Pendergrass, A. G., Lehner, F., Sanderson, B. M. & Xu, Y. Does extreme precipitation intensity depend on the emissions scenario? *Geophys. Res. Lett.* **42**, 8767–8774 (2015).

51. Williamson, D., Blaker, A. T., Hampton, C. & Salter, J. Identifying and removing structural biases in climate models with history matching. *Clim. Dyn.* **45**, 1299–1324 (2015).

52. Cumming, J. & Goldstein, M. Bayes linear uncertainty analysis for oil reservoirs based on multiscale computer experiments. In *Oxford Handbook of Applied Bayesian Analysis*, Oxford: Oxford University Press, (eds. O'Hagan, A., & West, M.) pp. 241–270 (2010).

53. McNeall, D. J., Challenor, P. G., Gattiker, J. R. & Stone, E. J. The potential of an observational data set for calibration of a computationally expensive computer model. *Geosci. Model Dev. Discuss.* **6**, 1715–1728 (2013).

54. Salter, J. M. & Williamson, D. A comparison of statistical emulation methodologies for multi-wave calibration of environmental models. *Environmetrics* **27**, 507–523 (2016).

55. Rougier, J., Sexton, D. M. H., Murphy, J. M. & Stainforth, D. Analyzing the climate sensitivity of the HadSM3 climate model using ensembles from different but related experiments. *J. Clim.* **22**, 3540–3557 (2009).

56. Lee, L. A., Reddington, C. L. & Carslaw, K. S. On the relationship between aerosol model uncertainty and radiative forcing uncertainty. *Proc. Natl Acad. Sci. USA.* **113**, 5820–5827 (2016).

57. Edwards, T. L. et al. Revisiting Antarctic ice loss due to marine ice-cliff instability. *Nature* **566**, 58–64 (2019).

58. Castruccio, S. et al. Statistical emulation of climate model projections based on precomputed GCM runs. *J. Clim.* **27**, 1829–1844 (2014).

59. Tran, G. T. et al. Building a traceable climate model hierarchy with multi-level emulators. *Adv. Stat. Climatol. Meteorol. Oceanogr.* **2**, 17–37 (2016).

60. Shindell, D. T., Voulgarakis, A., Faluvegi, G. & Milly, G. Precipitation response to regional radiative forcing. *Atmos. Chem. Phys.* **12**, 6969–6982 (2012).

61. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **58**(1), 267–288 (1996).

62. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

63. GPy. *GPy: A gaussian process framework in python.* http://github.com/SheffieldML/GPy (2014).

64. Huntingford, C. & Cox, P. M. An analogue model to derive additional climate change scenarios from existing GCM simulations. *Clim. Dyn.* **16**, 575–586 (2000).

65. Harris, G. R. et al. Frequency distributions of transient regional climate change from perturbed physics ensembles of general circulation model simulations. *Clim. Dyn.* **27**, 357–375 (2006).

66. Ishizaki, Y. et al. Temperature scaling pattern dependence on representative concentration pathway emission scenarios. *Clim. Change* **112**, 535–546 (2012).

67. Gregory, J. M. et al. A new method for diagnosing radiative forcing and climate sensitivity. *Geophys. Res. Lett.* **31**, L03205 (2004).

68. Sanderson, M. G. et al. A multi-model study of the hemispheric transport and deposition of oxidised nitrogen. *Geophys. Res. Lett.* **35**, L17815 (2008).

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS

This work was initiated by A.V. L.A.M. carried out the analyses and wrote the manuscript. A.V. and P.J.N. supervised and contributed to writing. P.J.N. and R.G.E. advised on statistical methods. M.K. and B.C. performed the simulations.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41612-020-00148-5.

Correspondence and requests for materials should be addressed to L.A.M.

Reprints and permission information is available at http://www.nature.com/reprints