AeroGP: machine learning how aerosols impact regional climate

JGR Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2025JH000741

Key Points:

- Anthropogenic aerosols play a key role in global and regional climate change
- There is a need for more aerosol-focused climate modeling, as aerosols represent a major source of uncertainty in future climate change
- We use Gaussian processes (GPs) to accurately predict global spatial patterns of the temperature response to aerosol emission perturbations

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

M. Dewey,
maura.dewey@misu.su.se

Citation:

Dewey, M., Hansson, H.-C., Watson-Parris, D., Samset, B. H., Wilcox, L. J., Lewinschal, A., et al. (2025). AeroGP: Machine learning how aerosols impact regional climate. *Journal of Geophysical Research: Machine Learning and Computation*, *2*, e2025JH000741. https://doi.org/10.1029/2025JH000741

Received 16 APR 2025
Accepted 4 OCT 2025

Author Contributions:

Conceptualization: Maura Dewey, Hans-Christen Hansson, Duncan Watson-Parris, Bjørn H. Samset, Laura J. Wilcox, Annica M. L. Ekman
Formal analysis: Maura Dewey, Annica M. L. Ekman
Funding acquisition: Annica M. L. Ekman
Methodology: Maura Dewey, Duncan Watson-Parris, Annica M. L. Ekman
Project administration: Annica M. L. Ekman

AeroGP: Machine Learning How Aerosols Impact Regional Climate

Maura Dewey[1,2], Hans-Christen Hansson[2,3], Duncan Watson-Parris[4], Bjørn H. Samset[5], Laura J. Wilcox[6], Anna Lewinschal[1,2], Maria Sand[5], Øyvind Seland[7], Srinath Krishnan[5], and Annica M. L. Ekman[1,2]

[1]Department of Meteorology (MISU), Stockholm University, Stockholm, Sweden, [2]Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden, [3]Department of Environmental Science (ACES), Stockholm University, Stockholm, Sweden, [4]Scripps Institution of Oceanography and Halıcıoğlu Data Science Institute, UC San Diego, San Diego, CA, USA, [5]Center for International Climate Research (CICERO), Oslo, Norway, [6]Department of Meteorology, National Centre for Atmospheric Science, University of Reading, Reading, UK, [7]Norwegian Meteorological Institute, Oslo, Norway

Abstract Aerosol particles from both natural and anthropogenic sources play a critical role in the Earth's climate by interacting with solar radiation and clouds. Anthropogenic aerosol and precursor emissions have historically exerted a global cooling effect, which has partially offset the warming from concurrent greenhouse gas emissions. Recent reductions and shifts in aerosol and precursor emission patterns may reduce this offset and introduce spatially and temporarily varying climate impacts. Investigating aerosol-climate effects is typically done with computationally expensive Earth System Models, which include complex representations of physical, chemical, biological, and geological processes and their coupled interactions for the entire global climate system. In this study, we develop a machine-learning climate emulator using Gaussian processes, called *AeroGP*, that can be used to quickly assess, for example, the impact of different policy decisions on future climate mitigation strategies. The emulator is trained on a unique data set from the Norwegian Earth System Model (NorESM), analyzed as an ensemble here for the first time. AeroGP accounts for the joint spatial covariance of the output variables and captures the complex, heterogeneous impacts of aerosols on surface temperature using coregionalization. We believe this is the first time this method has been used to account for the spatial correlation of such climate data. We show that AeroGP retains the spatial complexity of NorESM at a fraction of the computational cost and demonstrate its usefulness to assess the sensitivity of temperature to idealized future aerosol emission scenarios.

Plain Language Summary Aerosols—airborne particles from both natural and human sources—play a crucial role in Earth's climate. Historically, anthropogenic aerosols have had a cooling effect, partially offsetting greenhouse gas-induced warming. However, reductions and geographic shifts in aerosol sources may weaken this offset and introduce regionally varied climate impacts. Climate models used to study these impacts are either highly detailed but computationally expensive or highly simplified and lacking in spatial complexity. This study presents *AeroGP*, a machine-learning emulator based on Gaussian processes that captures the spatial pattern of the temperature response to aerosols while significantly reducing computational costs. Such machine-learning tools can enable more efficient assessments of future aerosol emission scenarios, supporting climate mitigation policy decisions.

1. Introduction

Aerosol-cloud-climate interactions are one of the largest sources of uncertainty in future climate projections due to the complexity of processes involved, from microphysical to large-scale dynamics, and due to uncertainties in future aerosol emission pathways (Boucher et al., 2013; Forster et al., 2021; Polonik et al., 2021; Szopa et al., 2021; Watson-Parris & Smith, 2022). Aerosols are small particles suspended within the atmosphere that are emitted directly from natural and anthropogenic sources or formed from gases within the air. They interact with climate both locally and remotely: directly via the absorption and scattering of solar radiation and indirectly through interactions with clouds and the hydrological cycle, therefore impacting cloud radiative effects and precipitation patterns (Albrecht, 1989; Boucher et al., 2013; Hansen et al., 1997; Li et al., 2022; Stier et al., 2024; Twomey, 1977).

Resources: Bjørn H. Samset, Laura J. Wilcox, Anna Lewinschal, Maria Sand, Øyvind Seland, Srinath Krishnan
Software: Maura Dewey
Supervision: Annica M. L. Ekman
Visualization: Maura Dewey
Writing – original draft: Maura Dewey
Writing – review & editing: Maura Dewey, Hans-Christen Hansson, Duncan Watson-Parris, Bjørn H. Samset, Laura J. Wilcox, Anna Lewinschal, Maria Sand, Øyvind Seland, Srinath Krishnan, Annica M. L. Ekman

Unlike long-lived greenhouse gases, aerosols are not well mixed in the atmosphere. They produce heterogeneous radiative forcing patterns and regional climate responses which depend on both the type of aerosol and the emission location. For example: black carbon induces shortwave heating through the atmospheric column, which can impact monsoon variability close to emission locations and suppress precipitation globally (Richardson et al., 2018; Samset, 2022). Sulfur dioxide emissions in Asia have been shown to have teleconnections across the Northern Hemisphere, primarily via impacting circulation patterns and the resulting cloud fields (Lewinschal et al., 2019; L. Wilcox et al., 2019). The climate impact of aerosol processes across scales, from optical properties impacting microphysics to large-scale teleconnections (Persad, 2023; Stier et al., 2024), remains uncertain, and therefore aerosols are continually highlighted in climate assessment reports as a major source of uncertainty in climate projections (Boucher et al., 2013; Forster et al., 2021).

The process-level uncertainties in aerosol-cloud-climate interactions, radiative forcing, and dynamical responses are compounded by uncertainty in emissions policy decisions and therefore potential future emission scenarios. These decisions are driven by a range of factors including economic development, technological advancements, and societal choices ranging from international climate agreements to regional air quality regulations (Lund et al., 2019; Szopa et al., 2021). Recently, global anthropogenic aerosol emission rates have fallen, due to a combination of reductions resulting from, for example, air quality targets and climate goals (Aas et al., 2019; Crippa et al., 2016; Elguindi et al., 2020). There have also been geographical shifts in emissions from changes in industrial production and transport (Elguindi et al., 2020; Quaas et al., 2022). The global mean effect of anthropogenic aerosols in the industrial era has been a net cooling of −0.5°C (−0.22 to −0.95)°C (Forster et al., 2021), which has masked some of the warming from concurrent greenhouse gas (GHG) emissions (Forster et al., 2021; Szopa et al., 2021). A continued future reduction in aerosol and precursor emissions is likely to reduce this cooling in the global mean (Hodnebrog et al., 2024) and to produce spatially complex trends in temperature, precipitation, air quality, and extreme events (Dong et al., 2017; Persad, 2023; Persad & Caldeira, 2018; Samset et al., 2016; Westervelt et al., 2020). This study is motivated by the need to assess the climate response to these potential future emission scenarios, to improve understanding of regional patterns of aerosol-climate interactions, and to develop a tool for policymakers.

Typically, global climate is studied using physics-based Earth System Models (ESMs), which are composed of various submodels representing the different parts of the Earth system and how they interact: atmospheric and ocean models (for simulating, e.g., weather, ocean biogeochemistry, atmospheric chemistry, winds, ocean currents, and thermodynamics), land surface models (for vegetation, hydrology, and soil processes), ice models (for glaciers and sea ice), and geological models (for long-term geochemical processes). These models are designed to capture interactions, teleconnections, and feedbacks between these components, at as high a spatial and temporal resolution as possible, and to be stable for long-term simulations of past, current, and possible future climate scenarios. Due to this complexity, ESMs are computationally very costly and require specialized supercomputing resources for both running simulations and managing data. Therefore, a limited number of future scenarios can be simulated with fully coupled ESMs, and the models cannot be easily used by nonexpert stakeholders such as policymakers.

Machine learning (ML) *emulators* can address these resource limitations and therefore explore more of the scenario space than is possible with ESMs. Emulators do not explicitly resolve all the physical dynamics of an ESM but instead focus on learning the patterns driving a limited subset of variables, making them computationally efficient while retaining spatial and temporal complexity. Emulators can be used to explore more emission scenarios than are computationally feasible with ESMs, allowing a broader exploration of potential future pathways and of the range of uncertainty in future projections. In this study, we develop an emulator to predict the spatially resolved temperature response to regional perturbations of anthropogenic aerosols in order to facilitate a rapid assessment of the local and remote responses to regional policy and air quality decisions. We refer to this emulator as *AeroGP*.

To build AeroGP, we use Gaussian processes (GPs), which is a machine learning methodology that is well suited to Earth-system problems because it can incorporate prior physical knowledge and inherently produce uncertainty estimates (Camps-Valls et al., 2016). GPs have been used for global spatially resolved climate emulation for scenarios including both GHG and aerosol forcing (e.g., ClimateBench (Watson-Parris et al., 2022) or FairGP (Bouabid et al., 2024)) and for limited area spatio-temporal modeling such as predicting air quality or regional downscaling (Axen et al., 2022; Hamelijnck et al., 2021; Krock et al., 2023; Tazi et al., 2024). These studies typically had relatively poor performance when trying to predict the signal to aerosol-only perturbations, due to the dominance of the GHG response. Here, we predict the global spatially resolved temperature response to *only* aerosol perturbations, which previous ML-emulators have struggled to do. This is a more complex problem due to the short-lived nature of aerosols compared to GHGs, and due to spatial and temporal heterogeneity in the forcing and the rate of change of emissions as well as the sign of that change. To our knowledge, we are also the first to use coregionalization to account for the joint spatial covariance of surface temperature. We show that the performance of AeroGP is comparable to that of the parent model NorESM and then use it to assess the sensitivity to scales of potential future emission scenarios.

Our aim is to provide an example that machine-learning techniques can be useful within climate science seen from both the view of climate scientists and experts on machine learning. For those with a background in climate science, we show that our machine learning emulator can achieve comparable accuracy to traditional tools for a fraction of the computational cost and is therefore well suited to problems such as extending or constraining ESM results, exploring novel emission scenarios, and providing projections for impact and policy assessments. An inexpensive climate emulator which can provide spatially resolved projections such as AeroGP could be particularly useful for end-users of climate products, such as in policy development or insurance. For those with a machine learning background, we highlight the unique data structures and physical considerations to keep in mind when dealing with spatially and temporally resolved climate data and suggest some methods for working with a large amount of climate model output as training data.

In the remainder of this paper, we first give a brief overview of aerosol-climate interactions (aimed at nonexperts in climate science) in Section 2. We then describe the development of the training data set and give an overview of Gaussian processes in Section 3. We evaluate the performance of AeroGP and test it on potential future aerosol perturbations in Section 4 and finally present a discussion of our results in Section 5 and conclusions in Section 6.

2. Anthropogenic Aerosols and Their Climate Impacts

In this study, we are interested in the climate impact of anthropogenic emissions of three aerosol species or aerosol precursors which are known to impact climate: sulfur dioxide (SO_2, the precursor for sulfate, SO_4, aerosols), black carbon (BC), and organic carbon (OC). Sulfate aerosols form in the atmosphere via gas-phase oxidation and aqueous-phase (in-cloud) oxidation of SO_2 (Seinfeld & Pandis, 2016). SO_2 is primarily emitted from fossil fuel combustion but also has natural sources, for example, from the ocean and volcanoes (Boucher et al., 2013; Szopa et al., 2021). Sulfate influences climate via scattering incoming solar radiation as well as increasing cloud albedo and contributes to a majority of the anthropogenic aerosol forcing globally (Haywood & Boucher, 2000). The cooling impact of sulfate aerosols has masked a significant portion of the warming due to greenhouse gas emissions in the industrial era. As clean-air policies have improved and industrial production has shifted geographically, the resulting reduction in sulfate has begun to reveal this warming (Szopa et al., 2021).

BC is a primary aerosol emitted during incomplete combustion of, for example, fossil fuels and biomass. It impacts the climate primarily via absorbing incoming solar radiation and therefore acts to warm the surface through exerting positive radiative forcing. Additionally, however, BC also changes atmospheric heating rates, humidity, and cloudiness (so-called "rapid adjustments"), which impact both temperature and precipitation locally and remotely (Bond et al., 2013; Quaas et al., 2024; Samset et al., 2016). BC also contributes to indirect aerosol effects (i.e., changing cloud microphysical properties) as BC particles typically are coated by hydrophilic materials in the atmosphere, for example, sulfuric acid, and thereby contribute to the cloud condensation nuclei (CCN) population (Bond et al., 2013; Haywood & Boucher, 2000; Twomey, 1977). BC can also modify the surface albedo after deposition on the surface, which is especially important at high latitudes (Sand et al., 2013). The vertical distribution of BC also controls its climate impacts: Ban-Weiss et al., 2012 showed that BC in the lower troposphere warms the surface and increases precipitation, whereas BC in the upper troposphere cools the surface and suppresses precipitation. BC can also have nonlocal impacts via the transport of heat (Sand et al., 2020; Stjern et al., 2017).

OC is emitted as a primary aerosol both from anthropogenic combustion (fossil fuels and biomass) and natural sources (e.g., pollen and algae). It is also formed in the atmosphere (secondary production) via gas oxidation and subsequent condensation (Haywood & Boucher, 2000). OC impacts the climate mainly via scattering incoming solar radiation as well as via indirect aerosol-cloud interactions and generally cools the surface and reduces evaporation (Boucher et al., 2013).

Table 1
Summary of Aerosol Perturbation Experiments

MIP or project	Aerosol species	Experiments (baseline)	Years (ensemble members)	NorESM v.	Reference
DAMIP	SO_2, BC, OC	hist-aer, ssp245-aer, (piControl)	1850–2020 (3), 2021–2100 (1)	NorESM2-LM	Gillett et al. (2016), Seland et al. (2020)
AerChemMIP	SO_2, BC, OC	ssp370-lowNTCF, (SSP3-7.0)	2015–2055 (3)	NorESM2-LM	Collins et al. (2017)
RAMIP	SO_2, BC, OC	ssp370-126aer, ssp370-EAS126aer, ssp370-SAS126aer, ssp370-AFR126aer, ssp370-NAE126aer, (SSP3-7.0)	2015–2065 (10)	NorESM2-LM	L. J. Wilcox et al. (2023)
AeroGP (NorESM)	SO_2, BC	BC126-$SO_2$370, $SO_2$126-BC370, (SSP2-4.5)	2015–2055 (2)	NorESM2-LM	This work
Regional SO_2	SO_2	0xEU, EA, SA, NA; 7xEU; 5xNA; 5xEA; 10xSA; (year 2000)	150; 110 (1)	NorESM1	Lewinschal et al. (2019)
Regional BC	BC	5xEA; 10xEU; 10xSA; 10xNA; (year 2000)	80 (3)	NorESM1	Sand et al. (2020)
Global simultaneous removal	SO_2, BC, OC	0xGlobal, (year 2000)	50 (1)	NorESM1	Samset et al. (2018)
Global individual removal	SO_2, BC, OC	0xGlobal, (year 2008)	50 (2)	NorESM1	Baker et al. (2015)

Note. Columns are (from left to right): the model intercomparison project (MIP) or study, the aerosol species that are changed (with the baseline indicated in brackets), the region and amount of aerosol perturbation, the years available (in the case of transient experiments) or the number of years after spin-up (in the case of equilibrium experiments) (with the number of ensemble members in brackets), the NorESM model version, and the main reference. Region abbreviations are EU—Europe, EA—East Asia, SA—South Asia, NA—North America, AF—Africa, NAE—North America and Europe together. See the main text for all MIP abbreviations and experiment details.

In the global mean, anthropogenic aerosols have historically cooled the planet by −0.5°C (−0.22 to −0.95)°C over the industrial era (Forster et al., 2021). Future reductions in anthropogenic emissions could therefore cause an apparent warming of about the same magnitude if all emissions were stopped, see Figure 2 and (Samset et al., 2018). However, because aerosols generally have a short atmospheric residence time (days to a few years, depending on location within in the atmosphere (Prospero et al., 1983)), the effects of increases or reductions as well as geographic shifts in emissions actually result in a spatially heterogeneous forcing pattern (Persad & Caldeira, 2018; Westervelt et al., 2020), and the resulting temperature patterns are less well-understood than the global mean impacts and one of the key motivations for this study. A fast ML-model can facilitate the testing and exploration of many more regional perturbation scenarios than would be possible with a traditional physics-based ESM, allowing for the exploration of how future aerosol-induced warming or cooling may evolve.

3. Data and Methods

AeroGP is trained on a unique set of aerosol perturbation experiments from NorESM. This data set is compiled here for the first time from many different sources, including new simulations run specifically for this study, with the aim of capturing the broad state-space of potential emission scenarios and regional patterns of the response to aerosol forcing. In this section, we will first describe NorESM and the creation of the training data set (which is also summarized in Table 1). We then present an overview of GPs, how we apply this method to build AeroGP, and the metrics we use to validate its performance. In the following sections, we will show the results of the validation and some initial results using AeroGP for emulating climate impacts. This entire process is summarized in Figure 1.

3.1. NorESM Model Description

AeroGP is trained on output from the Norwegian Earth System model (NorESM), an ESM based on the Community Earth System Model (CESM) (Danabasoglu et al., 2020; Hurrell et al., 2013), extended with a custom ocean dynamics model (Bentsen et al., 2013; Seland et al., 2020) and relatively advanced aerosol-chemistry-cloud-radiation modules (Kirkevåg et al., 2013, 2018; Seland et al., 2020). In order to create a data set that encompasses a large range of plausible anthropogenic aerosol emission scenarios, we have compiled a set of experiments run with two versions: NorESM1-M (Bentsen et al., 2013; Iversen et al., 2013) and NorESM2.0-LM (Seland et al., 2020), both of which have relatively complex descriptions of aerosols and aerosol-cloud

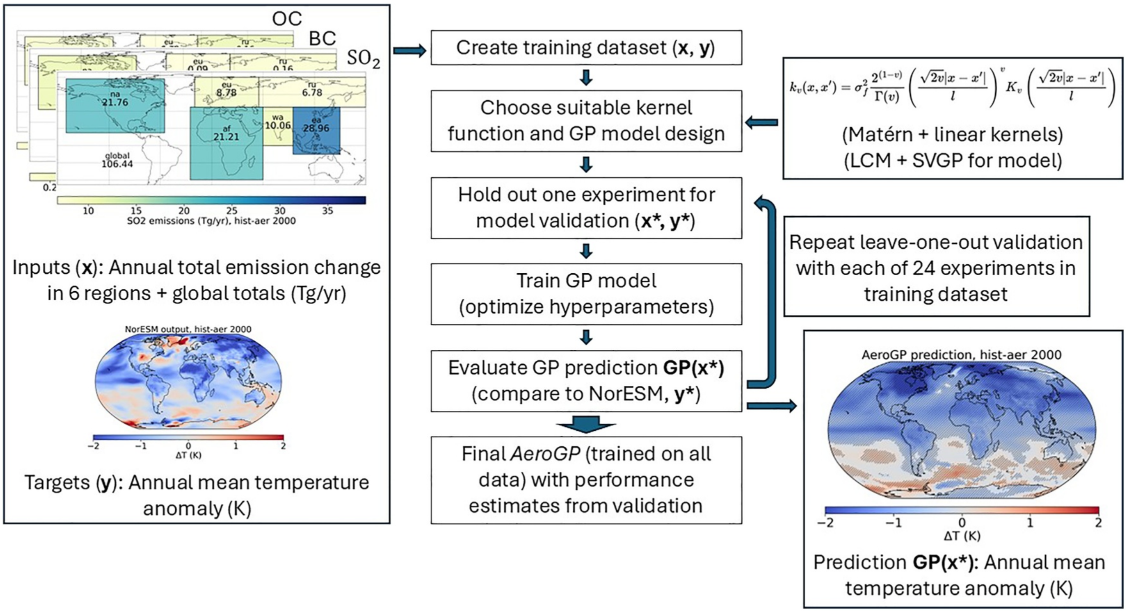
OC
BC
SO$_2$
na 21.76
eu 8.78
ru 6.78
wa 10.06
ea 28.96
af 21.21
global 106.44
SO2 emissions (Tg/yr), hist-aer 2000
Inputs (x): Annual total emission change in 6 regions + global totals (Tg/yr)
NorESM output, hist-aer 2000
ΔT (K)
Targets (y): Annual mean temperature anomaly (K)
Create training dataset (x, y)
Choose suitable kernel function and GP model design
Hold out one experiment for model validation (x*, y*)
Train GP model (optimize hyperparameters)
Evaluate GP prediction GP(x*) (compare to NorESM, y*)
Final AeroGP (trained on all data) with performance estimates from validation
$k_v(x,x') = \sigma_f^2 \frac{2^{(1-v)}}{\Gamma(v)} \left(\frac{\sqrt{2v}|x-x'|}{l}\right)^v K_v\left(\frac{\sqrt{2v}|x-x'|}{l}\right)$
(Matérn + linear kernels)
(LCM + SVGP for model)
Repeat leave-one-out validation with each of 24 experiments in training dataset
AeroGP prediction, hist-aer 2000
ΔT (K)
Prediction GP(x*): Annual mean temperature anomaly (K)

Figure 1. Workflow diagram showing the development of AeroGP.

interactions compared to other ESMs, which we outline below. In the sections following this model description, we will refer to both versions simply as NorESM.

NorESM1 was the model version used for the World Climate Research Program's Coupled Model Intercomparison Project Phase 5 (CMIP5) and the Intergovernmental Panel on Climate Change's 5th Assessment Report (IPCC AR5). It is based on the Community Climate System Model (CCSM4.0) (Gent et al., 2010) and the Community Earth System Model (CESM1.0.3) (Hurrell et al., 2013), with the latest code released in 2018. The ocean component is the Bergen version of the Miami Isopycnic Coordinate Ocean Model (MICOM), the land

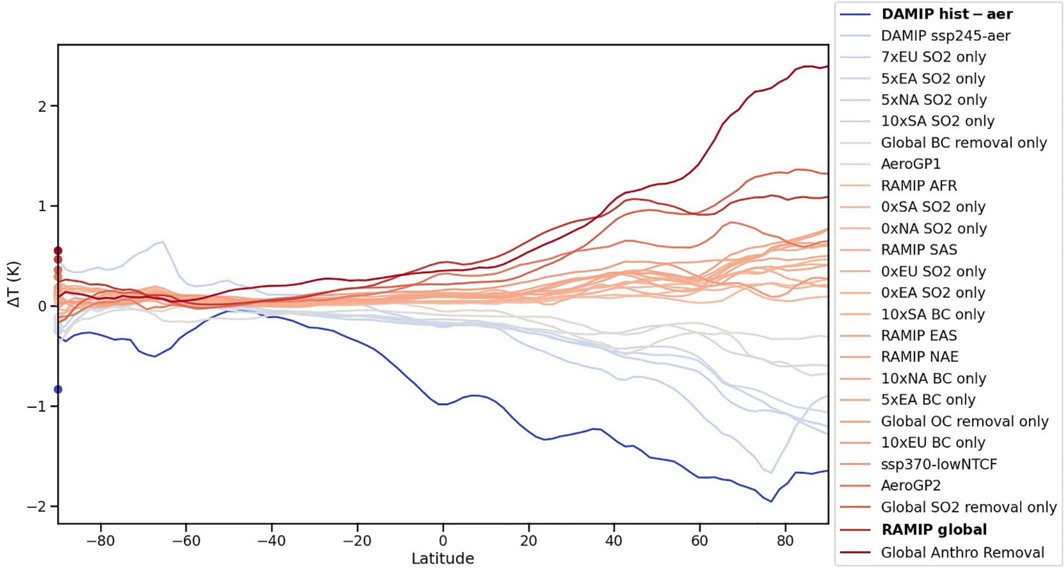

Figure 2. Zonal mean temperature change for the last 20 years of all experiments in the training data set compared to their respective baselines. Color indicates global mean temperature change and corresponds to the dots on the left-side axis, with red being the warmest experiment (complete removal of all anthropogenic aerosols, +0.56°C) and blue the coldest (historical aerosol and precursor emissions, −0.84°C). The two experiments where we explore the validation results in detail are highlighted in bold.

surface component is the Community Land Model version 4 (CLM4), and the sea ice component is CICE4 (Bentsen et al., 2013). The atmospheric component is CAM4-Oslo, which is based on the Community Atmosphere Model (CAM4: (Neale et al., 2013)) and extended with a custom aerosol module (OsloAero4.0) that includes prognostic double moment cloud microphysics, direct aerosol-radiation interactions, and aerosol-cloud interactions (Kirkevåg et al., 2013). The aerosol module calculates the mass concentrations of five different aerosols: sulfate, black carbon, organic matter, sea salt, and mineral dust. The primary aerosol emissions follow log-normal size distributions. There is also secondary aerosol formation through clear-sky and in-cloud gas phase and aqueous phase chemical reactions. Aerosol mass concentrations, size distributions, and optical properties are continuously updated during the simulations (and not required to remain log-normal) via lookup tables calculated offline with a size-resolving model which accounts for processes such as condensation, coagulation, hygroscopic growth, gas-phase chemistry, and cloud processing. NorESM1 produces a global aerosol effective radiative forcing (ERF: including both aerosol-radiation interactions (ARI) and aerosol-cloud interactions (ACI)) of -1.0 $\mathrm{W/m^2}$ in the year 2000 compared with preindustrial conditions, which is slightly weaker than the CMIP5 multimodel mean of $-1.17 \pm 0.30\mathrm{W/m^2}$ (Kirkevåg et al., 2018; Zelinka et al., 2014). NorESM1 has an equilibrium climate sensitivity of around 2.9 K, which is average compared to other CMIP5 models (Iversen et al., 2013).

NorESM2 is the second generation model including updates to the parent model (CESM2.1 (Danabasoglu et al., 2020)), the ocean component (BLOM (Seland et al., 2020) and iHAMOCC (Tjiputra et al., 2020)), the atmospheric component (CAM6-Nor (Seland et al., 2020)), and the aerosol module (OsloAero6 (Kirkevåg et al., 2018; Seland et al., 2020)). Pertinent updates to the aerosol treatments compared to NorESM1 include improved emissions, nucleation, and coagulation processes, and updated aerosol-cloud interactions via new cloud schemes in CAM6 (Seland et al., 2020). There are also key improvements for the production of secondary organic aerosols, the microphysical properties of black carbon and mineral dust, and wind-driven sea salt and dimethyl sulfide emissions (Kirkevåg et al., 2018). CAM6-Nor produces a global mean aerosol ERF of -1.36 $\mathrm{W/m^2}$ for 2014 compared to preindustrial conditions, which is slightly stronger than the CMIP6 multimodel mean for 2014 ($-1.01 \pm 0.23\mathrm{W/m^2}$) (Kirkevåg et al., 2018; Smith et al., 2020). NorESM2 produces an equilibrium climate sensitivity of 2.5 K, which is toward the low end of CMIP6 estimates and is primarily due to a slow long-term ocean response (Bock & Lauer, 2024; Seland et al., 2020).

The aerosol module in NorESM (OsloAero) is a production-tagged mode, which includes an online life-cycling component and a set of offline size-resolved lookup tables for size distribution parameters and interpolations related to aerosol-radiation and aerosol-cloud interactions (called AeroTab). Aerosol tracers are divided into "background" tracers, which are primary aerosol emissions with log-normal modes, and "process" tracers which modify the shape and chemical composition of the initial background modes. Examples of background tracers include accumulation-mode sulfate and nucleation-, Aitken-, and accumulation-mode BC. Examples of process tracers include sulfate condensate and sulfate from cloud processing (aqueous-phase chemistry within cloud droplets). A full list of the aerosol tracers used can be found in Kirkevåg et al. (2018). Once process tracers are applied to an initial distribution, the resulting mixture is not required to remain log-normal. The mass of the mixtures is tracked and the resulting aerosol size distributions and optical properties are derived from the lookup tables.

Both NorESM versions used here have an atmospheric horizontal resolution of 1.9° (latitude) by 2.5° (longitude). NorESM1 has 26 vertical pressure levels up to 2.9 hPa (Bentsen et al., 2013), while NorESM2 has 32 vertical pressure levels up to 3.6 hPa (Seland et al., 2020). NorESM2 has reduced troposphere and near-surface temperature biases and improved spatial precipitation bias compared with its predecessor (Seland et al., 2020). We use experiments from both versions in our training data ensemble to ensure that we have enough diversity of aerosol perturbation experiments to cover the required state-variable-space for training. We find that the advantage of having more training data and better coverage of the input space outweighs the potential disadvantage of differences in aerosol treatment (chemistry, transport, etc.) and the resulting forcing between the two NorESM versions.

3.2. Data Set Description

We created the training data set from a diverse set of NorESM aerosol perturbation experiments, which span a large range of possible anthropogenic emissions scenarios: from complete global removal to as much as 10x current emission levels in some regions (see Figures S2 and S3 in Supporting Information S1). These results are from seven previously published aerosol perturbation studies and two additional experiments simulated specifically for this study, all of which are summarized in Table 1 as well as Figure 2 and Figure S2 in Supporting Information S1. The experiments were chosen because (a) they have perturbation experiments varying only anthropogenic aerosol and precursor emissions, while other anthropogenic forcings (such as GHG emissions) are kept constant relative to a baseline experiment and (b) they use a version of NorESM where the atmospheric module is coupled to a dynamical ocean module. The first requirement means we can regress out any nonaerosol forcing and learn only aerosol impacts on climate, and the second means that the full range of climate system feedbacks is included (and therefore realistic remote impacts and teleconnection patterns).

Four sets of experiments in our training data set are transient runs, meaning that the emissions are continuously perturbed throughout the experiment so the climate system follows a plausible scenario. The scenarios are based on the Shared Socioeconomic Pathways (SSPs) defined in the sixth IPCC assessment report (AR6), which outline different possible future anthropogenic emission pathways, depending on different political and socioeconomic policies for greenhouse gas emission reductions (Forster et al., 2021; Gidden et al., 2019; O'Neill et al., 2014, 2017). This includes two sets of experiments from the Detection and Attribution Model Intercomparison Project (DAMIP) (Gillett et al., 2016): *hist-aer,* which is forced with anthropogenic aerosol and precursor emissions for the historical period (1850–2020), and *ssp245-aer*, which is an extension of the *hist-aer* simulation through the 21st century (2021–2100) forced with SSP2-4.5 (an intermediate forcing scenario) emissions. Both perturbations are relative to the preindustrial control (*piControl*) baseline, which is a quasi-equilibrium simulation representative of conditions before widespread industrialization (Eyring et al., 2016). From the Aerosol Chemistry Model Intercomparison Project (AerChemMIP) (Collins et al., 2017), we include the *ssp370-lowNTCF* experiment, which reproduces strong levels of air quality control measures in the beginning of the 21st century (2015–2055) and has SSP3-7.0 (strong warming scenario) as a baseline which is a future scenario without strong climate mitigation policies. We include Tier 1 transient experiments performed with NorESM from the Regional Aerosol Model Intercomparison Project (RAMIP), which have a baseline of SSP3-7.0 with aerosol and precursor emissions reduced following SSP1-2.6 globally (*ssp370-126aer*) or in specific regions (i.e., East Asia: *ssp370-EAS126aer*, South Asia: *ssp370-SAS126aer*, Africa and the Middle East: *ssp370-AFR126aer*, and North American and Europe together: *ssp370-NAE126aer*) for the first half of the 21st century (2015–2065) (L. J. Wilcox et al., 2023). The final sets of transient experiments were run with NorESM2 specifically for this study (referred to as AeroGP (NorESM) in Table 1) in order to have a set of experiments in which SO_2 and BC emissions vary in opposite directions. These experiments were conducted with SSP2-4.5 as the baseline scenario; one with SO_2 increasing according to a continuation of current policies while BC decreased according to maximum mitigation and one with the opposite configuration.

The other four sets of experiments included in our data set are equilibrium simulations, meaning that emissions are changed abruptly and then held constant while the model is run to quasi-equilibrium with the new forcing. We use all eight experiments from Lewinschal et al. (2019) where anthropogenic SO_2 emissions in a specific region (Europe, East Asia, South Asia, or North America, as defined by Task Force on Hemispheric Transport of Air Pollution (HTAP) (Janssens-Maenhout et al., 2015)) are either removed completely or increased in order to achieve a global radiative forcing of $-0.45\ \mathrm{W/m^2}$, compared to a baseline of constant year 2000 emissions. From Sand et al. (2020), we use four experiments where BC emissions are increased in the four HTAP regions to give a global direct radiative forcing of $\sim 1\ \mathrm{W/m^2}$, also compared to a year 2000 baseline. We also have two sets of global perturbation equilibrium experiments: complete simultaneous global removal of anthropogenic aerosols in 1.5°C warming world (with 430 ppm CO_2) from Samset et al. (2018) and experiments where global anthropogenic emissions of each of the three aerosol species are removed individually from Baker et al. (2015).

These experiments span a wide range of plausible future changes in anthropogenic aerosols and the corresponding temperature responses, from completely stopping all emissions and inducing warming globally (*Global Anthro Removal* in Figure 2) to maximum historical levels of industrial emissions causing cooling globally (*hist-aer* in Figure 2). Current emissions are in between these two and are being reduced (at differing rates) in most regions of the world (Quaas et al., 2022). Globally, this trend is likely to continue; however, there may be regional increases as manufacturing and industrial production increases in regions such as Africa and South-East Asia (Lund et al., 2019; Myhre et al., 2017). Including experiments with both increasing and decreasing regional emissions for individual aerosol species, which cause more regionally heterogeneous temperature impacts, is also important for spanning the potential emission-space and for accurate training of the emulator. In the simulations included here, the temperature response is much more pronounced in the Northern Hemisphere than in the south (Figure 2), especially in the midlatitudes (where the emission changes primarily happen) and in the Arctic.

While there are substantial differences between the two generations of NorESM used here (see Section 3.1), we find it more important to span as large a range as possible of potential emission scenarios and in general find improved performance of the emulator as more diverse data are included in the training data set. Indeed, this could be considered similar to including multiple models, which is a common method for increasing the representation of climate internal variability.

The aerosol emission data and NorESM model output are preprocessed into a consistent format for training the emulator. The inputs (**X**) to our emulator are the annual cumulative emissions of SO_2, BC, and OC in Tg/yr for six box regions (see Figure 1), as well as the global total. We do not use spatial information as inputs (latitude and longitude) because of the number of perturbation experiments we use. If the input data for all three aerosol species were used at native NorESM spatial resolution, and latitude and longitude were included as inputs, the size of the input data set would be intractable. The spatial resolution of NorESM is roughly 2° (96 × 144, latitude by longitude), resulting in 13,824 data pixels for one global map, and therefore using every pixel at annual resolution would result in roughly 83 million data points, which is intractable for GP methods. Previous studies have used data reduction methods such as empirical orthogonal functions (EOFs) to reduce the input data sets for global spatially resolved emulators (e.g., Watson-Parris et al., 2022) of climate scenarios primarily forced by greenhouse gas emission changes. Because we use aerosol and precursor emissions which are much more heterogeneous, and include many smaller regional perturbations in our training data set, we found there was no small set of spatial EOFs which described the variance of the training data set. We also wanted to retain regional forcing information, so therefore we sum the gridded emissions into eight regions: Europe, Russia, East Asia, South Asia, North America, and Africa (Figure 1). We also include the global sum as a training input and therefore end up with seven inputs for each aerosol species and a total of 21 input data dimensions for each year.

The target data for emulation (**Y**) is the annual mean change in temperature induced by the emission perturbation, calculated as the difference in surface temperature in degrees Kelvin between the perturbation experiment and the corresponding baseline simulation, at the native spatial resolution of NorESM. By regressing out the baseline experiment, we remove the impact of greenhouse gases such as CO_2 on temperature and emulate only the response driven by changes in aerosols. This is possible because the temperature response to single-forcing experiments tend to combine linearly, so subtracting the baseline from the perturbation experiment reveals the single-forcing response (Bône et al., 2023; Marvel et al., 2015). We leave the targets at the native NorESM spatial resolution and predict the temperature response at each grid point. For the transient experiments, an ensemble of NorESM outputs is available. In these cases, we average all ensemble members together for each experiment and predict the ensemble mean response. For the equilibrium experiments, we treat the years after spin-up as the ensemble and predict the mean response. Therefore, the posterior prediction (**Y**) of the emulator is a roughly 2° (96 × 144, latitude by longitude) resolution map of the temperature change for a given year. The resulting training data set consists of 621 input-output (perturbation-response) pairs.

3.3. Gaussian Process Model Description

Here, we give a brief overview of the theory behind GPs (Rasmussen & Williams, 2006) and the specific setup used in AeroGP. GPs are a probabilistic and nonparametric machine learning approach for modeling functions, well suited to climate science problems: they involve noisy data with strong temporal and spatial (co-)variability driven by systems for which we have some underlying prior knowledge (such as physical equations) as well as complex sources of uncertainty. GPs have been used for many years in geostatistics where the method is known as *kriging* (Chilès & Desassis, 2018; Cressie & Wikle, 2011; Matheron, 1963) and have been applied to a wide range of nonlinear regression and emulation studies within climate science (Camps-Valls et al., 2016; Glassmeier et al., 2019; Hamelijnck et al., 2021).

In general, a GP model is a distribution of functions such that the joint distribution of every finite subset of function values is also multivariate Gaussian. The model can be fully described by a mean function,

$$\mathbb{E}[f(x)] = \mu(x) \tag{1}$$

and a kernel (or covariance matrix),

$$\mathbb{E}[(f(x)-\mu(x))(f(x')-\mu(x'))] = \mathrm{Cov}[f(x),f(x')] = k(x,x') \tag{2}$$

such that

$$f(x) \sim \mathcal{GP}(\mu(x), k(x,x')) \tag{3}$$

The kernel describes the similarity between two values in the input space x, x' and therefore the covariance between the output GP values at those inputs $(f(x), f(x'))$. The mean $\mu(x)$ can be set to any function which describes the average behavior of the system; however, it is common to center in the training data and set the prior mean to zero, such that

$$f(x) \sim \mathcal{GP}(0, k(x,x')) \tag{4}$$

This does not restrict the mean of the posterior process or predictions of the model to zero, and uncertainty about the mean function can be taken into account by adding a noise term to the kernel. In the case of a zero mean function, the physical knowledge about the system, and therefore the behavior of the GP, is determined by the choice of kernel. The kernel describes the correlation between any two output values $(f(x), f(x'))$ and therefore restricts the properties of the functions which are possible under a given GP prior, for example, their smoothness or stationarity. A commonly used type of kernel in machine learning is the Matérn family, which we will use in this study, and is given by

$$k_\nu(x,x') = \sigma_f^2 \frac{2^{(1-\nu)}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\,|x-x'|}{l} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\,|x-x'|}{l} \right) \tag{5}$$

where l is the length-scale hyperparameter, σ_f^2 is the variance hyperparameter, K_ν is a modified Bessel function, Γ is the gamma function, and ν is a smoothness parameter which is typically $\nu \in \left(\frac{1}{2}, \frac{3}{2}, \frac{5}{2}\right)$ such that the kernel becomes the product of an exponential function and a polynomial. Kernels can be defined over multidimensional inputs (where x is then a vector), either sharing hyperparameters (such as l) across dimensions or with different length scales for each input dimension in which case the kernel is anisotropic. In AeroGP, we use an anisotropic Matérn$_{3/2}$ kernel for the regional aerosol response to each aerosol species, with a different length scale for each emission region. A GP which uses the Matérn$_{3/2}$ kernel is continuous and once differentiable, which balances smoothness and local variability.

Training a GP model is done by conditioning the GP prior on a given set of training data: inputs $X = \{x_N\}$ and outputs (or targets) $y = \{y_N\}$. If the target data are created by a noisy process, then

$$y = f(x) + \epsilon \tag{6}$$

where $\epsilon \sim \mathcal{N}(0, \sigma_y^2 I)$ is independent Gaussian noise, and $f(x)$ is modeled with a Gaussian prior as in Equation 4 with kernel hyperparameters θ, then for a finite set of input data X the marginal likelihood, which is the probability density of the data given the parameter is

$$p(y|X,\theta) \sim \mathcal{N}(y|\mu, K_{XX} + I\sigma_y^2) \tag{7}$$

where K_{XX} is the covariance matrix evaluated at training inputs X. The kernel hyperparameters θ and the noise variance σ_y^2 can be learned by maximizing the log marginal likelihood,

$$\arg\max_\theta \log p(y|X,\theta) \tag{8}$$

Finally, the posterior predictive model over possible functions given the training data is derived via Bayes theorem:

$$p(f|X,y) \propto p(y|f,X)p(f|X) \tag{9}$$

The posterior predictive distribution for new test inputs X_* is

$$p(f_*|X_*,X,y,\theta) = \mathcal{N}(\mu_*,K_*) \tag{10}$$

and the posterior mean and covariance can be calculated analytically as

$$\mu_* = \mu + K_{X_*X}\left[K_{XX} + I\sigma^2\right]^{-1}(y-\mu) \tag{11}$$

$$K_* = K_{X_*X_*} - K_{X*X}\left[K_{XX} + I\sigma^2\right]^{-1}K_{XX_*} \tag{12}$$

where K_{X_*X} is the cross-covariance between training and new inputs, and $K_{X_*X_*}$ is the covariance matrix evaluated at the new inputs. The above is the general case for a single output GP, which can also be driven by multidimensional inputs.

For climate-science problems, we are often interested in modeling variables which depend on spatial locations as well as some physical driving parameters. Typically for geospatial GP modeling, that means latitude, longitude, and sometimes other physical variables are used as inputs to the model. This works well for limited area studies (e.g., modeling air quality in a city (Hamelijnck et al., 2021) or precipitation in a particular basin (Lalchand et al., 2023)), where the Euclidean distance metric $|x - x'|$ used in most kernel functions is an approximately valid metric between locations. However, this does not work for global problems because it is not a valid distance metric on a sphere, and additionally it assumes that correlation decays with distance and therefore does not necessarily capture teleconnections. The solution is to either transform the coordinates to some alternative space, define a new distance metric which is valid for spherical coordinates, or simply not use latitude and longitude as predictors. Here, we choose the latter, which also keeps the input data set at a reasonable size. We do retain regional forcing information by using regional emission totals as our inputs directly.

However, we do not want to ignore the spatial nature of our data entirely, and we are especially interested in the spatial covariability of surface temperature. Previous spatially resolved global emulation studies (Bouabid et al., 2024; Watson-Parris et al., 2022) assume each grid point is independent and model them with independent samples from the GP prior for each location. Here, we extend the basic GP model described above to a multioutput case which considers the correlation between outputs, called the intrinsic model of coregionalization (IMC) (Journel & Huijbregts, 1976; van der Wilk et al., 2020). We model the covariability of each grid point by expressing each output as a combination of shared latent GPs, weighted by a coregionalization matrix, which dictates how the latent functions influence the outputs. This setup enables the modeling of complex dependencies among outputs and takes into account the fact that temperatures at each grid point will be in some way correlated with those around them, rather than treating each grid point as independent. Mathematically, this is equivalent to weighting the kernel function by a coregionalization matrix **B** such that

$$\mathbf{k}(\mathbf{x},\mathbf{x}') = k(x,x') \otimes \mathbf{B}\mathbf{B}^T \tag{13}$$

Standard GP regression can become computationally prohibitive for ICM, especially with large data sets or many outputs (such as in our case), because of the need to invert the kernel **k**. Therefore, we use the sparse variational GP (SVGP, (Hensman et al., 2013, 2015)) method to implement IMC for our model. SVGP approximates the full GP by using a subset of the full training data set, called inducing points $Z = z_m$ with corresponding inducing variables $u = f(Z)$ where M is the number of inducing points, N is the number of original training data points, and $M \ll N$. These act as a compressed representation of the GP prior so that the full function $f(x)$ is expressed conditionally on u as

$$p(f|u,\theta) = \mathcal{N}\left(K_{NM}K_{MM}^{-1}u, K_{NN} - K_{NM}K_{MM}^{-1}K_{MN}\right) \tag{14}$$

The posterior distribution over u is given by a variational distribution

$$q(u) = \mathcal{N}(m, S) \tag{15}$$

such that the posterior over the function values becomes

$$q(f) = \int p(f|u, \theta)q(u)du \tag{16}$$

which is optimized to be as close as possible to the true posterior (Titsias, 2009; Titsias & Lawrence, 2010). The variational parameters m, S, the kernel hyperparameters θ, and the inducing point locations Z are all learned during optimization, which is done by maximizing the evidence lower bound (ELBO), given by

$$L = \mathbb{E}_{q(f)}[logp(y|f)] - D_{KL}[q(u)\| p(u)] \tag{17}$$

which balances the likelihood of the observed data given the inducing points and a regularization term (D_{KL}, the Kullback-Leibler divergence) that measures how closely the variational distribution matches the GP prior.

Here, we find a good balance between training time and emulator performance with 100 inducing points, which are taken in the time dimension (so this is a randomized subset of 100 years from the training data set). The locations of the inducing variables within the input data space are treated as latent variables and learned during optimization. The ELBO is maximized using natural gradients (Salimbeni et al., 2018) to optimize the variational parameters and the Adam algorithm to optimize the kernel hyperparameters (Kingma & Ba, 2017).

Here, we use one latent GP, with a kernel that is a sum of linear kernels for the global response and Matérn$_{3/2}$ kernels for the regional dependence of each aerosol species. Combinations of linear and Matérn kernels have been shown to work well in other atmospheric modeling contexts, that is, Lamminpää et al. (2025). Temperature can be thought of as a diffusion process in that closer locations should be more highly correlated than more distant locations, but with the potential for remote teleconnections. Adding kernels together allows for these nonlocal interactions and improved extrapolation away from the training data because the additive combination can account for covariance between function values which are similar in any subset of dimensions. We also add a white noise kernel for modeling internal variability and set the initial likelihood variance (or nugget) to 0.5, which is also optimized during training. AeroGP is implemented with the Python package GPflow (Matthews et al., 2017; van der Wilk et al., 2020).

3.4. Metrics for Emulator Validation

We evaluate the performance of AeroGP by comparing the predictions to that of NorESM in a set of leave-one-out experiments, where the exact same emulator architecture is trained on the full set of training data except one experiment, which is then used for validation.

We use four common metrics to evaluate the performance of the emulator. In all equations in this section, x is the posterior mean prediction from AeroGP and y is the target NorESM prediction. The global mean is denoted by <>, which is weighted by latitude to account for the decreasing grid size toward the poles by

$$\langle x_{i,j} \rangle = \frac{1}{N_{lat}N_{lon}} \sum_{i}^{N_{lat}} \sum_{j}^{N_{lon}} cos(lat(i))x_{i,j} \tag{18}$$

where i and j are the indices for latitude and longitude, respectively.

The first metric is the normalized, global mean root mean square error (NRMSE), calculated as

$$\text{NRMSE}(x, y) = \sqrt{\left\langle \left(x_{i,j} - y_{i,j,n} | \right)^2 \right\rangle} / \langle \sigma_y \rangle \tag{19}$$

which is the global mean RMSE normalized by the standard deviation of the NorESM prediction. An NRMSE < 1 implies that the error between AeroGP and NorESM is less than the internal variability in NorESM, and therefore the AeroGP prediction could be considered to have some skill compared to natural internal variability. An NRMSE > 1 would imply the prediction error between AeroGP and NorESM is larger than the natural variability within NorESM itself.

The second metric is the global mean bias, defined as

$$\mathrm{BIAS}(x,y) = \langle x_{i,j} - y_{i,j} \rangle \tag{20}$$

This is simply the global mean of the difference between the AeroGP and NorESM predictions, and therefore a lower score (closer to zero) indicates better performance.

For the transient experiments, these metrics are calculated using the average across the last 20 years of the experiment for both AeroGP and the target NorESM output (which itself is an average across all available ensemble members). For the equilibrium runs, the target NorESM output is averaged across all years after model spin-up (as defined in the original publications) and compared to the single posterior map predicted by the emulator.

The third metric is the generalized continuous ranked probability score (CRPS) (Gneiting et al., 2005; Wilks, 2019), which can be considered as a probabilistic extension of the RMSE that compares the posterior distribution from AeroGP to the NorESM target. CRPS is defined as

$$\mathrm{CRPS}(F,y) = \int \left(\langle F_{i,j}(x)\rangle - \langle F_{i,j}(y)\rangle\right)^2 dx \tag{21}$$

where $F(x)$ is the cumulative distribution function (CDF) of the prediction (x) or target (y). This measures the area between the two CDFs so that smaller values indicate better performance. The CDFs can be approximated over finite ensembles using quadrature or direct integration if the PDFs can be assumed to be Gaussian. CRPS is often used in evaluation of probabilistic weather forecasts (e.g., Hersbach, 2000) and a better score (closer to zero) indicates both an accurate prediction of the mean as well as good calibration of the variance.

The fourth metric is the confidence interval-based Expected Calibration Error (ECE) (Kuleshov et al., 2018), which checks the calibration of the posterior variance by calculating which percentage of the predictions falls within a set of predictive intervals. It is defined as the mean of the absolute difference between the nominal and empirical coverage in each of the M bins:

$$\mathrm{ECE} = \frac{1}{M}\sum_{m} \left| cov_{emp}(\alpha_m) - \alpha_m \right| \tag{22}$$

where the empirical coverage in each bin is given by

$$cov_{emp}(\alpha_m) = \frac{1}{N}\sum_{i} \left[y_i \epsilon PI_i(\alpha_m) \right] \tag{23}$$

and where μ_i is the AeroGP posterior mean, σ_i is the AeroGP posterior standard deviation, y_i is the NorESM target, α is the nominal coverage level (i.e., 0.9 for 90%), and

$$PI_i(\alpha_m) = \left[\mu_i - z_\alpha \sigma_i, \mu_i + z_\alpha \sigma_i \right] \tag{24}$$

is the predictive interval where z_α is the quantile that corresponds to the central α coverage. A lower ECE score (closer to zero) indicates better calibration of the posterior variance.

In addition to these single-valued metrics, we also present maps comparing the spatial patterns of the NorESM target and the AeroGP posterior mean prediction. In some regions, and especially for small aerosol perturbations, the resulting temperature response signal may be obscured by internal variability within NorESM (Tebaldi & Knutti, 2007), making it difficult to predict. We therefore measure the significance of the response in the original NorESM training data by comparing the perturbation experiment to the baseline using the two-sided Welch's *t*-

test (Welch, 1947), which is a version of the Student's *t*-test that allows the two populations to have different variances and sample sizes. The NorESM signal is considered significant in grid boxes where $p < 0.05$. For AeroGP, we define the significant response according to the posterior credible interval, which does not include zero:

$$\text{significant} = \mathbf{1}_A(\mu - z\sigma > 0 \textbf{ or } \mu + z\sigma < 0), \quad z = \phi^{-1}\left(\frac{1 - ci}{2}\right) \tag{25}$$

A significant result indicates that the posterior predictive mean is distinct from zero at the given percentage and the signal is unlikely to be due to internal variability alone. For example, if zero lies outside the 95% credible interval, the posterior probability that the mean change equals zero is less than 5%, which is the Bayesian analog of rejecting the null hypothesis used in the *t*-test at the $p < 0.05$ level. When comparing NorESM to AeroGP, we also test for significance with the two-sided Welch's *t*-test and use the same threshold ($p < 0.05$). Whenever we use the *t*-test, we control the false discovery rate with the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). In all maps, hatching shows where the response is significant according to these thresholds.

4. Results

4.1. Emulator Validation

To validate AeroGP, we conduct leave-one-out tests, where the emulator is trained on the full training data set excluding one scenario, which is then used for validation. This means that when AeroGP is evaluated with a certain scenario, it has not been seen in the training data. Good performance, especially across multiple scenarios, implies that the emulator architecture is well suited to the problem (predicting aerosol-temperature impacts) and that the training data used capture the required range of regional patterns of the response to aerosol forcing. Poor performance on a particular scenario could mean that the target signal is weak or that the experiment is at the edge of the input state-space where the emulator may struggle to extrapolate. The results for all scenarios are summarized in Figure 4 and a full table of performance metrics is included in Supporting Information S1 (see Figure S1). In general, we find that AeroGP performs well (NRMSE < 1.0) on a wide range of emission scenarios including both increasing and decreasing aerosol and precursor emissions, changes in only one, simultaneous changes in multiple aerosol species, and global as well as regional perturbations. It performs less well on experiments where there is a weak temperature anomaly in NorESM. In the remainder of this section, we will evaluate the performance of AeroGP in more detail focusing on two transient global experiments (RAMIP *ssp370-126aer* and DAMIP *hist-aer*), which together cover increasing and decreasing anthropogenic aerosol emission perturbations and either end of the spectrum of surface temperature responses (cf. Figure 2).

AeroGP performs best when tested on the RAMIP global scenario *ssp370-126aer*, where anthropogenic aerosol and precursor emissions are reduced according to an aggressive mitigation scenario and the global mean surface temperature increases (NRMSE = 0.47, ECE = 0.35, and the mean global temperature response is +0.44°C in AeroGP and +0.47°C in NorESM averaged over the last 20 years, see Figure S1 in Supporting Information S1). We also evaluate AeroGP's performance on DAMIP *hist-aer* where emissions increase from preindustrial to present-day levels and the global mean surface temperature decreases (NRMSE = 0.52, ECE = 0.11, and the global mean temperature response is −0.67°C in AeroGP and −0.84°C in NorESM averaged over the last 20 years, see Figure S1 in Supporting Information S1). AeroGP underestimates the global mean response slightly as compared to NorESM in both cases; however, the difference between them is not significant for much of the globe, especially over land surfaces. AeroGP produces a significant signal over the much of the same regions as NorESM and predicts realistic spatial patterns of response for both experiments (Figure 3).

In the case of RAMIP *ssp370-126aer*, aerosol and precursor emissions are reduced globally, primarily through large reductions in East and South Asia, and to a lesser extent in Africa and North America. The result is large local temperature anomalies in those areas, and additional nonlocal impacts in both North America and the Arctic via teleconnections. AeroGP creates realistic patterns of temperature anomalies for this scenario and therefore is able to emulate both these local and nonlocal effects, especially over land. The only region of significant difference between AeroGP and NorESM is in the North Atlantic where AeroGP underestimates both the magnitude of cooling in the "warming hole" (Keil et al., 2020) and the anomaly pattern across the North Atlantic. The combination results in an underestimation of the North Atlantic mean response: between 40° and 65° latitude and

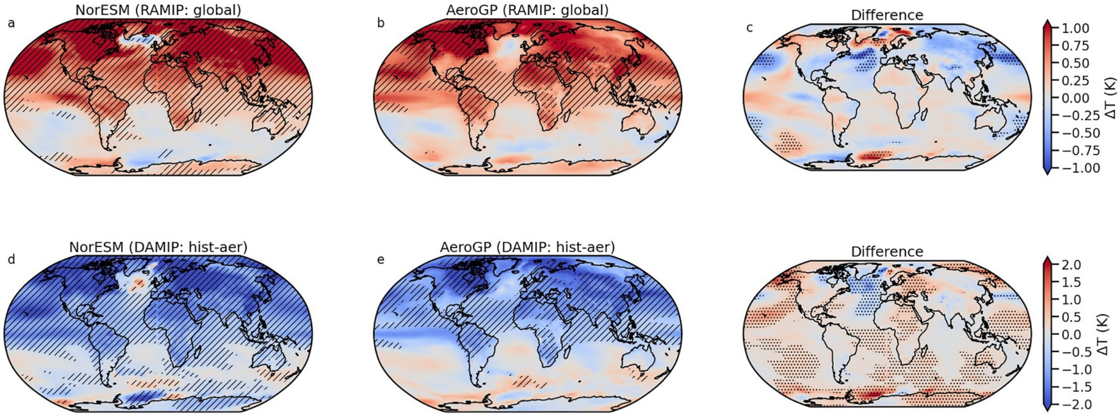

Figure 3. Time mean temperature response for *RAMIP global ssp370-126aer* (2045–2064) and *DAMIP hist-aer* (1995–2014). The left column shows the NorESM target (a, d), the center shows the AeroGP posterior mean prediction (b, e), and the right column shows the difference between the two (c, f). For NorESM and the differences, the shading indicates a significant response according to the Welch's t-test ($p < 0.05$). Shaded areas are where the AeroGP posterior mean is credibly different from zero at the 95% level for *DAMIP hist-aer*, corresponding to $\sim 2\sigma$ and at the 68% level for *RAMIP global ssp370-126aer*, corresponding to $\sim 1\sigma$ (see Section 3.4 for the calculation of credible intervals).

300° to 360° longitude, the mean response in NorESM is +0.48°C and in AeroGP is +0.38°C. AeroGP also underestimates the magnitude of the response in some parts of the North Pacific.

In the case of DAMIP *hist-aer*, aerosol and precursor emissions are increased according to a historical industrial emission scenario, while other climate forcers such as GHGs remain at preindustrial levels. Emissions increase primarily in Europe, Asia, and North America, and the result is a global cooling, particularly in the Northern Hemisphere over land. AeroGP is able to emulate the resulting spatial pattern in temperature anomalies and again performs best over land, although there are significant differences between AeroGP and NorESM in parts of Central Africa and Eastern Europe. The strongest temperature anomalies in AeroGP are found in the higher northern latitudes, particularly in North America, Russia, Scandinavia, and the Arctic, which is similar to the pattern of response in NorESM. AeroGP again underestimates the magnitude of the temperature anomaly in the North Atlantic and predicts a slightly weaker global mean response than NorESM.

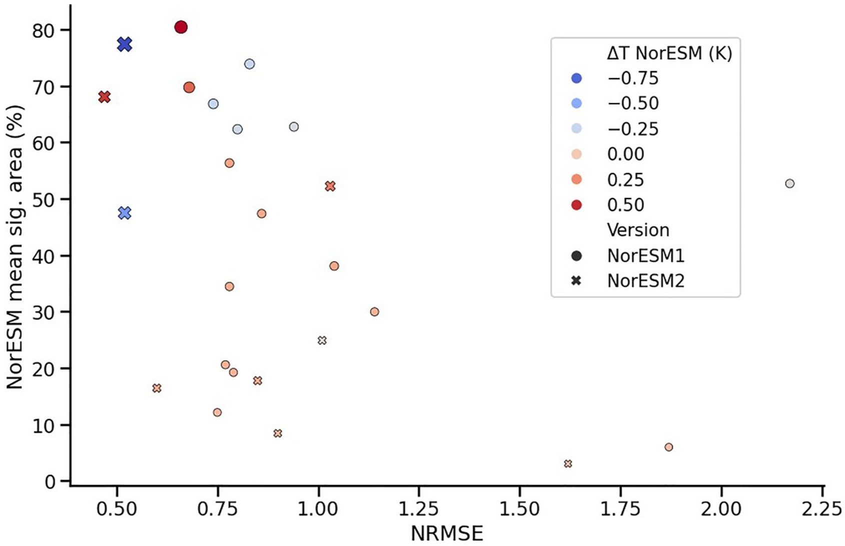

Figure 4. Normalized root mean square error (NRMSE) between NorESM targets and AeroGP predictions (using all years for equilibrium experiments and the last 20 years of all ensemble members for transient experiments), compared to how much of the globe has a significant signal in the NorESM target (in % so that 100% means a significant response everywhere, normalized by the grid area), with color of the markers indicating the global mean temperature change in the NorESM targets and the shape of the marker indicating the version of NorESM.

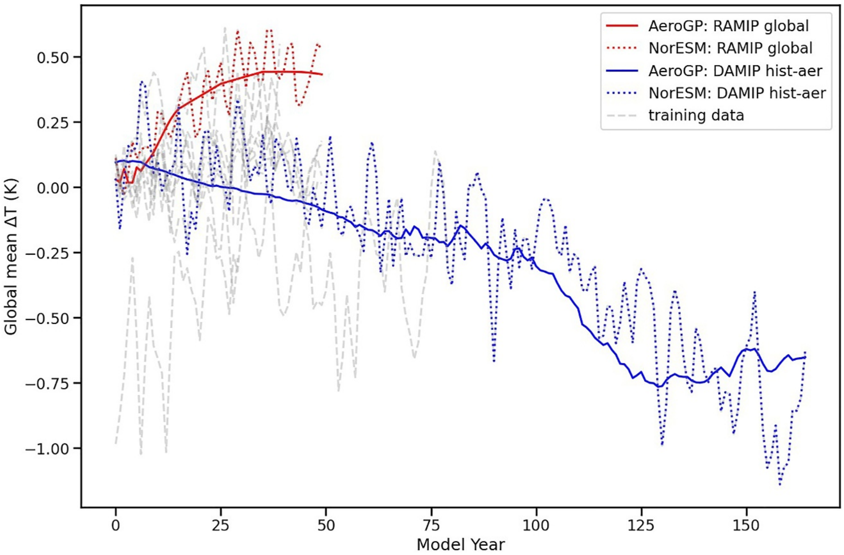

Figure 5. Global mean temperature change for the top two leave-one-out tests. AeroGP posterior mean predictions in solid lines, NorESM targets in dotted, and the other training data in dashed.

Five of the top six best-performing experiments are all global perturbations (emissions change by large amounts over most of the globe, see Figures S2 and S3 in Supporting Information S1), although they include both increasing and decreasing aerosol emission trends which differ regionally, as well as individual removal (e.g., the sixth best NRMSE is for the global removal of only SO_2: NRMSE = 0.68). Generally, we find that AeroGP emulates the temperature response more accurately when there is a strong signal to learn: when there is a significant response in the original NorESM targets and when the magnitude of the global mean temperature response is larger. A significant target signal in NorESM is a result of both a strong aerosol forcing and enough simulation years or ensemble members to have a significant signal (Figure 4). AeroGP has comparable performance to NorESM (NRMSE below 1.0) for 17 out of 24 experiments, with only one experiment having an NRMSE above 2.0. The performance does not seem to depend on the particular version of NorESM (Figure 4).

Both RAMIP *ssp370-126aer* and DAMIP *hist-aer* are transient experiments, meaning that aerosol and precursor emissions are changing throughout the simulation. We make the assumption that the annual mean temperature responses can be considered independently, and therefore AeroGP predicts each annual mean response independently from the previous years and does not take previous time steps into account. However, it is able to capture the long-term trend in the global mean temperature change for both experiments and produce some year-to-year variability around that trend, although less so compared to the internal variability in NorESM and in the real climate system (Figure 5).

In most experiments in our training data, the emission perturbations and resulting temperature changes are in the Northern Hemisphere, and there is generally a weak and nonsignificant response in the Southern Hemisphere. AeroGP captures this zonally dependent behavior so that variability in the mean response and the predicted standard deviation both increase toward the north and show evidence of Arctic amplification in the temperature response, as shown in Figure 6.

4.2. Sensitivity to Scaled Emissions

We now explore the sensitivity of the emulator to the magnitude and spatial extent of the aerosol perturbations by conducting three sensitivity experiments to test if the pattern of the temperature response differs depending on the magnitude and location of the emissions. We choose cases that are relevant for policy assessment and where AeroGP showed relatively good performance (see Figure S1 in Supporting Information S1): DAMIP *hist-aer*, RAMIP *EAS126aer* (where East Asian aerosols are reduced), and *5xEA* SO_2 (where East Asian SO_2 is increased). AeroGP is used to predict the temperature change from four idealized scaled versions of the input emissions: 0.1x, 0.5x, 1x, and 2x. By scaling the aerosol emission change in DAMIP *hist-aer*, we test the sensitivity to increasing

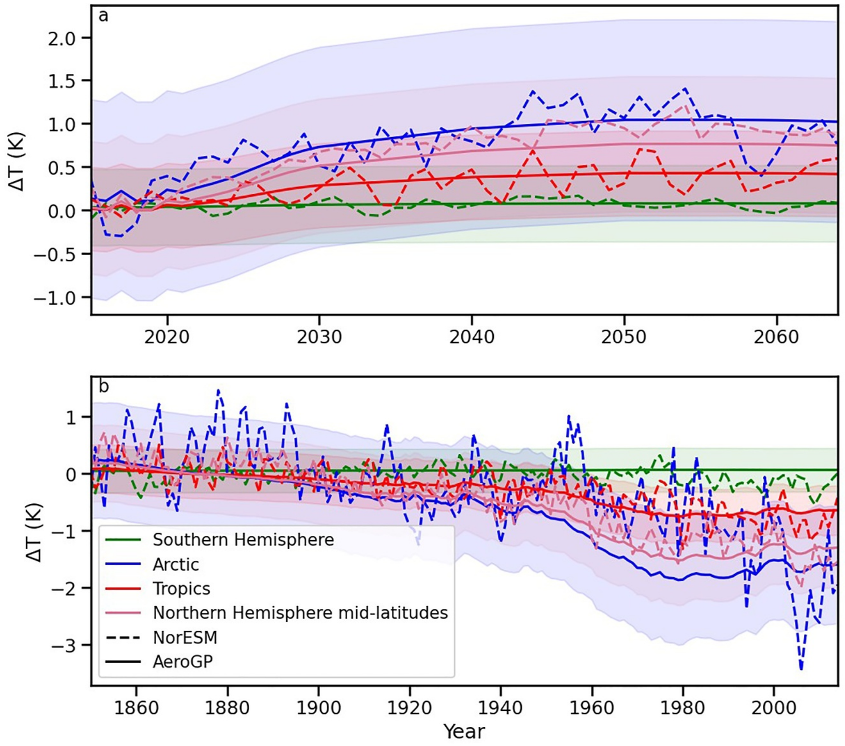

Figure 6. Zonal mean temperature response for RAMIP ssp370-126aer (top) and DAMIP hist-aer (bottom). The Southern Hemisphere in green (−90° to −30°), the Tropics in yellow (−30° to +30°), the Northern Hemisphere midlatitudes in red (30° to 60°), and the Arctic in blue (60° to 90°). AeroGP posterior mean predictions are in solid lines and NorESM targets are in dashed lines, with shading showing the posterior standard deviation (i.e., AeroGP predicted variability).

emissions globally, particularly in North America and Europe before 1975, and in Asia and Africa after 1980. Scaling RAMIP *EAS126aer* and *5xEA* SO_2 tests the sensitivity to a regional perturbation that is either a reduction in all aerosol and precursor emissions or an increase in one particular aerosol. It is important to know the scale of the emission change that is required to produce a significant signal for the motivation and assessment of potential emission policy changes. In each case, we use the validation version of AeroGP where the original experiment is held out of the training data set.

The rate and location at which the signal emerges depend on the magnitude of the emission perturbation. Figures 7a–7d show the posterior predictions for scaled DAMIP *hist-aer* emission changes. A significant response is predicted for much of the Northern Hemisphere at 0.5x emissions, with the strongest response in eastern North America and the Scandinavian Arctic. At the original level of emission change, the AeroGP pattern is similar to that predicted by NorESM (see Figure 3) and a significant response is found over all continents. At 2x original emissions, the response is strong (cooler than −3.0°C) across most of the Northern Hemisphere landmasses and a significant response is found almost globally, including a strong pattern in the Southern Ocean and the tropical Pacific. Figures 7e–7h show the posterior predictions for RAMIP *EAS126aer*, where all three aerosol species are reduced only in East Asia. When the emissions are scaled down (0.5 × the original emission change), there is no significant response predicted by AeroGP. At the original level of emission change, a significant response is predicted across the Arctic, Eastern Europe, East Asia, and eastern North America. At 2× the emission perturbation (very strong reductions in East Asia), there is a significant response across the Northern Hemisphere and the temperature impact is large (>+0.5°C) on most northern continents including Europe as well as in the Arctic. Figures 7i–7l show the posterior predictions for *5xEA* SO_2, where Eastern Asia sulfur dioxide emissions are increased 5x compared to the year 2000. Again, no significant response is predicted when the emission change is scaled down (0.1x and 0.5x). At the original level, there is only a significant response in a small part of Asia and some of the Tropics. At 2x emissions, there is a much stronger and significant response over Europe, Asia, and most of North America. Although both *hist-aer* and *5xEA* SO_2 produce cooling in the Northern Hemisphere, AeroGP distinguishes the difference in effect between increasing all three aerosol species globally in *hist-aer* and

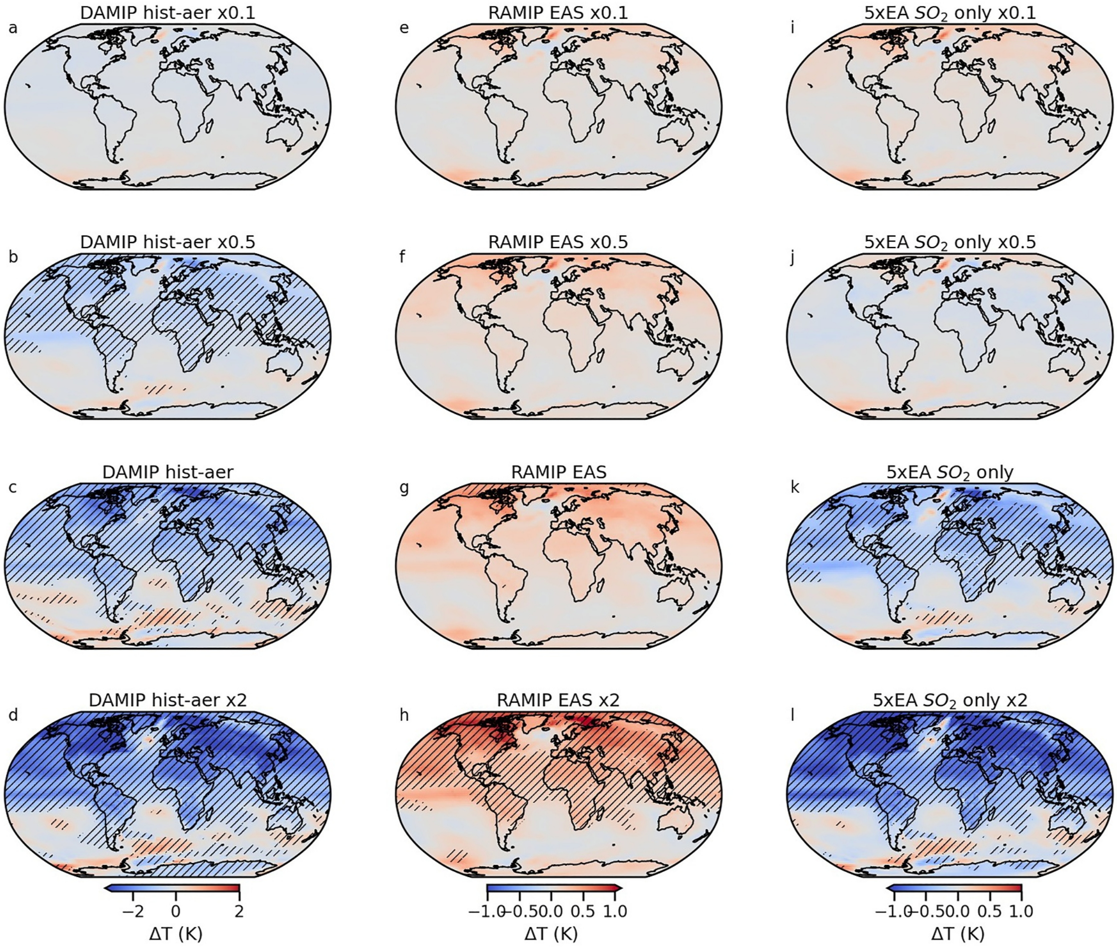

Figure 7. AeroGP predicted temperature response for scaled emission perturbations. Shaded areas are where the AeroGP posterior mean is credibly different from zero at the 68% level, corresponding to $\sim 1\sigma$.

SO_2 only in *5xEA SO_2*. The pattern and the magnitude of the response in each are distinct: the global mean response in *hist-aer* is −0.66°C, whereas the global mean response in *5xEA SO_2* is −0.15°C. We also find that the posterior prediction from AeroGP is not the same response pattern scaled up or down as the input emissions are scaled up or down, but rather a distinct spatial pattern that depends on the magnitude of the emission perturbation (See Figures S4–S6 in Supporting Information S1, which show the differences between the predicted responses to the scaled inputs and the baseline response scaled by the same amount).

4.3. Arctic Aerosol Impacts

The Arctic is a very climatically sensitive region, warming at more than two times the global average (Boucher et al., 2013; Constable et al., 2022). It is also sensitive to aerosol forcing, both from remote temperature responses (i.e., teleconnections) (Conley et al., 2018; Sand et al., 2013, 2020; Westervelt et al., 2020; von Salzen et al., 2022) and long-range transport of aerosols into the Arctic (Backman et al., 2021). These impacts are of particular relevance to policymakers globally, including those from Arctic-adjacent countries who may have direct economic and political consequences from a warming Arctic as well as those from other regions who are concerned about mitigating climate change and Arctic amplification. Understanding the impact from remote regional emission changes on a region like the Arctic is an excellent use case for AeroGP, which can resolve the spatial heterogeneity of the temperature response.

We investigate the Arctic response to an increase in aerosol and precursor emissions by doubling the year 2025 emissions of each aerosol type individually in all seven regions used as input to AeroGP. This tests the Arctic sensitivity to both the magnitude and location of the perturbations; for example, doubling current European SO_2

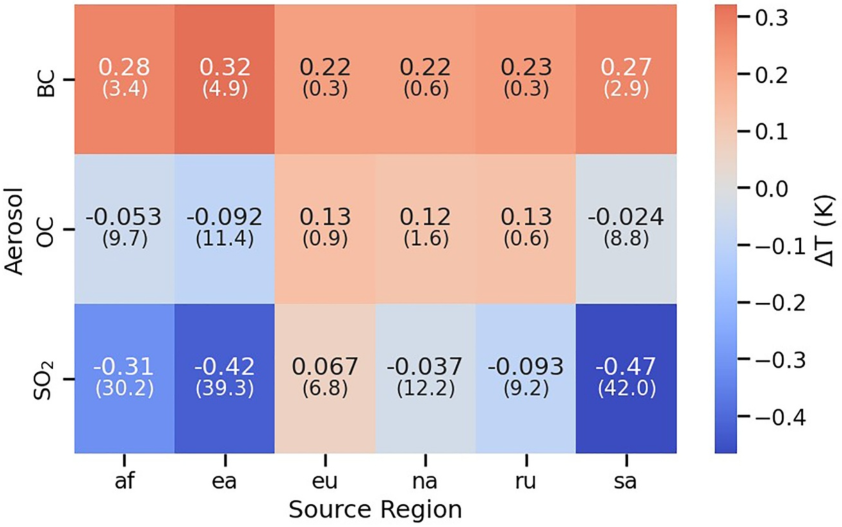

Figure 8. Area-weighted mean temperature change above 60°N for individual aerosol perturbations of 2x 2025 emissions in each region. Smaller numbers in brackets show the total regional perturbation of each aerosol type in teragrams. Regional labels are af = Africa, ea = East Asia/China, eu = Europe, na = North America, ru = Russia, sa = South Asia/India.

emissions corresponds to an increase of +6.8 Tg, whereas doubling current South Asia emissions corresponds to an increase of +42.0 Tg. Figure 8 shows the mean temperature response above 60°N for each regional aerosol perturbation. Increasing BC emissions produces Arctic warming no matter where the emissions originate, however, the strongest Arctic warming coming from increased African (+3.4 Tg) and Asian (+4.9 Tg and +2.9 Tg) emissions, that is, the regions with the largest emission changes. This is likely primarily due to remote temperature responses from heat transport (Sand et al., 2013, 2020), although there is some evidence for long-range transport of BC through Central Asia and Europe which could cause local radiative forcing (Backman et al., 2021; Liu et al., 2015). Increasing OC emissions produces a weak warming response for source regions, which are directly adjacent to the Arctic (North America, Europe, and Russia), although the response is nonsignificant in the whole Arctic regardless of the source region or magnitude of perturbation (+0.6 to +11.4 Tg). Increasing SO_2 emissions produces strong cooling in the Arctic when the emissions are from Africa (+30.2 Tg) and Asia (+39.3 and +42.0 Tg), most likely due to long-range teleconnections (Lewinschal et al., 2019), and a weak cooling when emissions are from North America (+12.2 Tg) and Russia (+9.2 Tg). Increasing SO_2 emissions from Europe (+6.8 Tg) produces a nonsignificant warming response, although this is due to a persistent warming feature east of Greenland, which washes out the cooling response in the European Arctic (see Figure 9).

Figure 9 shows the spatial pattern of the temperature response to emission perturbations in Europe and South Asia (primarily India). The response to European emissions is warming for all three aerosols, but the response is not significantly different from zero across the entire region, even using a very relaxed credible interval of half a standard deviation. Other studies have found a strong Arctic cooling response to a much larger magnitude of increasing European SO_2 emissions, for example, Lewinschal et al., 2019 find significant Arctic cooling with an increase of 7xEU SO_2 compared to the year 2000, (roughly +100 Tg), so it is likely that the perturbation here is too small to produce a significant response. The mean Arctic response is also dominated by a persistent warming feature present east of Greenland. The response to increased Asian emissions is much stronger because doubling emissions result in a much larger total perturbation. The response to increased Asian SO_2 is a significant cooling across the whole Arctic. The response to an increase in OC emissions is a weak warming that is not significant anywhere. The response to increased BC emissions is a warming which is only significant in the Canadian Arctic archipelago.

5. Discussion

Accurate regional climate projections are critical for quantifying near-term climate change and informing adaptation and mitigation efforts. Understanding the impact of anthropogenic aerosol emission changes is a key

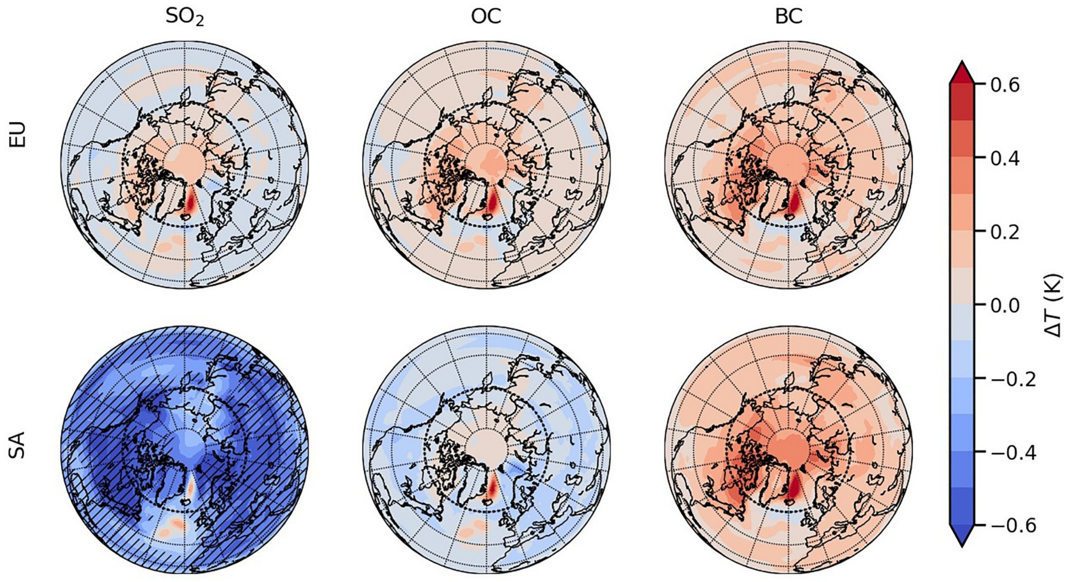

Figure 9. Maps of the Arctic temperature change for individual aerosol perturbations of 2x 2025 emissions in Europe (top row) and in South Asia/India (bottom row). Shaded areas are where the AeroGP posterior mean is credibly different from zero at the 38% level, corresponding to $\sim 0.5\sigma$.

part of understanding regional climate change because of aerosol impacts on atmospheric processes (Persad et al., 2022; L. Wilcox et al., 2019), the hydrological cycle (e.g., impacting Asian monsoon dynamics (Bartlett et al., 2018; L. J. Wilcox et al., 2020; Xie et al., 2020) and European precipitation patterns (Lopez-Romero et al., 2021)), and ocean heat and circulation (Hassan et al., 2021; Wang et al., 2024). It is also critical to understand the relationship between clean air policies and the resulting climate forcing for informing policy decisions and for impact assessment and attribution studies. Here, we describe the development of a machine learning model (AeroGP) for predicting the spatially resolved temperature response to regional aerosol emission changes. Our model produces accurate estimates of the global mean response and the spatial distribution of the response for a wide range of global and regional emission scenarios.

Non-machine learning climate emulators and simplified climate models have been a key part of understanding climate change for many years; before the fifth IPCC assessment report (AR5), they were the main tool used to assess the climate response to potential future emission scenarios (Randall et al., 2007), and emulators have continued to complement the results from ESMs participating in CMIPs by, for example, exploring and constraining climate scenarios beyond the SSPs and determining the attribution of individual forcers to observed and future warming (Nicholls et al., 2021; Pirani et al., 2024; van Vuuren et al., 2011). Typically, climate model emulators take anthropogenic emissions as inputs and produce predictions of the resulting global mean radiative forcing or temperature (Meinshausen et al., 2011; Millar et al., 2017). These models often underestimate (or do not even include) aerosol forcing (Harmsen et al., 2015; Schwarber et al., 2019; van Vuuren et al., 2011), and in particular often miss the complex regional heterogeneity of aerosol impacts (Persad et al., 2022). Here, AeroGP predicts the spatial patterns of the temperature response to aerosol emission changes directly from training data, without imposed assumptions of the nature of the spatial pattern of the response, and without the dominant and confounding influence of including GHGs.

Because GHGs are long-lived and well mixed in the atmosphere, the temperature response to anthropogenic climate change is dominated by GHG forcing and the aerosol response is difficult to untangle. This is because of the combination of local rapid responses and remote impacts through transport, teleconnections, and aerosol-cloud interactions, which are unique to aerosol-climate forcing (Forster et al., 2021). Our results show that AeroGP can learn these impacts, for example, reproducing significant teleconnections from East Asian SO$_2$ emissions in regions such as Europe and the Arctic. Machine learning methods can learn the outcomes of such complex combination of processes for much less computational cost than a traditional ESM. The initial training of the ML model is possible on a standard laptop computer and once trained, inference is possible in a matter of seconds. Nevertheless, the quality of AeroGP depends on the availability of a comprehensive training data set and therefore does not replace the need for physics-based models but rather builds upon the foundational data provided by ESMs.

AeroGP also provides a significant improvement over traditional ESMs in terms of the computational resources required. The two simulations run with NorESM2 for this study (referred to as AeroGP in Table 1) required on average 1060 core hours per simulated year, performed on the supercomputer Tetralith from the National Academic Infrastructure for Supercomputing in Sweden (NAISS). Extrapolating, roughly 2 million core hours were used to create the entire training data set used in this study. In comparison, AeroGP takes less than an hour to train on a conventional PC laptop and is almost instantaneous to run.

The computational expense is part of the motivation for constructing the training data set for AeroGP from an ensemble of opportunity, primarily from preexisting ESM experiments conducted for other studies. This approach allows us to leverage a wide diversity of aerosol perturbations without incurring the computational cost of running a dedicated suite of simulations. However, this does introduce heterogeneity in the experiment design across the data set, for example, differences in perturbation location and scales, ensemble size, and background climate state for both transient and equilibrium runs. These choices in the experimental design are often motivated by the desire to have a clear signal distinct from natural internal variability; for example, idealized equilibrium perturbation experiments (e.g., Lewinschal et al., 2019) typically have very large emission changes and long simulation periods after model spin-up which are averaged together, whereas transient experiments (e.g., L. J. Wilcox et al., 2023) require multiple ensemble members disentangle natural variability from aerosol forcing. We consistently find that AeroGP learns the pattern and scale of the response better for experiments with larger signal-to-noise ratio (see Figure 4).

Internal climate variability can also be accounted for by using a multimodel ensemble; however, AeroGP is trained and tested using output from multiple generations of a single climate model, NorESM. This choice has its advantages, for example, NorESM2 incorporates relatively advanced aerosol treatment compared to other ESMs (Kirkevåg et al., 2018), and by only using one model, we retain a good understanding of the underlying physics which AeroGP learns from. There are differences between the two generations of NorESM, but we find it advantageous to have a larger and more diverse training data set that spans more of the range of potential emission scenarios; indeed if AeroGP is trained only on the available experiments from NorESM2, it performs worse, for example, at predicting the *RAMIP global ssp370-126aer* scenario. Including two generations of NorESM also increases the climate variability represented in the training data set, compared to using a single version. The results of our leave-one-out validation indicate that experiments on the edge of the input state-space are important for learning; when these unique experiments are left out of the training data set, the performance notably decreases (see Figure S1 in Supporting Information S1). For example, we have only one experiment which perturbs BC emissions only and only one experiment which perturbs African emissions. Including more parent models could broaden the training state space, but may at the same time wash out certain physical signals due to model structural differences such as differences in numerical schemes, physical modules, and parameter uncertainty, which are especially important for aerosol-related processes across CMIP6 models (Pathak et al., 2023; L. Wilcox, 2025). Differences across CMIP6 models include one-moment versus two-moment microphysics schemes, whether the aerosol-cloud interactions include both the first and second indirect effects, and whether chemistry is interactive or prescribed. This diversity means that the multimodel signal for small perturbations like regional aerosol changes is more washed out and harder to distinguish from the baseline climatology (which may also be different across different models). In addition, not all experiments included in our training data set were conducted by multiple models; multimodel output is mainly available for transient simulations which perturb multiple aerosols simultaneously (see Table 1). The choice between single-model and multimodel training therefore depends on prioritizing model-specific process fidelity and input-space sampling, compared to potentially undersampling natural variability, and here we choose the former as a first step.

The set of aerosol species represented in the training data also constrains AeroGP's applicability. The present data set includes perturbations to sulfate, black carbon, and organic carbon, and omits other potentially important aerosol species such as nitrate and ammonium. Nitrate aerosols in particular have radiative and cloud-interaction properties similar to sulfate and are projected to become increasingly important contributors to anthropogenic aerosol forcing by the end of the 21st century; however, they are not currently included in many ESMs including both NorESM1 and NorESM2 (Kirkevåg et al., 2013; Seland et al., 2020; L. Wilcox, 2025). AeroGP predictions therefore cannot account for their potential climatic impact. Inclusion of these aerosols in future training data sets would improve the physical completeness of the emulator and extend its applicability to more policy-relevant emission scenarios involving coemitted species.

The spatial distribution of aerosol perturbation regions and the choice of background climate state are also important considerations in designing a bespoke training data set for an aerosol-climate emulator such as AeroGP. To maximize policy relevance across the globe, such a data set should include emission perturbations across different regions at different stages of industrial development and consider how this heterogeneity is expected to change into the future. For example, including both industrialized regions which have undergone strong air quality improvements that result in reduced emissions and increased warming, such as East Asia (Samset et al., 2025), as well regions currently undergoing increased industrialization such as Africa, that may see increases in aerosol forcing before similar transitions occur (Lund et al., 2019). Such a data set should also address systematic biases such as the consistent undersampling of the Southern Hemisphere. It would also be interesting to consider sector-based sources, natural sources, and the interactions between natural and anthropogenic aerosol emissions, for example, as forest fires and biomass burning are expected to become increasingly prominent sources of aerosol emissions under climate change (Allen et al., 2024). These interactions also depend on the background climate state, as do smaller scale aerosol processes such as aging, transport, and deposition. To minimize potential sources of uncertainty in forcing in a bespoke data set, the background climate should be the same across all experiments and should reflect reasonable assumptions about future emission scenarios (Riahi et al., 2017).

In order to include a wide range of emission scenarios in the training data set but still keep the total number of inputs manageable, the spatial and temporal resolution of the input data was limited: we sum emissions over six large regions and consider only annual total emission changes. Therefore, AeroGP cannot resolve spatial emission perturbations that are located within the same region (e.g., the difference between emission changes in Scandinavia and in Spain), and it does not predict rapid (subannual) responses to emission perturbations. It predicts annual mean temperature responses based on annual total emissions for that year. Given the residence time of aerosols in the atmosphere, this temporal averaging should account for most aerosol effects, but we cannot say anything about the persistence of the response to a temporary emission increase or decrease, or about the different timescales of local versus nonlocal responses. For example, we may miss preconditioning impacts on the monsoon (Dong et al., 2019) or decadal response lags for the AMOC (Robson et al., 2022). Additionally, while our results suggest that large perturbations in emissions lead to more significant signals, it remains unclear whether the temperature response saturates as the scale of the emission change increases in AeroGP. From physics, we expect a nonlinear response, with maximum sensitivity to aerosols occurring for intermediate perturbation magnitudes, beyond which the impact of additional aerosol emission changes may diminish due to the saturation of radiative forcing (Bellouin et al., 2020).

Looking ahead, there are several directions for future work. One promising extension is using multioutput GP methods to jointly learn temperature and precipitation responses, which would broaden the applicability of AeroGP for policy assessment and impact studies. Additionally, incorporating a heteroskedastic likelihood, which could vary by latitude, could improve both spatial and temporal internal variability. Non-Gaussian likelihoods could also help in modeling precipitation, where the distributions are skewed and we are particularly interested in the impact of aerosol perturbations on extreme events, such as heatwaves or heavy rainfall.

6. Conclusions

In this study, we have developed a machine learning emulator (AeroGP) using Gaussian processes (GP) to predict the spatially resolved temperature response to regional aerosol-only emission perturbations, trained on output from the Norwegian Earth System Model (NorESM). AeroGP is uniquely capable of predicting the spatially resolved temperature response at ESM-resolution, which allows for a more detailed exploration of regional temperature impacts than is possible with most climate emulators that provide global mean responses, and AeroGP can additionally provide uncertainty estimates of that response. We have shown that AeroGP captures distinct patterns of temperature change influenced by the scale, aerosol type, magnitude, and location of emissions. Our results show that AeroGP performs comparably to NorESM in a series of leave-one-out validation experiments, demonstrating its sensitivity to realistic emission changes and ability to predict both global and regional temperature responses effectively. We have also demonstrated multiple potential use cases for AeroGP, including testing the global sensitivity to scaled emissions and investigating the Arctic response to regional increases in individual aerosol species. Detailed spatially resolved temperature predictions are a particular improvement over spatially averaged values for end-users of climate information, for example, policy-making, impact assessment, and economic evaluations such as for insurance.

We present a novel approach from a GP-architecture perspective, as we believe this is the first time that coregionalization has been used as a method to account for the spatial correlation of climate data. We hope this provides a step toward using machine learning methods for studying the global impacts of regional climate change and for developing tools that can be used for the evaluation of policy-relevant climate scenarios.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

Data and software availability: All processed training data and the code used to create it as well as the code required to build, train, and test the emulator are available at Dewey (2025). AeroGP is implemented with GPflow version 2.9.1 (Matthews et al., 2017; van der Wilk et al., 2020). The original NorESM model output is available according to the references in Table 1, or available to freely download from ESGF (https://aims2.llnl.gov/search).

Acknowledgments

Maura Dewey received funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreement No 860100 (iMIRACLI). Maria Sand and Srinath Krishnan received funding from the Norwegian Research Council through Grant 315195 (ACCEPT). Data storage and processing resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725. Parts of the simulations were performed and data exchanged on resources provided by UNINETT Sigma2—the National Infrastructure for High Performance Computing and Data Storage in Norway (nn9188k and ns9188k). Bjørn H. Samset and Laura J. Wilcox received funding from the Norwegian Research Council through Grant 324182 (CATHY). We also acknowledge support by the Center for Advanced Study in Oslo, Norway, which funded and hosted the HETCLIF center during the academic year of 2023/24. Laura J. Wilcox also received funding from the Natural Environment Research Council (NERC through Grant NE/W004895/1, TerraFIRMA).

References

Aas, W., Mortier, A., Bowersox, V., Cherian, R., Faluvegi, G., Fagerli, H., et al. (2019). Global and regional trends of atmospheric sulfur. *Scientific Reports*, *9*(1), 953. https://doi.org/10.1038/s41598-018-37304-0

Albrecht, B. A. (1989). Aerosols, cloud microphysics, and fractional cloudiness. *Science*, *245*(4923), 1227–1230. https://doi.org/10.1126/science.245.4923.1227

Allen, R. J., Samset, B. H., Wilcox, L. J., & Fisher, R. A. (2024). Are northern hemisphere boreal forest fires more sensitive to future aerosol mitigation than to greenhouse gas–driven warming? *Science Advances*, *10*(13), 4007. https://doi.org/10.1126/sciadv.adl4007

Axen, S. D., Gessner, A., Sommer, C., Weitzel, N., & Tejero-Cantero, Á. (2022). Spatiotemporal modeling of European paleoclimate using doubly sparse Gaussian processes. In *2022 NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems (GPSMDMS)*. Retrieved from http://arxiv.org/abs/2211.08160

Backman, J., Schmeisser, L., & Asmi, E. (2021). Asian emissions explain much of the arctic black carbon events. *Geophysical Research Letters*, *48*(5), e2020GL091913. https://doi.org/10.1029/2020GL091913

Baker, L. H., Collins, W. J., Olivié, D. J., Cherian, R., Hodnebrog, Myhre, G., & Quaas, J. (2015). Climate responses to anthropogenic emissions of short-lived climate pollutants. *Atmospheric Chemistry and Physics*, *15*(14), 8201–8216. https://doi.org/10.5194/acp-15-8201-2015

Ban-Weiss, G. A., Cao, L., Bala, G., & Caldeira, K. (2012). Dependence of climate forcing and response on the altitude of black carbon aerosols. *Climate Dynamics*, *38*(5–6), 897–911. https://doi.org/10.1007/s00382-011-1052-y

Bartlett, R., Bollasina, M., Booth, B., Dunstone, N., Marenco, F., Messori, G., & Bernie, D. (2018). Do differences in future sulfate emission pathways matter for near-term climate? A case study for the Asian monsoon. *Climate Dynamics*, *50*(5–6), 1863–1880. https://doi.org/10.1007/s00382-017-3726-6

Bellouin, N., Quaas, J., Gryspeerdt, E., Kinne, S., Stier, P., Watson-Parris, D., et al. (2020). Bounding global aerosol radiative forcing of climate change. *Reviews of Geophysics*, *58*(1), e2019RG000660. https://doi.org/10.1029/2019RG000660

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bentsen, M., Bethke, I., Debernard, J. B., Iversen, T., Kirkevåg, A., Seland, et al. (2013). The Norwegian Earth system model, NorESM1-M—Part 1: Description and basic evaluation of the physical climate. *Geoscientific Model Development*, *6*(3), 687–720. https://doi.org/10.5194/gmd-6-687-2013

Bock, L., & Lauer, A. (2024). Cloud properties and their projected changes in CMIP models with low to high climate sensitivity. *Atmospheric Chemistry and Physics*, *24*(3), 1587–1605. https://doi.org/10.5194/acp-24-1587-2024

Bond, T. C., Doherty, S. J., Fahey, D. W., Forster, P. M., Berntsen, T., Deangelo, B. J., et al. (2013). Bounding the role of black carbon in the climate system: A scientific assessment. *Journal of Geophysical Research: Atmospheres*, *118*(11), 5380–5552. https://doi.org/10.1002/jgrd.50171

Bône, C., Gastineau, G., Thiria, S., Gallinari, P., & Mejia, C. (2023). Detection and attribution of climate change using a neural network. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2022MS003475. https://doi.org/10.1029/2022MS003475

Bouabid, S., Sejdinovic, D., & Watson-Parris, D. (2024). FaIRGP: A Bayesian energy balance model for surface temperatures emulation. *Journal of Advances in Modeling Earth Systems*, *16*(6), e2023MS003926. https://doi.org/10.1029/2023MS003926

Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., et al. (2013). In T. F. Stocker, et al. (Eds.), *Clouds and aerosols* (pp. 571–657). https://doi.org/10.1017/CBO9781107415324.016

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

Chilès, J.-P., & Desassis, N. (2018). Fifty years of kriging. In B. Daya Sagar, Q. Cheng, & F. Agterberg (Eds.), *Handbook of mathematical geosciences: Fifty years of IAMG* (pp. 589–612). Springer International Publishing. https://doi.org/10.1007/978-3-319-78999-6_29

Collins, J. W., Lamarque, J. F., Schulz, M., Boucher, O., Eyring, V., Hegglin, I. M., 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

Conley, A. J., Westervelt, D. M., Lamarque, J.-F., Fiore, A. M., Shindell, D., Correa, G., et al. (2018). Multimodel surface temperature responses to removal of U.S. sulfur dioxide emissions. *Journal of Geophysical Research: Atmospheres*, *123*(5), 2773–2796. https://doi.org/10.1002/2017JD027411

Constable, A., Harper, S., Dawson, J., Holsman, K., Mustonen, T., Piepenburg, D., & Rost, B. (2022). Cross-chapter paper 6: Polar regions. In *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. https://doi.org/10.1017/9781009325844.023

Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons.

Crippa, M., Janssens-Maenhout, G., Dentener, F., Guizzardi, D., Sindelarova, K., Muntean, M., et al. (2016). Forty years of improvements in European air quality: Regional policy-industry interactions with global impacts. *Atmospheric Chemistry and Physics*, *16*(6), 3825–3841. https://doi.org/10.5194/acp-16-3825-2016

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*, e2019MS001916. https://doi.org/10.1029/2019MS001916

Dewey, M. (2025). mauradewey/aerogp: Final submission to jgr: Ml and computation. *Zenodo*. https://doi.org/10.5281/zenodo.17099941

Dong, B., Sutton, R. T., & Shaffrey, L. (2017). Understanding the rapid summer warming and changes in temperature extremes since the mid-1990s over western Europe. *Climate Dynamics*, *48*(5–6), 1537–1554. https://doi.org/10.1007/s00382-016-3158-8

Dong, B., Wilcox, L. J., Highwood, E. J., & Sutton, R. T. (2019). Impacts of recent decadal changes in Asian aerosols on the east Asian summer monsoon: Roles of aerosol–radiation and aerosol–cloud interactions. *Climate Dynamics*, *53*(5), 3235–3256. https://doi.org/10.1007/s00382-019-04698-0

Elguindi, N., Granier, C., Stavrakou, T., Darras, S., Bauwens, M., Cao, H., et al. (2020). Intercomparison of magnitudes and trends in anthropogenic surface emissions from bottom-up inventories, top-down estimates, and emission scenarios. *Earth's Future*, *8*(8), e2020EF001520. https://doi.org/10.1029/2020EF001520

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., Colins, W., Dufresne, J.-L., Frame, D., et al. (2021). The Earth's energy budget, climate feedbacks and climate sensitivity. In V. Masson-Delmotte, et al. (Eds.), *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 923–1054). Cambridge University Press. https://doi.org/10.1017/9781009157896.009

Gent, P. R., Yeager, S. G., Neale, R. B., Levis, S., & Bailey, D. A. (2010). Improvements in a half degree atmosphere/land version of the CCSM. *Climate Dynamics*, *34*(6), 819–833. https://doi.org/10.1007/s00382-009-0614-8

Gidden, M. J., Riahi, K., Smith, S. J., Fujimori, S., Luderer, G., Kriegler, E., et al. (2019). Global emissions pathways under different socioeconomic scenarios for use in CMIP6: A dataset of harmonized emissions trajectories through the end of the century. *Geoscientific Model Development*, *12*(4), 1443–1475. https://doi.org/10.5194/gmd-12-1443-2019

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

Glassmeier, F., Hoffmann, F., Johnson, J. S., Yamaguchi, T., Carslaw, K. S., & Feingold, G. (2019). An emulator approach to stratocumulus susceptibility. *Atmospheric Chemistry and Physics*, *19*(15), 10191–10203. https://doi.org/10.5194/acp-19-10191-2019

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

Hamelijnck, O., Wilkinson, W. J., Loppi, N. A., Solin, A., & Damoulas, T. (2021). Spatio-temporal variational Gaussian processes. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*. Retrieved from http://arxiv.org/abs/2111.01732

Hansen, J., Sato, M., & Ruedy, R. (1997). Radiative forcing and climate response. *Journal of Geophysical Research*, *102*(D6), 6831–6864. https://doi.org/10.1029/96JD03436

Harmsen, M. J. H. M., van Vuuren, D. P., van den Berg, M., Hof, A. F., Hope, C., Krey, V., et al. (2015). How well do integrated assessment models represent non-CO_2 radiative forcing? *Climatic Change*, *133*(4), 565–582. https://doi.org/10.1007/s10584-015-1485-0

Hassan, T., Allen, R. J., Liu, W., & Randles, C. A. (2021). Anthropogenic aerosol forcing of the Atlantic meridional overturning circulation and the associated mechanisms in CMIP6 models. *Atmospheric Chemistry and Physics*, *21*(8), 5821–5846. https://doi.org/10.5194/acp-21-5821-2021

Haywood, J., & Boucher, O. (2000). Estimates of the direct and indirect radiative forcing due to tropospheric aerosols: A review. *Reviews of Geophysics*, *38*(4), 513–543. https://doi.org/10.1029/1999RG000078

Hensman, J., de, G., Matthews, A. G., & Ghahramani, Z. (2015). Scalable variational Gaussian process classification. In *Proceedings of AISTATS*.

Hensman, J., Fusi, N., & Lawrence, N. D. (2013). Gaussian processes for big data. Retrieved from https://arxiv.org/abs/1309.6835

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015⟨0559:DOTCRP⟩2.0.CO;2

Hodnebrog, Ø., Myhre, G., Jouan, C., Andrews, T., Forster, P. M., Jia, H., et al. (2024). Recent reductions in aerosol emissions have increased Earth's energy imbalance. *Communications Earth & Environment*, *5*(1), 166. https://doi.org/10.1038/s43247-024-01324-8

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., et al. (2013). The community Earth system model: A framework for collaborative research. *Bulletin of the American Meteorological Society*, *94*, 1339–1360. https://doi.org/10.1175/BAMS-D-12-00121.1

Iversen, T., Bentsen, M., Bethke, I., Debernard, J. B., Kirkevåg, A., Seland, et al. (2013). The Norwegian Earth system model, NorESM1-M—Part 2: Climate response and scenario projections. *Geoscientific Model Development*, *6*(2), 389–415. https://doi.org/10.5194/gmd-6-389-2013

Janssens-Maenhout, G., Crippa, M., Guizzardi, D., Dentener, F., Muntean, M., Pouliot, G., et al. (2015). Htap v2.2: A mosaic of regional and global emission grid maps for 2008 and 2010 to study hemispheric transport of air pollution. *Atmospheric Chemistry and Physics*, *15*(19), 11411–11432. https://doi.org/10.5194/acp-15-11411-2015

Journel, A. G., & Huijbregts, C. J. (1976). *Mining geostatistics*. Academic Press.

Keil, P., Mauritsen, T., Jungclaus, J., Hedemann, C., Olonscheck, D., & Ghosh, R. (2020). Multiple drivers of the north Atlantic warming hole. *Nature Climate Change*, *10*(7), 667–671. https://doi.org/10.1038/s41558-020-0819-8

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. Retrieved from https://arxiv.org/abs/1412.6980

Kirkevåg, A., Grini, A., Olivié, D., Seland, Ø., Alterskjær, K., Hummel, M., et al. (2018). A production-tagged aerosol module for earth system models, OSLOAERO5.3—Extensions and updates for CAM5.3-OSLO. *Geoscientific Model Development*, *11*(10), 3945–3982. https://doi.org/10.5194/gmd-11-3945-2018

Kirkevåg, A., Iversen, T., Seland, Hoose, C., Kristjánsson, J. E., Struthers, H., et al. (2013). Aerosol–climate interactions in the Norwegian Earth system model—NorESM1-M. *Geoscientific Model Development*, *6*(1), 207–244. https://doi.org/10.5194/gmd-6-207-2013

Krock, M. L., Kleiber, W., Hammerling, D., & Becker, S. (2023). Modeling massive highly multivariate nonstationary spatial data with the basis graphical lasso. *Journal of Computational & Graphical Statistics*, *32*(4), 1472–1487. https://doi.org/10.1080/10618600.2023.2174126

Kuleshov, V., Fenner, N., & Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 2796–2804). PMLR. Retrieved from https://proceedings.mlr.press/v80/kuleshov18a.html

Lalchand, V., Tazi, K., Cheema, T. M., Turner, R. E., & Hosking, S. (2023). Kernel learning for explainable climate science. Retrieved from https://arxiv.org/abs/2209.04947

Lamminpää, O., Susiluoto, J., Hobbs, J., McDuffie, J., Braverman, A., & Owhadi, H. (2025). Forward model emulator for atmospheric radiative transfer using Gaussian processes and cross validation. *Atmospheric Measurement Techniques*, *18*(3), 673–694. https://doi.org/10.5194/amt-18-673-2025

Lewinschal, A., Ekman, A. M., Hansson, H. C., Sand, M., Berntsen, T. K., & Langner, J. (2019). Local and remote temperature response of regional SO_2 emissions. *Atmospheric Chemistry and Physics*, *19*(4), 2385–2403. https://doi.org/10.5194/acp-19-2385-2019

Li, J., Carlson, B. E., Yung, Y. L., Lv, D., Hansen, J., Penner, J. E., et al. (2022). Scattering and absorbing aerosols in the climate system. *Nature Reviews Earth & Environment*, *3*(6), 363–379. https://doi.org/10.1038/s43017-022-00296-7

Liu, D., Quennehen, B., Darbyshire, E., Allan, J. D., Williams, P. I., Taylor, J. W., et al. (2015). The importance of Asia as a source of black carbon to the European Arctic during springtime 2013. *Atmospheric Chemistry and Physics*, *15*(20), 11537–11555. https://doi.org/10.5194/acp-15-11537-2015

Lopez-Romero, J. M., Pedro Montávez, J., Jerez, S., Lorente-Plazas, R., Palacios-Peña, L., & Jiménez-Guerrero, P. (2021). Precipitation response to aerosol-radiation and aerosol-cloud interactions in regional climate simulations over Europe. *Atmospheric Chemistry and Physics*, *21*(1), 415–430. https://doi.org/10.5194/acp-21-415-2021

Lund, M. T., Myhre, G., & Samset, B. H. (2019). Anthropogenic aerosol forcing under the shared socioeconomic pathways. *Atmospheric Chemistry and Physics*, *19*(22), 13827–13839. https://doi.org/10.5194/acp-19-13827-2019

Marvel, K., Schmidt, G. A., Shindell, D., Bonfils, C., LeGrande, A. N., Nazarenko, L., & Tsigaridis, K. (2015). Do responses to different anthropogenic forcings add linearly in climate models? *Environmental Research Letters*, *10*(10), 104010. https://doi.org/10.1088/1748-9326/10/10/104010

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, *58*(8), 1246–1266. https://doi.org/10.2113/gsecongeo.58.8.1246

Matthews, A. G. D. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., et al. (2017). GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, *18*(40), 1–6. Retrieved from http://jmlr.org/papers/v18/16-537.html

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., Nicholls, Z. R., Friedlingstein, P., & Allen, M. R. (2017). A modified impulse-response representation of the global near-surface air temperature and atmospheric concentration response to carbon dioxide emissions. *Atmospheric Chemistry and Physics*, *17*(11), 7213–7228. https://doi.org/10.5194/acp-17-7213-2017

Myhre, G., Aas, W., Cherian, R., Collins, W., Faluvegi, G., Flanner, M., et al. (2017). Multi-model simulations of aerosol and ozone radiative forcing due to anthropogenic emission changes during the period 1990–2015. *Atmospheric Chemistry and Physics*, *17*(4), 2709–2720. https://doi.org/10.5194/acp-17-2709-2017

Neale, R. B., Richter, J., Park, S., Lauritzen, P. H., Vavrus, S. J., Rasch, P. J., & Zhang, M. (2013). The mean climate of the community atmosphere model (CAM4) in forced SST and fully coupled experiments. *Journal of Climate*, *26*(14), 5150–5168. https://doi.org/10.1175/JCLI-D-12-00236.1

Nicholls, Z., Meinshausen, M., Lewis, J., Corradi, M. R., Dorheim, K., Gasser, T., et al. (2021). Reduced complexity model intercomparison project phase 2: Synthesizing Earth system knowledge for probabilistic climate projections. *Earth's Future*, *9*(6), e2020EF001900. https://doi.org/10.1029/2020EF001900

O'Neill, B. C., Kriegler, E., Ebi, K. L., Kemp-Benedict, E., Riahi, K., Rothman, D. S., et al. (2017). The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century. *Global Environmental Change*, *42*, 169–180. https://doi.org/10.1016/j.gloenvcha.2015.01.004

O'Neill, B. C., Kriegler, E., Riahi, K., Ebi, K. L., Hallegatte, S., Carter, T. R., et al. (2014). A new scenario framework for climate change research: The concept of shared socioeconomic pathways. *Climatic Change*, *122*(3), 387–400. https://doi.org/10.1007/s10584-013-0905-2

Pathak, R., Dasari, H., Ashok, K., & Hoteit, I. (2023). Effects of multi-observations uncertainty and models similarity on climate change projections. *npj Climate and Atmospheric Science*, *6*(1), 144. https://doi.org/10.1038/s41612-023-00473-5

Persad, G. G. (2023). The dependence of aerosols' global and local precipitation impacts on the emitting region. *Atmospheric Chemistry and Physics*, *23*(6), 3435–3452. https://doi.org/10.5194/acp-23-3435-2023

Persad, G. G., & Caldeira, K. (2018). Divergent global-scale temperature effects from identical aerosols emitted in different regions. *Nature Communications*, *9*(1), 3289. https://doi.org/10.1038/s41467-018-05838-6

Persad, G. G., Samset, B. H., & Wilcox, L. J. (2022). Aerosols must be included in climate risk assessments. *Nature*, *611*(7937), 662–664. https://doi.org/10.1038/d41586-022-03763-9

Pirani, A., Fuglestvedt, J. S., Byers, E., O'Neill, B., Riahi, K., Lee, J.-Y., et al. (2024). Scenarios in IPCC assessments: Lessons from AR6 and opportunities for AR7. *npj Climate Action*, *3*(1), 1. https://doi.org/10.1038/s44168-023-00082-1

Polonik, P., Ricke, K., & Burney, J. (2021). Paris agreement's ambiguity about aerosols drives uncertain health and climate outcomes. *Earth's Future*, *9*(5), e2020EF001787. https://doi.org/10.1029/2020EF001787

Prospero, J. M., Charlson, R. J., Mohnen, V., Jaenicke, R., Delany, A. C., Moyers, J., et al. (1983). The atmospheric aerosol system: An overview. *Reviews of Geophysics*, *21*(7), 1607–1629. https://doi.org/10.1029/RG021i007p01607

Quaas, J., Andrews, T., Bellouin, N., Block, K., Boucher, O., Ceppi, P., et al. (2024). Adjustments to climate perturbations—Mechanisms, implications, observational constraints. *AGU Advances*, *5*(5), e2023AV001144. https://doi.org/10.1029/2023AV001144

Quaas, J., Jia, H., Smith, C., Albright, A. L., Aas, W., Bellouin, N., et al. (2022). Robust evidence for reversal of the trend in aerosol effective climate forcing. *Atmospheric Chemistry and Physics*, *22*(18), 12221–12239. https://doi.org/10.5194/acp-22-12221-2022

Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., et al. (2007). Climate models and their evaluation. In S. Solomon, et al. (Eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (Chap. 8)*. University Press. Cambridge, United Kingdom and New York, NY, USA: Cambridge.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Riahi, K., van Vuuren, D. P., Kriegler, E., Edmonds, J., O'Neill, B. C., Fujimori, S., et al. (2017). The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environmental Change*, *42*, 153–168. https://doi.org/10.1016/j.gloenvcha.2016.05.009

Richardson, T. B., Forster, P. M., Andrews, T., Boucher, O., Faluvegi, G., Fläschner, D., et al. (2018). Drivers of precipitation change: An energetic understanding. *Journal of Climate*, *31*(23), 9641–9657. https://doi.org/10.1175/JCLI-D-17-0240.1

Robson, J., Menary, M. B., Sutton, R. T., Mecking, J., Gregory, J. M., Jones, C., et al. (2022). The role of anthropogenic aerosol forcing in the 1850–1985 strengthening of the AMOC in CMIP6 historical simulations. *Journal of Climate*, *35*(20), 6843–6863. https://doi.org/10.1175/JCLI-D-22-0124.1

Salimbeni, H., Eleftheriadis, S., & Hensman, J. (2018). Natural gradients in practice: Non-conjugate variational inference in Gaussian process models. Retrieved from https://arxiv.org/abs/1803.09151

Samset, B. H. (2022). Aerosol absorption has an underappreciated role in historical precipitation change. *Communications Earth & Environment*, *3*(1), 242. https://doi.org/10.1038/s43247-022-00576-6

Samset, B. H., Myhre, G., Forster, P. M., Hodnebrog, Andrews, T., Faluvegi, G., et al. (2016). Fast and slow precipitation responses to individual climate forcers: A PDRMIP multimodel study. *Geophysical Research Letters*, *43*(6), 2782–2791. https://doi.org/10.1002/2016GL068064

Samset, B. H., Sand, M., Smith, C. J., Bauer, S. E., Forster, P. M., Fuglestvedt, J. S., et al. (2018). Climate impacts from a removal of anthropogenic aerosol emissions. *Geophysical Research Letters*, *45*(2), 1020–1029. https://doi.org/10.1002/2017GL076079

Samset, B. H., Wilcox, L. J., Allen, R. J., Stjern, C. W., Lund, M. T., Ahmadi, S., et al. (2025). East Asian aerosol cleanup has likely contributed to the recent acceleration in global warming. *Communications Earth & Environment*, *6*(1), 543. https://doi.org/10.1038/s43247-025-02527-3

Sand, M., Berntsen, T. K., Ekman, A. M., & Lewinschal, A. (2020). Surface temperature response to regional black carbon emissions: Do location and magnitude matter-. *Atmospheric Chemistry and Physics*, *20*(5), 3079–3089. https://doi.org/10.5194/acp-20-3079-2020

Sand, M., Berntsen, T. K., Øyvind, S., & Kristjánsson, J. E. (2013). Arctic surface temperature change to emissions of black carbon within arctic or midlatitudes. *Journal of Geophysical Research: Atmospheres*, *118*(14), 7788–7798. https://doi.org/10.1002/jgrd.50613

Schwarber, A. K., Smith, S. J., Hartin, C. A., Vega-Westhoff, B. A., & Sriver, R. (2019). Evaluating climate emulation: Fundamental impulse testing of simple climate models. *Earth System Dynamics*, *10*(4), 729–739. https://doi.org/10.5194/esd-10-729-2019

Seinfeld, J., & Pandis, S. (2016). *Atmospheric chemistry and physics: From air pollution to climate change*. Wiley.

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

Smith, C. J., Kramer, R. J., Myhre, G., Alterskjr, K., Collins, W., Sima, A., et al. (2020). Effective radiative forcing and adjustments in CMIP6 models. *Atmospheric Chemistry and Physics*, *20*(16), 9591–9618. https://doi.org/10.5194/acp-20-9591-2020

Stier, P., van den Heever, S. C., Christensen, M. W., Gryspeerdt, E., Dagan, G., Saleeby, S. M., et al. (2024). Multifaceted aerosol effects on precipitation. *Nature Research*, *17*(8), 719–732. https://doi.org/10.1038/s41561-024-01482-6

Stjern, C. W., Samset, B. H., Myhre, G., Forster, P. M., Hodnebrog, I., Andrews, T., et al. (2017). Rapid adjustments cause weak surface temperature response to increased black carbon concentrations. *Journal of Geophysical Research: Atmospheres*, *122*(21), 11462–11481. https://doi.org/10.1002/2017JD027326

Szopa, S., Naik, V., Adhikary, B., Artaxo, P., Berntsen, T., Collins, W., et al. (2021). In V. Masson-Delmotte, et al. (Eds.), *Short-lived climate forcers* (pp. 817–921). Cambridge University Press. https://doi.org/10.1017/9781009157896.008

Tazi, K., Orr, A., Hernandez-González, J., Hosking, S., & Turner, R. E. (2024). Downscaling precipitation over high-mountain Asia using multi-fidelity Gaussian processes: Improved estimates from ERA5. *Hydrology and Earth System Sciences*, *28*(22), 4903–4925. https://doi.org/10.5194/hess-28-4903-2024

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *365*(1857), 2053–2075. https://doi.org/10.1098/rsta.2007.2076

Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In D. van Dyk & M. Welling (Eds.), *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics* (Vol. 5, pp. 567–574). PMLR. Retrieved from https://proceedings.mlr.press/v5/titsias09a.html

Titsias, M., & Lawrence, N. D. (2010). Bayesian gaussian process latent variable model. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Vol. 9, pp. 844–851). PMLR. Retrieved from https://proceedings.mlr.press/v9/titsias10a.html

Tjiputra, J. F., Schwinger, J., Bentsen, M., Morée, A. L., Gao, S., Bethke, I., et al. (2020). Ocean biogeochemistry in the Norwegian Earth system model version 2 (NorESM2). *Geoscientific Model Development*, *13*(5), 2393–2431. https://doi.org/10.5194/gmd-13-2393-2020

Twomey, S. (1977). The influence of pollution on the shortwave albedo of clouds. *Journal of the Atmospheric Sciences*, *34*(7), 1149–1152. https://doi.org/10.1175/1520-0469(1977)034⟨1149:TIOPOT⟩2.0.CO;2

van der Wilk, M., Dutordoir, V., John, S., Artemev, A., Adam, V., & Hensman, J. (2020). A framework for interdomain and multioutput Gaussian processes. Retrieved from https://arxiv.org/abs/2003.01115

van Vuuren, D. P., Lowe, J., Stehfest, E., Gohar, L., Hof, A. F., Hope, C., et al. (2011). How well do integrated assessment models simulate climate change? *Climatic Change*, *104*(2), 255–285. https://doi.org/10.1007/s10584-009-9764-2

von Salzen, K., Whaley, C. H., Anenberg, S. C., Dingenen, R. V., Klimont, Z., Flanner, M. G., et al. (2022). Clean air policies are key for successfully mitigating Arctic warming. *Communications Earth & Environment*, *3*(1), 222. https://doi.org/10.1038/s43247-022-00555-x

Wang, H., Zheng, X. T., Cai, W., Han, Z. W., Xie, S. P., Kang, S. M., et al. (2024). Atmosphere teleconnections from abatement of China aerosol emissions exacerbate Northeast Pacific warm blob events. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(21), e2313797121. https://doi.org/10.1073/pnas.2313797121

Watson-Parris, D., Rao, Y., Olivié, D., Seland, 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*(10), e2021MS002954. https://doi.org/10.1029/2021MS002954

Watson-Parris, D., & Smith, C. J. (2022). Large uncertainty in future warming due to aerosol forcing. *Nature Climate Change*, *12*, 1111–1113. https://doi.org/10.1038/s41558-022-01516-0

Welch, B. L. (1947). The generalization of "student's" problem when several different population varlances are involved. *Biometrika*, *34*(1–2), 28–35. https://doi.org/10.1093/biomet/34.1-2.28

Westervelt, D. M., Mascioli, N. R., Fiore, A. M., Conley, A. J., Lamarque, J. F., Shindell, D. T., et al. (2020). Local and remote mean and extreme temperature response to regional aerosol emissions reductions. *Atmospheric Chemistry and Physics*, *20*(5), 3009–3027. https://doi.org/10.5194/acp-20-3009-2020

Wilcox, L. (2025). Opinion: The role of AerChemMIP in advancing climate and air quality research. (Data review of CMIP6 models across DeCK, AerChemMIP, and ScenarioMIP). https://doi.org/10.17605/OSF.IO/8FWJ3

Wilcox, L., Dunstone, N., Lewinschal, A., Bollasina, M., Ekman, A., & Highwood, E. (2019). Mechanisms for a remote response to Asian anthropogenic aerosol in boreal winter. *Atmospheric Chemistry and Physics*, *19*(14), 9081–9095. https://doi.org/10.5194/acp-19-9081-2019

Wilcox, L. J., Allen, R. J., Samset, B. H., Bollasina, M. A., Griffiths, P. T., Keeble, J., et al. (2023). The regional aerosol model intercomparison project (RAMIP). *Geoscientific Model Development*, *16*(15), 4451–4479. https://doi.org/10.5194/gmd-16-4451-2023

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*(20), 11955–11977. https://doi.org/10.5194/acp-20-11955-2020

Wilks, D. S. (2019). Chapter 9–Forecast verification. In D. S. Wilks (Ed.), *Statistical methods in the atmospheric sciences* (4th ed., pp. 369–483). Elsevier. https://doi.org/10.1016/B978-0-12-815823-4.00009-2

Xie, X., Myhre, G., Liu, X., Li, X., Shi, Z., Wang, H., et al. (2020). Distinct responses of Asian summer monsoon to black carbon aerosols and greenhouse gases. *Atmospheric Chemistry and Physics*, *20*(20), 11823–11839. https://doi.org/10.5194/acp-20-11823-2020

Zelinka, M. D., Andrews, T., Forster, P. M., & Taylor, K. E. (2014). Quantifying components of aerosol-cloud-radiation interactions in climate models. *Journal of Geophysical Research: Atmospheres*, *119*(12), 7599–7615. https://doi.org/10.1002/2014JD021710