JAMES Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE

10.1029/2021MS002954

Special Section:
Machine learning application to Earth system modeling

Key Points:

- We introduce the first benchmark 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

Correspondence to:
D. Watson-Parris,
duncan.watson-parris@physics.ox.ac.uk

Citation:
Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls, G., et al. (2022). ClimateBench v1.0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, *14*, e2021MS002954. https://doi.org/10.1029/2021MS002954

Received 19 DEC 2021
Accepted 1 SEP 2022

Author Contributions:
Conceptualization: D. Watson-Parris, Y. Rao, P. Nowack
Data curation: D. Watson-Parris, D. Olivié, Ø. Seland
Formal analysis: D. Watson-Parris, S. Bouabid, M. Dewey, E. Fons, J. Gonzalez, P. Harder, K. Jeggle, J. Lenhardt, P. Manshausen, M. Novitasari, L. Ricard, C. Roesch
Funding acquisition: P. Stier

ClimateBench v1.0: A Benchmark 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], and 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, 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, 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

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—the first benchmarking framework based on a suite of Coupled Model Intercomparison Project, AerChemMIP and Detection-Attribution Model Intercomparison Project 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 on simple approximations of the global mean temperature response to a given scenario. Here we present ClimateBench—the first 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.

Project Administration: D. Watson-Parris
Software: D. Watson-Parris
Visualization: D. Watson-Parris
Writing – original draft: D. Watson-Parris
Writing – review & editing: Y. Rao, D. Olivié, Ø. Seland, P. Nowack, G. Camps-Valls, P. Stier, S. Bouabid, M. Dewey, E. Fons, J. Gonzalez, P. Harder, K. Jeggle, J. Lenhardt, P. Manshausen, M. Novitasari, L. Ricard, C. Roesch

1. 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 (Meinshausen et al., 2011; Nicholls et al., 2020; Smith et al., 2018) are physically interpretable and can capture the general non-linear behavior 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., Tebaldi & Arblaster, 2014) 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 & 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 data set 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 90th 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 (CO_2), sulfur dioxide (SO_2; a precursor to sulfate aerosol), black carbon (BC) and methane (CH_4). To enable spatially accurate emulators ClimateBench includes (annual mean): spatial distributions of emissions for the short-lived aerosol species (SO_2 and BC), globally averaged emissions of CH_4, and global cumulative emissions of CO_2.

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 harmonized data set. 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 data set still allows us to explore both scenario uncertainty and internal variability. Further, since even very simple models are able to capture a variety of forcing responses (Smith et al., 2021), there is reason to believe that the response of the models to a given forcing is more consistent than the range of responses (e.g., Richardson et al., 2019). We thus suggest that an emulator that works best for NorESM2 will also have the tendency to perform better in emulating other CMIP models, mainly because the data characteristics are by design similar (CMIP models represent the same physical system). In contrast, variations in the structure of learning algorithms vary more significantly and follow entirely different ways of building a regression model.

As a demonstration of the variety of possible approaches to tackle this benchmark we also introduce three distinct baseline emulators trained and evaluated against ClimateBench. These constitute the first data driven models for the projection of multiple climatic variables and show promising skill in both the global-mean and spatial responses. We discuss the merits and challenges in using each class of (regression) model and hope these provide a useful starting point for researchers wishing to develop more advanced emulators.

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

2. 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 in the data availability statement).

We use a selection of complementary simulations in order to provide as large a training data set 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 data set. 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 data set 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.

ClimateBench also includes a selection of more idealized 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-4x*CO_2 and *1pct*CO_2, respectively. As the name suggests, the *abrupt-4x*CO_2 includes an abrupt quadrupling of CO_2 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 contribution 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 *1pct*CO_2 simulation gradually increases the atmospheric concentration of CO_2 by 1% per year, again with other forcing agents unchanged. While potentially very useful, they are not used in the training of the emulators presented in this work. Two other idealized simulations performed as part of the Detection-Attribution Model Intercomparison Project (DAMIP; Gillett et al., 2016) represent the historical period forced by only CO_2 and other long-lived greenhouse gases (*hist-GHG*), or only anthropogenic aerosol (*hist-aer*). These provide opportunities to train emulators in regions of the input (emissions) space that are at the limits of plausible future scenarios and were used in training the emulators described in Section 4.

Table 1
Details of Post-Processed Simulations Provided as Part of the ClimateBench Data Set

Protocol	Experiment	Period	Notes
ScenarioMIP (O'Neill et al., 2016)	ssp126	2015–2100	A high ambition scenario designed to produce significantly less than 2° warming by 2100.
	ssp245	2015–2100	Designed to represent a medium forcing future scenario. This is the test scenario to be held back for evaluation
	ssp370	2015–2100	A medium-high forcing scenario with high emissions of near-term climate forcers (NTCF) such as methane and aerosol
	ssp370-lowNTCF	2015–2054	Variation of SSP370 with lower emissions of aerosol and their precursors
	ssp585	2015–2100	This scenario represents the high end of the range of future pathways in the IAM literature and leads to a very large forcing of 8.5 $\mathrm{Wm^{-2}}$ in 2100
CMIP6 (Eyring et al., 2016)	historical	1850–2014	A simulation using historical emissions of all forcing agents designed to recreate the historically observed climate
	abrupt-4xCO_2[a]	500 years	Idealized simulation in which CO_2 is abruptly quadrupled. Other forcing agents remain unchanged
	1pctCO_2[a]	150 years	Idealized simulation in which CO_2 is gradually increased by 1%/year. Other forcing agents remain unchanged
	piControl[a]	500 years	Baseline simulation in which all forcing agents remain unchanged
DAMIP (Gillett et al., 2016)	hist-GHG	1850–2014	A historical simulation with varying concentrations for CO_2 and other long-lived greenhouse-gases (only)
	hist-aer	1850–2014	A historical simulation only forced by changes in anthropogenic aerosol
	ssp245-aer	2015–2100	A medium forcing scenario with only changes in anthropogenic aerosol, which provides an alternative test scenario for emulator evaluation

[a] Ancillary data that, while potentially useful, are not used in training the baseline emulators presented here.

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 years) simulation also enables a robust estimation of internal variability of the climate system for those emulators which are able to represent it in future work, as discussed further in Section 5.1.

2.1. Input Variables

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 CO_2 and CH_4 for each of the realistic (historical, ScenarioMIP and DAMIP) experiments from the checksum files provided by the Community Emissions Data System (CEDS) data set (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 CO_2. Some historical and future periods are only provided in 5 yearly

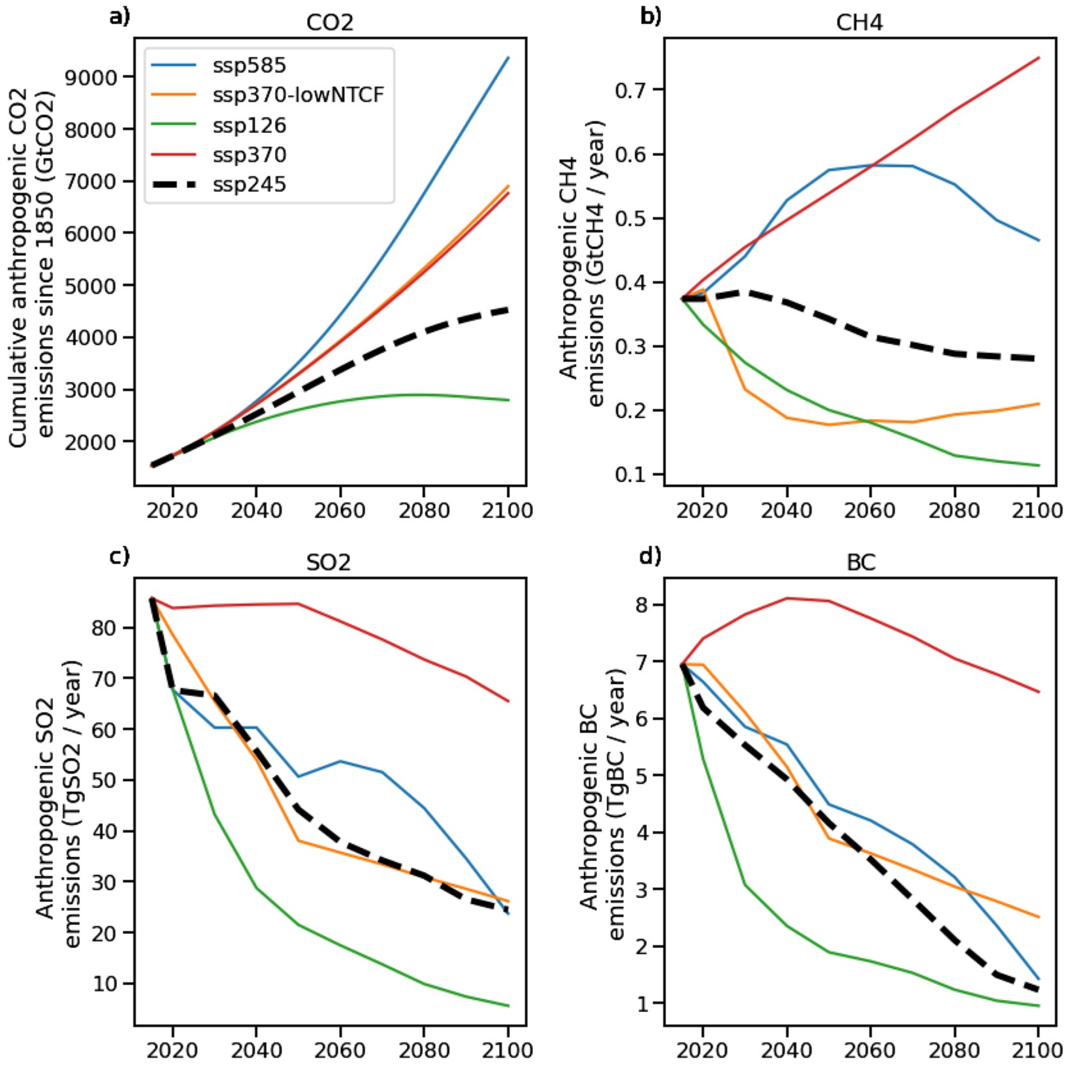

Figure 1. (a) Time series of cumulative anthropogenic carbon dioxide (CO_2) emissions since 1850, (b) emissions of methane (CH_4) (c) global mean emissions of sulfur dioxide (SO_2) and (d) black carbon (BC) derived from NorESM2 ScenarioMIP simulations available within ClimateBench, including the SSP245 test scenario (shown in black).

increments, so we linearly interpolate to yearly values for consistency. The CO_2 emissions are summed cumulatively since, for realistic scenarios, a compensation between forcing efficiency and ocean uptake means the temperature response to CO_2 is approximately linear in the cumulative emissions (Allen et al., 2009; Matthews & Caldeira, 2008). 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.

The aerosol (precursor) emissions are derived from the latest version of the spatially resolved CEDS data set and again summed over sectors and months to produce maps of annual total emissions, as shown in Figure 2 for SO_2 in different years. While the spatial distribution clearly evolves over the historical period and into the future scenarios, the emissions are fairly localized around industrialized 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.

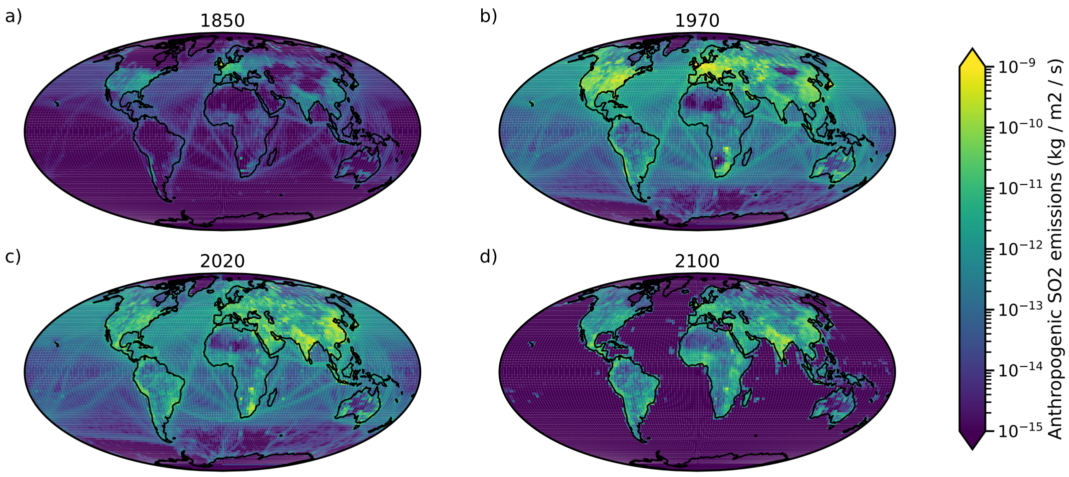

Figure 2. (a) Maps showing the evolution of the spatial distribution of anthropogenic sulfur dioxide (SO_2) emissions in the pre-industrial era represented by 1850, (b) the peak emissions era of 1970, (c) current emissions, (d) and future emissions under Shared Socioeconomic Pathways 245.

For the idealized CMIP simulations (*abrupt-4x*CO_2 and *1pct*CO_2) no emissions files are used and so the cumulative anthropogenic CO_2 emissions are calculated from the difference in the diagnosed CO_2 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).

2.2. Target ESM

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 CO_2) of 2.5 K, particularly compared to the 5.3 K 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). This is, nevertheless, well within the assessed likely range of ECS (90% probability between 2 and 5°C; Forster et al., 2021) and makes the emulation task harder than it might be for other CMIP6 models since the warming signal is weaker at the end of SSP245 (by which point CO_2 is the dominant forcing), as shown in Figure A4. The combination of a weak ECS with a relatively strong aerosol forcing (-1.36 W m^{-2} for 1850 to 2014), likely accounts for the somewhat anomalous cooling between 1950 and 1980 in the historical simulations (Seland et al., 2020), although it has been noted that the combined anthropogenic response in NorESM is realistic (Gillett et al., 2021).

2.3. Output Variables

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 of the difference in daily maximum and minimum surface air temperatures: $|\mathrm{tasmax} - \mathrm{tasmin}|_{\mathrm{ann}}$. The PR90 is calculated as the 90th percentile of the daily precipitation in each year. The annual mean baseline values (from the full *piControl* simulation) 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 data set might choose to do this with certain emulators. Many of the NorESM2 simulations include 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. These are included to allow (optional) emulation of internal variability.

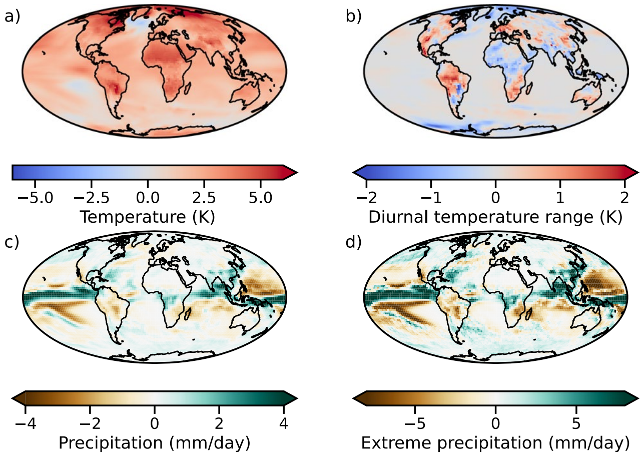

Figure 3. Maps of target outputs from the SSP245 held-back test scenario at 2100 (as an anomaly to the pre-industrial control run) performed by NorESM2: (a) Annual mean surface temperature; (b) annual mean diurnal surface temperature range; (c) annual mean precipitation; and (d) 90th percentile of the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

Samples of these output fields from the target ssp245 data set are shown in Figure 3. The relative increase in warming in the northern polar regions (known as Arctic amplification) is clearly seen in Figure 3a, as well as the north Atlantic warming hole (Drijfhout et al., 2012; Manabe & Stouffer, 1993; Woollings et al., 2012), 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 Figures 3c and 3d 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).

Also included in the data set 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 normalizing the inputs by their efficacy or developing more physically interpretable emulators that derive the climate response via the forcing, but these are not used in the present study.

3. Benchmark Task

The task defined by ClimateBench is the prediction of the output variables described in Section 2.3 using only the inputs available from Section 2.1 under the chosen test scenario—ssp245. Emulators may choose to use as much or as little of the data presented in Table 1 in order to train their models as appropriate for a given approach. They may also choose to predict the contemporaneous response to emissions (as used in our RF and GP baseline emulators), account for a lagged response (as in our baseline NN emulator), or even predict the full time-series simultaneously.

3.1. Evaluation Metrics

The evaluation criteria are a crucial aspect to any benchmark data set 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 global mean changes in temperature and precipitation are key climatic variables but the spatial characteristics of the outputs in this task also need to be considered if the emulators are to be used for regional projections. As a primary metric we choose to combine the normalized, global mean root-mean square error (NRMSE_s) and the NRMSE in the global mean (NRMSE_g), calculated following:

$$\mathrm{NRMSE}_s = \sqrt{\left\langle \left(|x_{i,j,t}|_t - |y_{i,j,t,n}|_{n,t} \right)^2 \right\rangle} / | \langle y_{i,j} \rangle |_{t,n} \tag{1}$$

$$\mathrm{NRMSE}_g = \sqrt{|(\langle x_{i,j,t} \rangle - \langle |y_{i,j,t,n}|_n \rangle)^2|_t} / | \langle y_{i,j} \rangle |_{t,n} \tag{2}$$

$$\mathrm{NRMSE}_t = \mathrm{NRMSE}_s + \alpha * \mathrm{NRMSE}_g, \tag{3}$$

where the global mean denoted $\langle x_{i,j} \rangle$ includes a weighting function that accounts for the decreasing grid-cell area toward the poles and is defined as: $\langle x_{i,j} \rangle = \frac{1}{N_{\mathrm{lat}} N_{\mathrm{lon}}} \sum_i^{N_{\mathrm{lat}}} \sum_j^{N_{\mathrm{lon}}} \cos(\mathrm{lat}(i)) x_{i,j}$, and α is a coefficient empirically chosen to be 5 so that each component provides roughly equal weight.

Combining these commonly used metrics in this way provides a single number summarizing the mismatch between the predictions (x) and the target variables (y). By squaring the difference, the RMSE also weighs large discrepancies more heavily, penalizing larger errors. We average the target variables over the three available ensemble members (n) and a relatively long period of the target scenario (2080–2100) in order to minimize the contribution of internal variability. We choose the final years of the century since the start of the *ssp245* is quite similar to some of the training scenarios. We normalize the RMSEs so that the metrics are broadly comparable across the target variables.

Estimates of this internal variability can be very valuable for climate projections however and since ClimateBench includes three ensemble members for each training data set 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):

$$\mathrm{CRPS} = \int_{x=-\infty}^{x=\infty} (\langle F_{i,j,t}(x)\rangle - \langle F_{i,j,t}(y)\rangle)^2 dx, \tag{4}$$

where $F(x)$ and $F(y)$ 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. It should be noted that the relatively low number of ensemble members available in ClimateBench will likely underestimate full internal variability and a larger ensemble (e.g., 100 members in Rodgers et al., 2021) should be used for robust estimation. Indeed, the formulation above only includes variability in the global mean since such small ensembles are unlikely to capture regional variability. Methods to calculate both metrics based on the *climpred* (Brady & Spring 2021) package are provided in the example notebooks included with the data set. 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.

3.2. Baseline Evaluation

Before evaluating some baseline statistical emulators, it is useful to consider two cases with which we hope to place the skill of the data-driven approaches in a broader context. 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. This is estimated as the standard deviation across the three NorESM2 ensemble members in *ssp245*. 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 second is a comparison against the inter-model spread encountered within CMIP6 for the variables of interest which, despite (as discussed above) not providing a robust model uncertainty, represents a lower bound on the accuracy we would like our emulators to achieve.

Additionally, we introduce a linear pattern scaling model which uses independent linear regressions of each of the output variables at each model grid cell given the global mean temperature response to the emissions (e.g., Tebaldi & Arblaster, 2014). This approach is somewhat simpler than the other data driven models since it assumes access to an accurate impulse response (or box) model to determine the global mean temperature but provides a useful baseline. We train the regression models using the same training output data as the other emulators (described in the next section) but the only input is the global mean temperature. We assume this is available at prediction time as well so that this constitutes a "perfect" pattern scaling approach.

4. Baseline Emulators

Three baseline emulators are developed to demonstrate various potential approaches to tackling the machine learning problem this data set poses. These are performed using the Earth System Emulator (ESEm; 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 Section 5). All three emulators are trained using all the available training data: the historical data; ssp126; ssp370; ssp585; and the historical data with aerosol (hist-aer) and greenhouse gas (hist-GHG) forcings only, leading to 754 training/validation points (which are nevertheless not fully independent). More details on emulator specific data pre-processing, training procedure and results are described in each of the following subsections.

The emulators all perform skilfully, as summarized in Table 2 and Figure 4. 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. They also tend to overpredict northern-hemisphere warming while

Table 2

The Spatial, Global and Total NRMSE of the Different Baseline Emulators for the Years 2080–2100 Against the ClimateBench Task of Estimating Key Climate Variables Under Future Scenario SSP245

	NRMSE surface air temperature (1)			NRMSE diurnal temperature range (1)			NRMSE precipitation (1)			NRMSE 90th percentile precipitation (1)		
	Spatial	Global	Total	Spatial	Global	Total	Spatial	Global	Total	Spatial	Global	Total
Gaussian Process	0.109	0.074	0.478	9.207	2.675	22.582	2.341	0.341	4.048	**2.556**	0.429	4.702
Neural Network	**0.107**	**0.044**	**0.327**	9.917	**1.372**	**16.778**	**2.128**	**0.209**	**3.175**	2.610	**0.346**	**4.339**
Random Forest	0.108	0.058	0.400	**9.195**	2.652	22.457	2.524	0.502	5.035	2.682	0.543	5.399
Pattern Scaling	0.080	0.048	0.320	8.083	2.327	19.719	2.006	0.331	3.662	2.400	0.412	4.461
Variability	0.052	0.072	0.414	2.513	1.492	9.973	1.350	0.268	2.691	1.757	0.457	4.043
CMIP6	0.258	0.177	1.141	1.962	0.799	5.958	1.994	0.389	3.940	–	–	–

Note. The best (lowest) emulator scores for each task are highlighted in bold. The normalized standard deviation in each variable over 22 different CMIP6 models and across the NorESM ensemble members are also included as indications of inter-model and internal variability, respectively. Bold figures represent the lowest value (best performance) for each metric.

underpredicting warming elsewhere. This might suggest that these particular changes are driven by different climate forcers or longer time-scale changes than modeled in this study. A direct comparison of the emulator predictions and NorESM is shown in Figure A2. Overall, the neural network performs the best in predicting temperature and precipitation changes, but is also the most complex emulator.

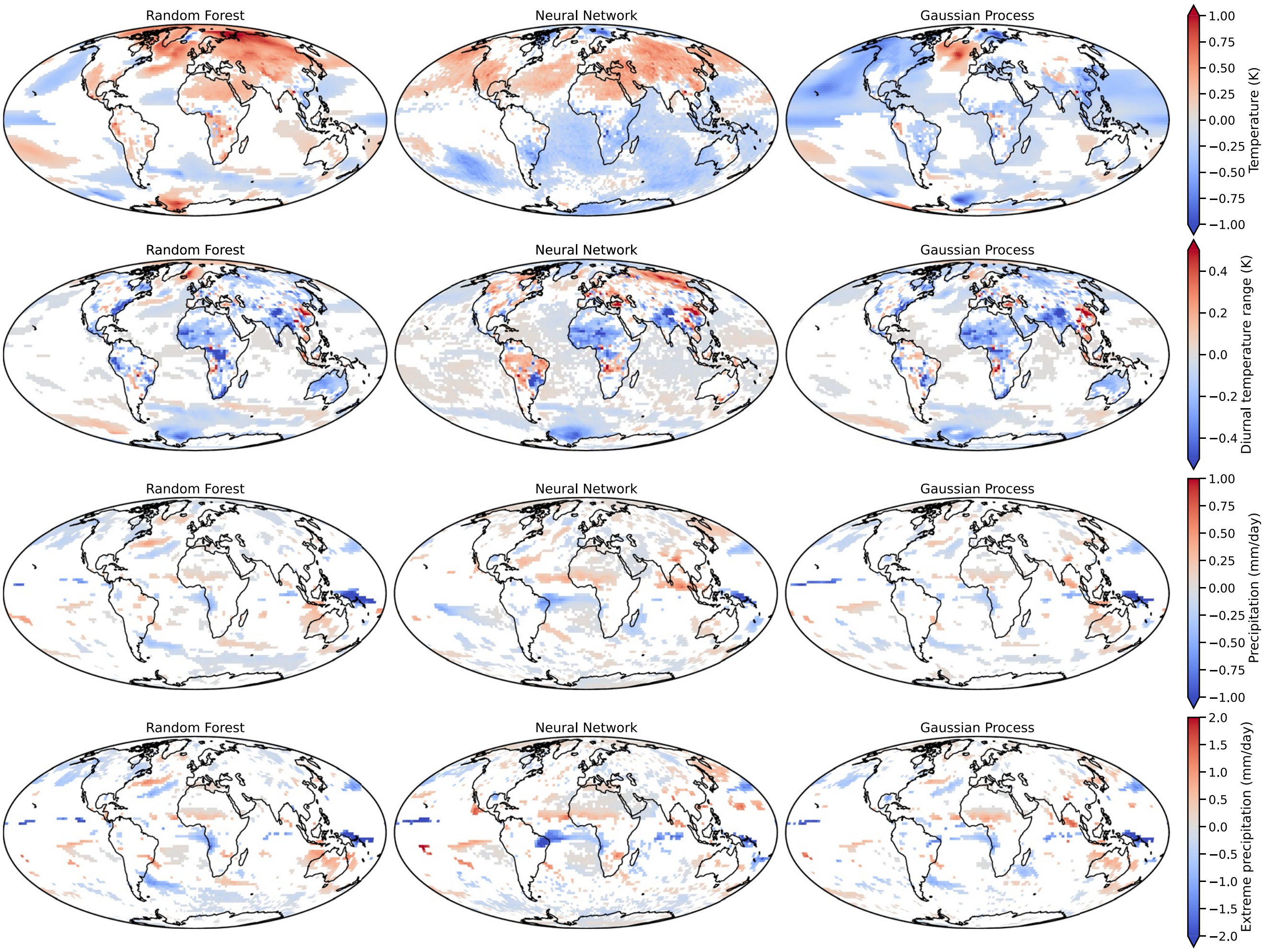

Figure 4. Maps of the mean difference in the ClimateBench target variables for each baseline emulator against the target NorESM values under the test ssp245 scenario averaged between 2080 and 2100. Differences insignificant at the $p < 5\%$ level are masked from the plots.

4.1. Gaussian Process Regression

Gaussian processes (GPs) (Rasmussen & Williams, 2005) are probabilistic models which assume predictions can be modeled 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: (a) infer the emulator parameters (typically only the noise variance and the kernel parameters) by maximizing the log-likelihood of the observations under the evidence; and then (b) allow to obtain its posterior probability distribution that is used to make predictions over unseen inputs.

To prepare the input samples, the dimensionality of the SO_2 and BC emission maps are reduced with principal component analysis, and we only use the five first principal components of each as inputs, corresponding to 96% and 98% of the explained variance, respectively. 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 SO_2 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 constant 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, once differentiable function; details are provided in Section A2. The mean value, kernels parameters and internal variability variance are jointly tuned against the training data by marginal likelihood maximization with the limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) optimisation algorithm. The emulators used have 18 parameters in total: Five lengthscale and one variance parameter for each aerosol kernel; one lengthscale and one variance for each of the GHG kernels; one mean and one likelihood variance.

As reported in Table 2, the total NRMSE of the mean predictions with the GPs are competitive with the neural network for all the variables. This is remarkable given the limited number of parameters that are learned. 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 CO_2 emissions generally influence all predictions, and unequivocally dominate the predictions for surface air temperature and diurnal temperature range. CH_4 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 A2). While we are unable to compare these scores with the other baseline methods the similarity to the global NRMSE indicates that the GP is also predicting the internal variability accurately (otherwise it would be penalized in the CPRS relative to the NRMSE).

4.2. Random Forests

Random forests aggregate predictions of multiple decision trees (Breiman, 2001; Ho, 1995). 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 modeling 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.

Analogously to the GP emulator, the dimensionality of aerosol emission maps is reduced with principal component analysis. The first five principal components of SO_2 and BC together with the global emission maps of CO_2 and CH_4 form the input features of the model. Separate random forest emulators are trained for the four target variables. The following hyperparameters are tuned using random search of the training data without replacement: 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 A3.

As shown in Table 2, the spatial NRMSE scores of the random forest regressors are comparable to the performance of the other emulators for all variables but the global NRMSE is significantly worse for temperature and precipitation (as can also be seen in Figure 6). Discontinuities in the predicted global mean temperature change time series over this period (not shown) perhaps indicate a deeper tree structure is required. 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 CO_2 concentrations dominate the predictions. For temperature predictions the other featsures are negligible. SO_2 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 *ssp245* test scenario. The regional influences may be more significant however and this will be explored separately.

4.3. Neural Networks

Artificial Neural Networks (ANNs) are algorithms inspired by the biological neural networks of human brains that have shown outstanding success in areas like Computer Vision and Natural Language Processing. Two major ANN architectures are Convolutional Neural Networks (CNNs) (Le Cun et al., 1990), that are able to model spatial dependencies, and Recurrent Neural Networks (RNNs), that are able to process time series and sequential data. ANNs have recently been employed to tackle a variety of problems in earth system science (Camp-Valls et al., 2021). CNNs are helpful for modeling climate data with a spatial structure, for instance, precipitation

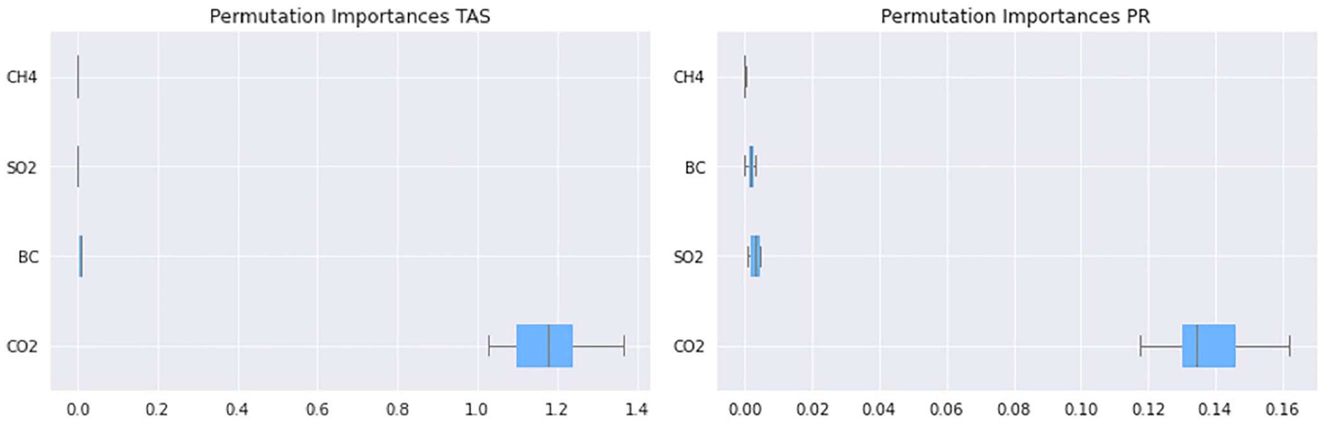

Figure 5. Permutation importances for the most important component of each variable in predicting global mean temperature (TAS) and precipitation (PR). Each emulator input variable is shuffled in turn to determine the relative contribution to prediction skill. Note that these average estimates do not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

patterns or satellite imagery, and are frequently applied in climate science and weather forecasting (Harder et al., 2020; Trebing et al., 2021). Long short-term memory (LSTM) networks (Hochreiter & Schmidhuber, 1997), an advanced type of RNNs, have proven skillful for modeling climate time series, for example, for the prediction El Niño-Southern Oscillation (Broni-Bedaiko et al., 2019).

For time series of spatial variables, as in the ClimateBench data set, we can use the two types of networks in sequence to model both spatial and temporal dependencies. The chosen architecture consists of a CNN followed by an LSTM built with the Keras library. The CNN includes one convolutional layer with 20 filters, a filter size of 3, and a ReLU activation function. The 3 × 3 pixel filters scan the input images to detect spatial patterns and feed these patterns to the next layer. These next layers are average pooling layers that reduce the spatial dimensionality ahead of the LSTM layer. The LSTM uses 25 units (i.e., the output dimension of each LSTM cell) and a ReLU activation function. The LSTM is followed by a dense layer and reshaping layer to (96, 144), that is, the (latitude, longitude) dimension of the output variables.

The training data time-series is segmented into 10-year chunks, using a moving-time window in one-year increments, leading to 754 training samples of shape (10, 96, 144, 4) corresponding to the number of years, latitude, longitude and then number of variables. We trained four different emulators for the four different output variables. Each emulator is trained for 30 epochs, using a batch size of 16. For this baseline approach, we chose not to do any hyperparameter optimization, and all the parameters were chosen manually.

RMSE scores obtained with the CNN-LSTM architecture are somewhat better than those achieved with the other methods, particularly in the global-mean. This might be because the LSTM is able to better capture the temporal autocorrelation than the other emulators which treat the prediction instantaneously. The CNN-LSTM architecture also captures spatial changes in temperature well (e.g., the Arctic amplification), even though warming at the poles is somewhat underestimated. In general, warming in the Northern hemisphere is overestimated while it is underestimated in the Southern Hemisphere. Given the overestimated temperature response in the *ssp245-aer* simulations shown in Figure A3, this may be due to an overestimation of the effect of aerosol on the temperature by this emulator. The diurnal temperature range is well predicted, with a lower performance over land. The CNN-LSTM also captures spatio-temporal changes in precipitation (e.g., the ICTZ shift) quite well.

5. Discussion

5.1. 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 suited 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 Figure 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 idealized 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 data set is the relatively high dimensional inputs and outputs (96 × 144). Most approaches to emulating the regional temperature response to a CO_2 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 toward 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., 2020).

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 & Gal, 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).

As previously discussed, and shown in Figure A4, there is large inter-model variability in the projected climate variables in CMIP6, even across a single scenario. Future work should explore the ability of a given emulator to robustly recreate each of these model responses, and could allow a deeper understanding of their discrepancies.

5.2. 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.

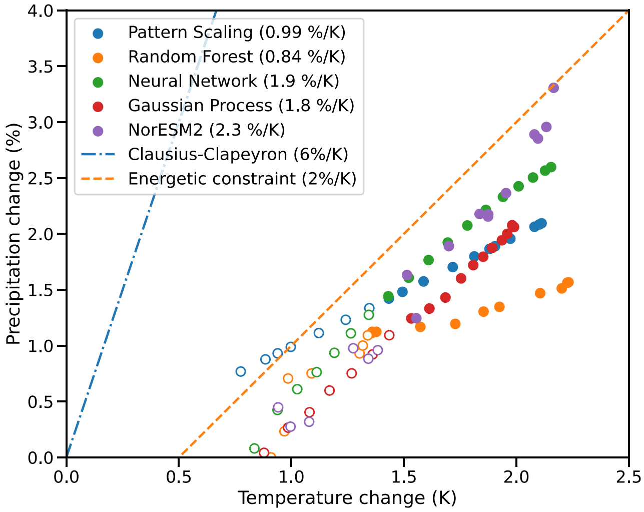

Figure 6. The relative change in global mean precipitation as a function of global mean temperature change in the baseline emulators and NorESM2 averaged in 5 year increments to reduce internal-variability. Hollow and solid points indicate years before and after 2050 respectively. The change predicted by the Clausius-Clapeyron relationship and energy conservation considerations are shown as dashed lines.

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 & Ingram, 2002; Dagan et al., 2019; Jeevanjee & Romps, 2018; Pendergrass & Hartmann, 2013). 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 over the range of their training data. 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.

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 (Barnes et al., 2020), while the latter provide mechanisms to determine post-hoc understanding (McGovern et al., 2019). While not as robust as physical laws, these techniques provide useful indications that such models are getting the right answer for the right reasons. Indeed, the physical ESMs currently considered the "gold standard" of climate modeling are often only interpretable or explainable by expert practitioners and it is hoped that (interpretable) ClimateBench emulators will be useful in analyzing and understanding the response of the underlying physical models themselves.

5.3. 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 modeling whereby statistical or ML based emulators embed physical equations, constraints or symmetries in order to improve accuracy, robustness and generalizability (Camps-Valls et al., 2021; Karpatne et al., 2017; Reichstein et al., 2019). 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 modeling 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 minimize 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 maximized), 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 minimize this risk.

5.4. 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 (Barnett et al., 2005; Otto et al., 2016; Shindell et al., 2009; Stott et al., 2010, 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 for example, approximate Bayesian computation, variational inference or Markov Chain Monte-Carlo sampling. Future work will investigate these possibilities.

As a simple demonstration of the potential of such an approach we have included a prediction by the emulators compared to the original NorESM2 simulations of the *ssp245-aer* DAMIP experiment in which only the aerosol species are emitted, shown in Figure A3. This is a more challenging scenario than the *ssp245* test case due to the much smaller total forcing, and the emulators do not perform as well (see NRMSE in Table A1). It is interesting to note that the emulators particularly struggle with temperature changes in the North Atlantic where slow ocean circulation changes (e.g., Dagan et al., 2020) may not be fully captured. They nevertheless capture the main features of the response and show promise for future work disentangling the forcings and feedbacks in NorESM2, other ESMs and ultimately observations.

6. 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 data set in order to galvanize 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).

We also introduce three quite distinct approaches for undertaking this challenge: a random forest; a Gaussian process; and a neural network model. These different models are based on different principles, have distinct assumptions and rely on quite different learning paradigms. Each has their strengths and weaknesses but all perform well in the evaluation metrics and generally reproduce the NorESM2 temperature and precipitation response well in a realistic (but unseen) future scenario, especially compared to CMIP6 inter-model diversity. The neural network model performs best overall and shows good skill both in the global mean and spatially. All the models perform less well in the aerosol only test, suggesting that they have not fully learned the distinct response due to each forcer and future emulators should aim to rectify this.

Current impact assessments are often based on simple emulators, which are then scaled to match modeled patterns, but which are unable to predict non-linear responses in for example, 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 data set 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.

Appendix A1: Additional Material

The following figures and two tables show additional details in support of the main text.

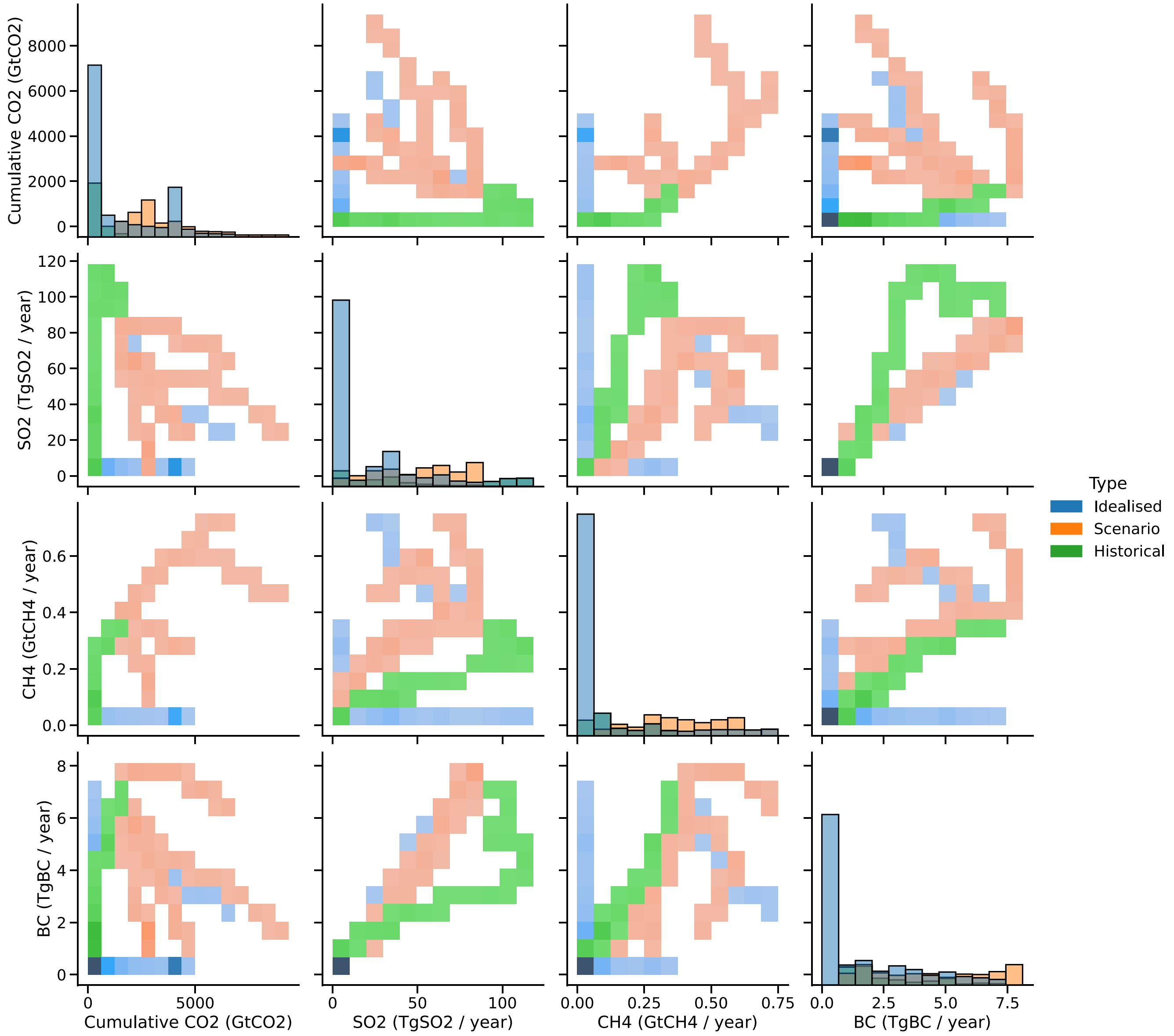

Figure A1. Joint and marginal distributions of annual global mean emissions and concentrations across the ClimateBench training data set. Input datasets are classified as Idealized (such as *1pct*CO_2 and *abrupt4x*CO_2, and including *ssp370-lowNTCF*), historical and scenario to demonstrate the contribution of each to sampling the full input space.

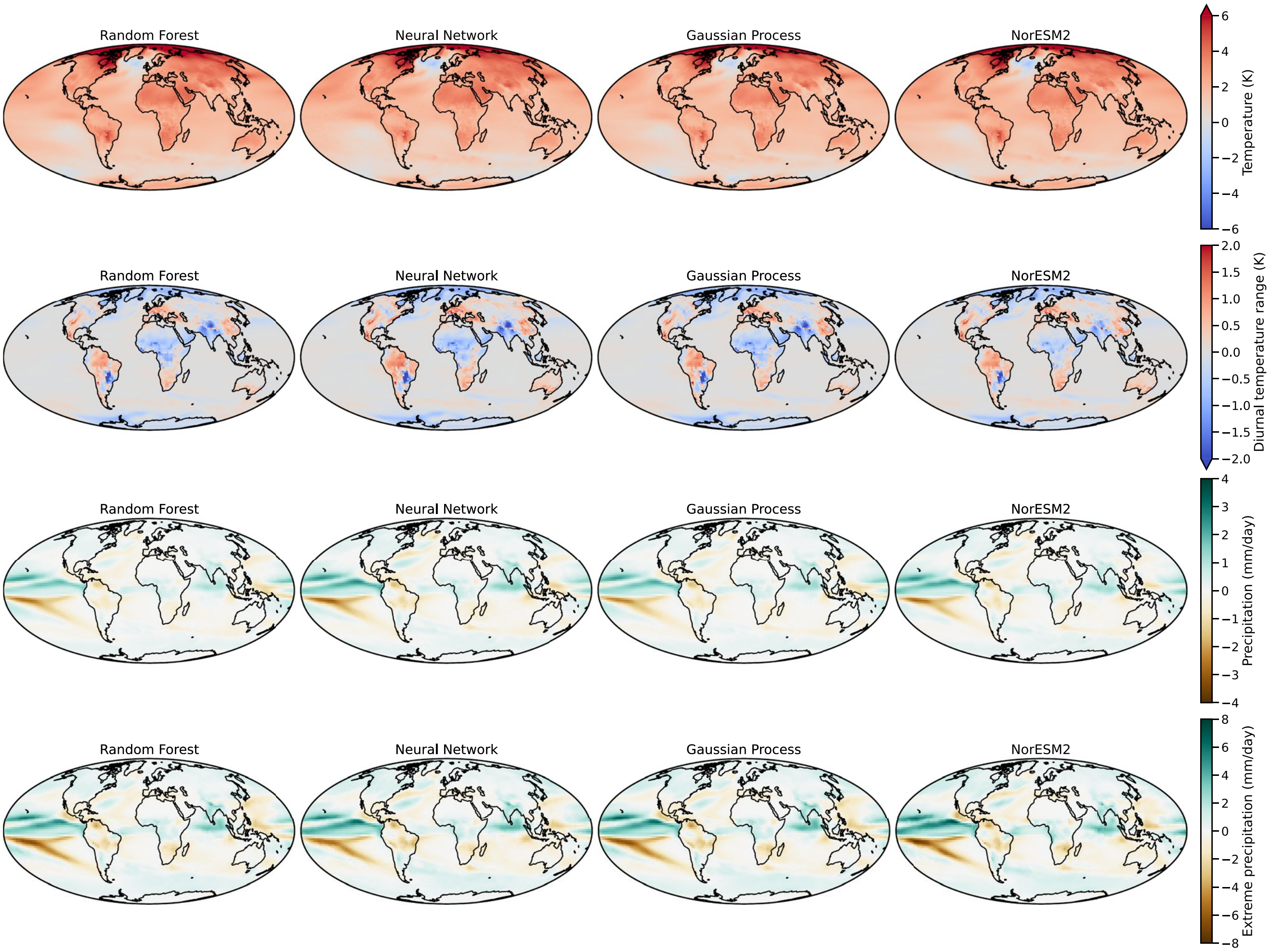

Figure A2. Maps of ClimateBench target variables for each baseline model and the target NorESM values under the test *ssp245* scenario averaged between 2080 and 2100.

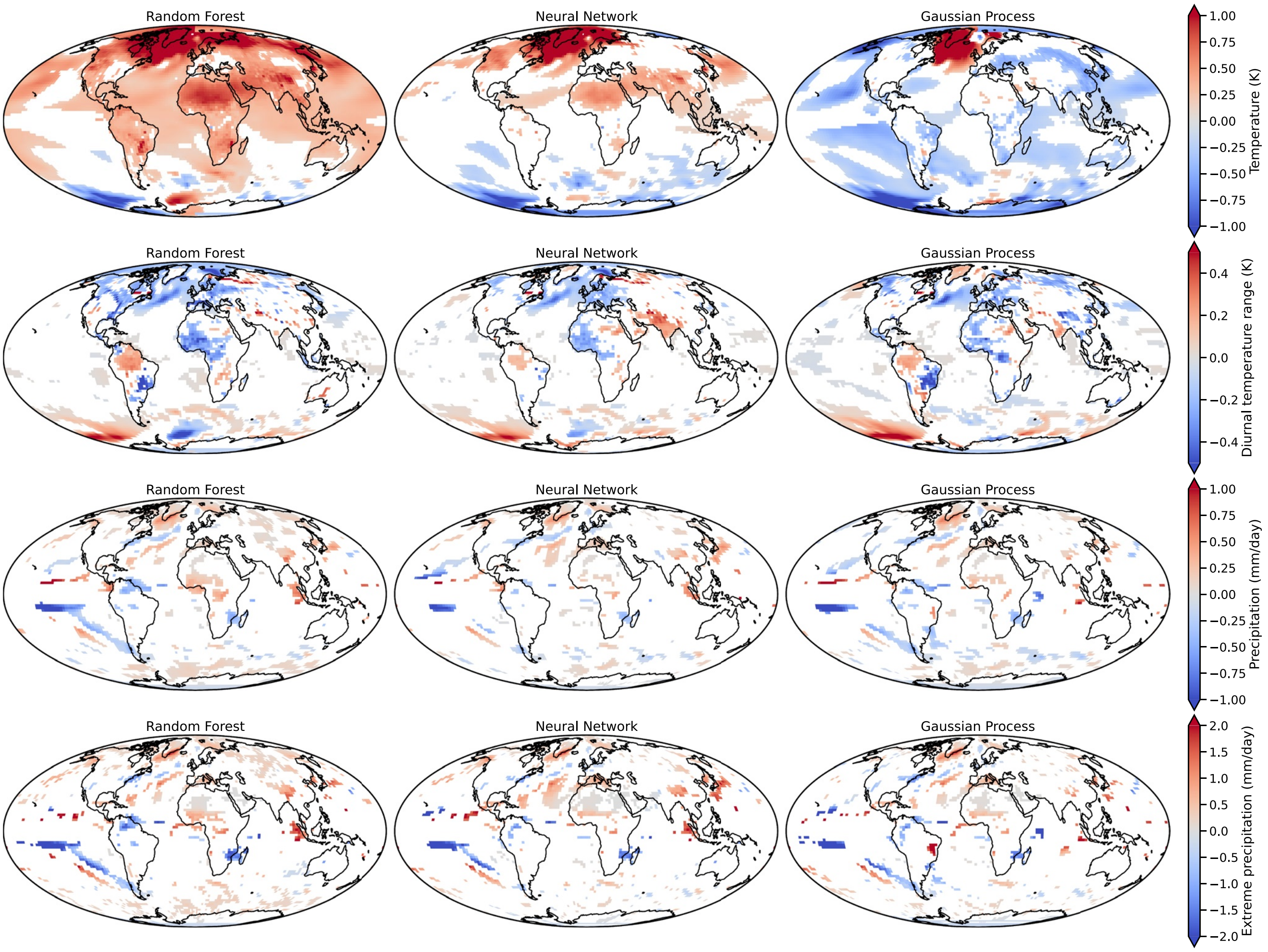

Figure A3. Maps of the mean difference in the ClimateBench target variables for each baseline emulator against the target NorESM values under the test ssp245-aer scenario averaged between 2080 and 2100. Differences insignificant at the $p < 5\%$ level are masked from the plots.

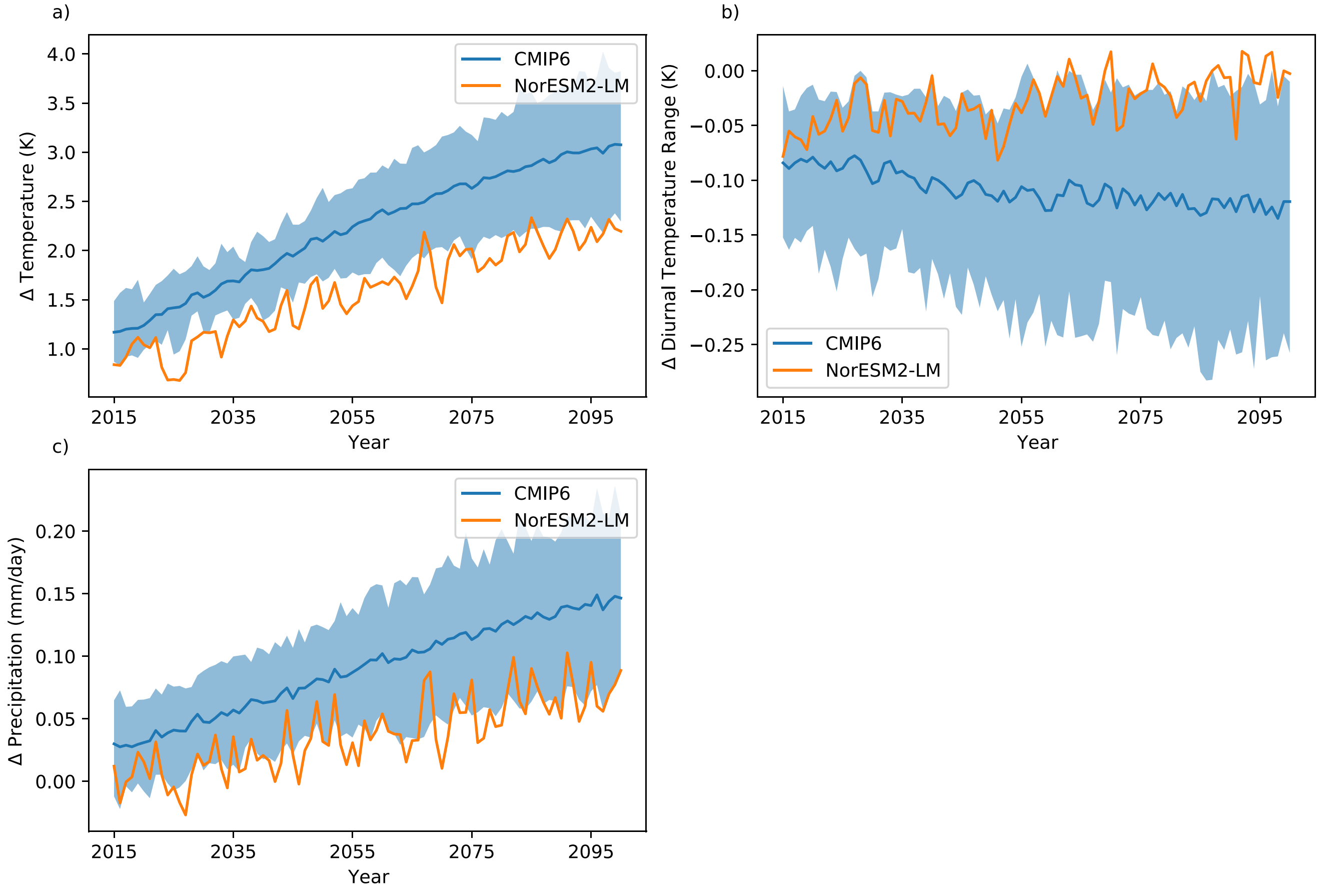

Figure A4. Global mean NorESM-LM projections under *ssp-245* as compared to all other available CMIP6 models for three of the target variables.

Table A1
The Spatial, Global and Total NRMSE of the Different Baseline Emulators for the Years 2080–2100 Against the ClimateBench Task of Estimating Key Climate Variables Under the Idealized Future Scenario SSP245-AER

	NRMSE surface air temperature (1)			NRMSE diurnal temperature range (1)			NRMSE precipitation (1)			NRMSE 90th percentile precipitation (1)		
	Spatial	Global	Total	Spatial	Global	Total	Spatial	Global	Total	Spatial	Global	Total
Gaussian Process	2.138	1.165	7.963	14.298	2.868	28.636	12.100	**0.933**	16.767	13.486	**1.353**	20.252
Neural Network	**2.116**	**1.011**	**7.173**	**12.387**	**2.200**	**23.386**	**10.316**	0.977	**15.199**	**12.224**	1.438	**19.414**
Random Forest	2.977	2.041	13.182	16.222	3.284	32.642	11.562	1.291	18.017	12.302	1.616	20.382

Note. Bold figures represent the lowest value (best performance) for each metric.

Table A2
The Continuous Ranked Probability Score for the Gaussian Process Emulator for the Years 2080–2100 Against the ClimateBench Task of Estimating Key Climate Variables Under Future Scenario SSP245

	CRPS surface air temperature (K)	CRPS diurnal temperature range (K)	CRPS precipitation (mm/day)	CRPS 90th percentile precipitation (mm/day)
Gaussian Process	0.4765	0.3601	1.0753	1.0029

Appendix A2: Gaussian Process Model Specifications

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

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

where k_{CO2} and k_{CH4} are kernels that respectively take as inputs $\mathrm{CO_2}$ and $\mathrm{CH_4}$ emissions. k_{BC} and k_{SO2} are kernels that take as inputs the 5 principal components of BC and $\mathrm{SO_2}$ emission maps respectively, each principal component being rescaled by an independent length scale term. We choose the Matérn-1.5 class of kernel,

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

where X is a general notation for $\mathrm{CO_2}$, $\mathrm{CH_4}$, BC or $\mathrm{SO_2}$, and $d(x, x')$ is a distance between inputs typically given by

$$d(x, x') = \sum_i \left|x_i - x_i'\right| / l_i$$

l_i is a length scale associated to the i^{th} coordinate x_i. Global $\mathrm{CO_2}$ and $\mathrm{CH_4}$ emissions are scalar inputs, hence the corresponding distances only involve one length scale parameter. The principal components decompositions of BC and $\mathrm{SO_2}$ 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 & Williams, 2005, Chapter 4 for more details on the Matérn kernel. Each kernel is multiplied by a variance term σ_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.

Appendix A3: Random Forest Model Specification

Hyperparameter	Number of trees	Min samples split	Min samples leaf	Maxdepth
Surface air temperature	250	5	7	5
Diurnal temperature range	150	15	8	40
Precipitation	250	15	12	25
90th percentile of precipitation	300	10	12	20

Appendix A4: Neural Network Model Specification

The parameters are the same for all four models.

Model Architecture			
Layer	Hyperparameter value (if not specified, the default parameters are used)	Output shape	Param #
Time distributed Conv2D	Number of filters: 20 Filter size: 3 Activation function: ReLu	(None, 10, 96, 144, 20)	740
Time distributed AveragePooling2D	Pool size: 2	(None, 10, 48, 72, 20)	0
Time distributed GlobalAveragePooling2D		(None, 10, 20)	0
LSTM	Number of units: 25 Activation function: ReLu	(None, 25)	4,600
Dense	Units: 96 × 144	(None, 13824)	359424
Activation	Activation function: linear	(None, 13824)	0
Reshape		(None, 1, 96, 144)	0

Model Training

Hyperparameter	Value
Batch size	16
Epochs	30
Optimizer	Rmsprop
Metric	MSE

Data Availability Statement

The baseline models, evaluation metrics and all code used to generate the plots in this paper are available here: https://doi.org/10.5281/zenodo.7064302. 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 for example,: https://esgf-index1.ceda.ac.uk/search/cmip6-ceda/.

Acknowledgments
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). YR was supported by NOAA through the Cooperative Institute for Satellite Earth System Studies under Cooperative Agreement NA19NES4320002. This research was supported, in part, by the National Science Foundation under Grant No. NSF PHY-1748958. 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). We would also like to thank the many other participants of the third NOAA AI Workshop hackathon who provided valuable feedback.

References

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

Allen, M. R., & Ingram, W. J. (2002). Constraints on future changes in climate and the hydrologic cycle. *Nature*, *419*(6903), 224–232. https://doi.org/10.1038/nature01092

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2020). Indicator patterns of forced change learned by an artificial neural network. *Journal of Advances in Modeling Earth Systems*, *12*(9). https://doi.org/10.1029/2020ms002195

Barnett, T., Zwiers, F., Hengerl, G., Allen, M., Crowly, T., Gillett, N., et al. (2005). Detecting and attributing external influences on the climate system: A review of recent advances. *Journal of Climate*, *18*(9), 1291–1314. https://doi.org/10.1175/jcli3329.1

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., & Gentine, P. (2021). Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, *126*(9), 098302. https://doi.org/10.1103/physrevlett.126.098302

Beusch, L., Gudmundsson, L., & Seneviratne, S. I. (2020). Emulating Earth system model temperatures with MESMER: From global mean temperature trajectories to grid-point-level realizations on land. *Earth Syst Dynam*, *11*(1), 139–159. https://doi.org/10.5194/esd-11-139-2020

Beusch, L., Nicholls, Z., Gudmundsson, L., Hauser, M., Meinshausen, M., & Seneviratne, S. I. (2021). 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 Development Discussions*, 1–26. https://doi.org/10.5194/gmd-2021-252

Bollasina, M. A., Ming, Y., Ramaswamy, V., Schwarzkopf, M. D., & Naik, V. (2014). Contribution of local and remote anthropogenic aerosols to the twentieth century weakening of the South Asian monsoon: Aerosols and South Asian monsoon. *Geophysical Research Letters*, *41*(2), 680–687. https://doi.org/10.1002/2013gl058183

Brady, R., & Spring, A. (2021). CLIMPRED: Verification of weather and climate forecasts. *Journal of Open Source Software*, *6*(59), 2781. https://doi.org/10.21105/joss.02781

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Broni-Bedaiko, C., Katsriku, F. A., Katsriku, F., Unemi, T., Atsumi, M., Abdulai, J.-D., et al. (2019). El Niño-Southern Oscillation forecasting using complex networks analysis of LSTM neural networks. *Artificial Life and Robotics*, *24*(4), 445–451. https://doi.org/10.1007/s10015-019-00540-2

Bueso, D., Piles, M., & Camps-Valls, G. (2020). Nonlinear PCA for spatio-temporal analysis of Earth observation data. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(8), 5752–5763. https://doi.org/10.1109/tgrs.2020.2969813

Cabré, M. F., Solman, S. A., & Nuñez, M. N. (2010). 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*(3–4), 449–469. https://doi.org/10.1007/s10584-009-9737-5

Camps-Valls, G., Verrelst, J., Munoz-Mari, J., Laparra, V., Mateo-Jimenez, F., Gomez-Dans, J., & Gomez-Dan, J. (2016). A survey on Gaussian processes for Earth-observation data analysis: A comprehensive investigation. *IEEE Transactions on Geoscience and Remote Sensing Magazine*, *4*(2), 58–78. https://doi.org/10.1109/mgrs.2015.2510084

Camp-Valls, G., Tula, D., Zhu, X. X., & Reichstein, M. (2021). Deep learning for the Earth sciences: A comprehensive approach to remote sensing, climate science and geosciences. Retrieved from https://onlinelibrary.wiley.com/doi/book/10.1002/9781119646181

Castruccio, S., Hu, Z., Sanderson, B., Karspeck, A., & Hammerling, D. (2019). Reproducing internal variability with few ensemble runs reproducing internal variability with few ensemble runs. *Journal of Climate*, *32*(24), 8511–8522. https://doi.org/10.1175/jcli-d-19-0280.1

Castruccio, S., McInerney, D. J., Stein, M. L., Crouch, F. L., Jacob, R. L., & Moyer, E. J. (2014). Statistical emulation of climate model projections based on precomputed GCM runs. *Journal of Climate*, *27*(5), 1829–1844. https://doi.org/10.1175/jcli-d-13-00099.1

Collins, W. J., Lamarque, J.-F., Schulz, M., Boucher, O., Eyring, V., Hegglin, M. I., et al. (2017). AerChemMIP: Quantifying the effects of chemistry and aerosols in CMIP6. *Geoscientific Model Development*, *10*(2), 585–607. https://doi.org/10.5194/gmd-10-585-2017

Dagan, G., Stier, P., & Watson-Parris, D. (2019). Contrasting response of precipitation to aerosol perturbation in the tropics and extratropics explained by energy budget considerations. *Geophysical Research Letters*, *46*(13), 7828–7837. https://doi.org/10.1029/2019gl083479

Dagan, G., Stier, P., & Watson-Parris, D. (2020). Aerosol forcing masks and delays the formation of the North Atlantic warming hole by three decades. *Geophysical Research Letters*, *47*(22), e2020GL090778. https://doi.org/10.1029/2020gl090778

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., et al. (2020). The community Earth system model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, *12*. https://doi.org/10.1029/2019ms001916

Drijfhout, S., van Oldenborgh, G. J., & Cimatoribus, A. (2012). Is a decline of AMOC causing the warming hole above the North Atlantic in observed and modeled warming patterns? *Journal of Climate*, *25*(24), 8373–8379. https://doi.org/10.1175/jcli-d-12-00490.1

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J. L., Frame, D., et al. (2021). The Earth's energy budget, climate feedbacks, and climate sensitivity.

Gettelman, A., Hannay, C., Bacmeister, J. T., Neale, R. B., Pendergrass, A. G., Danabasoglu, G., et al. (2019). High climate sensitivity in the community Earth system model version 2 (CESM2). *Geophysical Research Letters*, *46*(14), 8329–8337. https://doi.org/10.1029/2019gl083978

Gillett, N. P., Kirchmeier-Young, M., Ribes, A., Shiogama, H., Hegerl, G. C., Knutti, R., et al. (2021). Constraining human contributions to observed warming since the pre-industrial period. *Nature Climate Change*, *11*(3), 207–212. https://doi.org/10.1038/s41558-020-00965-9

Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., et al. (2016). The Detection and Attribution Model Intercomparison Project (DAMIP v1.0) contribution to CMIP6. *Geoscientific Model Development*, *9*(10), 3685–3697. https://doi.org/10.5194/gmd-9-3685-2016

Gjermundsen, A., Nummelin, A., Olivié, D., Bentsen, M., Seland, O., & Schulz, M. (2021). Shutdown of Southern Ocean convection controls long-term greenhouse gas-induced warming. *Nature Geoscience*, *14*(10), 724–731. https://doi.org/10.1038/s41561-021-00825-x

Gneiting, T., Raftery, A. E., Westveld, A. H., & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, *133*(5), 1098–1118. https://doi.org/10.1175/mwr2904.1

Hansen, J., Sato, M., & Ruedy, R. (1995). Long-term changes of the diurnal temperature cycle: Implications about mechanisms of global climate change. *Atmospheric Research*, *37*(1–3), 175–209. https://doi.org/10.1016/0169-8095(94)00077-q

Harder, P., Jones, W., Lguensat, R., Bouabid, S., Fulton, J., Quesada-Chacon, D., et al. (2020). NightVision: Generating night-time satellite imagery from infra-red observations. Retrieved from https://arxiv.org/abs/2011.07017

Hasselmann, K. (1993). Optimal fingerprints for the detection of time-dependent climate change. *Journal of Climate*, *6*(10), 1957–1971. https://doi.org/10.1175/1520-0442(1993)006<1957:offtdo>2.0.co

Ho, T. K. (1995). Random decision forests. *Proceedings of Third International Conference Document Analysis Recognition*, *1*, 278–282. https://doi.org/10.1109/icdar.1995.598994

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., et al. (2018). Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the Community Emissions Data System (CEDS). *Geoscientific Model Development*, *11*(1), 369–408. https://doi.org/10.5194/gmd-11-369-2018

Holden, P. B., & Edwards, N. R. (2010). Dimensionally reduced emulation of an AOGCM for application to integrated assessment modelling: Dimensionally reduced AOGCM emulation. *Geophysical Research Letters*, *37*(21). https://doi.org/10.1029/2010gl045137

Jeevanjee, N., & Romps, D. M. (2018). Mean precipitation change from a deepening troposphere. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(45), 11465–11470. https://doi.org/10.1073/pnas.1720683115

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., et al. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318–2331. https://doi.org/10.1109/tkde.2017.2720168

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

Kendall, A., & Gal, Y. (2017). 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).

Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters*, *40*(6), 1194–1199. https://doi.org/10.1002/grl.50256

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

Lee, K., Lee, K., Lee, H., & Shin, J. (2018). *A simple unified framework for detecting out-of-distribution samples and adversarial attacks*. Arxiv.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2020). *Fourier neural operator for parametric partial differential equations*. Arxiv.

Manabe, S., & Stouffer, R. J. (1993). Century-scale effects of increased atmospheric CO_2 on the ocean–atmosphere system. *Nature*, *364*(6434), 215–218. https://doi.org/10.1038/364215a0

Mansfield, L. A., Nowack, P. J., Kasoar, M., Everitt, R. G., Collins, W. J., & 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. https://doi.org/10.1038/s41612-020-00148-5

Matthews, H. D., & Caldeira, K. (2008). Stabilizing climate requires near-zero emissions. *Geophysical Research Letters*, *35*(4), L04705. https://doi.org/10.1029/2007gl032388

McGovern, A., Lagerquist, R. D. J. G., II., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin America Meteorology Social*, *100*, 2175–2199. https://doi.org/10.1175/bams-d-18-0195.1

Meinshausen, M., Raper, S. C. B., & Wigley, T. M. L. (2011). Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, MAGICC6—Part 1: Model description and calibration. *Atmospheric Chemistry and Physics*, *11*(4), 1417–1456. https://doi.org/10.5194/acp-11-1417-2011

Millar, R. J., Fuglestvedt, J. S., Friedlingstein, P., Rogelj, J., Grubb, M. J., Matthews, H. D., et al. (2017). Emission budgets and pathways consistent with limiting warming to 1.5°C. *Nature Geoscience*, *10*, 741–747. https://doi.org/10.1038/ngeo3031

Nicholls, Z. R. J., Meinshausen, M., Lewis, J., Gieseke, R., Dommenget, D., Dorheim, K., et al. (2020). Reduced complexity model intercomparison project phase 1: Introduction and evaluation of global-mean temperature response. *Geoscientific Model Development*, *13*(11), 5175–5190. https://doi.org/10.5194/gmd-13-5175-2020

O'Neill, B. C., Tebaldi, C., Vuuren, D. P., van Eyring, V., Friedlingstein, P., Hurtt, G., et al. (2016). The scenario model intercomparison project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, *9*, 3461–3482. https://doi.org/10.5194/gmd-9-3461-2016

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the Earth sciences. *Science*, *263*(5147), 641–646. https://doi.org/10.1126/science.263.5147.641

Otto, F. E. L., van Oldenborgh, G. J., Eden, J., Stott, P. A., Karoly, D. J., & Allen, M. R. (2016). The attribution question. *Nature Climate Change*, *6*(9), 813–816. https://doi.org/10.1038/nclimate3089

Pendergrass, A. G., & Hartmann, D. L. (2013). The atmospheric energy constraint on global-mean precipitation change. *Journal of Climate*, *27*(2), 130916120136005. https://doi.org/10.1175/jcli-d-13-00163.1

Pincus, R., Batstone, C. P., Hofmann, R. J. P., Taylor, K. E., & Glecker, P. J. (2008). Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. *Journal of Geophysical Research*, *113*(D14), D14209. https://doi.org/10.1029/2007jd009334

Pincus, R., Forster, P. M., & Stevens, B. (2016). The radiative forcing model intercomparison project (RFMIP): Experimental protocol for CMIP6. *Geoscientific Model Development*, *9*, 3447–3460. https://doi.org/10.5194/gmd-9-3447-2016

Rabanser, S., Günnemann, S., & Lipton, Z. C. (2018). *Failing loudly: An empirical study of methods for detecting dataset shift*. Arxiv.

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2018). Numerical Gaussian processes for time-dependent and nonlinear partial differential equations. *SIAM Journal on Scientific Computing*, *40*(1), A172–A198. https://doi.org/10.1137/17m1120762

Rasmussen, C. E., & Williams, C. K. I. (2005). Gaussian Processes for machine learning. https://doi.org/10.7551/mitpress/3206.001.0001

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, *12*(11). https://doi.org/10.1029/2020ms002203

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N. (2019). *Deep learning and process understanding for data-driven Earth system science* (pp. 195–204). Nature566. https://doi.org/10.1038/s41586-019-0912-1

Richardson, T. B., Forster, P. M., Smith, C. J., Maycock, A. C., Wood, T., Andrews, T., et al. (2019). Efficacy of climate forcings in PDRMIP models. *Journal of Geophysical Research: Atmospheres*, *124*(23), 12824–12844. https://doi.org/10.1029/2019jd030581

Rodgers, K. B., Lee, S.-S., Rosenbloom, N., Timmermann, A., Danabasoglu, G., Deser, C., et al. (2021). Ubiquity of human-induced changes in climate variability. *Earth System Dynamics*, *12*(4), 1393–1411. https://doi.org/10.5194/esd-12-1393-2021

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

Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., et al. (2020). Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations. *Geoscientific Model Development*, *13*(12), 6165–6200. https://doi.org/10.5194/gmd-13-6165-2020

Shindell, D. T., Faluvegi, G., Koch, D. M., Schmidt, G. A., Unger, N., & Bauer, S. E. (2009). Improved attribution of climate forcing to emissions. *Science*, *326*(5953), 716–718. https://doi.org/10.1126/science.1174760

Smith, C. J., Forster, P. M., Allen, M., Leach, N., Millar, R. J., Passerello, G. A., & Regayre, L. A. (2018). FAIR v1.3: A simple emissions-based impulse response and carbon cycle model. *Geoscientific Model Development*, *11*(6), 2273–2297. https://doi.org/10.5194/gmd-11-2273-2018

Smith, C. J., Harris, G. R., Palmer, M. D., Bellouin, N., Collins, W., Myhre, G., et al. (2021). Energy budget constraints on the time history of aerosol forcing and climate sensitivity. *Journal of Geophysical Research: Atmospheres*, *126*(13). https://doi.org/10.1029/2020jd033622

Stott, P. A., Christidis, N., Otto, F. E. L., Sun, Y., Vanderlinden, J., van Oldenborgh, G. J., et al. (2016). *Attribution of extreme weather and climate-related events* (Vol. 7, pp. 23–41). Wiley Interdisciplinary Reviews: Climate Change. https://doi.org/10.1002/wcc.380

Stott, P. A., Gillett, N. P., Hegerl, G. C., Karoly, D. J., Stone, D. A., Zhang, X., & Zwiers, F. (2010). *Detection and attribution of climate change: A regional perspective* (Vol. 1, pp. 192–211). Wiley Interdisciplinary Reviews: Climate Change. https://doi.org/10.1002/wcc.34

Tebaldi, C., & Arblaster, J. M. (2014). Pattern scaling: Its strengths and limitations, and an update on the latest model simulations. *Climatic Change*, *122*(3), 459–471. https://doi.org/10.1007/s10584-013-1032-9

Trebing, K., Stanczyk, T., & Mehrkanoon, S. (2021). SmaAt-Unet: Precipitation nowcasting usingSmall attention-UNet architecture. Retrieved from https://arxiv.org/abs/2007.04417

Watson-Parris, D. (2021). Machine learning for weather and climate are worlds apart. *Philosophical Transactions Royal Soc*, *379*(2194), 20200098. https://doi.org/10.1098/rsta.2020.0098

Watson-Parris, D., Williams, A., Deaconu, L., & Stier, P. (2021). Model calibration using ESEm v1.1.0—An open, scalable Earth system emulator. *Geoscientific Model Development*, *14*(12), 7659–7672. https://doi.org/10.5194/gmd-14-7659-2021

Wilcox, L. J., Liu, Z., Samset, B. H., Hawkins, E., Lund, M. T., Nordling, K., et al. (2020). Accelerated increases in global and Asian summer monsoon precipitation from future aerosol reductions. *Atmospheric Chemistry and Physics*, *20*, 11955–11977. https://doi.org/10.5194/acp-20-11955-2020

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