**ClimateBench: A benchmark dataset for data-driven climate projections**

D. Watson-Parris[1], Y. Rao[2], D. Olivié[3], Ø. Seland[3], P. Nowack[4], G. Camps-Valls[5], P. Stier[1], S. Bouabid[6], M. Dewey[7], E. Fons[8], J. Gonzalez[9], P. Harder[1,10], K. Jeggle[8], J. Lenhardt[9], P. Manshausen[1], M. Novitasari[11], L. Ricard[12], C. Roesch[13]

[1]Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, Oxford, UK

[2]North Carolina Institute for Climate Studies, North Carolina State University, Asheville, NC 28801, USA

[3]Norwegian Meteorological Institute, Oslo, Norway

[4]Climatic Research Unit, School of Environmental Sciences, Norwich, UK

[5] Image Processing Laboratory, Universitat de València, València, Spain

[6]Department of Statistics, University of Oxford, Oxford, UK

[7]Department of Meteorology, Stockholm University, Stockholm, Sweden

[8]Institute of Atmospheric and Climate Science, ETH Zurich, 8092 Zurich, Switzerland

[9]Institute for Meteorology, Universität Leipzig, Leipzig, Germany

[10]Fraunhofer ITWM, Kaiserslautern, Germany

[11] Department of Electronic and Electrical Engineering, University College London, London, UK

[12]Laboratory of Atmospheric Processes and their Impacts, School of Architecture, Civil & Environmental Engineering, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[13] School of Geosciences, University of Edinburgh, Edinburgh, UK

*Correspondence to*: Duncan Watson-Parris (duncan.watson-parris@physics.ox.ac.uk)

**Key Points:**

- We introduce a benchmark dataset for emulation of key spatially resolved climate variables derived from a full complexity Earth System Model
- Three baseline emulators are presented which are able to predict regional temperature and precipitation with varying skill
- Evaluation metrics and areas for future research are presented to encourage further development of trustworthy data-driven climate emulators

**Abstract**

Many different emission pathways exist that are compatible with the Paris climate agreement, and many more are possible that miss that target. While some of the most complex Earth System Models have simulated a small selection of Shared Socioeconomic Pathways, it is impractical to use these expensive models to fully explore the space of possibilities. Such explorations therefore mostly rely on one-dimensional impulse response models, or simple pattern scaling approaches to approximate the physical climate response to a given scenario. Here we present ClimateBench - a benchmarking framework based on a suite of CMIP, AerChemMIP and DAMIP simulations performed by a full complexity Earth System Model, and a set of baseline machine learning models that emulate its response to a variety of forcers. These emulators can predict annual mean global distributions of temperature, diurnal temperature range and precipitation (including extreme precipitation) given a wide range of emissions and concentrations of carbon dioxide, methane and aerosols, allowing them to efficiently probe previously unexplored scenarios. We discuss the accuracy and interpretability of these emulators and consider their robustness to physical constraints such as total energy conservation. Future opportunities incorporating such physical constraints directly in the machine learning models and using the emulators for detection and attribution studies are also discussed. This opens a wide range of opportunities to improve prediction, robustness and mathematical tractability. We hope that by laying out the principles of climate model emulation with clear examples and metrics we encourage engagement from statisticians and machine learning specialists keen to tackle this important and demanding challenge.

**Plain Language Summary**

Many different emission pathways exist that are compatible with the Paris climate agreement, and many more are possible that miss that target. While some of the most complex Earth System Models have simulated a small selection of possible futures, it is impractical to use these expensive models to fully explore the space of possibilities. Such explorations therefore mostly rely simple approximations of the global mean temperature response to a given scenario. Here we present ClimateBench - a benchmarking framework based on a suite of state-of-the-art simulations performed by a full complexity Earth System Model, and a set of baseline machine learning models that emulate its response to a variety of forcers. These emulators can predict annual mean global distributions of temperature, diurnal temperature range and precipitation (including extreme precipitation) given a wide range of emissions and concentrations of carbon dioxide, methane and aerosols, allowing them to efficiently probe previously unexplored scenarios. We also describe a set of evaluation metrics which we hope will entice statisticians and machine learning experts to tackle this important and demanding challenge.

# Introduction

Many different emission pathways exist that are compatible with the Paris Agreement of limiting global mean temperatures to "well below 2°C above pre-industrial levels and pursuing efforts to limit the temperature increase to 1.5 °C", and many more are possible that miss that target. Sampling possible emissions scenarios is therefore crucial for policy makers to weigh the economic cost and societal impact of different mitigation and adaptation strategies. While many of the most complex Earth System Models (ESMs) have simulated a small selection of 'Shared Socioeconomic Pathways' (SSPs; self-consistent emissions scenarios based on assumptions about future socio-economic changes and imperatives) it is impractical to use these expensive models to fully explore the space of possibilities (O'Neill et al., 2016). Therefore, such explorations mostly rely on one-dimensional impulse response models, or simple pattern scaling approaches to approximate the physical climate response to a given scenario (e.g., Millar et al., 2017).

Impulse response models (Smith et al., 2018; Meinshausen et al., 2011; Nicholls et al., 2020) are physically interpretable and can capture the general non-linear behaviour of the system, but are inherently unable to model regional climate changes, while pattern scaling approaches rely on a simple scaling of spatial distributions of temperature (e.g., Alexeeff et al., 2018) by global mean temperature changes. This approach breaks down when considering precipitation, however, because of the strong non-linearities in its response to temperature (e.g., Cabré et al., 2010). Statistical emulators of the regional climate have been developed although these have been quite bespoke (Castruccio et al., 2014) or focus on the relatively simple problem of emulating temperature (Holden and Edwards, 2010). These approaches also do not account for the influence of aerosol, which can be important for both regional temperature and precipitation (e.g. Kasoar et al. 2018, Wilcox et al. 2020). As has been noted recently (Watson-Parris, 2021), approaches including non-linear pattern scaling (Beusch et al., 2020) and Gaussian process (GP) regression of long-term climate responses (Mansfield et al., 2020) suggest the possibility of using modern machine learning (ML) tools to produce robust and general emulators of future scenarios. However, comparing and contrasting these approaches is currently hindered by the lack of a consistent benchmark.

ClimateBench defines a set of criteria and metrics for objectively evaluating such climate model emulation; aims to demonstrate the feasibility of such emulators; and provides a curated dataset that will allow, and hopefully encourage, broader engagement with this challenge in the same way WeatherBench (Rasp et al., 2020) has achieved for weather modeling. The target is to predict annual mean global distributions of temperature (T), diurnal temperature range (DTR), precipitation (PR) and the $90^{\text{th}}$ percentile of precipitation (PR90). These variables are chosen to represent a range of important climate variables which respond differently to each forcing and include extreme changes (PR90) that might not be expected to scale in the same way as the mean. For example, while T has

been shown to scale roughly linearly with global mean temperature changes (Castruccio et al., 2014), PR responds non-linearly, and DTR is more sensitive to aerosol perturbations than global mean temperature changes (Hansen et al., 1995). Four of the main anthropogenic forcing agents are provided as emulator inputs (predictors): carbon dioxide (CO2), sulphur dioxide (SO2; a precursor to sulfate aerosol), black carbon (BC) and methane (CH4). To enable spatially accurate emulators ClimateBench includes (annual mean): spatial distributions of emissions for the short-lived aerosol species (SO2 and BC), globally averaged emissions of CH4, and global total concentrations of CO2.

The training data which is provided in order to support such predictions is generated from the simulations performed by the second (and latest) version of the Norwegian Earth System Model (NorESM2; Seland et al., 2020) as part of the sixth coupled model intercomparison project (CMIP6; Eyring et al., 2016). The provided inputs are constructed from the same input data that is used to drive the original simulations. While we could have included simulations from multiple different models, only one model submitted all of the DECK (Diagnostic, Evaluation, and Characterization of Klima), historical, AerChemMIP (Collins et al., 2017) and ScenarioMIP (O'Neill et al., 2016) experiments required for our purposes, making it impossible to provide a harmonised dataset. Further, there is no agreed way of robustly combining multiple models, and while statistically combining multiple different models can lead to improved skill (Pincus et al., 2008) the resulting variance is not reliable since the models are not truly independent (Knutti et al., 2013). Nevertheless, this single model dataset still allows us to explore both scenario uncertainty and internal variability. Further, it is common with simple climate models to fit different emulators independently, allowing improved interpretability, and if an emulator is shown to have good skill in this task it seems reasonable to assume that it will perform similarly well for other models (or combinations of models) and so multi-model ensembles may be easily incorporated in the future.

The remainder of this paper describes the development of the dataset including the underlying ESM and all post-processing (Section 2), the evaluation metrics used to rank ClimateBench submissions (Section 3), a selection of baseline emulators that have been developed to demonstrate a variety of approaches to tackle ClimateBench (Section 4), a discussion of such approaches and future opportunities for diverse approaches (Section 5) before providing a few concluding remarks in Section 6.

## Data set description and preparation

The data provided as part of ClimateBench is a heavily curated version of that publicly available in the CMIP6 data archive. Here we describe the data extraction and processing steps, but the scripts used to perform this are also freely available (as described below).

We use a selection of complementary simulations in order to provide as large a training dataset as possible while attempting to avoid unnecessary redundancy. Table 1 details the full list of simulations included, the period they cover and a brief description of their purpose in this context. Given that the primary purpose of ClimateBench is to train emulators over different emission scenarios, ScenarioMIP simulations are a key component of the dataset. ScenarioMIP prescribes a limited set of possible future emissions pathways exploring different socio-economic scenarios representing plausible narratives. These scenarios are designed to span a range of mitigation scenarios (denoted by the first number in each scenario) and end-of-century forcing possibilities (denoted by the last two numbers in each scenario). We include all available simulations, including the AerChemMIP *ssp370-lowNTCF* variation of *ssp370* which includes lower emissions of near-term climate forcers (NTCFs) such as aerosol (but not methane). We choose *ssp245* as our test dataset against which all ClimateBench emulators are to be evaluated. This scenario represents a medium mitigation and medium forcing scenario, ensuring trained emulators are able to interpolate a solution rather than extrapolate (as discussed further in Section 5). The CMIP6 *historical* experiment is also included since it provides useful training data at low emissions values.

**Table 1: Details of post-processed simulations provided as part of the ClimateBench dataset**

| Protocol | Experiment | Period | Notes |
|---|---|---|---|
| ScenarioMIP (O'Neill et al., 2016) | ssp126 | 2015 - 2100 | A high ambition scenario designed to |
| | ssp245 | 2015 - 2100 | Designed to represent a medium forcin |
| | ssp370 | 2015 - 2100 | A medium-high forcing scenario with l |
| | ssp370-lowNTCF | 2015 - 2054 | Variation of SSP370 with lower emissi |
| | ssp585 | 2015 - 2100 | This scenario represents the high end |
| CMIP6 (Eyring et al., 2016) | historical | 1850 – 2014 | A simulation using historical emissions |
| | abrupt-4xCO2 | 500 years | Idealised simulation in which CO2 is a |
| | 1pctCO2 | 150 years | Idealised simulation in which CO2 is g |
| | piControl | 500 years | Baseline simulation in which all forcin |
| DAMIP (Gillett et al., 2016) | hist-GHG | 1850 – 2014 | A historical simulation with varying co |
| | hist-aer | 1850 – 2014 | A historical simulation only forced by |

ClimateBench also includes a selection of more idealised simulations which are intended to provide training data at the 'corners' of the four-dimensional input space, again helping reduce the chances of extrapolation in the resulting emulators (as demonstrated in Figure A1). Two simulations that are commonly used to diagnose the equilibrium and transient climate sensitivity are *abrupt-4xCO2* and *1pctCO2,* respectively. As the name suggests, the *abrupt-4xCO2* includes an abrupt quadrupling of CO2 over the pre-industrial concentrations while all other forcing agents remain unchanged. This level of concentration represents the high end of future scenarios, broadly in line with *ssp585* but with no contri-

bution from the other forcers, simplifying its interpretation. The abrupt nature of the forcing also allows the timescale of the responses to be determined which can be useful for emulators which account for this. The *1pctCO2* simulation gradually increases the atmospheric concentration of CO2 by 1% per year, again with other forcing agents unchanged. Two simulations performed as part of the Detection-Attribution Model Intercomparison Project (DAMIP; Gillett et al., 2016) represent the historical period forced by only CO2 and other long-lived greenhouse gases (*hist-GHG*), or only anthropogenic aerosol (*hist-aer*). Again, these provide opportunities to train emulators in regions of the input (emissions) space that are at the limits of plausible future scenarios.

Finally, the *piControl* simulation provides a baseline simulation with all forcings remaining unchanged from their pre-industrial values. All target variables are calculated as a change against this climatology to simplify the training and interpretation of the results. This long (500 year) simulation also enables a robust estimation of internal variability of the climate system for those emulators which are able to represent it, as discussed further in Section 5.1.

The input data for these simulations is prescribed by the experimental protocol and provided by the input4MIPS project (https://esgf-node.llnl.gov/search/input4mips/), which we collate and pre-process for ease of use. Specifically, we extract the provided global mean emissions of CO2 and CH4 for each of the realistic (historical, ScenarioMIP and DAMIP) experiments from the checksum files provided by the Community Emissions Data System (CEDS) dataset (Hoesly et al., 2018). We sum over each sector and each month in order to derive annual total emissions and convert from Kg to Gt of CO2. Some historical and future periods are only provided in 5-yearly increments, so we linearly interpolate to yearly values for consistency. The CO2 emissions are summed cumulatively since, for realistic scenarios, a compensation between forcing efficiency and ocean uptake means the temperature response to CO2 is approximately linear in the cumulative emissions (Matthews and Caldeira 2008; Allen et al. 2009). Figure 1 shows the global mean emissions of each of the forcing agents under different future emissions scenarios, showing a wide range of possible pathways.

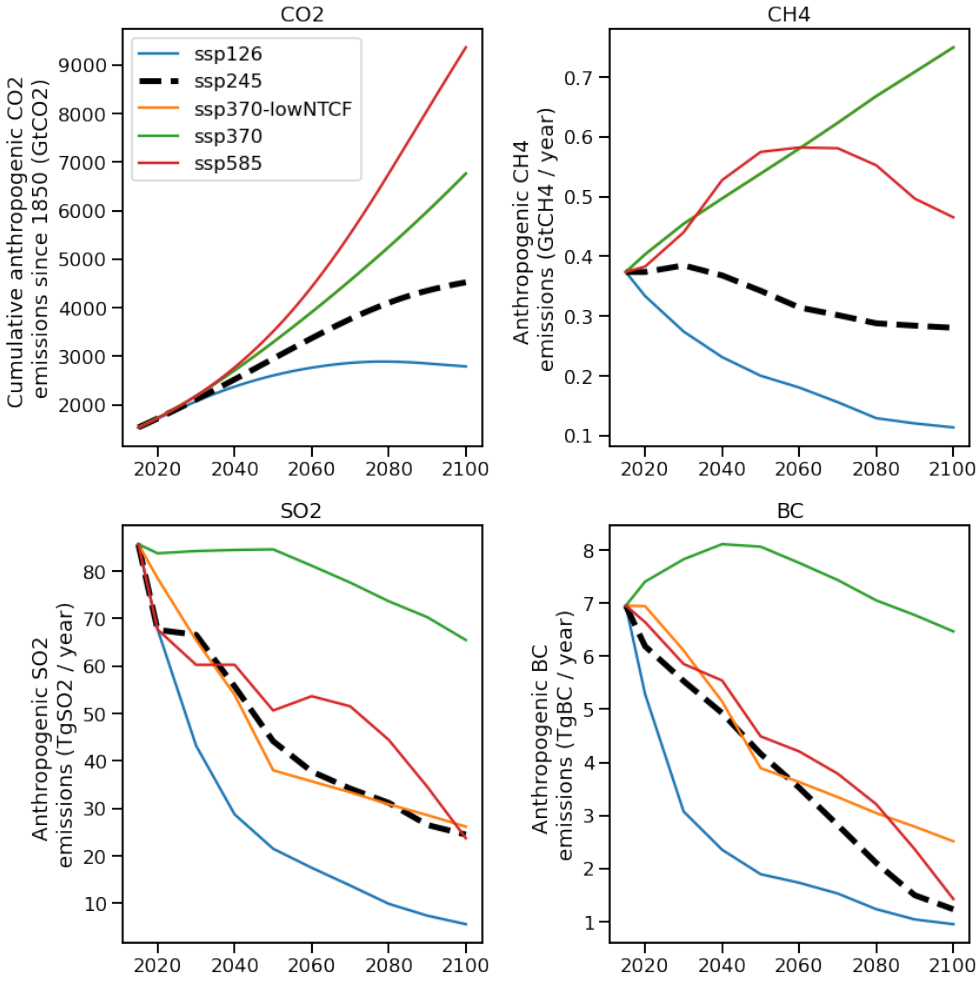


**Figure 1: Time series of cumulative anthropogenic CO2 emissions since 1850 (a); emissions of CH4 (b); global mean emissions of SO2 (c) and black carbon (BC; d) derived from NorESM2 ScenarioMIP simulations available within ClimateBench, including the SSP245 test scenario (shown in black).**

The aerosol (precursor) emissions are derived from the latest version of the spatially resolved CEDS dataset and again summed over sectors and months to produce maps of annual total emissions, as shown in Figure 2 for SO2 in different years. While the spatial distribution clearly evolves over the historical period and into the future scenarios, the emissions are fairly localised around industrialised regions and dimensionality reduction can be used to reduce the size of these input features (as discussed for the baseline emulators in Section 4). An area preserving interpolation is performed so that the emission data

are provided on the same spatial grid as the NorESM2 output fields to simplify its use in ML workflows. Again, as used for NorESM2 the 5-yearly data is interpolated to a yearly frequency for consistency.

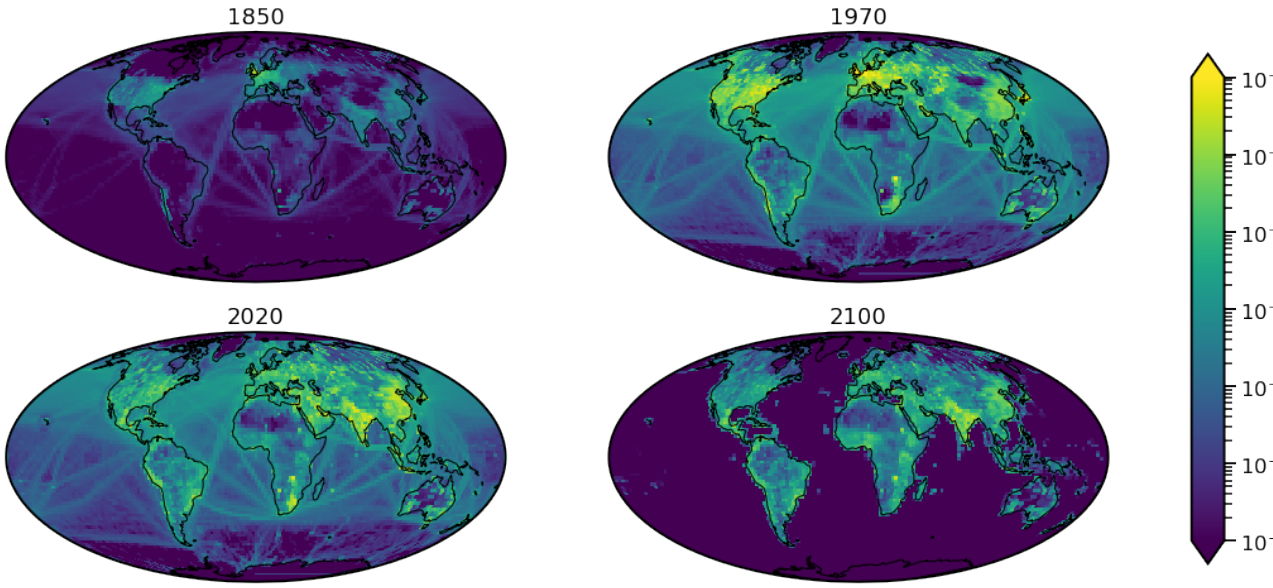


**Figure 2: Maps showing the evolution of the spatial distribution of anthropogenic SO2 emissions in the pre-industrial era represented by 1850 (a); the peak emissions era of 1970 (b); current emissions (c); and future emissions under SSP 245 (d).**

For the idealised CMIP simulations (*abrupt-4xCO2* and *1pctCO2*) no emissions files are used and so the cumulative anthropogenic CO2 emissions are calculated from the difference in the diagnosed CO2 atmospheric mass concentrations in these and the *piControl* experiment. Emissions of all other species are also provided but set to zero (as they represent no change since the pre-industrial).

## 2.1 Target ESM

As the model with the most relevant experiments completed, we use the output from simulations performed by the NorESM2 model in its low atmosphere-medium ocean resolution (LM) configuration (Seland et al., 2020). This model consists of a fully coupled earth system with online atmosphere, land, ocean, ice and biogeochemistry components. It shares many components with the Community Earth System Model Version 2 (Danabasoglu et al., 2020) but has a replaced aerosol and atmospheric chemistry scheme (including their interactions with clouds) and a different ocean model. It has a relatively low equilibrium climate sensitivity (ECS; equilibrium global mean temperature after a doubling of CO2) of 2.5 K, particularly compared to the 5.3K of CESM2 (Gettelman et

al., 2019), which has been attributed to ocean heat uptake and convective mixing in the Southern Ocean (Gjermundsen et al., 2021). Combined with a strong aerosol forcing ($-1.36\,\mathrm{W\,m^{-2}}$ for 1850 to 2014), this likely accounts for the somewhat anomalous cooling between 1950-1980 in the historical simulations.

The output of these simulations are aggregated to annual mean values but kept at their native spatial resolution (approximately 2°). The temperature (T) and precipitation (P) are exactly equivalent to the archived surface air temperature (tas) and total precipitation (pr) output variables respectively. The DTR is calculated as the annual mean difference in the daily maximum and minimum surface air temperatures (tasmax – tasmin). The PR90 is calculated as the $90^{\mathrm{th}}$ percentile of the daily precipitation in each year. The annual mean baseline values (from *piControl*) for each variable are then subtracted from each experiment so that they represent a difference from pre-industrial. Temperature changes under anthropogenic climate change are routinely reported in this way, and it also makes the downstream emulation task somewhat easier as it removes an offset. The values are not scaled to have unit variance, but users of the dataset might choose to do this with certain emulators. Samples of these output fields from the target ssp245 dataset are shown in Figure 3. The relative increase in warming in the northern polar regions (known as Arctic amplification) is clearly seen in Fig. 3a, as well as the north Atlantic warming hole (Woollings et al., 2012; Drijfhout et al., 2012; Manabe and Stouffer, 1993), the emergence of which is also affected by aerosol radiative forcing (Dagan et al., 2020). Figure 3b shows the strong land/sea contrast in DTR, since most of the change is confined to land, and largely caused by changes in aerosol (particularly sulfate) forcing. Most of the precipitation response shown in Figure 3c-d is due to the shift in the inter-tropical convergence zone (ITCZ) which results from a shift in the cross-equatorial energy balance under increased warming (Schneider et al., 2014), but some features, particularly in South-East Asia might be due to local aerosol effects (particularly due to BC; e.g., Bollasina et al. 2014, Wilcox et al. 2020, Mansfield et al. 2020).

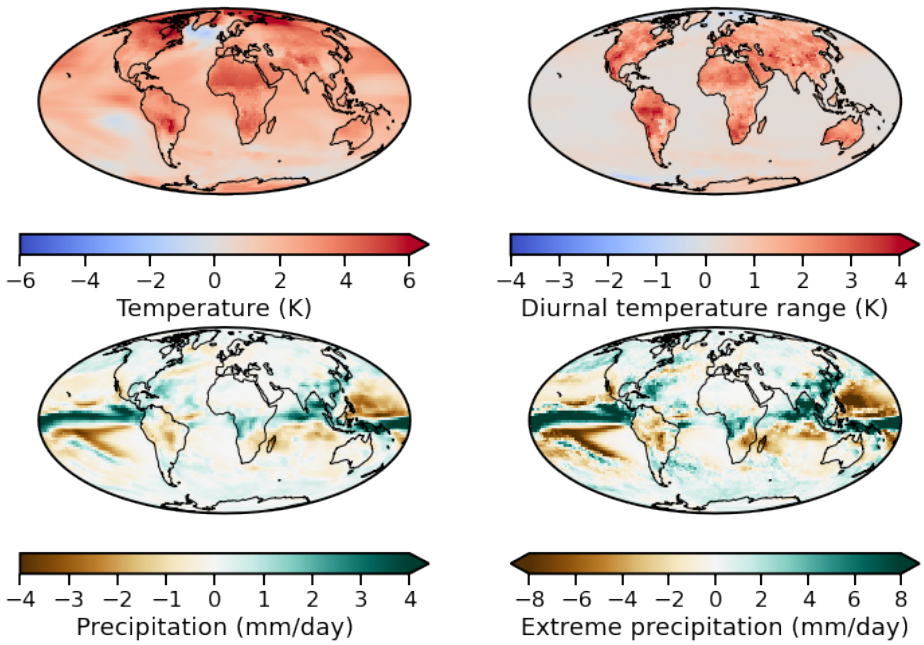


**Figure 3: Maps of target outputs from the SSP245 held-back test scenario at 2100 (as an anomaly to the pre-industrial control run) performed by NorESM2: (a) Annual mean surface temperature; (b) annual mean diurnal surface temperature range; (c) annual mean precipitation; and (d) 90th percentile of the daily precipitation.**

Also included in the dataset are the top-of-atmosphere Effective Radiative Forcings (ERFs) for this model for each forcing agent over the historical period. These are based on diagnostics of the fixed sea-surface temperature experiments of the Radiative Forcing Model Intercomparison Project (RFMIP; Pincus et al., 2016; Smith et al., 2021) and provide a more direct estimate of the radiative climate effect of each forcer over this period than simply emissions. It also allows an estimate of the efficacy of each forcer in this model (the temperature response per unit of forcing). This might be useful for normalising the inputs by their efficacy or developing more physically interpretable emulators that derive the climate response via the forcing.

## Evaluation

The evaluation criteria are a crucial aspect to any benchmark dataset and need to be concretely defined and accurately reflect the objectives of the machine learning task. Ideally, the criteria are also simple to implement such that they can be used as a target in any loss function that might be used to train emulators. The spatial characteristics of the outputs in this task also need to be

considered. As a primary metric we choose the area-weighted root-mean square error (RMSE), calculated following:

$$RMSE = \frac{1}{50} \sum_{y=2051}^{2100} \sqrt{\frac{1}{N_{\text{lat}} N_{\text{lon}}} \sum_{i}^{N_{\text{lat}}} \sum_{j}^{N_{\text{lon}}} L(i) \left( x_{i,j,y} - \left| x^{t}_{i,j,y,n} \right|_{n} \right)^2}, \quad (1)$$

where the weighting function $L$ accounts for the decreasing grid-cell area towards the poles and is defined as $L(i) = cos\left(\text{lat}(i)\right)$.

This commonly used metric provides a single number summarising the mismatch between the predictions ($x$) and the target variables ($x^t$). By squaring the difference, the RMSE also weighs large discrepancies more heavily, more heavily penalising larger errors. We average the target variables over the three available ensemble members ($n$) and the RMSE over a long period of the target scenario (2050 – 2100) in order to minimise the contribution of internal variability.

Estimates of this internal variability can be very valuable for climate projections however and since ClimateBench includes three ensemble members for each training dataset emulators are encouraged to include estimates of it if they are able. A natural extension of the RMSE for probabilistic estimates commonly used in weather forecasting is the Continuous Ranked Probability Score (CRPS):

$$CRPS = \frac{1}{50\ N_{\text{lat}} N_{\text{lon}}} \sum_{y=2051}^{2100} \sum_{i}^{N_{\text{lat}}} \sum_{j}^{N_{\text{lon}}} L(i) \int_{x=-\infty}^{x=\infty} \left( F_{i,j,y}(x) - F^{t}_{i,j,y}(x) \right)^2 dx, \quad (2)$$

where $F(x)$ and $F^t(x)$ are the cumulative distribution functions (CDFs) over the predicted and target ensembles respectively (Gneiting et al. 2005). This measures the area between the two CDFs so that smaller values are better and has the benefit of retaining a well-defined interpretation in the case of only a single target observation (whose CDF would be the Heaviside function). The CDFs can be approximated over finite ensembles using quadrature, or direct integration if the PDFs can be assumed to be Gaussian. Methods to calculate both metrics based on the climpred (Brady and Spring, 2021) package are provided in the example notebooks included with the dataset. While this metric is not included in the headline ranking of ClimateBench approaches, we include an example approach using GPs which is discussed in more detail in Section 4.1.

## Baseline evaluation

Before evaluating some baseline statistical emulators, it is useful to consider two cases with which we hope to bracket the data-driven approaches. The first is the internal variability of the NorESM2 target ensemble which provides an upper bound on the predictability of the scenario in the presence of the natural variability of the Earth system. The second is a comparison to another ESM

which also performed the test projection, in this case the UKESM1 model (Sellar et al., 2019). This provides an example of the inter-model spread encountered within CMIP6 and a lower bound on the accuracy we would like our emulators to achieve.

As noted previously, the NorESM2 *ssp245* projections included three ensemble members sampling internal variability by choosing different initial model states from the start of the piControl simulation at intervals of 30 model years apart. The average RMSE for each variable at each target time between ensemble members 1 and 2, and 1 and 3 are provided in Table 2 and provide an estimate of the best achievable skill over this period (since the members only differ by their internal state). In practice, the emulators can (and do) outperform this baseline because they target the mean over all three ensemble members, reducing the effect of internal variability.

The UKESM1 model performed the same *ssp245* experiment using the same emissions of climate forcers but due to different physical and structural assumptions produces quite a different climate. While a full comparison of the two models and their predictions is beyond the scope of this work, it is instructive to briefly discuss the key differences in order to place any emulator errors in the context of broader inter-model uncertainties. We see similar patterns of temperature change between the models, including Arctic amplification (see Fig. A2), although there is a much stronger mean response in UKESM1 compared to NorESM2, primarily due to its higher climate sensitivity (5.4 K). There is a very distinct difference in the modelled DTR between the models which cannot be explained by aerosol effects alone as the forcing is very similar between the models ($-1.45$ Wm$^{-2}$ in UKESM1 compared to $-1.36$ W m$^{-2}$ in NorESM2) and may be due to the different land models used; UKESM1 uses JULES (Harper et al. 2018) and NorESM2 uses CLM5 (Lawrence et al. 2019). Despite the large difference in temperature response, interestingly the precipitation response is broadly in agreement, suggesting quite distinct changes in the hydrological cycle. For example, the UKESM1 precipitation does not show a clear shift in the ITCZ and shows larger changes in the extra-tropics. The RMSE between UKESM1 and NorESM2 is correspondingly large for the temperature metrics and closer to the baseline approaches for precipitation, as shown in Table 2.

## Baseline emulators

Three baseline emulators are developed to demonstrate various potential approaches to tackling the machine learning problem this dataset provides. These are performed using the Earth System Emulator (Watson-Parris et al., 2021) to provide a simple interface for non-ML experts and permit sampling the emulators for potential use in detection and attribution workflows (as discussed in the Section 5). The three emulators all perform skillfully, as summarised in Table 2 and Figure 4 and discussed in more detail in each of the following subsections. The emulators also show broadly similar biases, particularly for precipitation

where they all slightly underestimate increases (decreases) in tropical (subtropical) rainfall in the western Pacific. This might suggest that these particular changes are driven by different climate forcers or longer time-scale changes than modelled in this study. A direct comparison of the emulator predictions and NorESM is shown in Figure A3.

**Table 2: The average root mean square error (RMSE) of the different baseline emulators for the years 2050-2100 against the ClimateBench task of estimating key climate variables under future scenario SSP245. Another state-of-the-art model (UKESM1) and the average RMSE between NorESM ensemble members as an estimate of internal variability are included for comparison.**

| **Emulator** | **RMSE in mean surface air temperature (K)** | **RMSE in mean diu** |
|---|---|---|
| Gaussian process regression | 0.36 (CRPS: 0.33) | 0.15 (CRPS: 0.12) |
| Neural network (CNN+LSTM) | 0.38 | 0.17 |
| Random Forest | 0.42 | 0.15 |
| UKESM1 | 2.20 | 1.28 |
| (variability) | 0.80 | 0.31 |

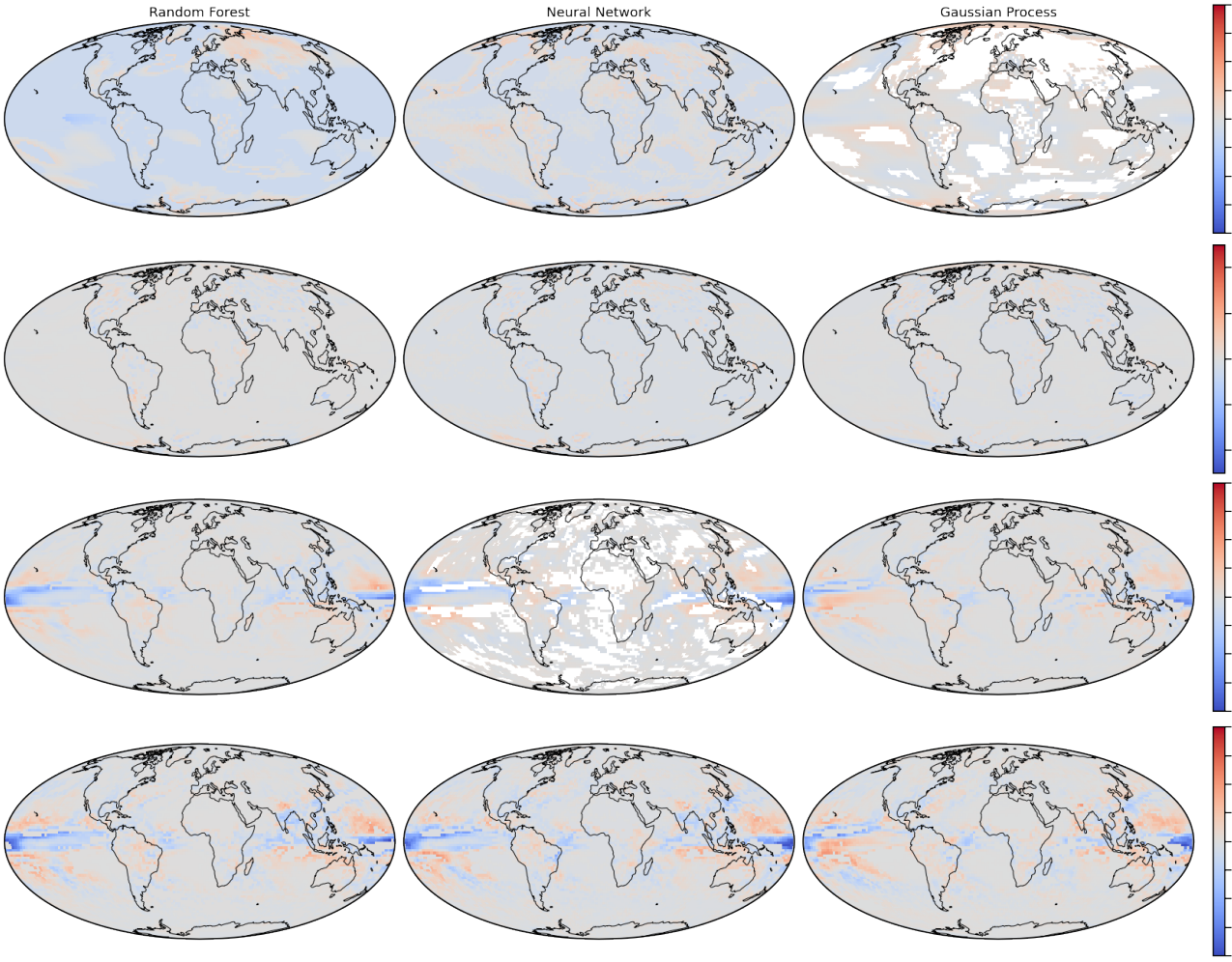


**Figure 4: Maps of the mean difference in the ClimateBench target variables for each baseline emulator against the target NorESM values under the test ssp245 scenario averaged between 2051-2100. Differences insignificant at the p<5% level are masked from the plots.**

## Gaussian process regression

Gaussian processes (GPs) (Rasmussen and Williams, 2005) are probabilistic models which assume predictions can be modelled jointly as normally distributed. GPs have been widely used for nonlinear and nonparametric regression problems in the geosciences (Camps-Valls et al., 2016).. A GP is fully determined by the expectation of individual predictions – referred to as the mean – and the covariance between pairs of predictions. Such covariance is

typically user-specified as a bivariate function of the input data called the kernel function. The choice of the kernel function allows to restrict the functional class the GP belongs to, offering, for example, control over functional smoothness. GPs for regression solve a supervised problem where the observed input-output sample pairs are used to: (1) infer the emulator parameters (typically the noise variance and the kernel parameters only) by maximising the log-likelihood of the observations under the gaussianity assumption (the so-called evidence); and then (2) allow to obtain its posterior probability distribution that is used to make predictions over unseen inputs.

To prepare the input samples, we average ensemble members for SO2 and BC emission maps and CO2 and CH4 global emissions. The dimensionality of each aerosol emission map is reduced with principal component analysis, restricting ourselves to the 5 first principal components. All input covariates and target outputs are standardised using training data mean and standard deviation.

The GP is set with a constant mean prior and separate kernels are devised for each species. Automatic relevance determination (ARD) kernels are used for SO2 and BC, allowing each principal component to be treated independently with its own lengthscale parameter. The GP covariance function is obtained by summing all kernels together, thus accounting for multiscale feature relations (see Camps-Valls et al., 2016 for several composite kernel constructions in remote sensing and geoscience problems). To account for internal variability between ensemble members, we consider an additional white noise term with homoscedastic variance over the output targets, which is also inferred from the training phase.

We use Matérn-1.5 kernels for each input. This guarantees the GP is a continuous function; details are provided in Section A1. The mean value, kernels parameters and internal variability variance are jointly tuned against the training data by marginal likelihood maximisation with the L-BFGS optimisation scheme. The emulators used have 16 parameters in total.

As reported in Table 2, the 2050-2100 averaged RMSE of the mean predictions with the GPs are the best of all the emulators when predicting surface air temperature, diurnal temperature range, precipitation rate and $90^{\text{th}}$ percentile of precipitation. This is remarkable given the limited number of parameters that are tuned. It suggests the GP prior is an adequate choice for the purposes of emulation. Study of the inferred kernel variance (not shown) suggests that cumulative CO2 emissions generally influence all predictions, and unequivocally dominate the predictions for surface air temperature and diurnal temperature range. CH4 and BC emissions on the other hand appear to have negligible influence on the predictions. Since the GP also provides posterior estimates of the variance (which will incorporate an estimate of internal variability) we also calculate the CPRS for this emulator (see Table 2). While we are unable to compare these scores with the other baseline methods the similarity to the RMSE indicates that the GP is also predicting the internal variability accurately (otherwise it would be penalised in the CPRS relative to the RMSE).

## Random Forests

Random forests aggregate predictions of multiple decision trees (Ho, 1995; Breiman, 2001). These trees repeatedly split data into subsets according to its features such that in-subset variance is low and between-subset variance is high. This makes decision trees good at modelling non-linear functions, in particular interactions between different variables. However, they are prone to overfitting (Ho, 1995). This problem is alleviated by ensemble methods which train a large number of different trees. Weak learners are combined to give strong learners. Bagging, used in Random Forests, describes training different trees on different subsets of the data or holding back some of the data dimensions for each individual tree. The Forest makes a prediction by averaging over the predictions of all individual trees.

Two main arguments support an ensemble method approach to climate model emulation: These methods are skillful at interpolation tasks, but by construction are unable to extrapolate (Breiman, 2001). However, for applications of climate model emulation, interesting predictions will likely lie inside the hypercube delimited by historical data, low-emissions (*ssp126*) and business-as-usual (*ssp585*) scenarios. A major advantage of ensemble methods over more complex ML methods such as neural networks (and even ESMs) is their interpretability. This is important as ultimately predictions should inform decision-making. Being able to provide explanations why a given input led to a prediction helps to understand the consequences of decisions about emission pathways.

To train the random forest emulator, we use the historical dataset and the socioeconomic pathways *ssp585*, *126*, and *370*. For each dataset, we take the average of its ensemble members and reduce the dimensionality of SO2 and BC emission maps to the first five principal components. Separate random forest emulators are trained for the four target variables. The following hyperparameters are tuned using random search: number of trees, tree depth, number of samples required to split a node and to be at each leaf node. The hyperparameters used for each emulator are indicated in Section A2.

As shown in Table 2, the mean RMSE scores for the years 2051-2100 of the random forest regressors are comparable to the performance of the other emulators for diurnal temperature range, precipitation and extreme precipitation. The temperature prediction is comparable but slightly worse than the predictions of the neural network architecture. To assess the impact of the four input features on the prediction, we calculate the permutation feature importance. It is defined as the decrease in a model score when a single feature value is randomly shuffled (Breiman, 2001). Figure 5 shows that CO2 concentrations dominate the predictions. For temperature predictions the other features are negligible. SO2 and BC aerosol emissions have a small impact on the global mean temperature and precipitation predictions. This is in line with the physical understanding that while anthropogenic aerosol can influence precipitation rates (both radiatively and through aerosol-cloud interactions), aerosol contributions play a negligible

role at the end of the century in the test scenario. The regional influences may be more significant however and this will be explored separately.

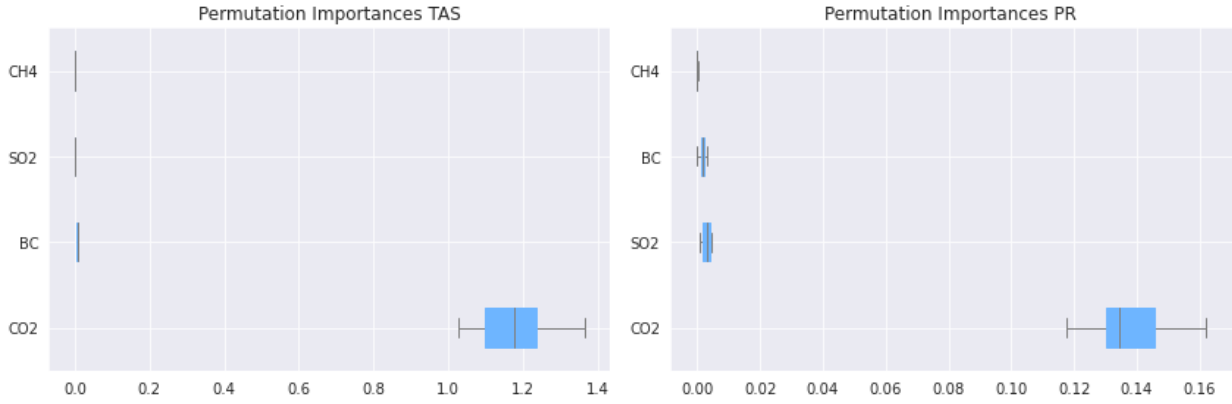


**Figure 5 : Permutation importances for the most important component of each variable in predicting global mean temperature (TAS) and precipitation (PR). Each emulator input variable is shuffled in turn to determine the relative contribution to prediction skill. Note that these average estimates do not account for potential regional contributions which may be particularly relevant for aerosol.**

## Neural Networks

Artificial Neural Networks (ANNs), algorithms inspired by the biological neural networks of human brains, have shown great success in areas like Computer Vision and Natural Language Processing. Two major architectures are Convolutional Neural Networks (CNNs) (LeCun et al., 1990), to model spatial dependencies, and Recurrent Neural Networks (RNNs), to process sequential data. Besides the traditional areas, ANNs have been recently employed to tackle a variety of problems in earth system science (Camp-Valls et al., 2021). Long short-term memory (LSTM) networks (Hochreiter et al., 1997), an advanced type of RNNs, are used for modelling time-series, for example for El Niño-Southern Oscillation prediction (Broni-Bedaiko et al., 2019). In cases where both input and target have a spatial structure, such as modelling of precipitation or changes in satellite imagery, a very commonly used CNN type is the U-Net (Ronneberger et al, 2015), which has been applied frequently in climate science and weather forecasting (Trebing et al., 2020, Harder et al., 2020).

We explored both a pure LSTM approach and a pure CNN approach, using a U-Net. A combination of both network types gave the best results, therefore we use an LSTM combined with a CNN for our example architecture. The CNN is used to extract spatial features before feeding our input in the LSTM. The CNN consists of one convolutional layer with a kernel size of 6, followed by a ReLU activation function and average pooling. The LSTM uses 25 units

and a ReLU activation function as well, which is followed by a dense layer and reshaping to the output dimension. To train the emulator we use *ssp126*, *370* and *585* scenarios and the historical data with a moving-time window size of 10 years (in one-year increments, leading to 570 training points). The emulator is trained for 20 epochs, using a batch size of 25 for T and DTR and 5 for PR and PR90. For this baseline approach we chose not to do any hyperparameter optimization.

RMSE scores obtained with the CNN-LSTM architecture are comparable to those obtained with the other methods. The CNN-LSTM architecture performs particularly well for temperature predictions, with average RMSE scores of 0.38 K over the second half of the 21st century. This might be because temperature has greater autocorrelation/less variability from one year to the next one compared to the other variables. Such autocorrelation would be well captured by a time-aware model like an RNN. Spatial patterns of temperature changes, such as the Arctic amplification, are reasonably well predicted, even though the coldest temperatures (e.g. in the North-Atlantic cold patch) are not as well captured (as shown in Figure 4). The CNN-LSTM performs slightly worse than the other emulators for diurnal temperature range and precipitation predictions. For precipitation, global patterns (e.g. the ITCZ shift) are well predicted by the emulator, but the relative changes are overestimated (too wet or too dry) in most places. Like for all other emulators showcased in this study, extreme precipitation proves the hardest variable to predict accurately.

1.

# Discussion

(a)

## Climate-specific challenges

The emulation of future climate states presents particular challenges for machine learning and other statistical approaches. Chiefly among those is the limited amount of training data that is typically available; current ML approaches are not prepared to learn such complex scenarios in small data regimes under a covariate shift. As pointed out, the complex ESMs that are trusted to model the future climate are extremely computationally expensive to run and the observational record cannot inform us about unseen future scenarios. By harnessing a large selection of simulations performed as part of CMIP6, ClimateBench attempts to alleviate this difficulty, but nevertheless only around 500 training points (years) represent realistic climate states, many of which are not independent (as shown in Fig. A1). This presents a challenge for deep learning approaches which typically require tens of thousands of training samples to avoid over-fitting. The inclusion of longer idealised simulations does provide

opportunities for pre-training however, particularly the 500-year long *piControl* simulations which could be used with contrastive learning to reduce the training samples required for neural network architectures.

The *piControl* simulation could also be used to inform emulators more explicitly about the internal variability of climate (as produced by NorESM2). The signal, particularly for the precipitation target variables, can be small compared to this variability and this proves challenging for some emulators to reproduce. An explicit model of the internal variability (Castruccio et al., 2019) could help to alleviate this.

Another challenge in applying statistical learning approaches to this dataset is the relatively high dimensional inputs and outputs (96 x 144). Most approaches to emulating the regional temperature response to a CO2 forcing have been carried out at, at most, dozens of locations, but accounting for the spatial correlations is something which CNNs can excel at and have recently been shown to produce accurate emulations of temperature across similar dimensionality (Beusch et al., 2020). Such approaches typically assume a regular spacing, however, and neglect the reducing area of each grid-cell towards the poles. While more traditional approaches of dimensionality reduction can also be used, such as (weighted) empirical orthogonal functions (EOFs), these may not be appropriate for the non-linear precipitation fields which might require kernel-based approximations (e.g., Bueso et al., 2019).

For practical purposes, an estimate of the uncertainty in any prediction would be extremely valuable. This uncertainty should encompass that due to the internal variability and the emulator approximation (and ideally that of the underlying physical model). In the ML community, these are known as the epistemic and the model uncertainties, and are being studied intensively (Kendall et al., 2017). Quantifying these two uncertainties would allow increased trust (a concept explored in the next section) in the prediction as well as quantitative comparison to other predictions. We encourage the estimation of uncertainty wherever possible, using the provided CRPS metric to evaluate such probabilistic projections. The ability to sample from such distributions would also permit the generation of so-called 'superensembles' which can provide very large ensembles of multiple models under given scenarios (Beusch et al., 2020).

## Emulator trustworthiness

For climate model emulators to be useful for policy decisions they must be trusted by their users. The trustworthiness of any model is a subjective concept that broadly represents one's belief that the model faithfully represents some underlying 'truth'. Model verification attempts to objectively assert this view (indeed the word derives from the Latin, *verus,* meaning true) but is formally impossible for an open system like the Earth (see e.g., Oreskes et al., 1994). While weather models can be regularly validated against observations, in the climate sciences we often instead resort to necessarily incomplete model evaluation and rely on underlying physical principles to provide reassurances of

broader validity. The ClimateBench emulators side-step this issue by aiming only to accurately reproduce an existing physical model which is assumed to already be well evaluated, and therefore attain trustworthiness through proxy. It would nevertheless be reassuring if the emulators could be demonstrated to respect some of the same physical constraints.

In this spirit, Figure 6 shows the relative change in global mean precipitation as a function of global mean temperature change (the hydrological sensitivity) of the baseline emulators and NorESM2. While locally precipitation can change in accordance with the Clausius-Clapeyron relationship (6-7% / K), energy conservation requires that the global changes in precipitation are balanced by radiative cooling and limited to 2-3% / K (Allen and Ingram, 2002; Pendergrass and Hartmann, 2013; Jeevanjee and Romps, 2018; Dagan et al., 2019). While the RF emulator underestimates the hydrological sensitivity of NorESM, it is clear that the emulators learn the physical relationship from the underlying model. Since the emulators were trained on the precipitation and temperature this is to be expected to some degree, but this demonstrates the principle that emulators trained correctly can retain the physical laws of the underlying models. Future efforts to introduce these invariances directly have the potential to significantly ease the training and improve the inference of climate model emulators (Beucler et al., 2021), ultimately improving their trustworthiness.

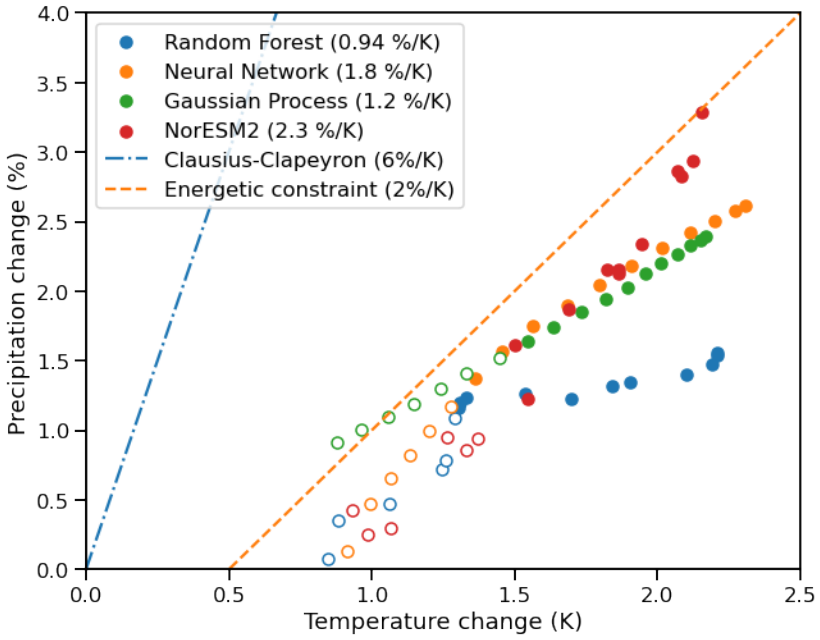


**Figure 6: The relative change in global mean precipitation as a function of global mean temperature change in the baseline emulators and NorESM2 averaged in 5-year increments to reduce internal-**

**variability. Hollow and solid points indicate years before and after 2050 respectively. The change predicted by the Clausius-Clapeyron relationship and energy conservation considerations are shown as dashed lines.**

There has been much attention recently given to 'interpretable' and 'explainable' machine learning models, the former of which are said to behave in a-priori understandable ways, while the latter provide mechanisms to determine post-hoc understanding. While these may be desirable properties in many settings they are very subjective concepts and much weaker foundations on which to build trust than physical laws and thorough evaluation. For example, in a GP the lengthscales that are inferred reflect the corresponding feature relevance and the regularization term accounts for observational noise, but this may not be so obvious for a climate scientist. Indeed, the physical ESMs currently considered the 'gold standard' of climate modelling are only interpretable or explainable by expert practitioners, and it is often part of their role to *explain* the behaviour of their models in response to different drivers. Indeed, given their computational efficiency it is hoped that ClimateBench emulators might be useful in analysing and understanding the response of the underlying physical models themselves.

## Research opportunities

While the challenges outlined above are mostly surmountable with modern architectures and carefully chosen workflows, there are also several broader opportunities ClimateBench presents to develop the state-of-the-art in climate model emulation.

As already mentioned, one area of particular interest is the use of hybrid modelling whereby statistical or ML based emulators embed physical equations, constraints or symmetries in order to improve accuracy, robustness and generalisability (Camps-Valls et al., 2021; Reichstein et al., 2019; Karpatne et al., 2017). One obvious way in which to apply such approaches to ClimateBench is to marry the simple impulse response models discussed in Section 1 with more complex methods to predict the spatial response. Such an approach has recently been demonstrated for temperature (Beusch et al., 2021) but could conceivably be extended to modelling each of the fields targeted in ClimateBench. A more unified, and ambitious, approach would be to model the ordinary differential equations of the response to a forcing directly in the statistical emulator using either numerical GPs (Raissi et al., 2018) or Fourier neural operators (Li et al., 2020).

Another important open question when using data-driven approaches to emulate the climate is how to ensure predictions are performed at locations within the distribution of the training data. In other words, how to ensure the emulator is being used to interpolate existing model simulations rather than extrapolating to completely unseen regions of input space. This can be easy to test for in low dimensions, but it becomes increasingly difficult in higher dimensions and while

the training and test data in ClimateBench have been chosen to minimise the risk of extrapolation broader use could be hindered by the risk of inadvertently asking for an out-of-distribution prediction. While the predictive variance of GPs provide such indications (out of the sample range the GP mean returns to the prior and the covariance is maximised), it is not so easy for other techniques and the use of modern techniques to detect such occurrences (e.g., Lee et al., 2018; Rabanser et al., 2018) could be of great value to minimise this risk.

### Application to detection and attribution

The use of an efficient and accurate way of estimating the climate impacts of different emission scenarios is not limited to exploring future pathways. We may also ask: 'What observed climate states and events can be attributed to anthropogenic emissions?'. A whole field, which started with the seminal work of Hasselmann (1993) has developed rapidly in the last decade (Stott et al., 2016; Barnett et al., 2005; Stott et al., 2010; Shindell et al., 2009; Otto et al., 2016) attempting to answer this question. A common approach is to use climate model (or ESM) simulations to determine optimal 'fingerprints' with which to test observations as well as the power of such a fingerprint under internal variability. These typically have to make fairly strong assumptions about the form of the climate response however (often relying on multiple linear regression) and can incorporate observations of only a few dimensions.

One possible application of the efficient emulators trained using ClimateBench could then be to allow the inference of higher dimensional attribution problems, incorporating more information (such as the DTR and PR) and potentially providing more confident assessments. It would be straightforward to implement such an approach using the ESEm package which provides a convenient interface for such inferences using e.g., ABC, variational inference or Markov Chain Monte-Carlo sampling. Future work will investigate these possibilities.

## Conclusions

The application of machine learning to the prediction of future climate states has, perhaps justifiably due to the challenges laid out above, been cautious to date. Particular applications however, with carefully chosen training data and objectives, can provide fruitful avenues for research and open exciting opportunities for improvement over the current state-of-the-art. This paper introduces the ClimateBench dataset in order to galvanise existing research in this area, provide a standard objective with which to compare approaches and also introduce new researchers to the challenge of climate emulation. It provides a diverse set of training data with clear objectives and challenging target variables, some of which have been extensively studied (surface air temperature) and some which have been somewhat neglected (diurnal temperature range and precipitation).

Current impact assessments are often based on simple emulators, which are

then scaled to match modelled patterns, but which are unable to predict non-linear responses in e.g. precipitation. A robust, trustworthy emulator which is able to provide such predictions could be immensely valuable in quantifying and understanding the changes and associated risks of different socio-economic pathways. Given the importance of faithfully and accurately reproducing the response of ESMs, we hope the challenge will also spur innovation in nascent physically informed ML techniques.

In order to meet these objectives, we have provided open, easy to access datasets and training notebooks which reproduce the results shown in this manuscript and demonstrate the use of the different baseline emulators. All software is open-source and readily available using commonly used package managers. We hope this dataset will provide a focus for climate and ML researchers to advance the field of climate model emulation and provide policy makers with the tools they require to make well informed decisions.

## Data and code availability

The baseline code is available on GitHub (https://github.com/duncanwp/ClimateBench) and a DOI for the specific version, including that used to generate the plots in this paper, will be made available on acceptance.

The benchmark data is available here: https://doi.org/10.5281/zenodo.5196512. The raw CMIP6 data used here are available through the Earth System Grid Federation and can be accessed through different international nodes e.g.: https://esgf-index1.ceda.ac.uk/search/cmip6-ceda/.

## Author Contributions

DWP conceptualised ClimateBench and performed the data-processing and initial analysis. YD and PN contributed to the definition and setup of the framework. DO and ØS performed the original NorESM2 simulations used for training. SB, MD, EF, PH, KJ, JL, PM, MN, LR, CR and JV developed the baseline emulators. DWP prepared the manuscript with contributions from all co-authors.

## Acknowledgements

DWP and PS acknowledge funding from NERC projects NE/P013406/1 (A-CURE) and NE/S005390/1 (ACRUISE). DWP, GCV, PS, SB, MD, EF, PH, KJ, JL, PM, MN, LR, CR and JV acknowledge funding from the European Union's Horizon 2020 research and innovation programme iMIRACLI under Marie Skłodowska-Curie grant agreement No 860100. PS additionally acknowledges support from the ERC project RECAP and the FORCeS project under the European Union's Horizon 2020 research programme with grant agreements

724602 and 821205. GCV was partly supported by the European Research Council (ERC) Synergy Grant "Understanding and Modelling the Earth System with Machine Learning (USMILE)" under the Horizon 2020 research and innovation programme (Grant agreement No. 855187).

The authors also gratefully acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modelling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF. In particular, DO and ØS acknowledge support from the Research Council of Norway funded project INES (270061). High-performance computing and storage resources for NorESM2 were provided by the Norwegian infrastructure for computational science (through projects NN2345K, NN9560K, NS2345K, NS9560K, and NS9034K).

# References

**Allen, M. R. and Ingram, W. J.: Constraints on future changes in climate and the hydrologic cycle, Nature, 419(6903), 224, doi:10.1038/nature01092, 2002.**

**Allen, M. R., Frame, D. J., Huntingford, C., Jones, C. D., Lowe, J. A., Meinshausen, M. and Meinshausen, N.: Warming caused by cumulative carbon emissions towards the trillionth tonne, Nature, 458(7242), 1163–1166, doi:10.1038/nature08019, 2009.**

**Alexeeff, S. E., Nychka, D., Sain, S. R., and Tebaldi, C.: Emulating mean patterns and variability of temperature across and within scenarios in anthropogenic climate change experiments, Climatic Change, 146, 319–333, https://doi.org/10.1007/s10584-016-1809-8, 2018.**

**Barnett, T., Zwiers, F., Hengerl, G., Allen, M., Crowly, T., Gillett, N., Hasselmann, K., Jones, P., Santer, B., Schnur, R., Scott, P., Taylor, K., and Tett, S.: Detecting and Attributing External Influences on the Climate System: A Review of Recent Advances, J Climate, 18, 1291–1314, https://doi.org/10.1175/jcli3329.1, 2005.**

**Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P. and Gentine, P.: Enforcing Analytic Constraints in Neural Networks Emulating Physical Systems, Phys Rev Lett, 126(9), 098302, doi:10.1103/physrevlett.126.098302, 2021.**

**Beusch, L., Gudmundsson, L., and Seneviratne, S. I.: Emulating Earth system model temperatures with MESMER: from global mean temperature trajectories to grid-point-level realizations on land,**

Earth Syst Dynam, 11, 139–159, https://doi.org/10.5194/esd-11-139-2020, 2020.

Beusch, L., Nicholls, Z., Gudmundsson, L., Hauser, M., Meinshausen, M., and Seneviratne, S. I.: From emission scenarios to spatially resolved projections with a chain of computationally efficient emulators: MAGICC (v7.5.1) – MESMER (v0.8.1) coupling, Geoscientific Model Dev Discuss, 2021, 1–26, https://doi.org/10.5194/gmd-2021-252, 2021.

Bollasina, M. A., Ming, Y., Ramaswamy, V., Schwarzkopf, M. D. and Naik, V.: Contribution of local and remote anthropogenic aerosols to the twentieth century weakening of the South Asian Monsoon: AEROSOLS AND SOUTH ASIAN MONSOON, Geophys Res Lett, 41(2), 680–687, doi:10.1002/2013gl058183, 2014.

Brady, R. and Spring, A.: climpred: Verification of weather and climate forecasts, J Open Source Softw, 6, 2781, https://doi.org/10.21105/joss.02781, 2021.

Breiman, L.: Random Forests, Mach Learn, 45(1), 5–32, doi:10.1023/a:1010933404324, 2001.

Broni-Bediako, Clifford & Katsriku, Ferdinand & Unemi, Tatsuo & Atsumi, Masayasu & Abdulai, Jamal-Deen & Shinomiya, Norihiko & Owusu, Ebenezer Owusu. (2019). El Niño-Southern Oscillation forecasting using complex networks analysis of LSTM neural networks. Artificial Life and Robotics. 24. 10.1007/s10015-019-00540-2.

Bueso, D., Piles, M. and Camps-Valls, G.: Nonlinear PCA for Spatio-Temporal Analysis of Earth Observation Data, Ieee T Geosci Remote, 58(8), 5752–5763, doi:10.1109/tgrs.2020.2969813, 2020.

Cabré, M. F., Solman, S. A., and Nuñez, M. N.: Creating regional climate change scenarios over southern South America for the 2020's and 2050's using the pattern scaling technique: validity and limitations, Climatic Change, 98, 449–469, https://doi.org/10.1007/s10584-009-9737-5, 2010.

Camps-Valls, G., Verrelst, J., Munoz-Mari, J., Laparra, V., Mateo-Jimenez, F., Gomez-Dans, J. and Gomez-Dan, J.: A Survey on Gaussian Processes for Earth-Observation Data Analysis: A Comprehensive Investigation, Ieee Geoscience Remote Sens Mag, 4(2), 58–78, doi:10.1109/mgrs.2015.2510084, 2016.

Camp-Valls G., Tula D., Zhu X. X. and Reichstein M.: Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences, https://onlinelibrary.wiley.com/doi/book/10.1002/978111 2021

**Castruccio, S., McInerney, D. J., Stein, M. L., Crouch, F. L., Jacob, R. L., and Moyer, E. J.: Statistical Emulation of Climate Model Projections Based on Precomputed GCM Runs*, J Climate, 27, 1829–1844, https://doi.org/10.1175/jcli-d-13-00099.1, 2014.**

**Castruccio, S., Hu, Z., Sanderson, B., Karspeck, A., and Hammerling, D.: Reproducing Internal Variability with Few Ensemble Runs Reproducing Internal Variability with Few Ensemble Runs, J Climate, 32, 8511–8522, https://doi.org/10.1175/jcli-d-19-0280.1, 2019.**

**Collins, W. J., Lamarque, J.-F., Schulz, M., Boucher, O., Eyring, V., Hegglin, M. I., Maycock, A., Myhre, G., Prather, M., Shindell, D., and Smith, S. J.: AerChemMIP: quantifying the effects of chemistry and aerosols in CMIP6, Geosci Model Dev, 10, 585–607, https://doi.org/10.5194/gmd-10-585-2017, 2017.**

**Dagan, G., Stier, P. and Watson-Parris, D.: Contrasting Response of Precipitation to Aerosol Perturbation in the Tropics and Extratropics Explained by Energy Budget Considerations, Geophys Res Lett, 46(13), 7828–7837, doi:10.1029/2019gl083479, 2019.**

**Dagan, G., Stier, P. and Watson-Parris, D.: Aerosol Forcing Masks and Delays the Formation of the North Atlantic Warming Hole by Three Decades, Geophys Res Lett, 47(22), e2020GL090778, doi:10.1029/2020gl090778, 2020.**

**Danabasoglu, G., Lamarque, J. -F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M. M., Large, W. G., Lauritzen, P. H., Lawrence, D. M., Lenaerts, J. T. M., Lindsay, K., Lipscomb, W. H., Mills, M. J., Neale, R., Oleson, K. W., Otto-Bliesner, B., Phillips, A. S., Sacks, W., Tilmes, S., Kampenhout, L., Vertenstein, M., Bertini, A., Dennis, J., Deser, C., Fischer, C., Fox-Kemper, B., Kay, J. E., Kinnison, D., Kushner, P. J., Larson, V. E., Long, M. C., Mickelson, S., Moore, J. K., Nienhouse, E., Polvani, L., Rasch, P. J., and Strand, W. G.: The Community Earth System Model Version 2 (CESM2), J Adv Model Earth Sy, 12, https://doi.org/10.1029/2019ms001916, 2020.**

**Drijfhout, S., Oldenborgh, G. J. van and Cimatoribus, A.: Is a Decline of AMOC Causing the Warming Hole above the North Atlantic in Observed and Modeled Warming Patterns?, J Climate, 25(24), 8373–8379, doi:10.1175/jcli-d-12-00490.1, 2012.**

**Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci Model Dev, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.**

**Gettelman, A., Hannay, C., Bacmeister, J. T., Neale, R. B., Pendergrass, A. G., Danabasoglu, G., Lamarque, J. -F., Fasullo, J. T., Bailey, D. A., Lawrence, D. M., and Mills, M. J.: High Climate Sensitivity in the Community Earth System Model Version 2 (CESM2), Geophys Res Lett, 46, 8329–8337, https://doi.org/10.1029/2019gl083978, 2019.**

**Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., Santer, B. D., Stone, D., and Tebaldi, C.: The Detection and Attribution Model Intercomparison Project (DAMIP v1.0) contribution to CMIP6, Geosci Model Dev, 9, 3685–3697, https://doi.org/10.5194/gmd-9-3685-2016, 2016.**

**Gneiting, T., Raftery, A. E., III, A. H. W. and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Mon Weather Rev, 133(5), 1098–1118, doi:10.1175/mwr2904.1, 2005.**

**Hansen, J., Sato, M., and Ruedy, R.: Long-term changes of the diurnal temperature cycle: implications about mechanisms of global climate change, Atmos Res, 37, 175–209, https://doi.org/10.1016/0169-8095(94)00077-q, 1995.**

**Harder, P., Jones, W., Lguensat, R., Bouabid, S., Fulton, J., Quesada-Chacon, D., Marcolongo, A., Stefanovic, S., Rao, Y., Manshausen, P. and Watson-Parris, Duncan: NightVision: Generating Night-time Satellite Imagery from Infra-Red Observations, https://arxiv.org/abs/2011.07017, 2020**

**Harper, A. B., Wiltshire, A. J., Cox, P. M., Friedlingstein, P., Jones, C. D., Mercado, L. M., Sitch, S., Williams, K., and Duran-Rojas, C.: Vegetation distribution and terrestrial carbon cycle in a carbon cycle configuration of JULES4.6 with new plant functional types, Geosci. Model Dev., 11, 2857–2873, https://doi.org/10.5194/gmd-11-2857-2018, 2018.**

**Hasselmann, K.: Optimal Fingerprints for the Detection of Time-dependent Climate Change, J Climate, 6, 1957–1971, https://doi.org/10.1175/1520-0442(1993)006<1957:offtdo>2.0.co;2, 1993.**

**Ho, T. K.: Random decision forests, Proc 3rd Int Conf Document Analysis Recognit, 1, 278–282 vol.1, doi:10.1109/icdar.1995.598994, 1995.**

**Hochreiter, S. and Schmidhuber, J. : Long Short-Term Memory. Neural computation, 9(8):1735–1780, 1997.**

**Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., Seibert, J. J., Vu, L., Andres, R. J.,**

**Bolt, R. M., Bond, T. C., Dawidowski, L., Kholod, N., Kurokawa, J., Li, M., Liu, L., Lu, Z., Moura, M. C. P., O'Rourke, P. R., and Zhang, Q.: Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the Community Emissions Data System (CEDS), Geosci Model Dev, 11, 369–408, https://doi.org/10.5194/gmd-11-369-2018, 2018.**

**Holden, P. B. and Edwards, N. R.: Dimensionally reduced emulation of an AOGCM for application to integrated assessment modelling: DIMENSIONALLY REDUCED AOGCM EMULATION, Geophys Res Lett, 37, n/a-n/a, https://doi.org/10.1029/2010gl045137, 2010.**

**Kasoar, M., Shawki, D. and Voulgarakis, A.: Similar spatial patterns of global climate response to aerosols from different regions, npj Clim Atmospheric Sci, 1(1), 12, doi:10.1038/s41612-018-0022-z, 2018.**

**Kendall, A. and Gal, Y.: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 5580-5590), 2017.**

**Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, Geophys Res Lett, 40, 1194–1199, https://doi.org/10.1002/grl.50256, 2013.**

**Jeevanjee, N. and Romps, D. M.: Mean precipitation change from a deepening troposphere., P Natl Acad Sci Usa, 115(45), 11465–11470, doi:10.1073/pnas.1720683115, 2018.**

**Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., Collier, N., Ghimire, B., Kampenhout, L., Kennedy, D., Kluzek, E., Lawrence, P. J., Li, F., Li, H., Lombardozzi, D., Riley, W. J., Sacks, W. J., Shi, M., Vertenstein, M., Wieder, W. R., Xu, C., Ali, A. A., Badger, A. M., Bisht, G., Broeke, M., Brunke, M. A., Burns, S. P., Buzan, J., Clark, M., Craig, A., Dahlin, K., Drewniak, B., Fisher, J. B., Flanner, M., Fox, A. M., Gentine, P., Hoffman, F., Keppel-Aleks, G., Knox, R., Kumar, S., Lenaerts, J., Leung, L. R., Lipscomb, W. H., Lu, Y., Pandey, A., Pelletier, J. D., Perket, J., Randerson, J. T., Ricciuto, D. M., Sanderson, B. M., Slater, A., Subin, Z. M., Tang, J., Thomas, R. Q., Martin, M. V. and Zeng, X.: The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty, J Adv Model Earth Sy, 11(12), 4245–4287, doi:10.1029/2018ms001583, 2019.**

**Le Cun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W. , Jackel, L.D. et al. : Hand-written digit recognition with a back-propagation network. In Advances in neural information processing systems, 1990.**

**Lee, K., Lee, K., Lee, H., and Shin, J.: A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks, Arxiv, 2018.**

**Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A.: Fourier Neural Operator for Parametric Partial Differential Equations, Arxiv, 2020.**

**Matthews, H. D. and Caldeira, K.: Stabilizing climate requires near-zero emissions, Geophys Res Lett, 35(4), doi:10.1029/2007gl032388, 2008.**

**Manabe, S. and Stouffer, R. J.: Century-scale effects of increased atmospheric C02 on the ocean–atmosphere system, Nature, 364(6434), 215–218, doi:10.1038/364215a0, 1993.**

**Mansfield, L. A., Nowack, P. J., Kasoar, M., Everitt, R. G., Collins, W. J., and Voulgarakis, A.: Predicting global patterns of long-term climate change from short-term simulations using machine learning, npj Clim Atmospheric Sci, 3, 44, https://doi.org/10.1038/s41612-020-00148-5, 2020.**

**Meinshausen, M., Raper, S. C. B., and Wigley, T. M. L.: Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, MAGICC6 – Part 1: Model description and calibration, Atmos Chem Phys, 11, 1417–1456, https://doi.org/10.5194/acp-11-1417-2011, 2011.**

**Millar, R. J., Fuglestvedt, J. S., Friedlingstein, P., Rogelj, J., Grubb, M. J., Matthews, H. D., Skeie, R. B., Forster, P. M., Frame, D. J., and Allen, M. R.: Emission budgets and pathways consistent with limiting warming to 1.5 °C, Nat Geosci, 10, 741, https://doi.org/10.1038/ngeo3031, 2017.**

**Nicholls, Z. R. J., Meinshausen, M., Lewis, J., Gieseke, R., Dommenget, D., Dorheim, K., Fan, C.-S., Fuglestvedt, J. S., Gasser, T., Golüke, U., Goodwin, P., Hartin, C., Hope, A. P., Kriegler, E., Leach, N. J., Marchegiani, D., McBride, L. A., Quilcaille, Y., Rogelj, J., Salawitch, R. J., Samset, B. H., Sandstad, M., Shiklomanov, A. N., Skeie, R. B., Smith, C. J., Smith, S., Tanaka, K., Tsutsui, J., and Xie, Z.: Reduced Complexity Model Intercomparison Project Phase 1: introduction and evaluation of global-mean temperature response, Geosci Model Dev, 13, 5175–5190, https://doi.org/10.5194/gmd-13-5175-2020, 2020.**

**O'Neill, B. C., Tebaldi, C., Vuuren, D. P. van, Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J.-F., Lowe, J., Meehl, G. A., Moss, R., Riahi, K., and Sanderson, B. M.: The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6,**

Geosci Model Dev, 9, 3461–3482, https://doi.org/10.5194/gmd-9-3461-2016, 2016.

Oreskes, N., Shrader-Frechette, K. and Belitz, K.: Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences, Science, 263(5147), 641–646, doi:10.1126/science.263.5147.641, 1994.

Otto, F. E. L., Oldenborgh, G. J. van, Eden, J., Stott, P. A., Karoly, D. J., and Allen, M. R.: The attribution question, 6, 813–816, https://doi.org/10.1038/nclimate3089, 2016.

Pendergrass, A. G. and Hartmann, D. L.: The Atmospheric Energy Constraint on Global-Mean Precipitation Change, J Climate, 27(2), 130916120136005, doi:10.1175/jcli-d-13-00163.1, 2013.

Pincus, R., Batstone, C. P., Hofmann, R. J. P., Taylor, K. E., and Glecker, P. J.: Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models, J Geophys Res, 113, https://doi.org/10.1029/2007jd009334, 2008.

Pincus, R., Forster, P. M., and Stevens, B.: The Radiative Forcing Model Intercomparison Project (RFMIP): experimental protocol for CMIP6, Geosci Model Dev, 9, 3447–3460, https://doi.org/10.5194/gmd-9-3447-2016, 2016.

Rabanser, S., Günnemann, S. and Lipton, Z. C.: Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift, Arxiv, 2018.

Raissi, M., Perdikaris, P., and Karniadakis, G. E.: Numerical Gaussian Processes for Time-Dependent and Nonlinear Partial Differential Equations, Siam J Sci Comput, 40, A172–A198, https://doi.org/10.1137/17m1120762, 2018.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian Processes for Machine Learning, , doi:10.7551/mitpress/3206.001.0001, 2005.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N.: WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting, J Adv Model Earth Sy, 12, https://doi.org/10.1029/2020ms002203, 2020.

Ronneberger, O., Fischer, P. and Brox, T. : U-Net: Convolutional Networks for Biomedical Image Segmentation. LNCS. 9351. 234-241. 10.1007/978-3-319-24574-4_28, 2015.

Schneider, T., Bischoff, T., and Haug, G. H.: Migrations and dynamics of the intertropical convergence zone, 513, https://doi.org/10.1038/nature13636, 2014.

Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., Debernard, J. B., Gupta, A. K., He, Y.-C., Kirkevåg, A., Schwinger, J., Tjiputra, J., Aas, K. S., Bethke, I.,

**Fan, Y., Griesfeller, J., Grini, A., Guo, C., Ilicak, M., Karset, I. H. H., Landgren, O., Liakka, J., Moseid, K. O., Nummelin, A., Spensberger, C., Tang, H., Zhang, Z., Heinze, C., Iversen, T., and Schulz, M.: Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations, Geosci Model Dev, 13, 6165–6200, https://doi.org/10.5194/gmd-13-6165-2020, 2020.**

**Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., O'Connor, F. M., Stringer, M., Hill, R., Palmieri, J., Woodward, S., Mora, L., Kuhlbrodt, T., Rumbold, S. T., Kelley, D. I., Ellis, R., Johnson, C. E., Walton, J., Abraham, N. L., Andrews, M. B., Andrews, T., Archibald, A. T., Berthou, S., Burke, E., Blockley, E., Carslaw, K., Dalvi, M., Edwards, J., Folberth, G. A., Gedney, N., Griffiths, P. T., Harper, A. B., Hendry, M. A., Hewitt, A. J., Johnson, B., Jones, A., Jones, C. D., Keeble, J., Liddicoat, S., Morgenstern, O., Parker, R. J., Predoi, V., Robertson, E., Siahaan, A., Smith, R. S., Swaminathan, R., Woodhouse, M. T., Zeng, G., and Zerroukat, M.: UKESM1: Description and Evaluation of the U.K. Earth System Model, J Adv Model Earth Sy, 11, 4513–4558, https://doi.org/10.1029/2019ms001739, 2019.**

**Shindell, D. T., Faluvegi, G., Koch, D. M., Schmidt, G. A., Unger, N., and Bauer, S. E.: Improved Attribution of Climate Forcing to Emissions, Science, 326, 716–718, https://doi.org/10.1126/science.1174760, 2009.**

**Smith, C. J., Forster, P. M., Allen, M., Leach, N., Millar, R. J., Passerello, G. A., and Regayre, L. A.: FAIR v1.3: a simple emissions-based impulse response and carbon cycle model, Geosci Model Dev, 11, 2273–2297, https://doi.org/10.5194/gmd-11-2273-2018, 2018.**

**Smith, C. J., Harris, G. R., Palmer, M. D., Bellouin, N., Collins, W., Myhre, G., Schulz, M., Golaz, J. -C., Ringer, M., Storelvmo, T., and Forster, P. M.: Energy Budget Constraints on the Time History of Aerosol Forcing and Climate Sensitivity, J Geophys Res Atmospheres, 126, https://doi.org/10.1029/2020jd033622, 2021.**

**Stott, P. A., Gillett, N. P., Hegerl, G. C., Karoly, D. J., Stone, D. A., Zhang, X., and Zwiers, F.: Detection and attribution of climate change: a regional perspective, Wiley Interdiscip Rev Clim Change, 1, 192–211, https://doi.org/10.1002/wcc.34, 2010.**

**Stott, P. A., Christidis, N., Otto, F. E. L., Sun, Y., Vanderlinden, J., Oldenborgh, G. J. van, Vautard, R., Storch, H. von, Walton, P., Yiou, P., and Zwiers, F. W.: Attribution of extreme weather and climate-related events, Wiley Interdiscip Rev Clim Change, 7, 23–41, https://doi.org/10.1002/wcc.380, 2016.**

**Trebing, K., Stanczyk T. and Mehrkanoon, S.: SmaAt-Unet: Precipitation Nowcasting usingSmall Attention-UNet Architecture, https://arxiv.org/abs/2007.04417, 2021.**

**Watson-Parris, D.: Machine learning for weather and climate are worlds apart, Philosophical Transactions Royal Soc, 379, 20200098, https://doi.org/10.1098/rsta.2020.0098, 2021.**

**Watson-Parris, D., Williams, A., Deaconu, L., and Stier, P.: Model calibration using ESEm v1.0.0 – an open, scalable Earth System Emulator, Geoscientific Model Dev Discuss, 2021, 1–24, https://doi.org/10.5194/gmd-2021-267, 2021.**

**Wilcox, L. J., Liu, Z., Samset, B. H., Hawkins, E., Lund, M. T., Nordling, K., Undorf, S., Bollasina, M., Ekman, A. M. L., Krishnan, S., Merikanto, J., and Turner, A. G.: Accelerated increases in global and Asian summer monsoon precipitation from future aerosol reductions, Atmos Chem Phys, 20, 11955–11977, https://doi.org/10.5194/acp-20-11955-2020, 2020.**

**Woollings, T., Gregory, J. M., Pinto, J. G., Reyers, M. and Brayshaw, D. J.: Response of the North Atlantic storm track to climate change shaped by ocean–atmosphere coupling, Nat Geosci, 5(5), 313–317, doi:10.1038/ngeo1438, 2012.**

# Appendix 1

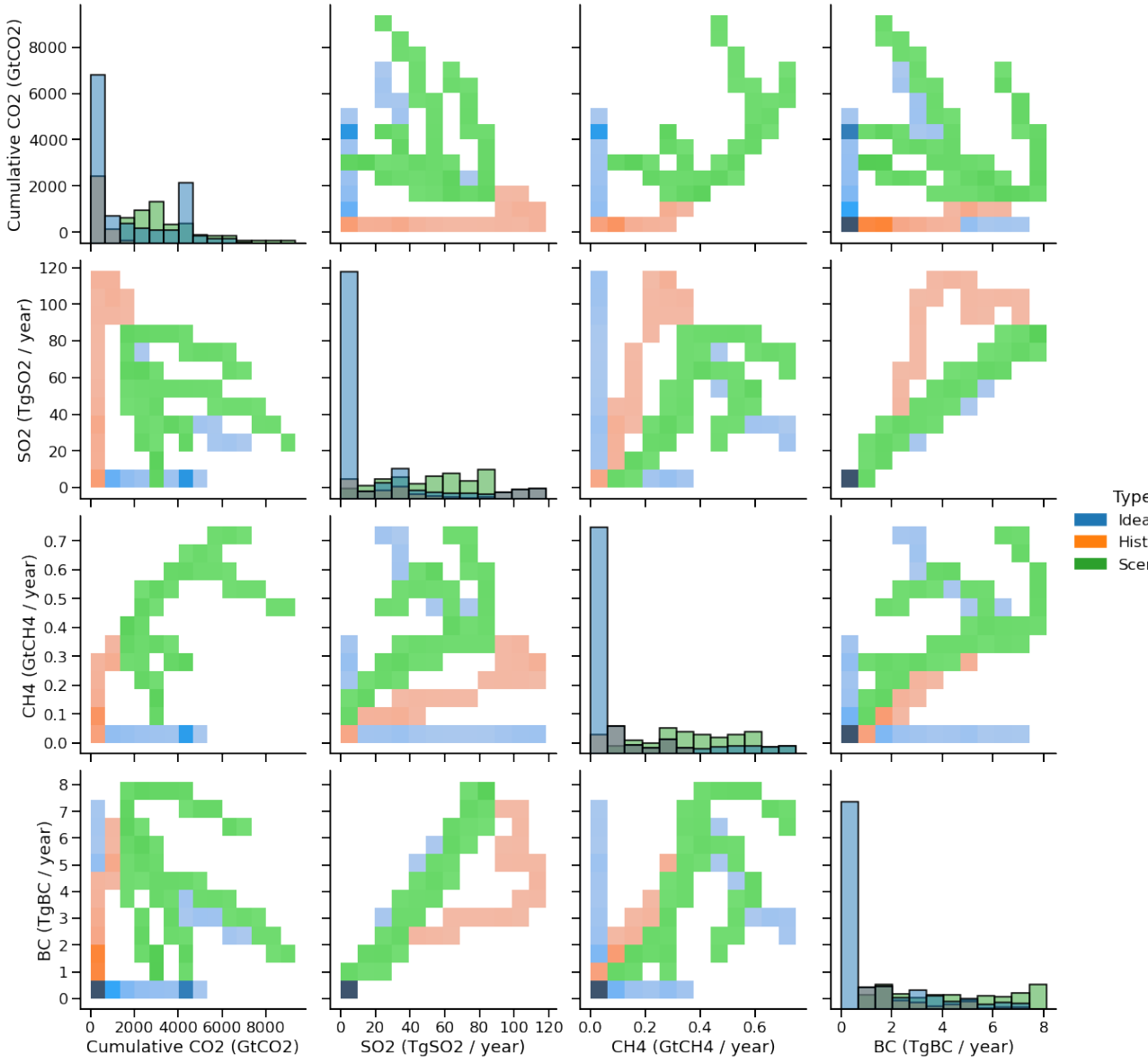


**Figure A1: Joint and marginal distributions of annual global mean emissions and concentrations across the ClimateBench training dataset. Input datasets are classified as Idealised (such as *1pctCO2* and *abrupt4xCO2*, and including *ssp370-lowNTCF*), Historical and Scenario to demonstrate the contribution of each to sampling the full input space.**

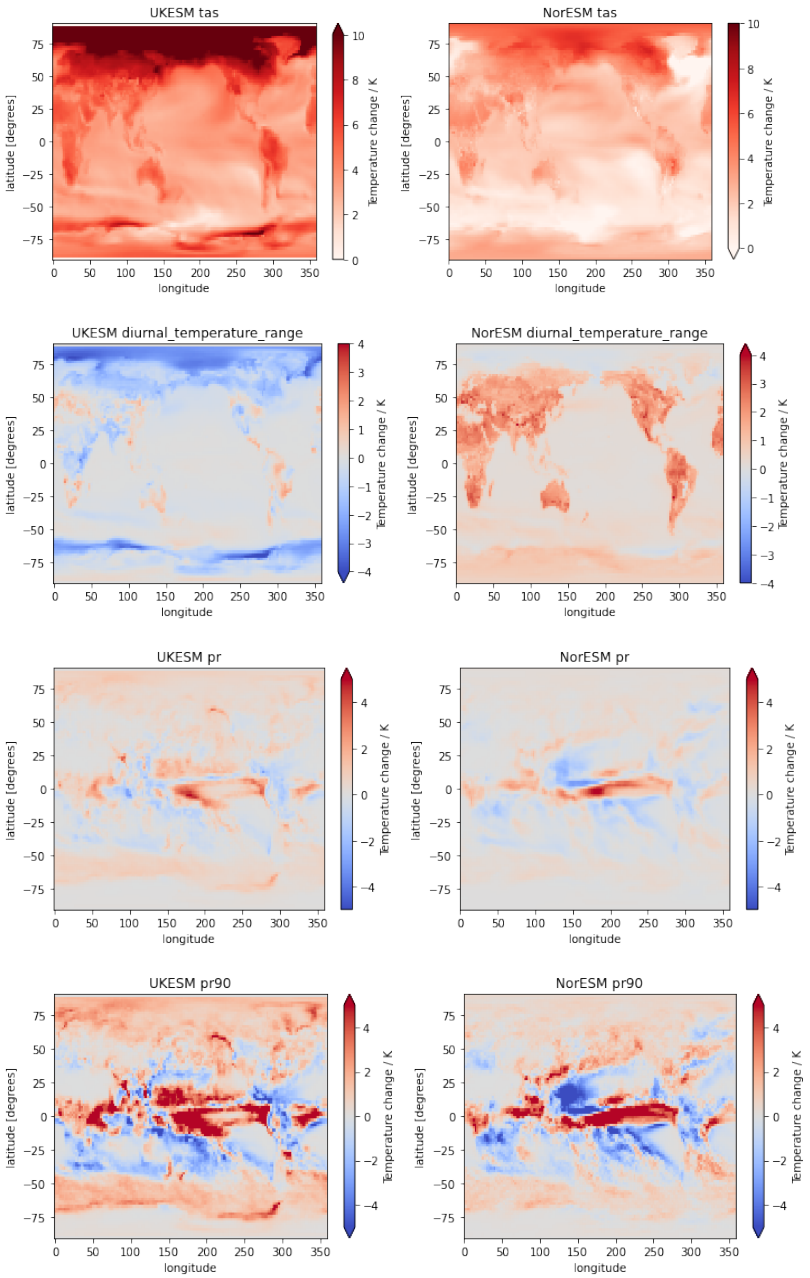


**Figure A2: Comparison of predicted changes in surface air temperature (a); diurnal temperature range (b); precipitation (c); and the 90th percentile of precipitation between UKESM and NorESM2 under SSP245**

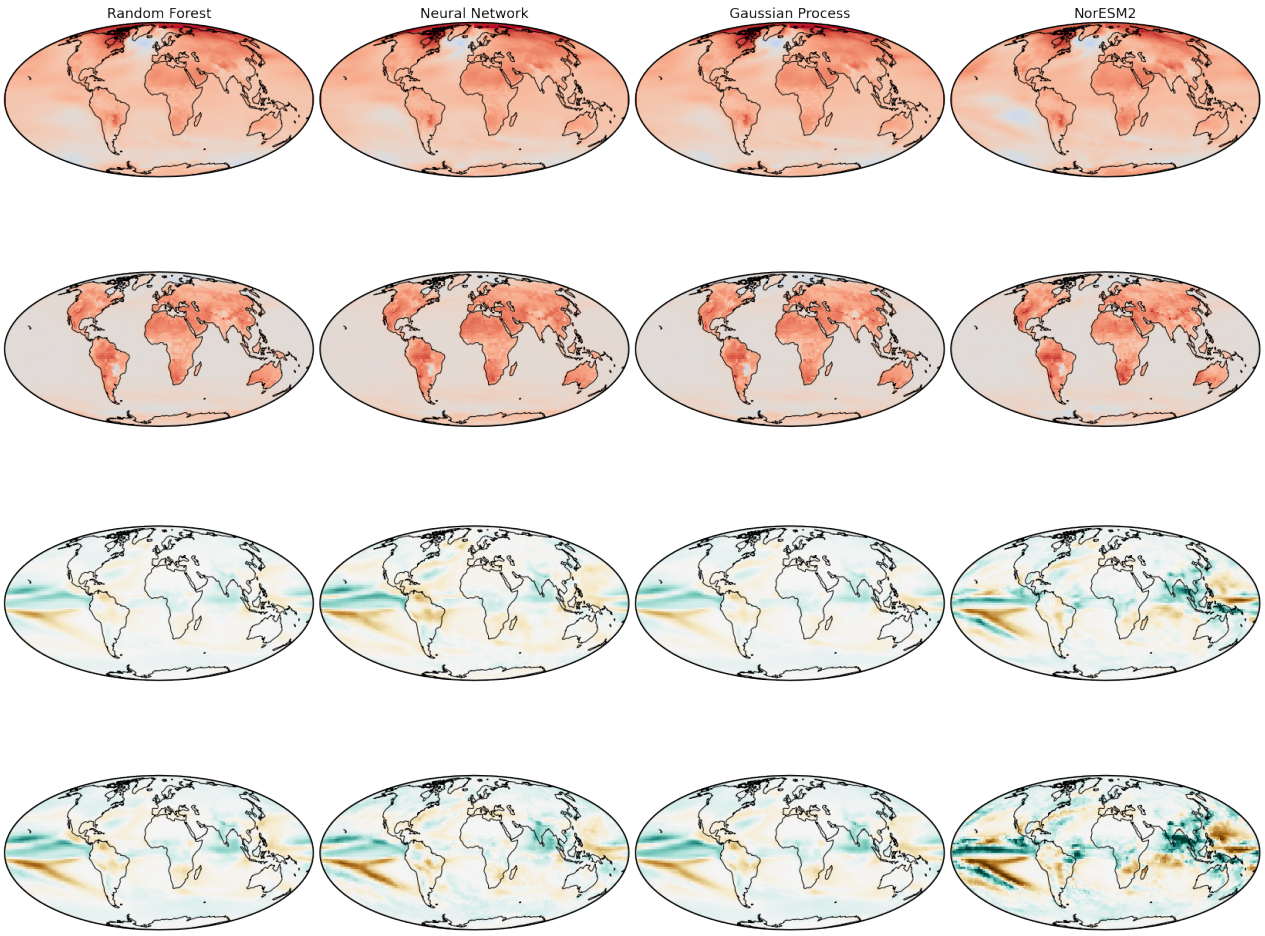


**Figure A3: Maps of ClimateBench target variables for each baseline model and the target NorESM values under the test *ssp245* scenario averaged between 2050-2100.**

## A.1 Gaussian process models specifications

The GP models kernel $k$ have the same form for all four climate response variables

$$k = k_{CO2} + k_{CH4} + k_{\text{BC}} + k_{SO2}$$

where $k_{CO2}$ and $k_{CH4}$ are kernels that respectively take as inputs CO2 and CH4 emissions. $k_{\text{BC}}$ and $k_{SO2}$ are kernels that take as inputs the 5 principal components of BC and SO2 emission maps respectively, each principal component

being rescaled by an independent length scale term. We choose the Matérn-1.5 class of kernel,

$k_X(x,\ x')\ = (1 + \sqrt{3}\ d(x, x'))exp(-\sqrt{3}\ d(x, x'))\ ,$

where $X$ is a general notation for CO2, CH4, BC or SO2, and $d(x, x')$ is a distance between inputs typically given by

$d(x, x')\ =\ \sum_i |x_i - x'_i|/l_i.$

$l_i$ is a length scale associated to the $i^{\text{th}}$ coordinate $x_i$. Global CO2 and CH4 emissions are scalar inputs, hence the corresponding distances only involve one length scale parameter. The principal components decompositions of BC and SO2 emission maps both have 5 coordinates, hence we set each principal component to be a different coordinate with its own length scale parameter. The Matérn-1.5 kernel guarantees that the corresponding GP lies in a space of continuous functions, hence providing regularity to the climate response predictions. We refer the reader to Rasmussen and Williams, 2005, Chapter 4 for more details on the Matérn kernel. Each kernel is multiplied by a variance term ${\sigma_X}^2$ , which rescales the kernel in the above sum and allows to balance relative features importance. Variances and length scales are tuned during the optimization step.

**Table A2 : Hyperparameters for Random Forest Models**

| Hyperparameter | number of trees | min samples split | min samples leaf | maxdepth |
|---|---|---|---|---|
| Surface air temperature | 200 | 10 | 4 | 30 |
| Diurnal temperature range | 250 | 15 | 4 | 45 |
| Precipitation | 250 | 25 | 8 | 55 |
| $90^{\text{th}}$ percentile of precipitation | 300 | 25 | 12 | 85 |