

Transforming kernel-based learners to incorporate domain knowledge from climate science



Shahine Bouabid
Lady Margaret Hall
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary 2024

Statement of Originality

I hereby declare that except where specific reference is made to the work of others, the content of this thesis is my own work and has not been submitted in whole or in parts for any other degree or qualification. This thesis is my own work unless otherwise stated in the authorship form at the end of the chapters.

Shahine Bouabid

Hilary 2024

I dedicate this thesis to my grand parents Jean-Pierre and Najat.
Thank you for the love. I needed it.

Acknowledgements

I want to thank my supervisor Dino Sejdinovic for being such a humane individual throughout my DPhil. Your unconditional support in everything I undertook (even when it was, in hindsight, terrible), your guidance, and the intellectual freedom you have allowed me have enabled me to grow while feeling supported despite the distance. I am proud and grateful to have been able to learn from you. My gratitude also goes to Duncan Watson-Parris, who, without knowing it, was an inspiration to me even before I started this thesis. Your mentorship and support have been so important during my time in Oxford and have helped me to find myself where I am today.

I extend my appreciation to Gustau Camps-Valls and Krikamol Muandet for welcoming me to Valencia and Saarbrücken. My time with you has allowed me to mature the way I envision research and given me new stimulating ideas. I also want to thank Tom Rainforth for kindly welcoming me into his group and for his consistently sharp and insightful guidance.

Next, I want to take a moment to express my appreciation for my two friends and collaborators, Alan Chau and Jake Fawkes. Alan, thank you for being such a caring and easy presence; you were a huge part of my Oxford days. Coucou Jake, my alternative tuck shop date, I'm so grateful our paths crossed and we got to establish a new standard for the smoothest possible collaboration.

I am blessed with long-lasting friendships I cherish with Adil, Othmane, and Van, even though we don't get to see each other nearly as often as I would like to. Lucile la cabane, I am so grateful we got to create this bond and share an office over the past four years. Ali, Karim, Kenza, Mikou, I am proud of what we have accomplished with Nechfate and thankful I get to share it with you.

I want to have a few words for the new bonds I have made in Oxford, ones I intend to hold onto for a long time. To Laura, my partner, how were we strangers to each other for so long? I feel so lucky to grow alongside someone as caring, intelligent, and brave person as you. Thank you. I love you.

A heartfelt mention goes to Peter for just being you, Peter, the most loving housemate I have ever had, and to Sofija for being a truly precious and inspiring encounter of my first year in the UK. I also want to give a shout-out for the bonds I have made within the department and beyond, without whom, this whole journey would not have been the same. A special thanks to Amitis, Alyson, Badr, Chris, Dan,

Eugenio, Francesco, Jef, Kamelia, Linying, Lorenzo, Maybritt, Paul, Otis, Sahra, Tim, Tyler, Veit, Vik, Xi, and the entire iMiracli crew.

Finally, I would like to thank the two beings closest to my heart: Malika and Mansour. You are perfect, thank you for being you, that's all I will ever need from you. My parents, Amine and Maija, I know you love me. I love you too.

Abstract

In the face of persistent modelling and observational challenges in climate science, which hinder our understanding of the climate system, statistical machine learning has emerged as a potential ally in recent years. Modern machine learning methods promise to leverage the vast volumes of data from climate model simulations, satellite imagery, or in-situ measurements to advance our understanding of the climate system and, thereby, our ability to anticipate the adverse consequences of climate change. However, without concerted efforts to align the use of machine learning with the needs of the climate science community, this promise may lead to disappointment due to wasted resources and unmet expectations.

In this thesis, we propose a set of guiding principles for the design of machine learning models that help align with the expectations of the climate science community. These guidelines include using models meaningfully-specified for a problem, a preference for mathematically transparent models, emphasising probabilistic modelling, and incorporating domain knowledge. We then choose to focus on kernel-based learners, a particular class of machine learning algorithms based on similarity measures between data points, which fits well with these guidelines, and provide relevant background on their application to regression tasks.

The core of this thesis contributes three studies demonstrating how kernel-based learners can help address challenges in climate science. In Chapter 3 we take on a methodological angle and contribute a framework and theoretical guarantees for the incorporation in regression of a particular kind of domain knowledge that can arise in climate science: knowledge of the causal structure underlying the data generating process. In Chapter 4, we consider an applied modelling challenge in climate science: the development of cheap surrogates of computationally expensive climate models, called climate model emulators. We demonstrate how incorporating Gaussian processes (GPs) modelling in a physically-motivated energy balance model allows us to formulate a simple probabilistic emulator of surface temperatures. This emulator, we call FaIRGP, can learn from data and outperform purely process-based emulators, while retaining the robustness from the incorporated energy balance model. In Chapter 5, we consider an applied observational challenge in climate science: obtaining global estimates of aerosol vertical profiles. We propose a Bayesian model, based on GPs and heuristics from satellite aerosol retrieval algorithms, that infers aerosol vertical

extinction profiles from aerosol optical depth measurements and vertically-resolved meteorological data.

Finally, we conclude in Chapter 6 by reflecting on the work presented, discussing limitations, and directions for future work.

Contents

1	Introduction	1
1.1	Motivation	1
1.1.1	Machine learning in climate science research	4
1.1.2	Machine learning <i>for</i> climate science research	6
1.2	Thesis outline and contributions	11
2	Background	14
2.1	Notations and Regression setup	14
2.2	Learning with kernels	15
2.2.1	Distilling infinite into finite dimensional learning	16
2.2.2	RKHS and the representer theorem	18
2.2.3	Specifying a RKHS through a positive definite kernel	19
2.2.4	Transparency of the RKHS induced by a kernel	20
2.2.5	Algorithmic transparency of kernel ridge regression	24
2.3	Bayesian learning with kernels	24
2.3.1	A Bayesian prior over functions	25
2.3.2	Transparency on the prior induced by a kernel	26
2.3.3	Gaussian process regression	27
2.4	Transforming kernel-based learners	28
3	Returning the Favour: When Regression Benefits from Probabilistic Causal Knowledge	31
4	FaIRGP: A Bayesian Energy Balance Model for Surface Temperatures Emulation	62
5	Aerosol Optical Depth Disaggregation: Toward Global Aerosol Vertical Profile	128

6	Conclusion and Outlooks	166
6.1	Summary	166
6.2	Limitations	167
6.3	Outlooks	170

Chapter 1

Introduction

This thesis follows the format of an integrated thesis and is composed of 6 chapters. The first, second, and last chapters consist of an introduction, background, and conclusion respectively. The remaining 3 chapters are formed of separately published or submitted papers, which each contain their own introduction and a literature review specific to the topics covered therein. In this introduction, we would like to give a more general overview and motivation of the work done throughout the DPhil.

1.1 Motivation

There exists increased scientific evidence the twentieth century has seen humanity enter a new geological epoch referred to as the *Anthropocene* (Lewis and Maslin, 2015). This new era is marked by a human footprint on the Earth system which threatens the fragile environmental equilibrium that has allowed life to thrive. The main responsibility of this footprint lies with the unprecedented scale of human extraction of natural resources, a dynamic bolstered by the capitalistic commodification of nature and imperialist colonialism over the nineteenth and twentieth centuries (Moore, 2017; Vergès, 2017). This has led, among things, to large-scale deforestation, destruction of biodiversity, soil degradation, but also to human rights abuses, and exacerbated inequalities in wealth distribution.

A defining feature of the Anthropocene is human-induced climate change, which has received increasing attention over the past decades. The unparalleled scale of ever-growing human activity since the industrial revolution has led to levels of atmospheric carbon dioxide unprecedented over the last 800,000 years (Eyring et al., 2021a). 2023 was reported to be the warmest year on record (World Meteorological Organization, 2024), with evidence that the annual global mean surface temperature anomaly may

have been for the first time 1.5°C above pre-industrial levels (Copernicus, 2024), exceeding the threshold set by the Paris Agreement.

This rise in global temperatures increases pressure on ecosystems and biodiversity, reduces inhabitable land surface, and leads to more frequent and severe extreme weather events (Pörtner et al., 2022; Seneviratne et al., 2021). It is estimated that nearly half of the world population lives today in contexts with high vulnerability to the adverse effects of climate change (Birkmann et al., 2022). The severity of the climate system response to a small change in global temperatures stems from its many complex and interdependent feedback loops. This means that even a minor increase in global mean temperature can trigger dramatic and far-reaching consequences.

The promise of Earth sciences is to disentangle this complex web of interactions in order to improve our understanding of the Earth’s system dynamics. Climate science, or climatology, is a vital branch of Earth sciences that investigates the dynamics of the atmosphere over extended periods of time. It builds upon universal laws of physics, along with chemical and biological principles, applied in the unique atmospheric conditions of our planet (Vallis, 2017; Seinfeld and Pandis, 2016). In the past fifty years, climate science has made tremendous progress, becoming an interdisciplinary field that merges the study of atmospheric phenomena with the influences of the oceans, land, ice, and biological processes (Hartmann, 2015). Nonetheless, despite these advancements, important modelling and observational challenges persist toward our understanding of the climate system.

Firstly, in the realm of climate modelling, the most significant achievement of the last third of the twentieth century was undoubtedly the advent of general circulation models (GCMs) (Phillips, 1956; Lorenz, 1967). They are advanced numerical models which incorporate principles of fluid dynamics, thermodynamics, and radiative transfer on a spatially discretised three-dimensional grid to simulate the complex dynamics and interactions between the atmosphere, oceans and land surface. Further developments have led to the state-of-the-art Earth System Models (ESMs), which go beyond GCMs by including chemical and biological processes (Scholze et al., 2012). Today, ESMs have become ubiquitous and indispensable tools to study past climate, project its response to human activity, and examine the underlying drivers of the climate system (Eyring et al., 2016; Chen et al., 2021).

Despite the core contributions to climate science they have pioneered, ESMs suffer from important shortcomings and may be running out of breath to facilitate further progress in climate research today (Balaji et al., 2022). One major limitation of ESMs is the high computational and energy cost incumbent to the resolution of

energy, mass, and momentum conservation equations across the atmosphere (Acosta et al., 2023). For example, simulating the hourly evolution of climate at 1° resolution over a single year using the CESM2 model requires about 3000 core hours on a supercomputer (Danabasoglu et al., 2020).

This is a computational burden which contributes a significant energy footprint (Acosta et al., 2023), and therefore only a handful of research institutions in the world can afford to cost of participating to international initiatives such as the Coupled Model Intercomparison Project (CMIP) (Eyring et al., 2016). As a result, it restricts various scientific fields dependent on climate projections (agriculture, energy infrastructure, and socio-economic modelling) to using precomputed projections that might not adequately address their specific needs. The high computational cost also poses a barrier to independent researchers and smaller teams, reinforcing the dominance of well-resourced institutions, primarily from the global north, in influential climate research (Tandon, 2021). Finally, running enough ESM simulations to fully explore climate uncertainties and potential socio-economic pathways is prohibitive, and therefore much of the future uncertainty phase space remains unexplored (Hawkins and Sutton, 2009; Beusch et al., 2020).

Another key issue is that the computational cost of solving coupled differential equations of fluid flow and thermal energy exchange in the atmosphere makes resolving climate processes at fine scales impractical for 100-year ensembles. As a result, ESMs are typically limited to simulating climate at a horizontal scale of order 50-100km (Eyring et al., 2021b). Whilst useful to study large-scale circulation patterns of the climate system, many crucial smaller-scale processes such as convection, cloud formation or radiative fluxes resulting from cloud patterns, require modelling at a much finer resolution of order at least 1-2km (Palmer, 2014). To bridge this gap, ESMs rely on parametrisation schemes based on empirical physics and heuristic reasoning to approximate the average effect of smaller-scale processes within a grid cell. However, these schemes fall short of providing an accurate representation of the complex interactions between sub-grid processes and large-scale circulations. The misrepresentation of the effect of sub-grid processes tends as a result to introduce biases in projections, and constitutes the primary source of uncertainty in climate change projections (Jakob, 2010; Sherwood et al., 2014; Caldwell et al., 2016; Terai et al., 2020).

Secondly, climate monitoring has also undergone major advances in the late twentieth and early twenty-first centuries. These advances have primarily been catalysed by the multiplication of remote sensing missions to monitor the Earth’s surface and

atmosphere. Today, orbiting satellite and radar systems provide a continuous flow of spatial, temporal and spectral imagery, contributing a colossal amount of climate observations, with daily transmission volumes that can exceed hundreds of terabytes (Ma et al., 2015; Agapiou, 2017). The role of such airborne imagery has been instrumental not only in the monitoring of the land surface and the cryosphere (Tomlinson et al., 2011; Tedesco, 2015), but also in enhancing our understanding of the climate system, in aspects such as the patterns of sea-level rise or the cooling effect resulting from increased aerosol emissions (Yang et al., 2013; Grosvenor et al., 2018).

Nonetheless, important observational challenges still lie ahead in climate monitoring, a particularly pressing one being precisely in the study of aerosol-cloud interactions. Aerosols are tiny particles in suspension that partially mask the global warming caused by greenhouse gases through a cooling effect, which is largely due to their modulation of cloud properties (Twomey, 1974; Albrecht, 1989). This interaction contributes the greatest uncertainty of anthropogenic climate forcing: the uncertainty on the effect of aerosol-cloud interactions could in practice either offset, or double warming levels (Bellouin et al., 2020; Forster et al., 2021). A major factor contributing to this uncertainty is the difficulty in obtaining accurate global measurements of aerosol concentrations in the atmosphere. Direct measurements are expensive and spatiotemporally sparse, therefore optical remote sensing data from satellites is widely used as a proxy for aerosol concentrations, which introduces substantial uncertainty in the observed magnitude of the cooling effect (Nakajima et al., 2001; Stier, 2016; Painemal et al., 2020).

1.1.1 Machine learning in climate science research

In recent years, data-driven methods from statistical machine learning have emerged as a potential ally to address challenges arising in climate science. Drawing from the colossal amounts of data collected by satellite imagery, climate model simulations and in-situ measurements, statistical machine learning methodologies have generated significant enthusiasm as they seek to address modelling and observational challenges in climate science (Camps-Valls et al., 2016; Reichstein et al., 2019; Irrgang et al., 2021; Beucler et al., 2023). By extracting insights and performing inference over the large amount of data collected, they stand as an opportunity to go beyond what process-based modelling can achieve alone, what computational constraints prohibit, and what observational limitations prevent.

For example, data-driven modelling can prove useful to address the prohibitive computational and energy costs of climate models by enabling the development of

cheaper surrogate versions of ESMs. These surrogates, known as *climate model emulators*, aim to mimic the outputs of ESMs with minimal computational resources by learning the patterns of climate responses from data (Beusch et al., 2020; Watson-Parris et al., 2022a; de Burgh-Day and Leeuwenburg, 2023). In Chapter 4, we consider the problem of developing a machine learning climate model emulator that builds upon robust physical modelling. Alternatively, considering that the resolution of coupled partial differential equations (PDEs) constitutes the primary computational bottleneck in ESMs, recent initiatives have aimed to use machine learning to develop surrogate PDE solvers specifically for atmospheric and oceanic circulation (Pathak et al., 2022; Bonev et al., 2023; Bire et al., 2023). Both these approaches are vibrant research directions, where the application of statistical machine learning techniques shows promise in knocking down the barriers of climate models’ computational cost.

Machine learning has also reported success in refining the representation of sub-grid processes such as convection and cloud formation in climate models. This is achieved by learning a parametrisation of small-scale processes that best mimics the behaviour of high-resolution climate models. Such machine learning parametrisations allow for a refined statistical representation of sub-grid processes at the grid level, which by improving over the heuristics representation of classical parametrisations, offers a pathway to mitigate the bias and the uncertainty the latter introduce (Rasp et al., 2018; Yuval and O’Gorman, 2020; Gentine et al., 2021; Watt-Meyer et al., 2024). Most recently, the emergence of climate models directly incorporating latent neural representations and learned physics marks a significant milestone, showcasing performances on par with the best climate models (Kochkov et al., 2023).

Beyond the purely practical considerations of modelling, data-driven methodologies have the potential to support advancements in understanding the physical mechanisms of the climate system. Domains such as equation discovery can leverage data to unveil new insights into physical processes underpinning facets of the climate system that are not yet fully understood such as cloud processes (Zanna and Bolton, 2020; Camps-Valls et al., 2023; Grundner et al., 2024). Causal discovery also offers the promise of helping identify causal relationships from data at various scales, from the micro (e.g. cloud-aerosol interactions) (Gryspeerdt et al., 2019; McCoy et al., 2020; Fons et al., 2023) to the macro (teleconnections between climate phenomena) (Kretschmer et al., 2016; Runge et al., 2019a,b). This in turn can guide the design of inference procedures that acknowledge existing causal relationships: in Chapter 3 we investigate how causal knowledge can serve as an inductive bias for regression. Statistical explainability methods in machine learning can also contribute valuable insights and help build

intuition on observed physical phenomena, albeit their conclusions might be less definitive (Bommer et al., 2023; Jeggle et al., 2023).

Lastly, machine learning can help overcome observational challenges currently faced in climate science by allowing to merge information from multiple sensors. Data fusion refers to a wide range of statistical approaches used to combine multiple streams of possibly multi-modal observations (Gustafsson, 2010; Cocchi, 2019). Modern statistical machine learning methods grant further flexibility in the design of data fusion models, allowing to merge information from heterogeneous datasets varying in time, location, resolution, fidelity or domain (Meng et al., 2020; Li et al., 2022). This greatly enhances our ability to enrich measurements beyond what any individual source could offer, with applications such as spatiotemporal super-resolution, data imputation, land-cover monitoring or bias correction (Gao et al., 2015; Deudon et al., 2020; Fulton et al., 2023), and constitutes one avenue for improving global measurements of aerosol concentrations (Arola et al., 2022). Some of this thesis’s contributions fall within this research thread: in Chapter 5, we propose a probabilistic model to combine satellite measurements of aerosol optical thickness with vertically resolved meteorological data to infer aerosol vertical extinction.

These use cases are solely meant to provide an overview of existing endeavours for machine learning in climate science. They constitute by no means a comprehensive review of how machine learning can be applied to address climate science challenges, as this is a vibrant and fast-growing field.

1.1.2 Machine learning *for* climate science research

Statistical machine learning appears as a powerful paradigm, which operates at such a high level of generality that it dangles the prospect of being able to learn arbitrarily complex functions that map any input to any output, provided enough data is available. This perception is reinforced by an unfortunate trend in machine learning research to make bold promises and sometimes overstate results in published works, which may not always reflect the realities faced by practitioners when they attempt to reproduce results (Hutson, 2018; Lipton and Steinhardt, 2019).

To some extent, this tendency to present machine learning as a “solve-it-all” has led to scepticism from practitioners and the climate science community. Indeed, whilst applying off-the-shelf machine learning algorithms to address climate tasks may be a scientifically valid approach, there is an underlying boldness in claiming these tools can solve global problems while dismissively overlooking centuries of climate science research. This calls for an essential shift in perspective to rethink machine

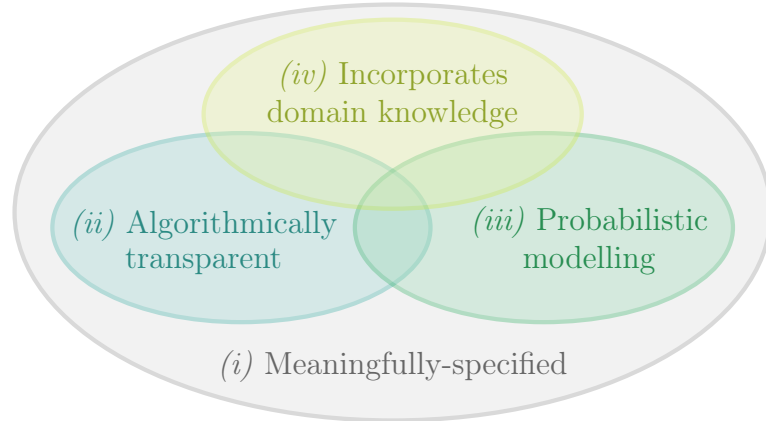


Figure 1.1: Illustrative overview of the guiding principles for designing relevant machine learning models for climate science. All models should be meaningfully-specified. Trust in the model is improved through algorithmic transparency, probabilistic modelling, and incorporation of domain knowledge.

learning tailored *for* climate science as its own field, as opposed to blindly applying existing machine learning *to* climate science. This shift begins by listening to, and understanding the climate science community’s concerns about the indiscriminate use of machine learning tools. We propose below guiding principles, illustrated in Figure 1.1, to assist machine learning practitioners in the design of relevant statistical machine learning models to address challenges in climate science.

(i) – Use meaningfully-specified models

Using meaningfully-specified models is crucial, and whilst obvious, it is a point worth stressing again. We refer to a machine learning model as *meaningfully-specified* if its design aligns with the specific problem it addresses, and reflects the structure and the assumptions about the data it uses.

As in any applied statistical modelling problem, it is essential to dedicate time to understanding the physical processes underpinning the climate tasks we seek to address — as opposed to frontally addressing any predictive problem as the regression of a response variable onto predictors. Understanding the nature of the problem not only helps grasp its broader relevance in climate science, but also helps design a statistical model that is both inspired by, and tailored to, the specific climate science context. This includes identifying physical variables relevant to the phenomenon of interest, accounting for existing spurious correlations in the design, or using an adequate spatial and temporal resolution to capture the processes’ variability.

Acknowledging these aspects conditions the design of a statistical model on problem-specific domain knowledge, which prevents the formulation of ad-hoc solutions that lack relevance to climate research practices. For example in regression, if the response variable and the predictors are selected carelessly, a model with enough parameters may still be able to learn a mapping that echoes patterns in the data without representing any real-world phenomena. Whilst this may be a valid statistical model, and even potentially a useful predictive model, it will likely lack relevance to the climate science community, should never be used to make inferences about the dynamics of the climate system, and drawing physical insights from their statistical representation may mislead climate research. Kretschmer et al. (2021) propose the following example which is useful to illustrate our point: precipitations in Denmark and the Mediterranean region display a negative correlation, therefore data on precipitations in Denmark can be useful to build a predictive statistical model of precipitations in the Mediterranean region. However, climate experts generally agree that this correlation arises because both areas are influenced by the North Atlantic storm track (Folland et al., 2009). As a result, such a predictive statistical model would draw from spurious dependencies and would not be a reliable tool to make conclusions about the climate system.

(ii) – Prefer algorithmic transparency

In disciplines like climate science where the same statistical analysis techniques have been used for decades, the community can be hesitant to adopt novel methods arising from machine learning. Such hesitance can be rightfully justified as many machine learning models tend to be opaque to practitioners, offering limited insights into the inner workings behind their predictions. Designing machine learning models that are transparent enough such that their predictions can be algebraically understood and considered trustworthy is crucial for making contributions that stand a chance of being embraced and making a meaningful impact on climate research. We refer to such models as *algorithmically transparent*, a term introduced by Lipton (2018) to refer to models where the learning algorithm is understood.

For example, deep neural networks, which have undoubtedly become the most popular machine learning modelling tool today when working with complex structured data, still suffer from limited algorithmic transparency since we do not understand why the heuristics of their optimisation procedures work. As a result, no matter how good their predictive performances are, the inability to understand the basis for their predictions undermines trust in such predictive machine learning models, as well as the confidence and usefulness of conclusions drawn from studies using them.

We advocate, when possible, preferring simple tools from statistical machine learning that have well-studied theoretical foundations, are established within methodological and applied machine learning literature, and grant hands-on control into their inner workings for practitioners. Under these considerations, examples of candidate tools are decision tree-based models, parametric models such as generalised linear models or non-parametric models such as kernel methods. Throughout this thesis, our preference will be with kernel-based learners and we provide in Chapter 2 relevant background on their theoretical foundations and algorithmic transparency.

(iii) – **Prefer probabilistic modelling**

Turning to statistical modelling often indicates a gap in our ability to model a phenomenon of interest because of a lack of understanding, missing variables, or difficulty obtaining precise measurements. Indeed, we resort to flexible statistical models hoping they can bridge this gap in the specification of a relationship by learning patterns from data. As in any modelling endeavour, it is always essential to quantify the modelling uncertainty arising from our lack of knowledge of the relationships we aim to model. However, we argue the need to acknowledge this uncertainty is even more critical with machine learning models. Unlike process-based models intended to represent specific physical phenomena, machine learning models aim to represent flexible distributional relationships, making them inherently flexible and not purpose-built. Therefore, additional caution must be taken in quantifying how accurately these models capture the phenomena of interest.

This can be achieved by preferring methods from probabilistic machine learning such as Monte Carlo methods, Bayesian modelling, or diffusion models (Murphy, 2023) that allow to account for the modelling uncertainty. In Chapter 4 and 5 for example, we use nonparametric Bayesian modelling to design predictive probabilistic machine learning models which provide a principled measure of uncertainty quantification over their predictions. Beyond the modelling uncertainty, it is also important to recall that other sources of uncertainty such as observational uncertainty, uncertainties in future scenarios, or the inherent stochasticity of the climate system itself must be accounted for, and it is only by accounting for all the sources of uncertainty that we can provide a nuanced and reliable foundation for decision making.

(iv) – **Incorporate domain knowledge**

Imagine wanting to build a wooden chair. One could cut pieces of wood, and stick them together into a chair using super glue. That may certainly work and ultimately,

once the glue has dried, the chair should be stable enough for someone to sit on it as long as they do not gesture too much. However, it may not inspire the same confidence in its stability as if the chair had been assembled with nails and a hammer, especially if making sudden movements. Nails are purpose-built tools designed for joining wooden planks, resulting in a sturdier chair, even if one were to gesture.

This analogy parallels the use of off-the-shelf machine learning models to address climate science tasks. Even when a model is meaningfully specified and algorithmically transparent, it still lacks the robust physical foundations provided by well-established process-based models, which ultimately may harm trust in the models. This is largely because machine learning models remain by construction interpolation machines, which learn to reproduce the pattern from the training data distribution. Whilst this may not be an issue when making predictions within the training data regime, limited guarantees exist regarding their capacity to extrapolate to unseen data regimes. This is a critical limitation in climate science, where the distribution of variables such as mean and extreme temperatures, precipitations and winds are constantly evolving due to human activity.

This does not mean machine learning models should never be used: once the chair has been assembled with nails, glue can help add a cushion or with finishing touches in a way that nails cannot. Similarly, statistical machine learning can build upon process-based modelling and go beyond, in a way that process-based models alone cannot. For this reason, we advocate for the incorporation of domain knowledge, or process-based modelling, into the design of statistical machine learning models. This can be framed as identifying an inductive (or learning) bias that reflects problem-specific domain knowledge, and which can be enforced onto the learning algorithm.

In climate science, domain knowledge often comes from physics. This has led to the development of a wide range of physics-informed machine learning methodologies (Reichstein et al., 2019; Beucler et al., 2023; Willard et al., 2020), which vary in the rigidity with which they encode physics into the machine learning modelling. Our work in Chapter 4 falls within this category since the climate emulator prior we propose is designed to satisfy the temperature response differential equation of an energy balance model. Causal knowledge constitutes another form of domain knowledge, and in Chapter 3 we introduce a methodology to incorporate in regression knowledge that the data admits a collider causal structure. Finally, heuristics that have passed the test of time can also constitute a valuable source of guidance for the design of statistical models. For example, in Chapter 5 we use the idealised vertical profile of satellite retrieval algorithms to guide the design of our prior.

1.2 Thesis outline and contributions

This thesis contains 3 published or submitted papers followed up with a conclusion and future work chapter. Even though not included, I have been fortunate enough to have been part of several other works which I list below.

- Deconditional Downscaling with Gaussian processes
Siu Lun Chau*, **Shahine Bouabid***, Dino Sejdinovic
Accepted at NeurIPS, 2021
(Chau et al., 2021)
- ClimateBench v1.0: A Benchmark for Data-Driven Climate Projections
Duncan Watson-Parris and co-authors including **Shahine Bouabid**
Published in Journal of Advances in Modeling Earth Systems, 2022
(Watson-Parris et al., 2022a)
- Domain Generalisation via Imprecise Learning
Anurag Singh, Siu Lun Chau, **Shahine Bouabid**, Krikamol Muandet
Accepted at ICML, 2024
(Singh et al., 2024)

The following three chapters are presented in this thesis.

- Returning the Favour: When Regression Benefits from Probabilistic Causal Knowledge
Shahine Bouabid*, Jake Fawkes*, Dino Sejdinovic
Accepted at ICML, 2023
(Bouabid et al., 2023)

In Chapter 3, we consider the problem of regressing a response variable Y onto predictors X , where we have knowledge of the causal structure underlying the data generating process. We ask if, and how, this form of domain knowledge could be encoded in the regression model. We find that once the causal structure has been used to discard uninformative predictors from X , the only remaining graphical structures that are informative for regression are *colliders* of the form $Y \rightarrow X_i \leftarrow X_j$. These structures occur when studying how independent latent climate variables, such as forcings from different agents, might affect observed warming trends. We then consider a simple collider structure, and derive from it a constraint that can be imposed on the regression model to acknowledge the presence of the collider. We prove that

encoding this information in regression can only benefit the generalisation of the model, and in particular, we obtain a guarantee of strictly positive generalisation benefit in expectation when working in a reproducing kernel Hilbert space.

- **FaIRGP: A Bayesian Energy Balance Model for Surface Temperatures Emulation**
Shahine Bouabid, Dino Sejdinovic, Duncan Watson-Parris
Published in Journal of Advances in Modeling Earth Systems, 2024
(Bouabid et al., 2024a)

In Chapter 4, we consider the problem of developing a climate model emulator able to mimic the outputs of climate models at a minimal computational cost. Existing climate emulators fall within two categories: physics-driven emulators, designed as lower-order representations of the climate system, and data-driven emulators, which result from statistical modelling of the climate response. We contribute in this work a novel hybrid emulator that combines the robust backbone of a physics-driven emulator with Bayesian modelling, thereby granting it additional flexibility to learn from data. We achieve this by treating the radiative forcing component of a simple climate model as an underspecified latent force modelled as a Gaussian process (GP), and derive the resulting temperature response. The posterior update of this model allows us to gain flexibility and go beyond what the simple climate model alone could do without sacrificing the robustness of the physics-driven modelling, which improves the extrapolation capacity of the model. Our emulator is fully mathematically tractable, making it a convenient and easy-to-use probabilistic tool for emulating surface temperatures.

- **Aerosol optical depth disaggregation: toward global aerosol vertical profiles**
Shahine Bouabid, Duncan Watson-Parris, Sofija Stefanović, Athanasios Nenes, Dino Sejdinovic
Published in Environmental Data Science, 2024
(Bouabid et al., 2024b)

In Chapter 5, we set out to address one of the observational challenges in climate science: improving the observational data we have on aerosol. Aerosol-cloud interactions constitute the largest source of uncertainty in assessments of anthropogenic climate change. This uncertainty is due in part to challenges in accurately measuring aerosols at different altitude levels in the atmosphere, with only sparse observations being available. We contribute in this work a hierarchical Bayesian model based on GPs that allows us to vertically disaggregate two-dimensional satellite measurements of aerosol

optical depth using vertical information we obtain from meteorological predictors such as temperature, relative humidity and updraft wind velocity.

Chapter 2

Background

Throughout this thesis, our focus will be on regression tasks tackled with kernel-based learners. We open this section by introducing notations for a standard least-square regression problem. We then propose a formal motivation to kernel methods for regression and provide an overview of Gaussian processes, often considered as the Bayesian counterpart of kernel methods. Lastly, we review how kernel-based learners behave under linear transformations.

2.1 Notations and Regression setup

Let $\mathcal{X}$ be a non-empty set and $\mathcal{Y} \subseteq \mathbb{R}$. We will denote X the $\mathcal{X}$ -valued random variable representing our predictors, and Y the $\mathcal{Y}$ -valued random variable representing our response variable. We denote $\mathbb{P}_{X,Y}$ the joint pushforward probability measure induced by (X, Y) . We are interested in the relationship between Y and X which¹ can always be decomposed following

$$Y = \mathbb{E}[Y|X] + (Y - \mathbb{E}[Y|X]). \quad (2.1)$$

When addressing a regression problem in statistical learning, the objective is to find a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ that best represents this relationship, therefore it is customary to rewrite the former as

$$Y = f^*(X) + \varepsilon, \quad (2.2)$$

where $f^* : \mathcal{X} \rightarrow \mathcal{Y}$ is such that $f^*(X) = \mathbb{E}[Y|X]$, and $\varepsilon = Y - \mathbb{E}[Y|X]$ is described as a residual random variable, which cannot be captured by X , and satisfies $\mathbb{E}[\varepsilon|X] = 0$.

With this problem formulation, and in the light of the orthogonal projection nature of the conditional expectation, it becomes apparent that finding f^* corresponds to

¹provided the existence of conditional expectations

finding the square-integrable function that minimises the *population risk*

$$R(f) = \mathbb{E}[(Y - f(X))^2] = \int_{\mathcal{X} \times \mathcal{Y}} (y - f(x))^2 \, d\mathbb{P}_{X,Y}(x, y). \quad (2.3)$$

In practice, tackling this problem over the whole of square-integrable functions is not amenable. Rather, it is customary to consider a hypothesis space $\mathcal{F}$ of functions $f : \mathcal{X} \rightarrow \mathcal{Y}$ which can be described in more practical terms. We will assume throughout that $\mathcal{F}$ has a Hilbert structure with inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$.

Further, we do not know the joint measure $\mathbb{P}_{X,Y}$ and therefore cannot evaluate the population risk $R(f)$. Instead, we only observe a finite number of independent samples drawn from this joint distribution $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$. These allow to replace the population risk objective with an unbiased estimate called the *empirical risk*

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2. \quad (2.4)$$

Finally, any function $f \in \mathcal{F}$ that satisfies $f(x_i) = y_i$ would be a minimiser of $\hat{R}_n(f)$. As a result, there might be infinitely many choices of minimisers, with many proving to be overly dependent on the training data and inadequate to capture the relationship between X and Y . Therefore, it is desirable to introduce additional restrictions to make the least-square criterion meaningful when $\mathcal{F}$ is infinite dimensional. A common approach, referred to as the *Ivanov regularisation*, consists in restricting the hypothesis space to a ball of radius $\delta > 0$ within $\mathcal{F}$, i.e.

$$\min_{f \in \mathcal{F}} \hat{R}_n(f) \text{ such that } \|f\|_{\mathcal{F}} \leq \delta. \quad (2.5)$$

It can be shown that this is equivalent to considering the *Tikhonov-regularised* or *ridge* empirical risk for a particular regularisation parameter $\lambda_\delta > 0$ (where we shall drop the δ dependency in what follows),

$$\hat{R}_{\lambda,n}(f) = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{F}}^2. \quad (2.6)$$

This is the setup we will consider in the next section.

2.2 Learning with kernels

Kernel methods are a fundamental part of machine learning, with strong theoretical foundations that still contribute to much of the machine learning research to date. In

this section, our exposition sets out to provide insights into the theoretical foundations and the algorithmic transparency of the ridge regression problem with kernels. We begin by offering a learning problem motivation to the theory of reproducing kernel Hilbert spaces (RKHSs). We then discuss how these spaces can be constructed, with tools offering control over their transparency, and finally present the ridge regression problem over these spaces.

Whilst our exposition focuses exclusively on a regression problem, kernel methods span a much broader spectrum of applications, including dimensionality reduction (Schölkopf et al., 1998; Fukumizu et al., 2004), probability distribution embedding (Smola et al., 2007; Muandet et al., 2017), hypothesis testing (Gretton et al., 2006; Sejdinovic et al., 2013), explainability (Chau, 2023), meta-learning (Ton, 2022), theory of neural networks (Bietti and Mairal, 2019) and many more (Schölkopf and Smola, 2002; Steinwart and Christmann, 2008a; Berlinet and Thomas-Agnan, 2011).

2.2.1 Distilling infinite into finite dimensional learning

Consider the objective of finding a function $f \in \mathcal{F}$ that minimises the following Tikhonov-regularised empirical least-square risk i.e.

$$\arg \min_{f \in \mathcal{F}} \hat{R}_{\lambda,n}(f). \quad (2.7)$$

In general, we may think of the hypothesis space $\mathcal{F}$ as an infinite dimensional space of non-linear functions. Therefore, without further assumptions on the class of functions described by $\mathcal{F}$ (e.g. parameterisation of functions in this space), it may be challenging to find a minimiser for $\hat{R}_{\lambda,n}$. However, if we make the slightly peculiar assumption that the evaluation of f at a point x can be written as an inner product, then there might be more to say. Suppose that for any $x \in \mathcal{X}$, there exists a function we denote $k_x \in \mathcal{F}$ such that

$$f(x) = \langle f, k_x \rangle_{\mathcal{F}}. \quad (2.8)$$

For the samples collected in our dataset, we can therefore define such functions $k_{x_1}, \dots, k_{x_n}$, and consider the subspace they span $\mathcal{F}_n := \text{Span}\{k_{x_1}, \dots, k_{x_n}\}$. Because this is a closed subspace, the hypothesis space $\mathcal{F}$ orthogonally decomposes into

$$\mathcal{F} = \mathcal{F}_n \oplus \mathcal{F}_n^\perp. \quad (2.9)$$

For any $f \in \mathcal{F}$, we can thus decompose f into its orthogonal projection over $\mathcal{F}_n$, and its orthogonal complement. Namely, there exists $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and $f^\perp \in \mathcal{F}_n^\perp$

such that $\langle f^\perp, k_{x_i} \rangle_{\mathcal{F}} = 0$ for any $i \in \{1, \dots, n\}$ and

$$f = \sum_{i=1}^n \alpha_i k_{x_i} + f^\perp. \quad (2.10)$$

Whilst simple in appearance, the existence of this orthogonal decomposition of f has strong implications for our ridge regression objective. Indeed, the empirical risk $\hat{R}_{\lambda,n}(f)$ involves the function f twice: through the evaluations $f(x_1), \dots, f(x_n)$ and through the norm $\|f\|_{\mathcal{F}}^2$ in the regularisation term.

First, consider the evaluations. Using the orthogonal decomposition of f , we have that for any $j \in \{1, \dots, n\}$

$$f(x_j) = \langle f, k_{x_j} \rangle_{\mathcal{F}} = \sum_{i=1}^n \alpha_i \langle k_{x_i}, k_{x_j} \rangle_{\mathcal{F}} + \underbrace{\langle f^\perp, k_{x_j} \rangle_{\mathcal{F}}}_{=0} = \sum_{i=1}^n \alpha_i \langle k_{x_i}, k_{x_j} \rangle_{\mathcal{F}}. \quad (2.11)$$

This implies that the first component of the empirical risk does not depend on $f^\perp$, but only on the coefficients $\alpha_1, \dots, \alpha_n$. Second, we have that the squared norm regularisation satisfies

$$\|f\|_{\mathcal{F}}^2 = \left\| \sum_{i=1}^n \alpha_i k_{x_i} \right\|_{\mathcal{F}}^2 + \|f^\perp\|_{\mathcal{F}}^2 \geq \left\| \sum_{i=1}^n \alpha_i k_{x_i} \right\|_{\mathcal{F}}^2, \quad (2.12)$$

where the norm decomposition follows from the orthogonality of k_{x_i} and $f^\perp$. This implies that if the Tikhonov regularisation admits a minimiser, such minimiser must satisfy $f^\perp = 0$ and only depend on the coefficients $\alpha_1, \dots, \alpha_n$.

From these two results, we can conclude that if $\hat{R}_{\lambda,n}(f)$ admits a minimiser in $\mathcal{F}$, then this minimiser must have no component in $\mathcal{F}_n^\perp$, and therefore take the form

$$f(x) = \sum_{i=1}^n \alpha_i \langle k_{x_i}, k_x \rangle_{\mathcal{F}}. \quad (2.13)$$

This a powerful result: by simply assuming that function evaluations can be decomposed as $f(x) = \langle f, k_x \rangle_{\mathcal{F}}$, we have transformed a complex statistical learning problem over $\mathcal{F}$, a potentially infinite dimensional space of non-linear functions, into simply learning a set of n scalar coefficients $\alpha_1, \dots, \alpha_n$. In other words, by slightly abusing notations we have

$$\min_{f \in \mathcal{H}} \hat{R}_{\lambda,n}(f) = \min_{\alpha \in \mathbb{R}^n} \hat{R}_{\lambda,n}(\alpha), \quad (2.14)$$

where $\alpha = [\alpha_1 \ \dots \ \alpha_n]^\top$. In the kernel method literature, this result is a particular case of the celebrated *representer theorem*. Motivated by this approach to statistical learning, we will now present the appropriate theoretical framework that allows function evaluations to be written as an inner product.

2.2.2 RKHS and the representer theorem

We present here the minimally sufficient mathematical structure that ensures the correctness of the previous derivations. First, let's introduce a few useful definitions.

Definition 2.2.1 (Bounded linear functional). Let $(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mathcal{F}})$ be a Hilbert space. A linear functional $A : \mathcal{F} \rightarrow \mathbb{R}$ is said to be *bounded* if there exists a finite constant $C > 0$ such that

$$|Af| \leq C\|f\|_{\mathcal{F}}, \quad \forall f \in \mathcal{F}. \quad (2.15)$$

The space of bounded linear functionals is denoted $\mathcal{F}^*$ and corresponds to the topological dual of $\mathcal{F}$.

Boundedness for linear operators is equivalent to continuity. Therefore, we can interchangeably talk about a continuous or a bounded linear operator. Furthermore, it is sufficient that an operator is continuous at 0 for it to be continuous everywhere.

Theorem 2.2.2 (Riesz representation theorem). *Let $(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mathcal{F}})$ be a Hilbert space and $A \in \mathcal{F}^*$. There exists a unique element $\varphi_A \in \mathcal{F}$, called the Riesz representer of A , such that*

$$Af = \langle f, \varphi_A \rangle_{\mathcal{F}}, \quad \forall f \in \mathcal{F}. \quad (2.16)$$

The Riesz representation theorem is a core result of Hilbert space analysis, which reflects the fact that Hilbert spaces are always reflexive Banach spaces, i.e. isomorphic to their topological dual. With these two definitions, we are now ready to introduce RKHSs as the minimally sufficient mathematical structure for our previous exposition to hold.

Definition 2.2.3 (Reproducing kernel Hilbert space, Paulsen and Raghupathi (2016)). Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ be a real-valued Hilbert space endowed with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Suppose that for any $x \in \mathcal{X}$, the evaluation functional $E_x : \begin{cases} \mathcal{H} & \longrightarrow \mathbb{R} \\ f & \longmapsto f(x) \end{cases}$ is a bounded linear operator. Then we say that $\mathcal{H}$ is a reproducing kernel Hilbert space.

It is worth noting that the evaluation functional $E_x : f \mapsto f(x)$ is always linear, and that the additional requirement made in a RKHS is its boundedness. Because it is bounded, it admits a Riesz representer which we denote $k_x \in \mathcal{H}$ such that for any $f \in \mathcal{H}$ we have

$$E_x f = f(x) = \langle f, k_x \rangle_{\mathcal{H}}, \quad (2.17)$$

which is exactly the property we require in the previous derivations. This property is a cornerstone of the RKHS theory², known as the *reproducing property*. The next result states the general version of the representer theorem, which allows transforming a complex learning problem over an infinite dimensional Hilbert space $\mathcal{H}$ into a finite dimensional learning problem over a set of n coefficients.

Theorem 2.2.4 (Representer theorem, Mairal and Vert (2018)). *Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ be a RKHS and let $\Psi : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ be a strictly increasing function with respect to its last variable. Then a solution to the optimisation problem*

$$\arg \min_{f \in \mathcal{H}} \Psi(f(x_1), \dots, f(x_n), \|f\|_{\mathcal{H}}), \quad (2.18)$$

must admit a representation of the form $f(x) = \sum_{i=1}^n \alpha_i \langle k_{x_i}, k_x \rangle_{\mathcal{H}}$.

This powerful result means that even for potentially complex choices of objective Ψ , performing statistical learning in a RKHS can still reduce to learning finite number of coefficients $\alpha_1, \dots, \alpha_n$. Note this result only provides a necessary condition for minimality, while existence and uniqueness may depend on particular properties of Ψ .

One aspect still requiring consideration is that, despite learning the coefficients $\alpha_1, \dots, \alpha_n$, the functional form taken by the minimiser also requires being able to evaluate the inner products $\langle k_{x_i}, k_x \rangle_{\mathcal{H}}$. This inner product defines a bivariate function

$$k(x, x') := \langle k_x, k_{x'} \rangle_{\mathcal{H}}, \quad \forall x, x' \in \mathcal{X}, \quad (2.19)$$

called the *reproducing kernel* of $\mathcal{H}$. Computing it via inner products is often impossible since we may not have an explicit form for $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. However, the reproducing kernel can be specified by practitioners directly as a bivariate function $k(x, x')$ that introduces a measure of similarity between elements of $\mathcal{X}$. The next section aims to outline how properties of RKHSs allow us to specify and evaluate this reproducing kernel.

2.2.3 Specifying a RKHS through a positive definite kernel

Reproducing kernels belong to a class of functions known as *positive semi-definite* (psd) kernels, which can be defined on their own, independently from any RKHS.

Definition 2.2.5 (Positive semi-definite kernel). A bivariate function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is said to be a psd kernel if it is symmetric and for any $x_1, \dots, x_n \in \mathcal{X}$, $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, $\sum_i \sum_j \alpha_i \alpha_j k(x_i, x_j) \geq 0$. When the inequality is strict, k is said positive definite.

²we limit our exposition to real-valued RKHSs, but it is also possible to consider complex-valued, or more generally vector-valued RKHSs (Schwartz, 1964).

Many popular functions define psd kernel, such as the minimum of two non-negative real numbers $k(x, x') = \min(x, x')$, the Euclidean inner product $k(x, x') = x^\top x'$, or the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|_2^2/\ell)$ between $\mathbb{R}^d$ vectors. Being psd is also a property which is stable under positive scaling, addition and multiplication such that psd kernels can be combined together. The following result provides an important characterisation of psd kernels.

Proposition 2.2.6 (Aronszajn (1950)). *$k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a psd kernel if and only if there exists a Hilbert space $(\mathcal{V}, \langle \cdot, \cdot \rangle_{\mathcal{V}})$ and a feature map $\phi : \mathcal{X} \rightarrow \mathcal{V}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{V}}$ for any $x, x' \in \mathcal{X}$.*

This result offers two key insights. First, it is always possible to construct a kernel from a feature map $\phi(x)$. Intuitively, the role of ϕ is to embed data points from $\mathcal{X}$ into a vector space $\mathcal{V}$ where we can evaluate inner products $\langle \cdot, \cdot \rangle_{\mathcal{V}}$, and where the complex structures of the data well captured. For example, if x is some image data, $\phi(x)$ can be its latent space representation learnt by a neural network.

Second, for any psd kernel $k(x, x')$, which we recall may be non-linear in x , there always exists a “linearisation” of the kernel with respect to a vector space $\mathcal{V}$ and an embedding $\phi : \mathcal{X} \rightarrow \mathcal{V}$. This ability to represent non-linearities as linearities in an embedding space $\mathcal{V}$ is an important characteristic of kernel methods, instrumental to their success in machine learning.

The above results are without any consideration of whether k is a reproducing kernel for some RKHS or not. The following fundamental result of RKHS theory provides a positive connection between psd kernels and reproducing kernels.

Theorem 2.2.7 (Moore-Aronszajn, Aronszajn (1950)). *Every psd kernel is the unique reproducing kernel of an RKHS, and reciprocally every RKHS has a unique reproducing kernel which defines a psd kernel.*

Practically, this means that as a machine learning practitioner, one can specify a RKHS by constructing a psd kernel $k(x, x')$ that provides some similarity measure on $\mathcal{X}$. Specifying the reproducing kernel endows the space with a way to evaluate the inner product $\langle k_x, k_{x'} \rangle_{\mathcal{H}}$, and therefore compute the solution to our empirical risk minimisation problem $f(x) = \sum_i \alpha_i \langle k_{x_i}, k_x \rangle_{\mathcal{H}} = \sum_i \alpha_i k(x_i, x)$.

2.2.4 Transparency of the RKHS induced by a kernel

Choosing a kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ automatically specifies a unique RKHS $\mathcal{H}$ of real-valued functions on $\mathcal{X}$. Since we intend to use $\mathcal{H}$ as the hypothesis space from

which we select the solution to our learning problem, it is important to understand how different choices of kernel might affect the nature of the RKHS they induce. We outline three different angles of transparency one can get on the RKHS induced by a kernel: reproducing transparency, practical transparency and theoretical transparency.

Reproducing transparency The class of functions described by $\mathcal{H}$ is intimately tied to the kernel, and in fact can be explicitly written in terms of the latter. This is what we refer to as *reproducing transparency* of the RKHS. Formally, for a kernel $k(x, x')$, the induced RKHS is given by

$$\mathcal{H} = \overline{\text{Span}} \{k(\cdot, x) : x \in \mathcal{X}\}, \quad (2.20)$$

where the closure is taken with respect to the norm in $\mathcal{H}$. This means that any function in $\mathcal{H}$ can be thought of as the limiting linear combination of k_x functions, which echoes the (finite) sum form taken by $f(x) = \sum_i \alpha_i k(x_i, x)$.

Furthermore, there always exists a feature map ϕ and vector space $\mathcal{V}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{V}}$. This alternative representation of the kernel provides another way to write its associated RKHS as

$$\mathcal{H} = \{x \mapsto \langle w, \phi(x) \rangle_{\mathcal{V}} : w \in \mathcal{V}\}. \quad (2.21)$$

This form allows us to clearly see that $\mathcal{H}$ contains functions which are potentially non-linear in x because of non-linearities in the feature map ϕ , but become linear when seen as functions of $\phi(x)$.

Practical transparency Properties of the kernel can translate into concrete structural or regularity properties on the class of function described by $\mathcal{H}$, and therefore can guide practitioners in defining a kernel that bestows the desired properties on $\mathcal{H}$. This is what we refer to as *practical transparency*.

For example, the Matérn family (Stein, 1999) is a widely used family of kernels in spatial statistics and machine learning, which we use throughout this thesis. It can be shown the RKHS they induce are equivalent to Sobolev spaces, and therefore bestow a controlled degree of differentiability. Therefore, by choosing an adequate kernel from the Matérn family, we can control the smoothness of our hypothesis space. Popular choices are the Matérn-1/2, Matérn-3/2 and Matérn-5/2 kernels which respectively induce RKHSs of continuous, differentiable and twice differentiable functions. The limiting case of the Matérn family corresponds to the Gaussian kernel, which induces a RKHS of infinitely differentiable functions. Many more elements of

practical transparency exist (e.g. periodic kernels induce RKHSs of periodic functions), with a rich literature already available, and we refer the reader to (Schölkopf and Smola, 2002) for more details.

As mentioned previously, being psd is a property which is stable under addition and multiplication. However, these operations also translate into concrete structural change for the hypothesis space. Namely, let k_1, k_2 denote two psd kernels and $\mathcal{H}(k_1), \mathcal{H}(k_2)$ their associated RKHS. Then $k_1 + k_2$ induces $\mathcal{H}(k_1 + k_2) = \{f + g : f \in \mathcal{H}(k_1), g \in \mathcal{H}(k_2)\}$ and $k_1 k_2$ induces $\mathcal{H}(k_1 k_2) = \{fg : f \in \mathcal{H}(k_1), g \in \mathcal{H}(k_2)\}$. This can be used to impose a desired additive and/or multiplicative structure on the hypothesis space. Duvenaud et al. (2013) provide details on how a desired structure for the hypothesis can be expressed through composition and combination of kernels.

Theoretical transparency Properties of the kernel may also translate into guarantees on the RKHS that are useful for theoretical reasoning, for instance in learning theory. This is what we refer to as *theoretical transparency*. For example, if the kernel we choose is bounded, every function in the RKHS will also be bounded.

An important category of theoretical guarantees we can have on a RKHS relates to its “functional richness”. For example, choosing $k(x, x') = x^\top x'$ induces a space of linear functions $\mathcal{H} = \{x \mapsto w^\top x : w \in \mathbb{R}^d\}$, which is a restrictive hypothesis space. On the other hand, it can be shown that the RKHS induced by the Gaussian kernel is dense in the space of square-integrable functions.

We say that a kernel is *universal* with respect to a space $\mathcal{F}$ if the RKHS it induces is dense in $\mathcal{F}$, which means it can arbitrarily well approximate functions from $\mathcal{F}$. This is a property satisfied by many popular kernels — such as the Matérn family — with respect to important function space such as continuous functions or square-integrable functions. Sriperumbudur et al. (2011) provide an in-depth discussion of universality, with a simple criterion to check whether a kernel is universal. For the purposes of theoretical analysis, working with a universal kernel may allow us to build intuition which should carry over to more general function spaces.

However, this is also an important practical consideration when tackling a learning problem. Indeed, we ideally want the optimal regressor f^* to be as close as possible to our chosen hypothesis space. Therefore, choosing $\mathcal{H}$ to be dense in large function spaces increases our chances to get arbitrarily close to f^* , thereby reducing what is referred to as the *approximation error*.

Example 2.2.8 (Transparency of the RKHS induced by the Gaussian kernel). We propose to illustrate these 3 levels of transparency with an example. When working

with kernel methods on $\mathcal{X} = \mathbb{R}$, the most popular and often surprisingly effective choice of kernel is the Gaussian kernel

$$k(x, x') = \exp\left(\frac{-(x - x')^2}{2\ell^2}\right). \quad (2.22)$$

We propose to flesh out the properties bestowed by this kernel over the RKHS it induces. First, from the angle of the reproducing transparency, we have that the RKHS can be written as $\mathcal{H} = \overline{\text{Span}}\{k(\cdot, x) : x \in \mathcal{X}\}$. Therefore, we already know that all functions in $\mathcal{H}$ arise from the limit when $n \rightarrow \infty$ of functions of the form

$$f_n(x) = \sum_{i=1}^n a_i \exp\left(\frac{-(x - x_i)^2}{2\ell^2}\right), \quad (2.23)$$

for some sequences of $(a_n)_{n \geq 1}$ and $(x_n)_{n \geq 1}$. Therefore, from the angle of practical transparency, it should be reasonable to expect that smoothness properties of the negative squared exponential function should be propagated into functions in $\mathcal{H}$.

Continuing with the reproducing transparency of the RKHS, let $\mathcal{V} = \ell^2(\mathbb{N})$ be the Hilbert space of square-summable sequences with inner product $\langle u, v \rangle_{\mathcal{V}} = \sum_{n=0}^{\infty} u_n v_n$. Further, define the feature map $\phi : \mathcal{X} \rightarrow \mathcal{V}$ given by

$$\phi : x \mapsto (b_n x^n e^{-x^2/\ell^2})_{n \in \mathbb{N}}, \quad (2.24)$$

where $b_n = \sqrt{2^n / \ell^{2n} n!}$. Using Taylor expansions, it is possible to show that the Gaussian kernel can be rewritten as $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{V}}$. As a result, we can derive an alternative expression of its RKHS following

$$\mathcal{H} = \{x \mapsto \langle w, \phi(x) \rangle_{\mathcal{V}} : w \in \mathcal{V}\} \quad (2.25)$$

$$= \left\{ x \mapsto \sum_{n=0}^{\infty} w_n x^n e^{-x^2/\ell^2} : \sum \left(\frac{w_n}{b_n} \right)^2 < \infty \right\}. \quad (2.26)$$

This form is of great interest since it provides a concrete representation of $\mathcal{H}$ as infinite polynomials multiplied by a negative squared exponential. From the angle of practical transparency, we can verify that any function of this form is infinitely differentiable. Therefore, using a Gaussian kernel bestows infinite smoothness on the hypothesis space we choose. From the angle of theoretical transparency, we can show using the Stone-Weierstrass theorem that any continuous function on a closed interval can be arbitrarily well approximated by this form of infinite sum. Therefore, the RKHS induced by the Gaussian kernel is dense in the larger space of continuous function on a closed interval — in reality $\mathcal{H}$ is dense in much more general spaces such as the space of square-integrable functions (Sriperumbudur et al., 2011). For many more details on the Gaussian kernel and its RKHS we refer to Steinwart and Christmann (2008b, Chapter 4, Section 4.4).

2.2.5 Algorithmic transparency of kernel ridge regression

Equipped with the RKHS theoretical background, we return for this last section to our motivating ridge regression problem. Let $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a psd kernel that we specify and fix, and denote $\mathcal{H}$ its associated RKHS of real-valued functions $f : \mathcal{X} \rightarrow \mathbb{R}$. By setting $\mathcal{F} = \mathcal{H}$ as our hypothesis space, the regression problem corresponds to what is commonly referred to as *kernel ridge regression* (KRR), i.e.

$$\arg \min_{f \in \mathcal{H}} \hat{R}_{\lambda,n}(f). \quad (2.27)$$

By the representer theorem, we know with slight notation abuse this is equivalent to the finite dimensional problem $\min_{\alpha \in \mathbb{R}^n} \hat{R}_{\lambda,n}(\alpha)$, where if a solution exists, it must take the form $f(x) = \sum_i \alpha_i k(x_i, x)$. The following result shows there exists a unique solution, which in addition can be obtained analytically.

Proposition 2.2.9 (Existence and uniqueness of the KRR solution). *The kernel ridge regression problem admits a unique minimiser in $\mathcal{H}$, and its coefficients are obtained in closed-form following*

$$\alpha = (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{y}, \quad (2.28)$$

where $\mathbf{y} = [y_1 \ \dots \ y_n]^\top$, $\mathbf{K} = [k(x_i, x_j)]_{1 \leq i, j \leq n}$, and $\mathbf{I}_n$ is the identity matrix.

This proposition can be easily verified, with proof provided, for example, in Mairal and Vert (2018). This is a powerful result because it guarantees that this learning problem in an infinite dimensional space $\mathcal{H}$, of potentially non-linear functions, admits a unique solution which can be determined analytically using standard matrix inversion techniques. This is another manifestation of how kernel methods allow the modelling of non-linear functions using tools from linear algebra. Such theoretical guarantees provide confidence that the learning algorithm will behave predictably and with analytical tractability on any dataset, and for any choice of kernel, which is a strong element of algorithmic transparency.

2.3 Bayesian learning with kernels

As far as kernel-based learning goes, Gaussian Processes (GPs) stand out as the other celebrated class of models in the machine learning literature. In a regression setting, they differ from kernel ridge regression by the probabilistic treatment they offer. They are often presented as the Bayesian counterpart of kernel methods, and in fact showcase striking similarities with RKHS methods. We refer the reader to the

work of Kanagawa et al. (2018) for an in-depth review of the connections between GPs and kernel methods.

In this section, we first introduce GPs and explain their relevance to specify a prior over a regression function. Then, we discuss the transparency of how the choice of covariance affects the nature of this prior, and finally we present how regression with uncertainty quantification is performed with GPs. Our exposition focuses on a regression setting and we refer the reader to Rasmussen and Williams (2005) for a comprehensive overview of GPs in machine learning.

2.3.1 A Bayesian prior over functions

Definition 2.3.1 (Gaussian process). A family of real-valued random variables $\{f(x) : x \in \mathcal{X}\}$ is called a Gaussian process if and only if any finite collection of elements from this family follows a multivariate normal distribution.

Concretely, this definition means that for any finite collection of input points $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$, if we denote $f(\mathbf{x}) = [f(x_1) \ \dots \ f(x_n)]^\top$, then there must exist a vector $\mathbf{m} \in \mathbb{R}^n$ and a positive definite matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ such that $f(\mathbf{x}) \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$.

For any GP $\{f(x) : x \in \mathcal{X}\}$, or simply f , we can associate a mean function $m : \mathcal{X} \rightarrow \mathbb{R}$ and a covariance function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ defined by

$$m(x) = \mathbb{E}[f(x)] \quad k(x, x') = \text{Cov}(f(x), f(x')). \quad (2.29)$$

The covariance function k defines a positive-definite kernel over $\mathcal{X}$, and for this reason, is often referred to as the *kernel* of the GP. An important property of GPs is that they are fully and uniquely characterised by their mean and covariance functions. For this reason, we classically denote

$$f \sim \text{GP}(m, k). \quad (2.30)$$

The mean and the kernel then allow to specify the distribution followed by any finite collection of elements from $\{f(x) : x \in \mathcal{X}\}$. In particular, if we define $\mathbf{m} = [m(x_1) \ \dots \ m(x_n)]^\top$ and $\mathbf{K} = [k(x_i, x_j)]_{1 \leq i, j \leq n}$, then we have

$$f(\mathbf{x}) \sim \mathcal{N}(\mathbf{m}, \mathbf{K}), \quad (2.31)$$

for any $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$.

A major interest of GPs in machine learning is that, given a function $m : \mathcal{X} \rightarrow \mathbb{R}$ and positive-definite kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, there exists a unique stochastic process $\{f(x) : x \in \mathcal{X}\}$ such that $f \sim \text{GP}(m, k)$. As a practitioner, this implies that by

selecting a mean and a kernel, we can completely define a GP. Thus, GPs can be specified and used as Bayesian non-parametric priors over functions, providing a powerful framework for probabilistic modelling.

This is particularly appealing in a regression problem, since it allows us to place a prior over the class of functions $f : \mathcal{X} \rightarrow \mathcal{Y}$ from which we seek our regressor. The next section discusses how the choice of mean and covariance function affects the nature of the prior we place over our regression function.

2.3.2 Transparency on the prior induced by a kernel

To place a GP prior over a function, two key components must be selected: a mean function m and a kernel k . The mean function m sets the average expected behaviour of the function. Using a zero mean for simplicity is customary, effectively focusing on the centred GP $\tilde{f} = f - m$. Alternatively, if we wish to parametrise the mean as a constant or a linear function, this can be built directly in the kernel. Therefore, from here onward we will assume $m = 0$.

The kernel k specifies how we encode correlations between the values the function takes at different locations. The intuition is that $f(x)$ and $f(x')$ should correlate if x and x' are close, where the appropriate notion of correlation is precisely what $k(x, x')$ specifies — which echoes the measure of similarity provided by the reproducing kernel in a RKHS.

Choosing the kernel allows the practitioner to transparently specify the kind of correlation the GP should capture, and therefore determines the form or prior uncertainty we place over f . For example, choosing a Dirac kernel $k(x, x') = \delta(x - x')$ encodes that $f(x)$ and $f(x')$ are independent unless $x = x'$. This encodes no correlation at all, i.e. even arbitrarily close points are uninformative of each other. With this kernel, f models a white noise. Conversely, choosing a constant kernel $k(x, x') = 1$ encodes that the values should correlate everywhere. Therefore arbitrarily distant points perfectly inform each other, and f can only model constant functions.

In between these extremes, standard choices of kernels rather specify a nuanced notion of correlation as a function of a distance between x and x' . For example, the Matérn family is a widely used class of covariances over $\mathcal{X} = \mathbb{R}^d$. The Matérn- ν kernel takes the general form

$$k_\nu(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{\|x - x'\|_2}{\ell} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{\|x - x'\|_2}{\ell} \right), \quad (2.32)$$

where Γ is the gamma function, K_ν is the modified Bessel function, and $\ell > 0$ is a bandwidth hyperparameter. This expression considerably simplifies for $\nu = p + 1/2$

where $p \in \mathbb{N}$, specifying a covariance that decays as the distance $\|x - x'\|$ increases. In addition, like in RKHSs, the Matérn family also allows to control the smoothness of the function modelled. As such, the Matérn-1/2, Matérn-3/2 and Matérn-5/2 kernels guarantee that the realisation from the GP will respectively be continuous, differentiable or twice differentiable³. The Gaussian (or Matérn- ∞) kernel bestows infinite differentiability on f .

For details on standard choices of kernels for GPs in machine learning, we refer the reader to (Rasmussen and Williams, 2005, Chapter 4).

2.3.3 Gaussian process regression

We return now to our introductory regression problem and seek a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ that best represents the relationship

$$Y = f^*(X) + \varepsilon, \quad (2.33)$$

where we recall that $f^*(X) = \mathbb{E}[Y|X]$ and $\varepsilon = Y - \mathbb{E}[Y|X]$ is the residual. We make the additional assumption that the residual can be modelled as an additive Gaussian noise term $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Instead of seeking f^* as the function minimising the population risk, we follow here a Bayesian route and place a GP prior $f \sim \text{GP}(0, k)$. Intuitively, this prior serves as an initial probability distribution over functions $f : \mathcal{X} \rightarrow \mathcal{Y}$ that are suitable candidates to model the conditional expectation f^* . The prior knowledge on the class of functions described is instilled through the choice of the kernel k .

To best reflect the observations we have in $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, we then seek to update this prior into a posterior given by $f | \mathcal{D}$. It is easy to show using Gaussian conjugacy that, under an additive Gaussian noise model, the posterior is given by another GP $f | \mathcal{D} \sim \text{GP}(\bar{m}, \bar{k})$ with a posterior mean and a posterior covariance available in closed-form as

$$f | \mathcal{D} \sim \text{GP}(\bar{m}, \bar{k}) \quad (2.34)$$

$$\bar{m}(x) = k(x, \mathbf{x}) (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{y} \quad (2.35)$$

$$\bar{k}(x, x') = k(x, x') - k(x, \mathbf{x}) (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} k(\mathbf{x}, x), \quad (2.36)$$

where $\mathbf{y} = [y_1 \ \dots \ y_n]^\top$, $k(x, \mathbf{x}) = [k(x, x_1) \ \dots \ k(x, x_n)]$ and $k(\mathbf{x}, x) = k(x, \mathbf{x})^\top$.

³continuity and differentiability being defined in terms of the square-integrable function norm, see Rasmussen and Williams (2005, Chapter 4, Section 4.2.1)

The posterior GP serves as a probabilistic model of f^* which can be used to make prediction over unseen inputs with principled uncertainty quantification. Indeed, a predictive probability distribution over any $x \in \mathcal{X}$ can be taken as

$$f(x) \mid \mathcal{D} \sim \mathcal{N}(\bar{m}(x), \bar{k}(x, x)), \quad (2.37)$$

where the posterior mean $\bar{m}(x)$ serves as a deterministic prediction like in KRR, and the posterior variance $\bar{k}(x, x)$ serves as a measure of uncertainty quantification. In particular, the posterior variance $\bar{k}(x, x)$ allows to reduce the uncertainty from the prior variance in areas of $\mathcal{X}$ where data has been observed, thereby acknowledging the information we have updated the prior with.

Interestingly, we notice that if we define $\boldsymbol{\alpha} = (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{y}$ then the posterior mean can be rewritten as

$$\bar{m}(x) = \sum_{i=1}^n \alpha_i k(x_i, x), \quad (2.38)$$

which is exactly the solution to KRR with a regularisation $\lambda = \sigma^2$. Besides highlighting yet another connection with kernel methods, this also shows that the amplitude of the noise ε plays a regularisation role in the posterior update.

Beyond the uncertainty quantification it offers over KRR, GP regression also proposes a principled strategy for hyperparameter tuning. Indeed, whilst KRR typically relies on cross-validation and grid search for hyperparameter selection, GPs allow tuning the hyperparameters through the principle of maximum marginal likelihood. This involves finding the hyperparameter values that maximise the observations' log-likelihood under the prior, and can be achieved for instance using a gradient-based routine.

2.4 Transforming kernel-based learners

Throughout this thesis, we dedicate particular attention to the incorporation of domain knowledge which can be framed as a transformation of the hypothesis $f \in \mathcal{F}$. Namely, starting from a hypothesis space $\mathcal{F}$ of functions $f : \mathcal{X} \rightarrow \mathcal{Y}$, the domain knowledge is represented by a transformation

$$A : \mathcal{F} \rightarrow \text{Range}(A), \quad (2.39)$$

such that $\text{Range}(A) = \{Af : f \in \mathcal{F}\}$ describes a restricted hypothesis space of functions that satisfy the domain knowledge we wish to incorporate. For example, the transformation $A : f \mapsto \exp \circ f$ can be used to restrict the hypothesis space to a space

of positive functions, and therefore encode the domain knowledge that predictions should be positive.

The transformation A may in principle take any form. For instance in Chapter 5, we consider a transformation of the form $Af(x, h) = \psi(f(x))e^{-h/L}$, where ψ is a non-linear mapping, h is the altitude, and L a lengthscale. A particular class of transformations however stands out by its simplicity and yet pervasiveness in seemingly unrelated tasks: linear transformations. For example, in Chapter 3 our probabilistic causal knowledge leads us to consider the linear transformation

$$Af(x_1, x_2) = f(x_1, x_2) - \mathbb{E}[f(X_1, X_2)|X_2 = x_2] \quad (2.40)$$

that maps onto the space of functions with zero conditional expectation. In Chapter 4, the resolution of a differential equation again yields a linear transformation of the form

$$Af(x, t) = \int_0^t f(x, s)e^{-(t-s)/d} ds \quad (2.41)$$

which corresponds to a convolution with an exponential function. Linear structures also naturally arise in the study of PDEs or geometric invariances (Cohen and Welling, 2016; Elesedy, 2021).

Given our emphasis on kernel-based learners, we would like to understand how kernel-based learners behave under linear transformations to encode domain knowledge. The following result states that if the hypothesis space is a RKHS $\mathcal{H}$, then under mild regularity conditions on the operator A , we can explicitly describe its range as another RKHS.

Proposition 2.4.1 (Linear transformation of a RKHS). *Let $\mathcal{H}$ be a real-valued RKHS with reproducing kernel k and let $A : \mathcal{H} \rightarrow \mathcal{H}$ be a bounded linear operator. Then $\text{Range}(A)$ is also a RKHS with reproducing kernel*

$$k_A(x, x') = \langle A^*k_x, A^*k_{x'} \rangle_{\mathcal{H}}, \quad (2.42)$$

where A^* denotes the adjoint of A .

Proof. Let $\mathcal{H}_A$ be the RKHS with reproducing kernel k_A . We start by showing that $\text{Range}(A) \subseteq \mathcal{H}_A$. Let $f \in \text{Range}(A)$, then it admits a pre-image $w_f \in \mathcal{H}$ such that $f = Aw_f$. Therefore, for any $x \in \mathcal{X}$ we have by the reproducing property that

$$f(x) = \langle f, k_x \rangle_{\mathcal{H}} = \langle Aw_f, k_x \rangle_{\mathcal{H}} = \langle w_f, A^*k_x \rangle_{\mathcal{H}}. \quad (2.43)$$

Hence, f writes as an element of the RKHS induced by the feature map $\phi : x \mapsto A^*k_x$ which corresponds exactly to $\mathcal{H}_A$. Therefore $\text{Range}(A) \subseteq \mathcal{H}_A$.

Reciprocally, let us now show that $\mathcal{H}_A \subseteq \text{Range}(A)$. Let $f \in \mathcal{H}_A$, then because $\mathcal{H}_A$ is induced by the feature map $\phi : x \mapsto A^*k_x$, there exists $w_f \in \mathcal{H}$ such that for any $x \in \mathcal{X}$,

$$f(x) = \langle w_f, \phi(x) \rangle_{\mathcal{H}} = \langle w_f, A^*k_x \rangle_{\mathcal{H}} = Aw_f(x), \quad (2.44)$$

which concludes the proof. $\square$

This proposition implies that, if the operator A can be evaluated, we can encode its domain knowledge by directly performing KRR in the RKHS with reproducing kernel k_A . This is relevant to the work presented in Chapter 3. It is worth noting this result requires two additional assumptions beyond the linearity of operator: first, that A is closed on the RKHS i.e. $\text{Range}(A) \subseteq \mathcal{H}$, and second, that A is a bounded (or continuous) linear operator. This last assumption, although common, excludes important classes of operators such as differentiation operators which are not continuous.

Turning now to GPs as the other category of kernel-based learners, recall that GPs can intuitively be interpreted as the infinite dimensional generalisation of multivariate normal distributions. It is well known that if $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ follows a d -dimensional multivariate normal distribution, then $\mathbf{AZ} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$ for any matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$. The following result shows this intuition carries over to linear transformation of GPs.

Proposition 2.4.2 (Linear transformation of a GP). *Let A be linear operator and $f \sim \text{GP}(m, k)$. Then Af is also a GP with mean and covariances given by*

$$\mathbb{E}[Af(x)] = Am(x) \quad (2.45)$$

$$\text{Cov}(Af(x), Af(x')) = (AkA^*)(x, x'), \quad (2.46)$$

where the notation AkA^* indicates that A and A^* respectively act on $k(x, x')$ as a function of its left and right variable.

The strength of this result is that it allows us to incorporate linear domain knowledge into a GP simply by defining a different GP prior over $\text{Range}(A)$. This involves adjusting both the mean and covariance functions in the new prior to reflect the transformation A . This is of particular relevance to the work presented in Chapter 4. In contrast to the RKHS version, this does not require the operator A to be continuous, and therefore holds with greater generality. Thus, provided differentiability of the mean and kernel, it applies to differentiation operators as well (Särkkä, 2011; Agrell, 2019)

Chapter 3

Returning the Favour: When Regression Benefits from Probabilistic Causal Knowledge

Bouabid, Shahine, Jake Fawkes, and Dino Sejdinovic. “Returning the Favour: When Regression Benefits from Probabilistic Causal Knowledge.” *International Conference on Machine Learning*. PMLR, 2023.

Returning The Favour: When Regression Benefits From Probabilistic Causal Knowledge

Shahine Bouabid^{*1} Jake Fawkes^{*1} Dino Sejdinovic²

Abstract

A directed acyclic graph (DAG) provides valuable prior knowledge that is often discarded in regression tasks in machine learning. We show that the independences arising from the presence of collider structures in DAGs provide meaningful inductive biases, which constrain the regression hypothesis space and improve predictive performance. We introduce *collider regression*, a framework to incorporate probabilistic causal knowledge from a collider in a regression problem. When the hypothesis space is a reproducing kernel Hilbert space, we prove a strictly positive generalisation benefit under mild assumptions and provide closed-form estimators of the empirical risk minimiser. Experiments on synthetic and climate model data demonstrate performance gains of the proposed methodology.

1. Introduction

Causality has recently become a main pillar of research in the machine learning community. Historically, machine learning has been used to help solve problems in the field of causal inference (Shalit et al., 2017; Zhang et al., 2012). But recently a different focus has emerged, asking what causality can do to return the favour to machine learning (Schölkopf et al., 2021). In this work we continue in this vein, and aim to answer whether the knowledge of a causal directed acyclic graph (DAG) underpinning the data generating process can assist and improve performance in regression tasks.

When a causal DAG is available, it constitutes a source of prior knowledge that is typically discarded when addressing a regression problem. It can however guide the setup of the regression problem. Classically, the structure of a DAG

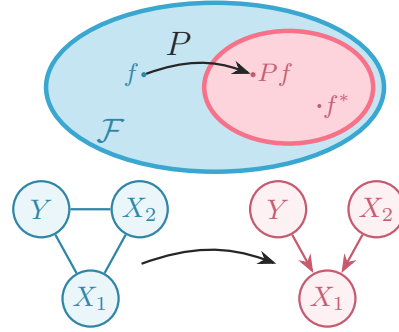


Figure 1. When performing regression in a hypothesis space $\mathcal{F}$ (blue), we implicitly assume that the data generating process could follow any DAG structure. The optimal regressor f^* lies in the subspace of function that satisfy the independence structure arising from the collider (pink), onto which the projection P maps.

informs on which predictors should be selected to regress a given response variable Y . This process, known as feature selection, is solved by selecting the predictors that are either adjacent to Y , or that influence children of Y . The resulting set of predictors is called the Markov boundary of Y (Pearl, 1987).

As we will see, the presence of a particular structure in a Markov boundary is typically overlooked in regression problems: colliders of the form $Y \rightarrow X_1 \leftarrow X_2$. In this work, we investigate how the conditional independence constraints arising due to colliders in the Markov boundary can be used to construct useful inductive biases in a regression problem and to guide the choice of the hypothesis space. We will see that the colliders are also unique in that regard: beyond colliders, the Markov boundary cannot contain any graphical structure implying a conditional independence with Y .

To understand the intuition behind colliders, consider this classic example: imagine we have a randomly timed sprinkler (X_2) and we want to infer whether it has rained (Y), having observed whether the sidewalk is wet (X_1). Although the sprinkler and the rain are marginally independent, knowing whether the sprinkler has been active is important for determining whether it has rained. Colliders arise naturally in many application domains. For example, in climate science, the objective may be to regress an environmental driver Y that, independently from human activity X_2 , influences observed global temperatures X_1 .

^{*}Equal contribution ¹Department of Statistics, University of Oxford, UK ²School of CMS & AIML, University of Adelaide, Australia. Correspondence to: Shahine Bouabid <shahine.bouabid@stats.ox.ac.uk>, Jake Fawkes <jake.fawkes@stats.ox.ac.uk>.

As illustrated in Figure 1, when performing least-square regression over a hypothesis space $\mathcal{F}$, only a subset of $\mathcal{F}$ will comply with the independences arising from the collider. By considering the projection operator P that maps onto this subspace, we propose a framework called *collider regression* to incorporate inductive biases arising from colliders into any regressor. We show that when the data generating process follows a collider, projecting any given regressor onto this subspace provides a positive generalisation benefit.

We then consider the specific case where the hypothesis space is a reproducing kernel Hilbert space (RKHS). Because RKHSs are rich functional spaces that also enjoy closed analytical solutions to the least-squares regression problem, they allow us to build intuition for the general case. We prove a strictly positive generalisation benefit from projecting the least-squares empirical risk minimiser in a RKHS, where the size of the generalisation gap increases with the complexity of the problem. We also show that for a RKHS, it is possible to solve the least-squares regression problem directly inside the projected hypothesis subspace and provide closed-form estimators.

We experimentally validate the effectiveness of our methodology on a synthetic dataset and on a real world climate science dataset. Results demonstrate that collider regression consistently provides an improvement in generalisation at test time in comparison with standard least-squares regressors. Results also suggest that collider regression is particularly beneficial when few training samples are available, but samples from the covariates can easily be obtained, i.e. in a semi-supervised learning setting.

2. Background

Regression notation Let Y be our target variable over $\mathcal{Y} \subseteq \mathbb{R}$ and X be our covariates over $\mathcal{X}$. Our goal is a standard regression task where we have access to a dataset $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\} \in (\mathcal{X} \times \mathcal{Y})^n$ of n samples $(x^{(i)}, y^{(i)})$ from (X, Y) . We aim to minimise the regularised empirical risk

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \left(y^{(i)} - f(x^{(i)}) \right)^2 + \lambda \Omega(f) \quad (1)$$

where $\mathcal{F}$ is a specified hypothesis space of functions $f: \mathcal{X} \rightarrow \mathcal{Y}$, $\lambda > 0$ and $\Omega(f) > 0$ is a regularisation term. This corresponds to finding a function $\hat{f}$ that best estimates the optimal regression function for the squared loss:

$$f^*(x) = \mathbb{E}[Y|X = x]. \quad (2)$$

For any two functions $h, h' \in \mathcal{F}$, the squared-error generalisation gap between h and h' is defined as the difference in their true risk:

$$\Delta(h, h') = \mathbb{E}[(Y - h(X))^2] - \mathbb{E}[(Y - h'(X))^2]. \quad (3)$$

Therefore if $\Delta(h, h') \geq 0$, it means that h' generalises better from the training data than h .

Reproducing kernel Hilbert spaces Let $\mathcal{X}$ be some non-empty space. A real-valued RKHS $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is a complete inner product space of functions $f: \mathcal{X} \rightarrow \mathbb{R}$ that admits a bounded evaluation functional. For $x \in \mathcal{X}$, the Riesz representer of the evaluation functional is denoted $k_x \in \mathcal{H}$ and satisfies the *reproducing property* $f(x) = \langle f, k_x \rangle_{\mathcal{H}}$, $\forall f \in \mathcal{H}$. The bivariate symmetric positive definite function defined by $k(x, x') = \langle k_x, k_{x'} \rangle_{\mathcal{H}}$ is referred to as the *reproducing kernel* of $\mathcal{H}$. Conversely, the Moore-Aronszajn theorem (Aronszajn, 1950) shows that any symmetric positive definite function k is the unique reproducing kernel of an RKHS. For more details on RKHS theory, we refer the reader to Berlinet & Thomas-Agnan (2011).

Conditional Mean Embeddings Conditional mean embeddings (CMEs) provide a powerful framework to represent conditional distributions in a RKHS (Fukumizu et al., 2004; Song et al., 2013; Muandet et al., 2016). Given random variables X, Z on $\mathcal{X}, \mathcal{Z}$ and an RKHS $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ with reproducing kernel $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, the CME of $\mathbb{P}(X|Z = z)$ is defined as

$$\mu_{X|Z=z} = \mathbb{E}[k_X|Z = z] \in \mathcal{H}. \quad (4)$$

It corresponds to the Riesz representer of the conditional expectation functional $f \mapsto \mathbb{E}[f(X)|Z = z]$ and can thus be used to evaluate conditional expectations by taking an inner product $\mathbb{E}[f(X)|Z = z] = \langle f, \mu_{X|Z=z} \rangle_{\mathcal{H}}$.

Introducing a second RKHS $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ with reproducing kernel $\ell: \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$, Grünewälder et al. (2012) propose an alternative view of CMEs as the solution to the least-square regression of canonical feature maps ℓ_Z onto k_X

$$\begin{cases} E^* = \arg \min_{C \in \mathcal{B}_2(\mathcal{G}, \mathcal{H})} \mathbb{E}[\|k_X - C\ell_Z\|_{\mathcal{H}}^2] \\ \mu_{X|Z=z} = E^* \ell_z \end{cases} \quad (5)$$

where $\mathcal{B}_2(\mathcal{G}, \mathcal{H})$ denotes the space of Hilbert-Schmidt operators¹ from $\mathcal{G}$ to $\mathcal{H}$. Given a dataset $\mathcal{D} = \{\mathbf{x}, \mathbf{z}\}$, this perspective allows to compute an estimate of the associated operator $E^*: \mathcal{G} \rightarrow \mathcal{H}$ as the solution to the regularised empirical least-squares problem as

$$\begin{cases} \hat{E}^* = \arg \min_{C \in \mathcal{B}_2(\mathcal{G}, \mathcal{H})} \frac{1}{n} \sum_{i=1}^n \|k_{x^{(i)}} - C\ell_{z^{(i)}}\|_{\mathcal{H}}^2 + \gamma \|C\|_{\mathcal{B}_2}^2 \\ = \mathbf{k}_x^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_z \\ \hat{\mu}_{X|Z=z} = \hat{E}^* \ell_z = \mathbf{k}_x^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_z(z) \end{cases} \quad (6)$$

where $\gamma > 0$, $\mathbf{L} = \ell(\mathbf{z}, \mathbf{z})$, $\mathbf{k}_x = k(\mathbf{x}, \cdot)$ and $\ell_z = \ell(\mathbf{z}, \cdot)$. We refer the reader to (Muandet et al., 2017) for a comprehensive review of CMEs.

¹i.e. bounded operators $A: \mathcal{G} \rightarrow \mathcal{H}$ such that $\text{Tr}(A^*A) < \infty$. $\mathcal{B}_2(\mathcal{G}, \mathcal{H})$ has a Hilbert space structure for the inner product $\langle A, B \rangle_{\mathcal{B}_2} = \text{Tr}(A^*B)$.

3. DAG inductive biases for regression

In this section, we aim to answer how knowledge of the causal graph of the underlying data generating process can help to perform regression. We start by reviewing the concept of Markov boundaries and how it is used for feature selection. We then show that even after feature selection has been performed, there is still residual information from colliders that is relevant for a regression problem.

3.1. Markov boundary for feature selection

Since we are focusing on regression, we are interested in how the DAG can inform us about $\mathbb{P}(Y|X)$. Suppose that for some vertex X_i , the DAG informs us that $Y \perp\!\!\!\perp X_i \mid X \setminus X_i$. Stated in terms of mutual information we have that² $I(Y; X) = I(Y; X \setminus X_i)$, therefore we can discard X_i from our set of covariates without any loss of probabilistic information for $\mathbb{P}(Y|X)$.

From a functional perspective, we can interpret this as incorporating the inductive bias that the regressor need only depend on $X \setminus X_i$, allowing us to learn simpler functions which should generalise better from the training set.

By repeating the process of removing features, we can iteratively construct a minimal set of necessary covariates that still retain all the probabilistic information about $\mathbb{P}(Y|X)$. This is known as feature selection (Dash & Liu, 1997).

Such a set, S , should satisfy $Y \perp\!\!\!\perp X \setminus S \mid S$ and we should not be able to remove a vertex from S without losing information about $\mathbb{P}(Y|X)$. A set of this form is known as the Markov boundary of Y (Statnikov et al., 2013), denoted by $\text{Mb}(Y)$. If the only independences in the distribution are those implied by the DAG structure³ then the Markov boundary is uniquely given by

$$\text{Mb}(Y) = \text{Pa}(Y) \cup \text{Ch}(Y) \cup \text{Sp}(Y), \quad (7)$$

where $\text{Pa}(Y)$ are the parents of Y , $\text{Ch}(Y)$ are the children of Y and $\text{Sp}(Y)$ are the spouses of Y , i.e. the children's other parents. In Figure 2 the Markov boundary of Y is highlighted in blue.

3.2. Extracting inductive bias for regression

By construction the Markov boundary of Y cannot contain independence relationships of the form $Y \perp\!\!\!\perp X_i \mid X \setminus X_i$. However, it can still contain unused independence statements that involve Y , and therefore provides useful information about the conditional distribution $\mathbb{P}(Y|X)$.

For example, the graphical structure in Figure 2 gives

²This follows from $I(Y; X) = I(Y; X \setminus X_i) + I(Y; X_i \mid X \setminus X_i)$ and the conditional independence gives $I(Y; X_i \mid X \setminus X_i) = 0$.

³An assumption known as faithfulness (Meek, 1995) which we take throughout.

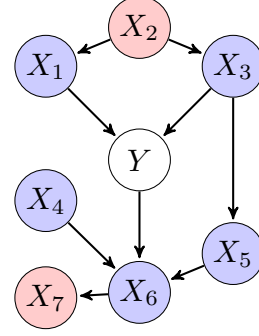


Figure 2. A causal graph with the Markov boundary of Y highlighted in blue and vertices outside the Markov boundary highlighted in red. Whilst Y and X_4 are marginally independent, the presence of the collider X_6 opens the path between Y and X_4 .

that $Y \perp\!\!\!\perp X_4$ and $Y \perp\!\!\!\perp X_5 \mid X_3$. This implies that $\mathbb{P}(Y|X_4) = \mathbb{P}(Y)$ and $\mathbb{P}(Y|X_3, X_5) = \mathbb{P}(Y|X_3)$ which by marginalisation constrains $\mathbb{P}(Y|X)$ and so gives us extra information about it. The presence of these independence relationships inside $\text{Mb}(Y)$ is only possible because a collider, X_6 , has allowed for the spouses X_4 and X_5 to be within the Markov boundary without being adjacent to Y .

Hence, the presence of collider structures within the Markov boundary of Y provides additional independence relationships involving Y . The following proposition shows that the presence of a collider is not only a sufficient condition, but also necessary.

Proposition 3.1. *The Markov boundary of Y contains a collider if and only if there exists $Z \in \text{Mb}(Y)$ and $S_Z \subset \text{Mb}(Y)$ such that $Y \perp\!\!\!\perp Z \mid S_Z$.*

Proof. We have a conditional independence between two variables if and only if they are not adjacent (Lemma 3.1, 3.2 Koller & Friedman (2009)) and $\text{Mb}(Y)$ contains a variable not adjacent to Y if and only if it contains a collider. $\square$

The collider structures are thus the only graphical structures that provide conditional independence statement relevant to $\mathbb{P}(Y|X)$ within the Markov boundary. To the best of our knowledge, this information is currently left unused when addressing a regression problem.

However, unlike for the feature selection process, we cannot simply use these independence statements to discard covariates and reduce the set of features. This is because while the spouses of Y are uninformative on their own, they become informative in the presence of other covariates. Namely in Figure 2, while $Y \perp\!\!\!\perp X_4$ we have $Y \not\perp\!\!\!\perp X_4 \mid X_6$ because X_6 is a collider. Therefore, we have that $I(Y; X) > I(Y; X \setminus X_4)$ and discarding X_4 would constitute a loss of information.

4. Collider Regression

In this section, we present a method for incorporating probabilistic inductive bias from a collider structure into a regression problem, and provide guarantees of improved generalisation error. For the sake of clarity, our exposition focuses on the simple collider structure depicted in Figure 3. We however emphasise this simplification does not harm the generality of our contribution and Section 5 shows how collider regression can be extended to more general DAGs.

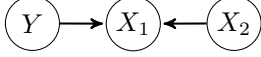


Figure 3. Simple collider structure

4.1. Simple collider regression setup

Let X_1, X_2, Y be random variables following the DAG structure in Figure 3 and taking values in $\mathcal{X}_1 \subseteq \mathbb{R}^{d_1}$, $\mathcal{X}_2 \subseteq \mathbb{R}^{d_2}$ and $\mathcal{Y} \subseteq \mathbb{R}$ respectively. Without loss of generality, we assume that $\mathbb{E}[Y] = 0$.

Under the squared loss, the optimal regressor is given by

$$f^*(x_1, x_2) = \mathbb{E}[Y | X_1 = x_1, X_2 = x_2]. \quad (8)$$

Since the collider gives the independence relationship $Y \perp\!\!\!\perp X_2$, we have that

$$\begin{aligned} \mathbb{E}[f^*(X_1, X_2) | X_2] &= \mathbb{E}[\mathbb{E}[Y | X_1, X_2] | X_2] \\ &= \mathbb{E}[Y | X_2] \\ &= \mathbb{E}[Y] \\ &= 0, \end{aligned} \quad (9)$$

where the second line comes from the tower property of the conditional expectation.

Hence, the optimal regressor f^* lies in the subspace of functions that have zero X_2 -conditional expectation. To incorporate the knowledge from the DAG into our regression procedure, we should therefore ensure that our estimate $\hat{f}$ lies within the same subspace of functions, i.e. we want to satisfy the zero conditional expectation constraint

$$\hat{f} \in \{f \in \mathcal{F} \mid \mathbb{E}[f(X_1, X_2) | X_2] = 0\}. \quad (\text{ZCE})$$

We propose to investigate how such a constraint can be enforced onto our hypothesis and how it benefits generalisation, starting by the general case of square-integrable functions. In what follows, we will use shorthand concatenated notations $X = (X_1, X_2)$, $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, $x = (x_1, x_2) \in \mathcal{X}$ and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{X}^n$.

4.2. Respecting the collider structure in the hypothesis

Let $L^2(X)$ denote the space of square-integrable functions with respect to the probability measure induced by X and

suppose $\mathcal{F} = L^2(X)$. Let $E : L^2(X) \rightarrow L^2(X)$ denote the conditional expectation operator defined by

$$Ef(x_1, x_2) = \mathbb{E}[f(X_1, X_2) | X_2 = \pi_2(x_1, x_2)], \quad (10)$$

where $\pi_2(x_1, x_2) = x_2$ is simply the mapping that discards the first component⁴.

The operator E classically defines an orthogonal projection over the subspace of X_2 -measurable functions. $L^2(X)$ thus orthogonally decomposes into its image, denoted $\text{Range}(E)$, and its null-space, denoted $\text{Ker}(E)$, as

$$L^2(X) = \text{Ker}(E) \oplus \text{Range}(E). \quad (11)$$

Using this notation, satisfying condition (ZCE) corresponds to having $\hat{f} \in \text{Ker}(E)$. Alternatively, if we denote

$$P = \text{Id} - E, \quad (12)$$

the orthogonal projection onto $\text{Ker}(E)$, then we want to take $\mathcal{F} = \text{Range}(P)$ as our hypothesis space.

In general, it may be hard to constrain the hypothesis space directly to be $\text{Range}(P)$. However, the solution to the empirical risk minimisation problem (1) will always orthogonally decompose within $L^2(X)$ as

$$\hat{f} = P\hat{f} + E\hat{f}, \quad (13)$$

where only $P\hat{f} \in \text{Range}(P)$ satisfies (ZCE). It turns out that discarding $E\hat{f}$ — the part that does not satisfy the constraint — will always yield generalisation benefits.

Proposition 4.1. *Let $h \in L^2(X)$ be any regressor from our hypothesis space. We have*

$$\Delta(h, Ph) = \|Eh\|_{L^2(X)}^2. \quad (14)$$

The generalisation gap is always greater than zero. Hence, for any given regressor $\hat{f}$, we can always improve its test performance by projecting it onto $\text{Range}(P)$.

In practice, a simple estimator of $P\hat{f}$ can be obtained by subtracting an estimate of $\mathbb{E}[\hat{f}(X_1, X_2) | X_2]$ as

$$\hat{P}\hat{f}(x_1, x_2) = \hat{f}(x_1, x_2) - \hat{\mathbb{E}}[\hat{f}(X_1, X_2) | X_2 = x_2] \quad (15)$$

by following the procedure outlined in Algorithm 1.

It is worth noting that the second step of Algorithm 1 does not require observations from Y . As such, it naturally fits a semi-supervised setup where additional observations $\mathcal{D}' = \{\mathbf{x}'_1, \mathbf{x}'_2\}$ are available, and can be used to produce a better estimate of the conditional expectation $\mathbb{E}[\hat{f}(X_1, X_2) | X_2]$.

⁴This notation emphasises that Ef is formally a function of (x_1, x_2) and belongs in $L^2(X)$

Algorithm 1 General procedure to estimate $P\hat{f}$

- 1: Regress $(X_1, X_2) \rightarrow Y$
to get $(x_1, x_2) \mapsto \hat{f}(x_1, x_2)$
- 2: Regress $X_2 \rightarrow \hat{f}(X_1, X_2)$
to get $x_2 \mapsto \hat{\mathbb{E}}[\hat{f}(X_1, X_2)|X_2 = x_2]$
- 3: Take $\hat{P}\hat{f}(x_1, x_2) = \hat{f}(x_1, x_2) - \hat{\mathbb{E}}[\hat{f}(X_1, X_2)|X_2 = x_2]$

4.3. Theoretical guarantees in a RKHS

RKHSs are mathematically convenient functional spaces and under mild assumptions on the reproducing kernel, they can be proven to be dense in $L^2(X)$ (Sriperumbudur et al., 2011). This makes them a powerful tool for theoretical analysis and building intuition which can be expected to carry over to more general function spaces. For this reason, in this section we study the case where the hypothesis space is a RKHS $\mathcal{F} = \mathcal{H}$. We denote its inner product by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and its reproducing kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$.

When solving the least-square regression problem in a RKHS, it is known that for Tikhonov regularisation $\Omega(f) = \|f\|_{\mathcal{H}}^2$, the solution to the empirical risk minimisation problem (1) in $\mathcal{H}$ enjoys a closed-form expression given by

$$\hat{f} = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{k}_x, \quad (16)$$

where $\mathbf{K} = k(\mathbf{x}, \mathbf{x})$ and $\mathbf{k}_x = k(\mathbf{x}, \cdot)$.

Therefore, if we now project $\hat{f}$ onto $\text{Range}(P)$ as previously, the projected empirical risk minimiser writes

$$P\hat{f} = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} P\mathbf{k}_x \quad (17)$$

with notation abuse $P\mathbf{k}_x = [Pk_{x(1)} \dots Pk_{x(n)}]^\top$.

Leveraging these analytical expressions, the following result establishes a strictly non-zero generalisation benefit from projecting $\hat{f}$. The proof techniques follows that of Elesedy (2021), but is adapted to our particular setup with relaxing assumptions about the projection orthogonality⁵ and the form of the data generating process.

Theorem 4.2. Suppose $M = \sup_{x \in \mathcal{X}} k(x, x) < \infty$ and $\text{Var}(Y|X) \geq \eta > 0$. Then, the generalisation gap between $\hat{f}$ and $P\hat{f}$ satisfies

$$\mathbb{E}[\Delta(\hat{f}, P\hat{f})] \geq \frac{\eta \mathbb{E}[\|\mu_{X|X_2}(X)\|_{L^2(X)}^2]}{(\sqrt{n}M + \lambda/\sqrt{n})^2} \quad (18)$$

where $\mu_{X|X_2} = \mathbb{E}[k_X|X_2]$ is the CME of $\mathbb{P}(X|X_2)$.

This demonstrates that in a RKHS, projecting the empirical risk minimiser is strictly beneficial in terms of generalisation

⁵ P is not necessarily orthogonal anymore as a projection of $\mathcal{H}$

⁵ E then corresponds to what is referred to as a conditional mean operator in the kernel literature (Fukumizu et al., 2004).

error. Specifically, if there exists a set with non-zero measure on which $Y \neq 0$ and $\mu_{X|X_2} \neq 0$ almost-everywhere, then the lower bound is strictly positive.

The magnitude of the lower bound depends on the variance of $\|\mu_{X|X_2}(X)\|_{L^2(X)}$ and the lower bound on $\text{Var}(Y|X)$. This indicates that problems with more complex conditional distributions $\mathbb{P}(X|X_2)$ and $\mathbb{P}(Y|X)$ should enjoy a larger generalisation gap.

The theorem also suggests that the lower bound on the generalisation benefit decreases at the rate $\mathcal{O}(1/n)$ as the number of samples n grows. Since for the well-specified kernel ridge regression problem, the excess risk upper bound also decreases at rate $\mathcal{O}(1/n)$ (Bach, 2021; Caponnetto & De Vito, 2007), we have that $\mathbb{E}[\Delta(\hat{f}, P\hat{f})] = \Theta(1/n)$.

In a RKHS, $P\hat{f}$ can be rewritten using CMEs as

$$Pf(x_1, x_2) = f(x_1, x_2) - \langle f, \mu_{X|X_2=x_2} \rangle_{\mathcal{H}}. \quad (19)$$

Therefore, introducing a kernel $\ell : \mathcal{X}_2 \times \mathcal{X}_2 \rightarrow \mathbb{R}$, the CME estimate from (6) allows to devise an estimator of $P\hat{f}$ as:

$$\hat{P}\hat{f} = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} (\mathbf{k}_x - \mathbf{K}(\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{\mathbf{x}_2}) \quad (20)$$

where $\mathbf{L} = \ell(\mathbf{x}_2, \mathbf{x}_2)$, $\ell_{\mathbf{x}_2} = \ell(\mathbf{x}_2, \cdot)$ and $\gamma > 0$.

4.4. Respecting the collider structure in a RKHS

Similarly to the $L^2(X)$ case, the solution to the empirical risk minimisation problem in $\mathcal{H}$ will also decompose as $\hat{f} = P\hat{f} + E\hat{f}$. Thus, we can proceed similarly by simply discarding $E\hat{f}$ to improve performance. However, it turns out that using elegant functional properties of RKHSs, it is possible to take a step further and directly take $\mathcal{F} = \text{Range}(P)$. In doing so, we can ensure that our hypothesis space only contains functions that satisfy constraint (ZCE).

Under assumptions detailed in Appendix C, we can view the projection P as a well-defined RKHS projection⁵ $P : \mathcal{H} \rightarrow \mathcal{H}$. In particular, an important assumption is that the kernel takes the form

$$k(x, x') = (r(x_1, x'_1) + 1) \ell(x_2, x'_2), \quad (21)$$

where $r : \mathcal{X}_1 \times \mathcal{X}_1 \rightarrow \mathbb{R}$ and $\ell : \mathcal{X}_2 \times \mathcal{X}_2 \rightarrow \mathbb{R}$ are also positive semi-definite kernels. This ensures that $\mathcal{H}$ contains functions that are constant with respect to x_1 . Thus, the conditional expectation mapping $(x_1, x_2) \mapsto \mathbb{E}[f(X_1, X_2)|X_2 = x_2]$ belongs to the same RKHS.

If these assumptions are met, we denote $\mathcal{H}_P = \text{Range}(P)$. The following result characterises $\mathcal{H}_P$ as a RKHS.

Proposition 4.3. Let P^* be the adjoint operator of P in $\mathcal{H}$. Then $\mathcal{H}_P$ is also a RKHS with reproducing kernel

$$k_P(x, x') = \langle P^*k_x, P^*k_{x'} \rangle_{\mathcal{H}} \quad (22)$$

with $P^*k_x = k_x - \mu_{X|X_2=\pi_2(x)}$.

Using the projected RKHS kernel k_P , it becomes possible to solve the least-square regression problem directly inside $\mathcal{F} = \mathcal{H}_P$. By taking $\Omega(f) = \|f\|_{\mathcal{H}_P}^2$, the empirical risk minimisation problem becomes a standard kernel ridge regression problem in $\mathcal{H}_P$ which admits closed-form solution

$$\hat{f}_P = \mathbf{y}^\top (\mathbf{K}_P + \lambda \mathbf{I}_n)^{-1} \mathbf{k}_{P,\mathbf{x}}, \quad (23)$$

where $\mathbf{K}_P = k_P(\mathbf{x}, \mathbf{x})$ and $\mathbf{k}_{P,\mathbf{x}} = k_P(\mathbf{x}, \cdot)$.

From a learning theory perspective, performing empirical risk minimisation inside $\mathcal{H}_P$ should provide tighter bounds on the generalisation error than on the entire space $\mathcal{H}$. This is because since $\mathcal{H}_P \subset \mathcal{H}$, the Rademacher complexity of $\mathcal{H}_P$ is smaller than that of $\mathcal{H}$.

It should be noted that k_P depends on the CME $\mu_{X|X_2=\pi_2(x)}$, which needs to be estimated. Therefore, in practice, our hypothesis will not lie in the true $\mathcal{H}_P$ but in an approximation of $\mathcal{H}_P$ and the approximation error will depend directly on the CME estimation error.

Algorithm 2 RKHS procedure to estimate $\hat{f}_P$

- 1: Let $\hat{P}^* k_x = k_x - \hat{\mu}_{X|X_2=\pi_2(x)}$
 - 2: Let $\hat{k}_P(x, x') = \langle \hat{P}^* k_x, \hat{P}^* k_{x'} \rangle_{\mathcal{H}}$
 - 3: Evaluate $\hat{\mathbf{K}}_P = \hat{k}_P(\mathbf{x}, \mathbf{x})$ and $\hat{\mathbf{k}}_{P,\mathbf{x}} = \hat{k}_P(\mathbf{x}, \cdot)$
 - 4: Take $\hat{f}_P = \mathbf{y}^\top (\hat{\mathbf{K}}_P + \lambda \mathbf{I}_n)^{-1} \hat{\mathbf{k}}_{P,\mathbf{x}}$
-

The estimation of (23) is again a two-stage procedure outlined in Algorithm 2. The distinction with the general $L^2(X)$ case is that we do not estimate the conditional expectation of any specific function. Instead, we estimate the conditional expectation operator through $\hat{\mu}_{X|X_2=\pi_2(x)}$, and then use it through $\hat{P}^*$ to constrain the hypothesis space. This is possible because in a RKHS, the estimation of the conditional expectation operator can be achieved independently from the function it is applied to. Due to the assumption on the kernel introduced in equation 21 there are now alternative estimators for $\hat{\mu}_{X|X_2=\pi_2(x)}$ which we provide details of in Appendix D.

The estimation of P^* in line 1 only requires observations from X_1, X_2 . Thus, like in the $L^2(X)$ case, additional observations $\mathcal{D}' = \{\mathbf{x}'_1, \mathbf{x}'_2\}$ can help better estimate CMEs, and thus better approximate the projected RKHS $\mathcal{H}_P$.

5. Collider Regression on a more general DAG

We now return to a general Markov boundary. Any Markov boundary may be partitioned following Figure 5, where X_1 contains all direct children of Y , X_3 contains all parents of Y and all other variables are grouped in X_2 . Furthermore, we assume that there exists no edge from a variable in X_1 to a variable in X_2 .

This provides us with the probabilistic information that $Y \perp\!\!\!\perp X_2 \mid X_3$ but $Y \not\perp\!\!\!\perp X_2 \mid X_3, X_1$, which implies in expectation that $\mathbb{E}[Y|X_3] = \mathbb{E}[Y|X_2, X_3]$.

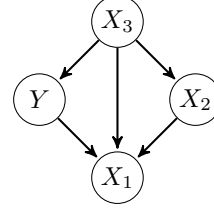


Figure 5. General Markov boundary collider structure.

If we now denote $X = (X_1, X_2, X_3)$ and $f_0(x) = \mathbb{E}[Y|X_3 = x_3]$, then the optimal least-square regressor $f^*(x) = \mathbb{E}[Y|X = x]$ satisfies

$$\begin{aligned} & \mathbb{E}[f^*(X) - f_0(X) \mid X_2, X_3] \\ &= \mathbb{E}[\mathbb{E}[Y|X] \mid X_2, X_3] - \mathbb{E}[\mathbb{E}[Y|X_3] \mid X_2, X_3] \\ &= \mathbb{E}[Y|X_2, X_3] - \mathbb{E}[\mathbb{E}[Y|X_2, X_3] \mid X_2, X_3] \\ &= 0. \end{aligned} \quad (24)$$

Therefore, if we center our hypothesis space on f_0 , then like in Section 4.1, we want our centered estimate $\hat{f} - f_0$ to lie within the following subspace:

$$\hat{f} - f_0 \in \{f \in \mathcal{F} \mid \mathbb{E}[f(X) \mid X_2, X_3] = 0\}. \quad (25)$$

When $\mathcal{F} = L^2(X)$, this space can again be seen as the range of an orthogonal projection, this time defined by

$$P' = \text{Id} - E' \quad (26)$$

where $E' : L^2(X) \rightarrow L^2(X)$ denotes the conditional expectation functional with respect to (X_2, X_3)

$$E' f(x_2, x_3) = \mathbb{E}[f(X) \mid X_2 = x_2, X_3 = x_3]. \quad (27)$$

While we focus in Section 4 on the simple collider structure for the sake of exposition, our result are stated for a general projection operator and still hold for P' — modulo a shift by f_0 . Hence, we can still apply the techniques we have presented to encode probabilistic information from the general DAG in Figure 5 into a regression problem, with similar guarantees on the generalisation benefits.

Proposition 5.1. *Let $h \in L^2(X)$ be any regressor from our hypothesis space. We have*

$$\Delta(h, f_0 + P'h) = \|E'h - f_0\|_{L^2(X)}^2. \quad (28)$$

This means that, for any given regressor $\hat{f}$, we can always improve its test performance by first projecting it onto $\text{Range}(P)$, and then shifting it by f_0 .

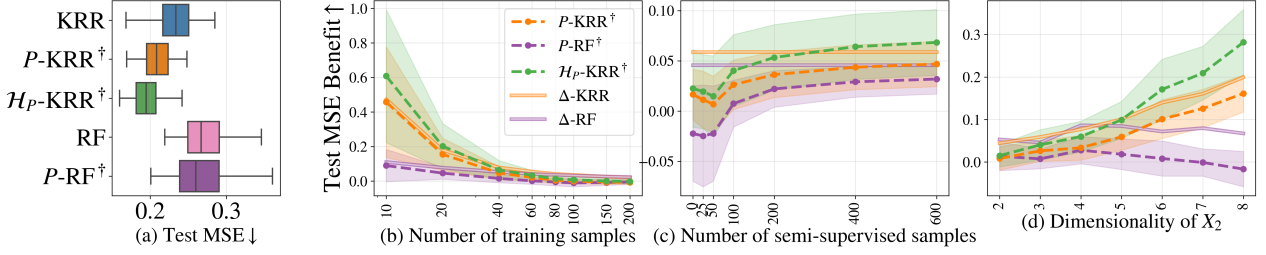


Figure 4. (a) : Test MSEs for the simulation experiment ; dataset is generated using $d_1 = 3, d_2 = 3, n = 50$ and 100 semi-supervised samples ; experiments is run for 100 datasets generated with different seeds ; statistical significance is confirmed in Appendix F ; (b, c, d) : Ablation study on the number of training samples, number of semi-supervised samples and dimensionality of X_2 ; experiments are run for 40 datasets generated with different seeds ; $\uparrow/\downarrow$ indicates higher/lower is better ; we report 1 s.d. ; $\dagger$ indicates our proposed methods.

In practice, the estimation strategies introduced in Section 4 can still be applied to obtain an estimate of $P'\hat{f}$. An additional procedure to estimate f_0 will however be needed. This can be achieved by regressing Y onto X_3 . We provide corresponding algorithms and estimators in Appendix E.

6. Experiments

This section provides empirical evidence that incorporating probabilistic causal knowledge into a regression problem benefits performance. First, we demonstrate our method on an illustrative simulation example. We conduct an ablation study on the number of training samples, the dimensionality of X_2 and the use of additional semi-supervised samples. Then, we address a challenging climate science problem that respects the collider structure. Our results underline the benefit of enforcing constraint (ZCE) onto the hypothesis. Code and data are made available⁶.

Models We compare five models:

1. *RF*: A baseline random forest model.
2. *P-RF*: The baseline RF model projected following Algorithm 1 and using a linear regression to estimate $\mathbb{E}[\hat{f}(X_1, X_2)|X_2 = x_2]$.
3. *KRR*: A baseline kernel ridge regression.
4. *P-KRR*: The KRR model projected following (20).
5. *H_P-KRR*: A kernel ridge regression model fitted directly in the projected RKHS following Algorithm 2.

For both KRR and RF, we use Proposition 4.1 to compute Monte Carlo estimates of the expected generalisation gap $\mathbb{E}[\Delta(\hat{f}, P\hat{f})]$, which we denote as Δ -KRR and Δ -RF respectively. This provides an indicator of the greatest achievable generalisation gain if we had access to the exact projection P . Hyperparameters are tuned using a cross-validated grid search and model details are specified in Appendix F.

⁶<https://github.com/shahineb/collider-regression>.

6.1. Simulation example

Data generating process We propose the following construction that follows the simple collider structure from Figure 3. Let $d_1, d_2 \geq 1$ denote respectively the dimensionalities of X_1 and X_2 . We first generate a fixed positive definite matrix Σ of size $(d_1 + d_2 + 1)$ which has zero off-diagonals on the $(d_1 + d_2)^{\text{th}}$ row and column. We then follow the generating process described in Algorithm 3 and generate a dataset of n observations $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}\}$. The zero off-diagonal terms in Σ ensure that we satisfy $Y \perp\!\!\!\perp X_2$ and g_1, g_2 are nontrivial mappings that introduce a non-linear dependence (details in Appendix F).

Algorithm 3 Data generating process simulation example

- 1: **Input:** $\Sigma \succcurlyeq 0, \sigma > 0, g_1: \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_1}, g_2: \mathbb{R}^{d_2} \rightarrow \mathbb{R}^{d_2}$
 - 2: $[X_1 \ X_2 \ Y]^\top \sim \mathcal{N}(0, \Sigma), \ \varepsilon \sim \mathcal{N}(0, \sigma^2)$
 - 3: $X_1 \leftarrow g_1(X_1) + \varepsilon$
 - 4: $X_2 \leftarrow g_2(X_2)$
 - 5: **return** X_1, X_2, Y
-

Results Figure 4(a) provides empirical evidence that, for both KRR and RF, incorporating probabilistic inductive biases from the collider structure in the hypothesis benefits the generalisation error.

In addition, Figure 4(b)(c)(d) shows that the empirical generalisation benefit is greatest when : fewer training samples are available, semi-supervised samples can be easily obtained and the dimensionality of X_2 is larger. This is in keeping with Theorem 4.2 which predicts the benefit will be larger when we have fewer labeled samples and a more complicated relationship between X_2 and X_1 .

Because the decision nodes learnt by RF largely rely on X_1 and the early dimensions of X_2 , increasing the dimensionality of X_2 has little to negative effect as shown in Figure 4(d).

6.2. Aerosols radiative forcing

Background The radiative forcing is defined as the difference between incoming and outgoing flux of energy in the Earth system. At equilibrium, the radiative forcing should be of 0 W m^{-2} . Carbon dioxide emissions from human activity contribute a positive radiative forcing of $+1.89 \text{ W m}^{-2}$ which causes warming of the Earth (Bellouin et al., 2020).

Aerosols are microscopic particles suspended in the atmosphere (e.g. dust, sea salt, black carbon) that contribute a negative radiative forcing by helping reflect solar radiation, which cools the Earth. However, the magnitude of their forcing represents the largest uncertainty in assessments of global warming, with uncertainty bounds that could offset global warming or double its effects. It is thus critical to obtain better estimate of the aerosol radiative forcing.

The carbon dioxide and aerosol forcings are independent factors⁷ that contribute to the observed global temperatures. Hence, by setting $Y = \text{“aerosol forcing”}$, $X_2 = \text{“CO}_2 \text{ forcing”}$ and $X_1 = \text{“global temperature”}$, this problem has a collider structure and observations from global temperature and CO_2 forcing can be used to regress the aerosol forcing.

Data generating process FaIR (for Finite amplitude Impulse Response) is a deterministic model that proposes a simplified low-order representation of the climate system (Millar et al., 2017; Smith et al., 2018). Surrogate climate models like FaIR — referred to as *emulators* — have been widely used, notably in reports of the Intergovernmental Panel on Climate Change (Masson-Delmotte et al., 2021), because they are fast and inexpensive to compute.

We use a modified version of FaIRv2.0.0 (Leach et al., 2021) where we introduce variability by adding white noise on the forcing to account for climate internal variability (Hasselmann, 1976; Cummins et al., 2020). To generate a sample, we run the emulator over historical greenhouse gas and aerosol emission data and retain scalar values for $y = \text{“aerosol forcing in 2020”}$, $x_2 = \text{“CO}_2 \text{ forcing in 2020”}$ and $x_1 = \text{“global temperature anomaly in 2020”}$. We perform this n times to generate dataset $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, y\}$.

Results Results are reported in Table 1. We observe that the incorporation of inductive bias from the collider resulted in consistently improved performance for both RF and KRRs. This shows that while the proposed methodology is only formulated in terms of squared error, it can also improve performance for other metrics.

⁷this is because whilst human activity can confound CO_2 and aerosol emissions, the timescale on which CO_2 and aerosol forcing operate (century vs week) are so different that the forcings at a given time can be considered independent.

Table 1. MSE, signal-to-noise ratio (SNR) and correlation on test data for the aerosol radiative forcing experiment ; $n = 50$ and 200 semi-supervised samples ; statistical significance is confirmed in Appendix F ; experiments is run for 100 datasets generated with different seeds ; $\uparrow/\downarrow$ indicates higher/lower is better ; we report 1 standard deviation ; $\dagger$ indicates our proposed methods.

	MSE $\downarrow$	SNR $\uparrow$	Correlation $\uparrow$
RF	0.90 ± 0.04	0.44 ± 0.19	0.32 ± 0.08
$P\text{-RF}^\dagger$	0.89 ± 0.03	0.49 ± 0.15	0.34 ± 0.07
KRR	0.88 ± 0.04	0.58 ± 0.17	0.37 ± 0.05
$P\text{-KRR}^\dagger$	0.86 ± 0.03	0.65 ± 0.13	0.40 ± 0.01
$\mathcal{H}_P\text{-KRR}^\dagger$	0.86 ± 0.03	0.65 ± 0.14	0.40 ± 0.01

7. Discussion and Related Work

Regression and Causal Inference Currently causal inference is most commonly used in regression problems when reasoning about invariance (Peters et al., 2016; Arjovsky et al., 2019). These methods aim to use the causal structure to guarantee the predictors will transfer to new environments (Gulrajani & Lopez-Paz, 2020) and recent work discusses how causal structure plays a role in the effectiveness of these methods (Wang & Veitch, 2022). Our work takes a complimentary route in asking how causal structure can benefit in regression, and, in contrast to prior work, focuses on a fixed environment.

Causal and Anti-causal learning Our work is closely related to work on anti-causal learning (Schölkopf et al., 2012) which argues that $\mathbb{P}(X)$ will only provide additional information about $\mathbb{P}(Y|X)$ if we are working in an anti-causal prediction problem $Y \rightarrow X$. This leads the authors to hypothesise that additional unlabelled semi-supervised samples will be most helpful in the anti-causal direction. In our work, we go further and prove a concrete generalisation benefit from using additional samples from $\mathbb{P}(X)$ when the data generating process follows a collider, a graphical structure which is inherently anti-causal as it relies on Y having shared children with another vertex.

Independence Regularisation and Fair Learning Our work is related to the large body of recent work aiming to force conditional independence constraints, either for fairness (Kamishima et al., 2011) or domain generalisation (Pogodin et al., 2022). However, it is important to note that if Y satisfies a conditional independence this does not mean that the optimal least-square regressor $\mathbb{E}[Y|X]$ will satisfy the same conditional independence. For example, let

$$\begin{cases} Y, X_2 \sim \mathcal{N}(0, 1) \text{ with } Y \perp\!\!\!\perp X_2 \\ X_1 = Y \mathbb{1}\{X_2 > 0\}. \end{cases} \quad (29)$$

Then we have $\mathbb{E}[Y|X_1, X_2] = X_1 \mathbb{1}\{X_2 > 0\}$, hence $\mathbb{E}[Y|X_1, X_2]$ is constant when $X_2 < 0$ but not otherwise. Therefore $\mathbb{E}[Y|X_1, X_2] \not\perp\!\!\!\perp X_2$, even though $Y \perp\!\!\!\perp X_2$.

Therefore, our methodology is more similar to ensuring independence in expectation. Specifically, the RKHS methodology is related to work on fair kernel learning (Pérez-Suay et al., 2017; Li et al., 2022b). However, in contrast to the work on fair kernel learning where regularisation terms for encouraging independence are proposed, we go further by enforcing the mean independence constraint directly onto the hypothesis space.

Availability of DAG as prior knowledge Our work is based on the premise of having exact knowledge of the DAG underlying the data generating structure. This knowledge typically comes from domain expertise, with examples in genetics (Day et al., 2016) or in the aerosol radiative forcing experiment we present. However, when domain expertise is insufficient, we may need causal discovery methods to uncover the causal relationships. These methods can be expensive to run at large scale and can provide a DAG with missing or extra edges when compared to the true DAG. If collider regression is run with a partially incorrect DAG, it is likely that it would degrade the performance, as such a setting would amount to introducing incorrect prior information in the model. However, if the estimated DAG is “close” to the true DAG in the sense of the independence relationships they induce, then there may still be benefit in the finite sample regime.

Generality of proposed method Two aspects of the methodology introduced in Section 5 need to be caveated. First, it is important to require there exists no edge from children of Y to spouses of Y , otherwise that would break the conditional independence $Y \perp\!\!\!\perp X_2|X_3$. Second, whilst this is a general procedure that provides a useful inductive bias and helps restrict the hypothesis class, this procedure may not account for all the possible inductive biases that arise from the DAG at its most granular level. The procedure accounts for the collider constraint that arises from grouping variables together, not for every collider structure that might exist in the DAG. Encoding more granular collider structure would require additional regression steps, and a systematic way to perform such additional steps remains an interesting avenue for further research.

8. Conclusion

In this work we have demonstrated that collider structures within causal graphs constitute a useful form of inductive bias for regression that benefits generalisation performance. Whilst we focused on least-square regression, we expect that the collider regression framework should benefit a wider range of machine learning problems that aim to make inferences about $\mathbb{P}(Y|X)$. For example, a natural extension of this work should investigate collider regression for classification or quantile regression tasks.

Acknowledgements

The authors would like to thank Bryn Elesedy, Dimitri Meunier, Siu Lun Chau, Jean-François Ton, Christopher Williams, Duncan Watson-Parris, Eugenio Clerico⁸ and Arthur Gretton for many helpful discussions and valuable feedbacks. Shahine Bouabid receives funding from the European Union’s Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 860100. Jake Fawkes receives funding from the EPSRC.

References

- Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.
- Aronszajn, N. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.
- Bach, F. Learning theory from first principles. *Draft of a book, version of Sept*, 6:2021, 2021.
- Bellouin, N., Davies, W., Shine, K. P., Quaas, J., Mülmenstädt, J., Forster, P. M., Smith, C., Lee, L., Regayre, L., Brasseur, G., Sudarchikova, N., Bouarar, I., Boucher, O., and Myhre, G. Radiative forcing of climate change from the copernicus reanalysis of atmospheric composition. *Earth System Science Data*, 12(3):1649–1677, 2020. doi: 10.5194/essd-12-1649-2020. URL <https://essd.copernicus.org/articles/12/1649/2020/>.
- Berlinet, A. and Thomas-Agnan, C. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.
- Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

⁸and Tyler Farghly

- Chau, S. L., Bouabid, S., and Sejdinovic, D. Deconditional downscaling with gaussian processes. *Advances in Neural Information Processing Systems*, 34:17813–17825, 2021.
- Cummins, D. P., Stephenson, D. B., and Stott, P. A. Optimal estimation of stochastic energy balance model parameters. *Journal of Climate*, 2020.
- Dash, M. and Liu, H. Feature selection for classification. *Intelligent data analysis*, 1(1-4):131–156, 1997.
- Day, F. R., Loh, P.-R., Scott, R. A., Ong, K. K., and Perry, J. R. A robust example of collider bias in a genetic association study. *The American Journal of Human Genetics*, 98(2):392–393, 2016.
- Elesedy, B. Provably strict generalisation benefit for invariance in kernel methods. *Advances in Neural Information Processing Systems*, 34:17273–17283, 2021.
- Fawkes, J., Hu, R., Evans, R. J., and Sejdinovic, D. Doubly robust kernel statistics for testing distributional treatment effects even under one sided overlap. *arXiv preprint arXiv:2212.04922*, 2022.
- Fukumizu, K., Bach, F. R., and Jordan, M. I. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 2004.
- Fukumizu, K., Song, L., and Gretton, A. Kernel bayes’ rule: Bayesian inference with positive definite kernels. *The Journal of Machine Learning Research*, 14(1):3753–3783, 2013.
- Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. Gpytorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems*, 2018.
- Grünewälder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., and Pontil, M. Conditional Mean Embeddings as Regressors. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.
- Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.
- Hasselmann, K. Stochastic climate models part i. theory. *Tellus*, 1976.
- Hsu, K. and Ramos, F. Bayesian deconditional kernel mean embeddings. In *International Conference on Machine Learning*, pp. 2830–2838. PMLR, 2019.
- Kamishima, T., Akaho, S., and Sakuma, J. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 643–650. IEEE, 2011.
- Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.
- Klebanov, I., Schuster, I., and Sullivan, T. J. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606, 2020.
- Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- Leach, N. J., Jenkins, S., Nicholls, Z., Smith, C. J., Lynch, J., Cain, M., Walsh, T., Wu, B., Tsutsui, J., and Allen, M. R. Fairv2.0.0: a generalized impulse response model for climate uncertainty and future scenario exploration. *Geoscientific Model Development*, 2021.
- Li, Z., Meunier, D., Mollenhauer, M., and Gretton, A. Optimal rates for regularized conditional mean embedding learning. *arXiv preprint arXiv:2208.01711*, 2022a.
- Li, Z., Perez-Suay, A., Camps-Valls, G., and Sejdinovic, D. Kernel dependence regularizers and gaussian processes with applications to algorithmic fairness. *Pattern Recognition*, pp. 108922, 2022b.
- Lun Chau, S., Hu, R., Gonzalez, J., and Sejdinovic, D. Rkhs-shap: Shapley values for kernel methods. *Advances in neural information processing systems*, 36, 2022.
- Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., et al. Climate change 2021: the physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*, 2, 2021.
- Meek, C. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 411–418, 1995.
- Millar, R. J., Nicholls, Z. R., Friedlingstein, P., and Allen, M. R. A modified impulse-response representation of the global near-surface air temperature and atmospheric concentration response to carbon dioxide emissions. *Atmospheric Chemistry and Physics*, 2017.
- Mollenhauer, M. and Koltai, P. Nonparametric approximation of conditional expectation operators. *arXiv preprint arXiv:2012.12917*, 2020.

- Mori, T. Comments on "a matrix inequality associated with bounds on solutions of algebraic riccati and lyapunov equation" by jm saniuk and ib rhodes. *IEEE transactions on automatic control*, 33(11):1088, 1988.
- Muandet, K., Sriperumbudur, B., Fukumizu, K., Gretton, A., and Schölkopf, B. Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 17, 2016.
- Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.
- Park, J. and Muandet, K. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in neural information processing systems*, 33:21247–21259, 2020.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems* 32. 2019.
- Paulsen, V. I. and Raghupathi, M. *An introduction to the theory of reproducing kernel Hilbert spaces*, volume 152. Cambridge university press, 2016.
- Pearl, J. Evidential reasoning using stochastic simulation of causal models. *Artificial intelligence*, 32(2):245–257, 1987.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.
- Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., and Camps-Valls, G. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 339–355. Springer, 2017.
- Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.
- Pogodin, R., Deka, N., Li, Y., Sutherland, D. J., Veitch, V., and Gretton, A. Efficient conditionally invariant representation learning. *arXiv preprint arXiv:2212.08645*, 2022.
- Rasmussen, C. and Williams, C. *Gaussian Processes for Machine Learning*, 2005.
- Särkkä, S. Linear operators and stochastic partial differential equations in gaussian process regression. In *International Conference on Artificial Neural Networks*, pp. 151–158. Springer, 2011.
- Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. M. On causal and anticausal learning. In *ICML*, 2012.
- Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.
- Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.
- Smith, C. J., Forster, P. M., Allen, M., Leach, N., Millar, R. J., Passerello, G. A., and Regayre, L. A. Fair v1.3: a simple emissions-based impulse response and carbon cycle model. *Geoscientific Model Development*, 2018.
- Song, L., Huang, J., Smola, A., and Fukumizu, K. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.
- Song, L., Gretton, A., Bickson, D., Low, Y., and Guestrin, C. Kernel belief propagation. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 707–715. JMLR Workshop and Conference Proceedings, 2011.
- Song, L., Fukumizu, K., and Gretton, A. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 2013.
- Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12 (7), 2011.
- Statnikov, A., Lytkin, N. I., Lemeire, J., and Aliferis, C. F. Algorithms for discovery of multiple markov boundaries. *Journal of machine learning research: JMLR*, 14:499, 2013.
- Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

- Szabó, Z. and Sriperumbudur, B. K. Characteristic and universal tensor product kernels. *J. Mach. Learn. Res.*, 18:233–1, 2017.
- Ton, J.-F., Lucian, C., Teh, Y. W., and Sejdinovic, D. Noise contrastive meta-learning for conditional density estimation using kernel mean embeddings. In *International Conference on Artificial Intelligence and Statistics*, pp. 1099–1107. PMLR, 2021.
- Wang, Z. and Veitch, V. A unified causal view of domain invariant representation learning. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022. URL <https://openreview.net/forum?id=-l9cpeEYwJJ>.
- Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

A. Notations and useful Results

A.1. Notations

Let $\mathcal{X}$ be a Borel space, $\mathcal{Y} \subseteq \mathbb{R}$ and let X and Y be random variables valued in $\mathcal{X}$ and $\mathcal{Y}$. We denote $(L^2(X), \langle \cdot, \cdot \rangle_{L^2(X)})$ the Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$ which are square-integrable with respect to the pushforward measure induced by X , i.e. $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$.

Let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be a RKHS of functions from $\mathcal{X}$ to $\mathbb{R}$ with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. We denote its canonical feature map as k_x for any $x \in \mathcal{X}$.

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, we denote $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ the smallest and largest eigenvalues of $\mathbf{A}$ respectively.

A.2. Useful results

Theorem A.1 (Theorem 3.11, (Paulsen & Raghupathi, 2016)). *Let $\mathcal{H}$ be a RKHS on $\mathcal{X}$ with reproducing kernel k and let $f : \mathcal{X} \rightarrow \mathbb{R}$. Then the following are equivalent:*

- (i) $f \in \mathcal{H}$
- (ii) *there exists $c \geq 0$ such that $c^2 k(x, y) - f(x)f(y)$ is kernel function*

Lemma A.2 (Corollary 5.5, (Paulsen & Raghupathi, 2016)). *Let $\mathcal{H}_1, \mathcal{H}_2$ be RKHS on $\mathcal{X}$ with reproducing kernels k_1, k_2 . If $\mathcal{H}_1 \cap \mathcal{H}_2 = \{0\}$, then $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ is a RKHS with reproducing kernel $k = k_1 + k_2$ and $\mathcal{H}_1, \mathcal{H}_2$ are orthogonal subspaces of $\mathcal{H}$.*

Proposition A.3. *Let $(\mathcal{V}, \langle \cdot, \cdot \rangle_{\mathcal{V}})$ be a Hilbert space, $\varphi : \mathcal{X} \rightarrow \mathcal{V}$ be a mapping function and*

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{V}}, \quad x, y \in \mathcal{X} \quad (30)$$

the kernel function induced by φ . Then the RKHS induced by k is given by

$$\mathcal{H} = \{x \mapsto \langle v, \varphi(x) \rangle_{\mathcal{V}} \mid v \in \mathcal{V}\}. \quad (31)$$

Proof. The proof follows from the application of the Pull-back Theorem [Theorem 5.7](Paulsen & Raghupathi, 2016) to the linear kernel $L : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}, (v, v') \mapsto \langle v, v' \rangle_{\mathcal{V}}$ composed with the feature map $\varphi : \mathcal{X} \rightarrow \mathcal{V}$. $\square$

Lemma A.4. *Suppose $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric, let Z be a random variable, $\mathbf{x} \in \mathbb{R}^n$ be a random vector.*

$$\mathbb{E}[\mathbf{x}^{\top} \mathbf{A} \mathbf{x} \mid Z] = \text{Tr}(\mathbf{A} \text{Var}(\mathbf{x} \mid Z)) + \mathbb{E}[\mathbf{x} \mid Z]^{\top} \mathbf{A} \mathbb{E}[\mathbf{x} \mid Z]. \quad (32)$$

Proof.

$$\mathbb{E}[\mathbf{x}^{\top} \mathbf{A} \mathbf{x} \mid Z] = \mathbb{E}[\text{Tr}(\mathbf{A} \mathbf{x} \mathbf{x}^{\top}) \mid Z] \quad (33)$$

$$= \text{Tr}(\mathbf{A} \mathbb{E}[\mathbf{x} \mathbf{x}^{\top} \mid Z]) \quad (34)$$

$$= \text{Tr}(\mathbf{A} (\text{Var}(\mathbf{x} \mid Z) + \mathbb{E}[\mathbf{x} \mid Z] \mathbb{E}[\mathbf{x} \mid Z]^{\top})) \quad (35)$$

$$= \text{Tr}(\mathbf{A} \text{Var}(\mathbf{x} \mid Z)) + \text{Tr}(\mathbf{A} \mathbb{E}[\mathbf{x} \mid Z] \mathbb{E}[\mathbf{x} \mid Z]^{\top}) \quad (36)$$

$$= \text{Tr}(\mathbf{A} \text{Var}(\mathbf{x} \mid Z)) + \mathbb{E}[\mathbf{x} \mid Z]^{\top} \mathbf{A} \mathbb{E}[\mathbf{x} \mid Z]. \quad (37)$$

$\square$

Lemma A.5 ((Mori, 1988)). *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ and suppose $\mathbf{A}$ symmetric and $\mathbf{B}$ positive semi-definite, then*

$$\text{Tr}(\mathbf{A} \mathbf{B}) \geq \lambda_{\min}(\mathbf{A}) \text{Tr}(\mathbf{B}). \quad (38)$$

Lemma A.6 (Lemma B.3, (Elesedy, 2021)). *Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, then*

$$\lambda_{\max}(\mathbf{A}) \leq n \max_{i,j} |\mathbf{A}_{ij}|. \quad (39)$$

B. Supporting proofs

B.1. Notations

We start by introducing measure-theoretic notations which will be of use in the supporting proofs.

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ denote a probability space, we denote $L^2(\Omega, \mathfrak{F}, \mathbb{P})$ the space of random variables with finite variance, which we will denote $L^2(\Omega)$ for conciseness when the σ -algebra is $\mathfrak{F}$. Endowed with inner product $\langle Z, Z' \rangle_{L^2(\Omega)} = \mathbb{E}[ZZ']$, $L^2(\Omega)$ has a Hilbert structure. For any random variable Z , we denote $\sigma(Z) \subset \mathfrak{F}$ the σ -algebra generated by Z .

B.2. Proofs of Proposition 4.1

Proposition 4.1. *Let $h \in L^2(X)$ be any regressor from our hypothesis space. We have*

$$\Delta(h, Ph) = \|Eh\|_{L^2(X)}^2. \quad (40)$$

Proof. The conditional expectation $\Pi : Z \in L^2(\Omega) \mapsto \mathbb{E}[Z|X_2]$ defines an orthogonal projection onto the space of X_2 -measurable random variables with finite variance $L^2(\Omega, \sigma(X_2), \mathbb{P})$. Thus, its range and null space are orthogonal in $L^2(\Omega)$.

Let $h \in L^2(X)$. We have $Eh(X) = \mathbb{E}[h(X)|X_2] = \Pi h(X)$ hence $Eh(X)$ is in the range of Π . On the other hand,

$$\mathbb{E}[Ph(X)|X_2] = \mathbb{E}[h(X)|X_2] - \mathbb{E}[Eh(X)|X_2] = \mathbb{E}[h(X)|X_2] - \mathbb{E}[h(X)|X_2] = 0, \quad (41)$$

therefore $Ph(X)$ is in the null space of Π . Finally, because $Y \perp\!\!\!\perp X_2$ we have $\mathbb{E}[Y|X_2] = \mathbb{E}[Y] = 0$ by assumption, therefore Y is also in the null space of Π .

Hence, adopting this random variable view, the desired result simply follows from $L^2(\Omega)$ orthogonality:

$$\begin{aligned} \Delta(h, Ph) &= \mathbb{E}[(Y - h(X))^2] - \mathbb{E}[(Y - Ph(X))^2] \\ &= \|Y - h(X)\|_{L^2(\Omega)}^2 - \|Y - Ph(X)\|_{L^2(\Omega)}^2 \\ &= \|Y - Ph(X) - Eh(X)\|_{L^2(\Omega)}^2 - \|Y - Ph(X)\|_{L^2(\Omega)}^2 \\ &= \|Y - Ph(X)\|_{L^2(\Omega)}^2 + \|Eh(X)\|_{L^2(\Omega)}^2 - \|Y - Ph(X)\|_{L^2(\Omega)}^2 \\ &= \mathbb{E}[Eh(X)^2] \\ &= \|Eh\|_{L^2(X)}^2. \end{aligned}$$

□

B.3. Proofs of Proposition 4.3

Proposition 4.3. *Let P^* be the adjoint operator of P in $\mathcal{H}$. Then $\mathcal{H}_P$ is also a RKHS with reproducing kernel*

$$k_P(x, x') = \langle P^*k_x, P^*k_{x'} \rangle_{\mathcal{H}} \quad (42)$$

with $P^*k_x = k_x - \mu_{X|X_2=\pi_2(x)}$.

Proof of Proposition 4.3. Let $\mathcal{H}_P$ denote the reproducing kernel with k_P . We start by showing that $P\mathcal{H} \subseteq \mathcal{H}_P$.

Let $f \in P\mathcal{H}$, then it admits a pre-image $w_f \in \mathcal{H}$ such that $f = Pw_f$. Hence for any $x \in \mathcal{X}$, we get that

$$f(x) = \langle f, k_x \rangle_{\mathcal{H}} = \langle Pw_f, k_x \rangle_{\mathcal{H}} = \langle w_f, P^*k_x \rangle_{\mathcal{H}}. \quad (43)$$

Hence, f writes as an element of the RKHS induced by the feature map $x \mapsto P^*k_x$ and by Proposition A.3 $f \in \mathcal{H}_P$.

Reciprocally, let us now show that $\mathcal{H}_P \subseteq P\mathcal{H}$. Let $f \in \mathcal{H}_P$, again by Proposition A.3 there exists $w_f \in \mathcal{H}$ such that for any $x \in \mathcal{X}$,

$$f(x) = \langle w_f, P^*k_x \rangle_{\mathcal{H}} = Pw_f(x). \quad (44)$$

This proves that $f \in P\mathcal{H}$ which concludes the proof. □

B.4. Proofs of Theorem 4.2

Theorem 4.2. Suppose $M = \sup_{x \in \mathcal{X}} k(x, x) < \infty$ and $\text{Var}(Y|X) \geq \eta > 0$. Then, the generalisation gap between $\hat{f}$ and $P\hat{f}$ satisfies

$$\mathbb{E}[\Delta(\hat{f}, P\hat{f})] \geq \frac{\eta \mathbb{E}[\|\mu_{X|X_2}(X)\|_{L^2(X)}^2]}{(\sqrt{n}M + \lambda/\sqrt{n})^2} \quad (45)$$

where $\mu_{X|X_2} = \mathbb{E}[k_X|X_2]$ is the CME of $\mathbb{P}(X|X_2)$.

Proof of Theorem 4.2. Let $\Pi = \mathbb{E}[\cdot|X_2]$ be the $L^2(\Omega)$ orthogonal projection onto the subspace of X_2 -measurable random variables. For any $h \in L^2(X)$, we verify that $Eh(X) = \mathbb{E}[h(X)|X_2] = \Pi[h(X)]$ hence $Eh(X) \in \text{Range}(\Pi)$. Furthermore, because $Y \perp\!\!\!\perp X_2$ we have $\Pi[Y] = \mathbb{E}[Y|X_2] = \mathbb{E}[Y] = 0$ by assumption, hence $Y \in \text{Ker}(\Pi)$.

Now let

$$\mathbf{x} = \begin{bmatrix} X^{(1)} \\ \vdots \\ X^{(n)} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} Y^{(1)} \\ \vdots \\ Y^{(n)} \end{bmatrix} \quad (46)$$

denote vectors of n independent copies of X and Y and let

$$j(x, x') = \langle Ek_x, Ek_{x'} \rangle_{L^2(X)} = \mathbb{E}[Ek_x(X)Ek_{x'}(X)] \quad \forall x, x' \in \mathcal{X}. \quad (47)$$

be the positive definite kernel induced by $L^2(X)$ inner product of Ek_x and

$$\mathbf{J} = j(\mathbf{x}, \mathbf{x}) = \left[j(X^{(i)}, X^{(j)}) \right]_{1 \leq i, j \leq n} \quad (48)$$

the resulting Gram-matrix.

Using notations from Section 4.3, we know the solution of the kernel ridge regression problem in $\mathcal{H}$ takes the form

$$\hat{f} = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{k}_\mathbf{x}. \quad (49)$$

Hence, by linearity of the projection, we have

$$E\hat{f} = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} E\mathbf{k}_\mathbf{x} \quad (50)$$

with notation abuse $E\mathbf{k}_\mathbf{x} = [Ek_{X^{(1)}} \dots Ek_{X^{(n)}}]^\top$.

Therefore, we can write

$$\Delta(\hat{f}, P\hat{f}) = \|E\hat{f}\|_{L^2(X)}^2 \quad (51)$$

$$= \mathbb{E}_X[E\hat{f}(X)^2] \quad (52)$$

$$= \mathbb{E}_X \left[\left(\mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} E\mathbf{k}_\mathbf{x}(X) \right)^2 \right] \quad (53)$$

$$= \mathbb{E}_X [\mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} E\mathbf{k}_\mathbf{x}(X) E\mathbf{k}_\mathbf{x}(X)^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{y}] \quad (54)$$

$$= \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbb{E}_X [E\mathbf{k}_\mathbf{x}(X) E\mathbf{k}_\mathbf{x}(X)^\top] (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{y} \quad (55)$$

$$= \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{J} (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{y}. \quad (56)$$

Let us now denote for conciseness $\mathbf{A} = (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{J} (\mathbf{K} + \lambda \mathbf{I}_n)^{-1}$. We have by Lemma XX,

$$\mathbb{E}_\mathbf{y}[\Delta(\hat{f}, P\hat{f}) | \mathbf{x}] = \mathbb{E}_\mathbf{y}[\mathbf{y}^\top \mathbf{A} \mathbf{y} | \mathbf{x}] \quad (57)$$

$$= \text{Tr}(\mathbf{A} \text{Var}(\mathbf{y}|\mathbf{x})) + \mathbb{E}[\mathbf{y}|\mathbf{x}]^\top \mathbf{A} \mathbb{E}[\mathbf{y}|\mathbf{x}] \quad \text{Lemma A.4} \quad (58)$$

$$\geq \text{Tr}(\mathbf{A} \text{Var}(\mathbf{y}|\mathbf{x})), \quad (59)$$

where the conditional variance is the diagonal matrix given by

$$\text{Var}(\mathbf{y}|\mathbf{x}) = \begin{bmatrix} \text{Var}(Y^{(1)}|X^{(1)}) & & \\ & \ddots & \\ & & \text{Var}(Y^{(n)}|X^{(n)}) \end{bmatrix} \quad (60)$$

because the copies of (X, Y) are mutually independent.

We therefore obtain,

$$\mathbb{E}_{\mathbf{y}}[\Delta(\hat{f}, P\hat{f}) | \mathbf{x}] \geq \text{Tr}(\mathbf{A} \text{Var}(\mathbf{y}|\mathbf{x})) \quad (61)$$

$$\geq \min_i \text{Var}(Y^{(i)}|X^{(i)}) \text{Tr}(\mathbf{A}) \quad (62)$$

$$\geq \eta \text{Tr}(\mathbf{A}) \quad (63)$$

$$= \eta \text{Tr}((\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{J} (\mathbf{K} + \lambda \mathbf{I}_n)^{-1}) \quad (64)$$

$$\geq \eta \lambda_{\min}((\mathbf{K} + \lambda \mathbf{I}_n)^{-1})^2 \text{Tr}(\mathbf{J}) \quad \text{Lemma A.5} \quad (65)$$

$$\geq \eta \frac{\text{Tr}(\mathbf{J})}{(Mn + \lambda)^2} \quad \text{Lemma A.6.} \quad (66)$$

Finally taking the expectation against $\mathbf{x}$, we get

$$\mathbb{E}[\Delta(\hat{f}, P\hat{f})] \geq \frac{\mathbb{E}_{\mathbf{x}}[\eta \text{Tr}(\mathbf{J})]}{(Mn + \lambda)^2} \quad (67)$$

$$= \frac{\eta \sum_{i=1}^n \mathbb{E}_{X^{(i)}}[j(X^{(i)}, X^{(i)})]}{(Mn + \lambda)^2} \quad (68)$$

$$= \frac{n\eta \mathbb{E}[j(X, X)]}{(Mn + \lambda)^2} \quad (69)$$

$$= \frac{\eta \mathbb{E}[j(X, X)]}{(M\sqrt{n} + \lambda/\sqrt{n})^2} \quad (70)$$

$$= \frac{\eta \mathbb{E}[\|Ek_X\|_{L^2(X)}^2]}{(M\sqrt{n} + \lambda/\sqrt{n})^2} \quad (71)$$

$$(72)$$

Now, for our particular choice of projection E , we have for any $x \in \mathcal{X}$ that

$$Ek_X(x) = \mathbb{E}_{X'}[k_X(X')|X_2 = \pi_2(x)] \quad (73)$$

$$= \mathbb{E}_{X'}[k(X, X')|X_2 = \pi_2(x)] \quad (74)$$

$$= \langle k_X, \mathbb{E}_{X'}[k_{X'}|X_2 = \pi_2(x)] \rangle_{\mathcal{H}} \quad (75)$$

$$= \langle k_X, \mu_{X|X_2=\pi_2(x)} \rangle_{\mathcal{H}} \quad (76)$$

$$= \mu_{X|X_2=\pi_2(x)}(X) \quad (77)$$

Therefore using the measure-theoretical CME notation from (Park & Muandet, 2020), we have

$$\|Ek_X\|_{L^2(X)}^2 = \mathbb{E}_{X'}[Ek_X(X')^2] = \mathbb{E}_{X'}[\mu_{X|X_2=\pi_2(X')}(X)^2] = \|\mu_{X|X_2}(X)\|_{L^2(X)}^2 \quad (78)$$

which concludes the proof. $\square$

B.5. Proof of Proposition 5.1

Proposition 5.1. *Let $h \in L^2(X)$ be any regressor from our hypothesis space. We have*

$$\Delta(h, f_0 + P'h) = \|E'h - f_0\|_{L^2(X)}^2. \quad (79)$$

Proof. This proof follows the same structure than the proof of Proposition 4.1.

Let $\Pi = \mathbb{E}[\cdot | X_2, X_3]$ be the $L^2(\Omega)$ orthogonal projection onto the subspace of (X_2, X_3) -measurable random variables with finite variance $L^2(\Omega, \sigma(X_2, X_3), \mathbb{P})$. We have that

$$\Pi[Y - f_0(X)] = \mathbb{E}[Y | X_2, X_3] - \mathbb{E}[f_0(X) | X_2, X_3] \quad (80)$$

$$= \mathbb{E}[Y | X_2, X_3] - \mathbb{E}[\mathbb{E}[Y | X_3] | X_2, X_3] \quad (81)$$

$$= \mathbb{E}[Y | X_2, X_3] - \mathbb{E}[\mathbb{E}[Y | X_2, X_3] | X_2, X_3] \quad (Y \perp\!\!\!\perp X_2 | X_3) \quad (82)$$

$$= 0, \quad (83)$$

therefore $Y - f_0(X) \in \text{Ker}(\Pi)$. On the other hand, we can easily verify that for any $h \in L^2(X)$, we have $E'h(X) \in \text{Range}(\Pi)$ and $P'h(X) \in \text{Ker}(\Pi)$.

Therefore, it follows by $L^2(\Omega)$ orthogonality that for any $h \in L^2(X)$

$$\begin{aligned} \|Y - h(X)\|_{L^2(\Omega)}^2 &= \|(Y - f_0(X)) - (h(X) - f_0(X))\|_{L^2(\Omega)}^2 \\ &= \|(Y - f_0(X)) - P'(h - f_0)(X)\|_{L^2(\Omega)}^2 + \|E'(h - f_0)(X)\|_{L^2(\Omega)}^2 \\ &= \|Y - (f_0(X) + P'h(X))\|_{L^2(\Omega)}^2 + \|E'h - f_0\|_{L^2(X)}^2 \quad (f_0 \in \text{Range}(E') = \text{Ker}(P')). \end{aligned}$$

Which allows to conclude that

$$\begin{aligned} \Delta(h, f_0 + P'h) &= \mathbb{E}[(Y - h(X))^2] - \mathbb{E}[(Y - (f_0(X) + P'h(X)))^2] \\ &= \|Y - h(X)\|_{L^2(\Omega)}^2 - \|Y - (f_0(X) + P'h(X))\|_{L^2(\Omega)}^2 \\ &= \|E'h - f_0\|_{L^2(X)}^2. \end{aligned}$$

□

C. Conditions for $P : \mathcal{H} \rightarrow \mathcal{H}$ to be well-defined

Let $\mathcal{H}$ be a RKHS of real-valued functions over $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. In this section, we discuss conditions under which the orthogonal projection $P : L^2(X) \rightarrow L^2(X)$ can be seen as a well-defined projection over $\mathcal{H} \subset L^2(X)$.

Formally, let $\iota : \mathcal{H} \rightarrow L^2(X)$ denote the inclusion operator that maps elements of the RKHS $\mathcal{H} \ni f \mapsto [f]_{\sim}$ to their equivalence class in $L^2(X)$. Saying that P is well-defined as a projection over $\mathcal{H}$ means that

$$P\iota f \in \iota\mathcal{H} \quad \forall f \in \mathcal{H}. \quad (84)$$

Such construction however raises two issues

1. Since $P = \text{Id} - E$ and $E\iota f : x \mapsto \mathbb{E}[\iota f(X) | X_2 = x_2]$ is a function of x_2 only, for $E\iota f$ to lie in RKHS it is necessary for $\mathcal{H}$ to contain functions that are constant with respect to x_1 .
2. If $f \in \mathcal{H}$, there is no guarantee that $E\iota f = \mathbb{E}[\iota f(X) | X_2 = \pi(\cdot)]$ will also lie in $\mathcal{H}$. In fact, this will often not be true — e.g. when $\mathcal{X}$ is a continuous domain (Song et al., 2009) — and we only have $P\iota\mathcal{H} \subset L^2(X)$.

In what follows, we permit ourselves to drop the ι notation.

C.1. Issue 1 : $\mathcal{H}$ must contain functions constant wrt x_1

In general, it is not guaranteed that a RKHS will contain constant functions. In fact, this is not the case for generic RKHSs such as the RKHSs induced by Gaussian or Matérn kernels (Steinwart & Christmann, 2008). To overcome this issue, we propose a particular form for the reproducing kernel that will ensure the RKHS contains constant functions with respect to x_1 .

Proposition C.1. *Let $r : \mathcal{X}_1 \times \mathcal{X}_1 \rightarrow \mathbb{R}$ and $\ell : \mathcal{X}_2 \times \mathcal{X}_2 \rightarrow \mathbb{R}$ be kernel functions. Then the RKHS with reproducing kernel*

$$k = (r + 1) \otimes \ell \quad (85)$$

contains functions that are constant with respect to the first variable x_1 .

Proof. Let $r : \mathcal{X}_1 \times \mathcal{X}_1 \rightarrow \mathbb{R}$ be a kernel function on $\mathcal{X}_1$ and consider the kernel defined by $r^+ = r + 1$ with RKHS $\mathcal{H}_{r^+}$. Let $c \in \mathbb{R}$ and consider the constant function $g(x_1) = c \quad \forall x_1 \in \mathcal{X}_1$.

Then for any $x_1, x'_1 \in \mathcal{X}_1$ we have

$$c^2 r^+(x_1, x'_1) - g(x_1)g(x'_1) = c^2 r(x_1, x'_1) + c^2 - c^2 \quad (86)$$

$$= c^2 r(x_1, x'_1) \quad (87)$$

which is a kernel function. By Theorem A.1 we conclude that $\mathcal{H}_{r^+}$ contains constant functions.

We now consider a second kernel $\ell : \mathcal{X}_2 \times \mathcal{X}_2 \rightarrow \mathbb{R}$ with RKHS $\mathcal{H}_\ell$ and we propose to take $\mathcal{H}$ as the tensor product RKHS

$$\mathcal{H} = \mathcal{H}_{r^+} \otimes \mathcal{H}_\ell, \quad (88)$$

which will have reproducing kernel

$$k = r^+ \otimes \ell. \quad (89)$$

Functions from $\mathcal{H}$ now contain functions which are the product of functions from $\mathcal{H}_{r^+}$ and $\mathcal{H}_\ell$ and are therefore allowed to be constant with respect to x_1 since $\mathcal{H}_{r^+}$ contains constant functions. $\square$

Note that while this structural assumption may appear to limit the generality of the proposed methodology, tensor product RKHSs are a widely used form of RKHS (Szabó & Sriperumbudur, 2017; Pogodin et al., 2022; Lun Chau et al., 2022) that preserve universality of kernels from individual dimension and provide a rich function space.

Recall now the expression of the finite sample P^* estimate used in (20),

$$\hat{P}^* = \text{Id} - \mathbf{k}_x^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{x_2}. \quad (90)$$

This allows to estimate the projected kernel k_P following

$$\begin{aligned} \hat{k}_P(x, x') &= \langle \hat{P}^* k_x, \hat{P}^* k_{x'} \rangle_{\mathcal{H}} \\ &= \left\langle k_x - \mathbf{k}_x^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{x_2}(x_2), k_{x'} - \mathbf{k}_x^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{x_2}(x'_2) \right\rangle_{\mathcal{H}} \\ &= k(x, x') \\ &\quad - \ell_{x_2}(x_2)^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \mathbf{k}_x(x') \\ &\quad - \ell_{x_2}(x'_2)^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \mathbf{k}_x(x) \\ &\quad - \ell_{x_2}(x_2)^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \mathbf{K} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{x_2}(x'_2). \end{aligned}$$

However, for the above derivation to be correct, we need that evaluations of the second kernel ℓ can be obtained by taking an inner product in $\mathcal{H}$. Namely, we need that

$$\ell(x_2, x'_2) = \langle \ell_{x_2}, \ell_{x'_2} \rangle_{\mathcal{H}_\ell} = \langle \ell_{x_2}, \ell_{x'_2} \rangle_{\mathcal{H}}. \quad (91)$$

The following result shows that a sufficient condition for this to hold is that $\mathcal{H}_r$ itself does not contain constant functions. As mentioned above, this is a condition satisfied by generic RKHSs such as the RKHSs of the Gaussian kernel or the Matérn kernels (Steinwart & Christmann, 2008) — which is the RKHS we work with in our experiments.

Proposition C.2. *Let $\mathcal{H} = \mathcal{H}_{r^+} \otimes \mathcal{H}_\ell$ where $r^+ = r + 1$. If $\mathcal{H}_r$ does not contain constant functions, then we have that $\ell(x_2, x'_2) = \langle \ell_{x_2}, \ell_{x'_2} \rangle_{\mathcal{H}}$.*

Proof. The kernel $r^+(x_1, x'_1) = r(x_1, x'_1) + 1$ here induces a RKHS $\mathcal{H}_{r^+}$ (of functions from $\mathcal{X}_1$ to $\mathbb{R}$) which does contain constant functions, e.g., $e \in \mathcal{H}_{r^+}$, where $e(x_1) = 1, \forall x_1 \in \mathcal{X}_1$.

This choice of kernel ensures that $\ell_{x_2} \in \mathcal{H}$ when viewed as a function on $\mathcal{X}_1 \times \mathcal{X}_2$, i.e. we can write it as $e \otimes \ell_{x_2}$, so it is clear that it belongs to $\mathcal{H} = \mathcal{H}_{r^+} \otimes \mathcal{H}_\ell$, since $e \in \mathcal{H}_{r^+}$ and $\ell_{x_2} \in \mathcal{H}_\ell$.

Furthermore, we have

$$\begin{aligned} \langle \ell_{x_2}, \ell_{x'_2} \rangle_{\mathcal{H}} &= \langle e \otimes \ell_{x_2}, e \otimes \ell_{x'_2} \rangle_{\mathcal{H}} \\ &= \langle e, e \rangle_{\mathcal{H}_{r^+}} \langle \ell_{x_2}, \ell_{x'_2} \rangle_{\mathcal{H}_\ell} \\ &= \langle e, e \rangle_{\mathcal{H}_{r^+}} \ell(x_2, x'_2). \end{aligned}$$

However,

$$\begin{aligned} \langle e, e \rangle_{\mathcal{H}_{r^+}} &= \langle e, e + r_{x_1} \rangle_{\mathcal{H}_{r^+}} - \langle e, r_{x_1} \rangle_{\mathcal{H}_{r^+}} \\ &= \langle e, r_{x_1}^+ \rangle_{\mathcal{H}_{r^+}} - \langle e, r_{x_1} \rangle_{\mathcal{H}_{r^+}} \\ &= e(x_1) - \langle e, r_{x_1} \rangle_{\mathcal{H}_{r^+}} \\ &= 1 - \langle e, r_{x_1} \rangle_{\mathcal{H}_{r^+}}. \end{aligned}$$

Now if $\mathcal{H}_r$ does not contain constant functions, we have $\text{Span}(\{e\}) \cap \mathcal{H}_r = \{0\}$. Hence, by Lemma A.2 we obtain that e and r_{x_1} are orthogonal in $\mathcal{H}_{r^+}$ which in turn gives that

$$\langle e, r_{x_1} \rangle_{\mathcal{H}_{r^+}} = 0 \Rightarrow \langle e, e \rangle_{\mathcal{H}_{r^+}} = 1. \quad (92)$$

Therefore, if $\mathcal{H}_r$ does not contain constant functions we have that

$$\langle \ell_{x_2}, \ell_{x'_2} \rangle_{\mathcal{H}} = \langle e, e \rangle_{\mathcal{H}_{r^+}} \ell(x_2, x'_2) = \ell(x_2, x'_2). \quad (93)$$

□

C.2. Issue 2 : P is not necessarily closed as an operator on $\mathcal{H}$

Too Long; Didn't Read We make the assumption that $\mathbb{E}[f(X)|X_2 = \cdot] \in \mathcal{H}$ for $f \in \mathcal{H}$.

Too Short; Want More It is possible to choose the reproducing kernel k such that $\mathcal{H}$ is dense in $L^2(X)$. This property is called L^2 -universality (Sriperumbudur et al., 2011). Whilst this might suggest that the assumption $Ef \in \mathcal{H}$ for $f \in \mathcal{H}$ could be reasonable when L^2 -universality is met, in practice no explicit case is provided in the literature where it is easy to verify that $\mathbb{E}[f(X)|X_2 = \cdot] \in \mathcal{H}$ for $f \in \mathcal{H}$.

In fact, a classic counter example given by Fukumizu et al. (2013) is the case where $\mathcal{H}$ is the RKHS of the Gaussian kernel on $\mathcal{X}$ and $X \perp\!\!\!\perp Z$. Then, $\mathbb{E}[f(X)|Z = \cdot]$ is constant for any $f \in \mathcal{H}$ but $\mathcal{H}$ does not contain constant functions (Steinwart & Christmann, 2008). In the context of our work, we do not have $(X_1, X_2) \perp\!\!\!\perp X_2$ but it nonetheless remains difficult to verify whether $\mathbb{E}[f(X)|X_2 = \cdot] \in \mathcal{H}$.

Efforts to study this nontrivial research direction must be highlighted : Mollenhauer & Koltai (2020) show that under denseness assumptions, it is possible to approximate the conditional expectation operator $E : L^2(X) \rightarrow L^2(X)$ with a Hilbert-Schmidt operator on $\mathcal{H}$ with arbitrary precision. Klebanov et al. (2020) propose a rigorous RKHS-friendly construction of E that only assumes that Ef lies a constant away from $\mathcal{H}$. Most recently, Li et al. (2022a) consider the weaker assumption that for $f \in \mathcal{H}$, Ef lies in an interpolation space between $\mathcal{H}$ and $L^2(X)$ and prove optimal learning rates for its estimator.

The theoretical intricacies of such considerations tend however to undermine more “practical”-driven work. For this reason, it is common to defer such consideration to theoretical research and make the assumption that $\mathbb{E}[f(X)|X_2 = \cdot] \in \mathcal{H}$ (Fukumizu et al., 2004; Song et al., 2011; Muandet et al., 2016; Hsu & Ramos, 2019; Ton et al., 2021; Chau et al., 2021; Fawkes et al., 2022). Since the RKHS theory is not central to our motivations but only a tool we use to demonstrate the benefits of collider regression, we propose to make a similar assumption and delegate this theoretical consideration for future work.

D. Collider Regression on a simple DAG: estimators

Let $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ and $\ell : \mathcal{X}_2 \times \mathcal{X}_2 \rightarrow \mathbb{R}$ be positive definite kernel. In what follows, we adopt notations from the Section 4.3. $\hat{f} = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{k}_\mathbf{x}$ denotes the solution to the kernel ridge regression problem in $\mathcal{H}$. We abuse notation and denote the pairwise inner product of feature maps as

$$\langle \mathbf{k}_\mathbf{x}, \mathbf{k}_\mathbf{x} \rangle_{\mathcal{H}} = [\langle k_{x_i}, k_{x_j} \rangle_{\mathcal{H}}]_{1 \leq i, j \leq n} = [k(x_i, x_j)]_{1 \leq i, j \leq n} = \mathbf{K}. \quad (94)$$

D.1. For a general choice of kernel k

D.1.1. ESTIMATING $\mu_{X|X_2=x_2}$

We are interested in estimating the CME $\mu_{X|X_2=x_2}$. Using the CME estimate from (6), we obtain

$$\hat{\mu}_{X|X_2=x_2} = \mathbf{k}_\mathbf{x}^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{\mathbf{x}_2}(x_2). \quad (95)$$

D.1.2. ESTIMATING $P\hat{f}$

Writing out

$$P\hat{f}(x_1, x_2) = \hat{f}(x_1, x_2) - \langle \hat{f}, \mu_{X|X_2=x_2} \rangle_{\mathcal{H}} \quad (96)$$

$$= \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{k}_\mathbf{x}(x_1, x_2) - \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \langle \mathbf{k}_\mathbf{x}, \mu_{X|X_2=x_2} \rangle_{\mathcal{H}}, \quad (97)$$

it appears we can obtain an estimate of $P\hat{f}$ by substituting $\mu_{X|X_2=x_2}$ with its estimate in the above. We obtain

$$\hat{P}\hat{f}(x_1, x_2) = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{k}_\mathbf{x}(x_1, x_2) - \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \underbrace{\langle \mathbf{k}_\mathbf{x}, \mathbf{k}_\mathbf{x} \rangle_{\mathcal{H}}}_{\mathbf{K}} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{\mathbf{x}_2}(x_2) \quad (98)$$

$$= \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} (\mathbf{k}_\mathbf{x}(x_1, x_2) - \mathbf{K} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{\mathbf{x}_2}(x_2)), \quad (99)$$

or in functional form

$$\hat{P}\hat{f} = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} (\mathbf{k}_\mathbf{x} - \mathbf{K} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{\mathbf{x}_2}). \quad (100)$$

D.2. When $k = (r + 1) \otimes \ell$

In Section 4.4, a sufficient assumption for the projection to be well-defined is that the kernel takes the form

$$k = (r + 1) \otimes \ell, \quad (101)$$

where $r : \mathcal{X}_1 \times \mathcal{X}_1 \rightarrow \mathbb{R}$ is a positive definite kernel. When we choose this particular form of kernel, alternative estimators can be devised.

In what follow, we denote $r^+ = r + 1$, $r_{\mathbf{x}_1}^+ = r^+(\mathbf{x}_1, \cdot)$ and $\mathbf{R}^+ = r^+(\mathbf{x}_1, \mathbf{x}_1)$.

D.2.1. ESTIMATING $\mu_{X|X_2=x_2}$

Going back to the definition of CMEs, we can write

$$\mu_{X|X_2=x_2} = \mathbb{E}[k_X | X_2 = x_2] = \mathbb{E}[r_{\mathbf{x}_1}^+ \otimes \ell_{X_2} | X_2 = x_2] = \mathbb{E}[r_{\mathbf{x}_1}^+ | X_2 = x_2] \otimes \ell_{x_2} = \mu_{X_1|X_2=x_2} \otimes \ell_{x_2}. \quad (102)$$

Therefore, it is sufficient to obtain an estimate of $\mu_{X_1|X_2=x_2}$, which we can get as

$$\hat{\mu}_{X_1|X_2=x_2} = \mathbf{r}_{\mathbf{x}_1}^{+\top} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{\mathbf{x}_2}(x_2), \quad (103)$$

and take as a CME estimator

$$\hat{\mu}_{X|X_2=x_2} = \left[\mathbf{r}_{\mathbf{x}_1}^{+\top} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{\mathbf{x}_2}(x_2) \right] \ell_{x_2}(\cdot) \quad (104)$$

D.2.2. ESTIMATING $P\hat{f}$

Following the similar derivations than in the general case, we obtain

$$\hat{P}\hat{f}(x_1, x_2) = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{k}_{\mathbf{x}}(x_1, x_2) - \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \langle \mathbf{r}_{\mathbf{x}_1}^+, \mathbf{r}_{\mathbf{x}_1}^+ \rangle_{\mathcal{H}_{r^+}} \langle \ell_{\mathbf{x}_2}, \ell_{x_2} \rangle_{\mathcal{H}_\ell} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{\mathbf{x}_2}(x_2) \quad (105)$$

$$= \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} [\mathbf{k}_{\mathbf{x}}(x_1, x_2) - \text{Diag}(\ell_{\mathbf{x}_2}(x_2)) \mathbf{R}^+ (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{\mathbf{x}_2}(x_2)], \quad (106)$$

where $\text{Diag}(\ell_{\mathbf{x}_2}(x_2))$ is the diagonal matrix that has the vector $\ell_{\mathbf{x}_2}(x_2) = \ell(\mathbf{x}_2, x_2)$ as its diagonal. Written in functional form we obtain

$$\hat{P}\hat{f} = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} [\mathbf{k}_{\mathbf{x}} - \text{Diag}(\ell_{\mathbf{x}_2}(\cdot)) \mathbf{R}^+ (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{\mathbf{x}_2}] \quad (107)$$

 D.2.3. ESTIMATING k_P

Writing out,

$$k_P(x, x') = \langle P^* k_x, P^* k_{x'} \rangle_{\mathcal{H}} \quad (108)$$

$$= \langle k_x - \mu_{X|X_2=x_2}, k_{x'} - \mu_{X|X_2=x'_2} \rangle_{\mathcal{H}} \quad (109)$$

$$= \langle k_x, k_{x'} \rangle_{\mathcal{H}} \quad (110)$$

$$- \langle \mu_{X_1|X_2=x_2} \otimes \ell_{x_2}, k_{x'} \rangle_{\mathcal{H}} \quad (111)$$

$$- \langle k_x, \mu_{X_1|X_2=x'_2} \otimes \ell_{x'_2} \rangle_{\mathcal{H}} \quad (112)$$

$$+ \langle \mu_{X_1|X_2=x_2} \otimes \ell_{x_2}, \mu_{X_1|X_2=x'_2} \otimes \ell_{x'_2} \rangle_{\mathcal{H}} \quad (113)$$

$$= r^+(x_1, x'_1) \ell(x_2, x'_2) \quad (114)$$

$$- \langle \mu_{X_1|X_2=x_2}, r_{x'_1}^+ \rangle_{\mathcal{H}_{r^+}} \ell(x_2, x'_2) \quad (115)$$

$$- \langle r_{x_1}^+, \mu_{X_1|X_2=x'_2} \rangle_{\mathcal{H}_{r^+}} \ell(x_2, x'_2) \quad (116)$$

$$+ \langle \mu_{X_1|X_2=x_2}, \mu_{X_1|X_2=x'_2} \rangle_{\mathcal{H}_{r^+}} \ell(x_2, x'_2) \quad (117)$$

$$= \ell(x_2, x'_2) \left[r^+(x_1, x'_1) - \langle \mu_{X_1|X_2=x_2}, r_{x'_1}^+ \rangle_{\mathcal{H}_{r^+}} - \langle r_{x_1}^+, \mu_{X_1|X_2=x'_2} \rangle_{\mathcal{H}_{r^+}} + \langle \mu_{X_1|X_2=x_2}, \mu_{X_1|X_2=x'_2} \rangle_{\mathcal{H}_{r^+}} \right] \quad (118)$$

Therefore, substituting $\mu_{X_1|X_2=x_2}$ with its estimate, we obtain

$$\hat{k}_P(x, x') = \ell(x_2, x'_2) \quad (119a)$$

$$\times [r^+(x_1, x'_1) \quad (119b)$$

$$- \ell_{\mathbf{x}_2}(x_2)^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \mathbf{r}_{\mathbf{x}_1}^+(x'_1) \quad (119c)$$

$$- \ell_{\mathbf{x}_2}(x'_2)^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \mathbf{r}_{\mathbf{x}_1}^+(x_1) \quad (119d)$$

$$- \ell_{\mathbf{x}_2}(x_2)^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \mathbf{R}^+ (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{\mathbf{x}_2}(x'_2)]. \quad (119e)$$

E. Collider Regression on a general DAG: algorithms and estimators

Let $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, $r : \mathcal{X}_1 \times \mathcal{X}_1 \rightarrow \mathbb{R}$ and $\ell : (\mathcal{X}_2 \times \mathcal{X}_3) \times (\mathcal{X}_2 \times \mathcal{X}_3) \rightarrow \mathbb{R}$ be psd kernels. We follow the same notation convention that in the case of a simple collider, except that now ℓ is a kernel over $\mathcal{X}_2 \times \mathcal{X}_3$. Define $f_0(x) = \mathbb{E}[Y|X_3 = x_3]$. Here $g^* = f^* - f_0$ must live in the appropriate subspace of functions which have zero conditional expectation on (X_2, X_3) .

E.1. Algorithms

Algorithm 4 General procedure to estimate $f_0 + P'\hat{g}$

- 1: Regress $X_3 \rightarrow Y$ to get $x_3 \mapsto \hat{f}_0(x_3)$
 - 2: Take $\tilde{Y} = Y - \hat{f}_0(X_3)$
 - 3: Regress $(X_1, X_2, X_3) \rightarrow \tilde{Y}$ to get $(x_1, x_2, x_3) \mapsto \hat{g}(x_1, x_2, x_3)$
 - 4: Regress $(X_2, X_3) \rightarrow \hat{g}(X_1, X_2, X_3)$ to get $(x_2, x_3) \mapsto \hat{\mathbb{E}}[\hat{g}(X_1, X_2, X_3)|X_2 = x_2, X_3 = x_3]$
 - 5: Take $\hat{P}'\hat{g}(x_1, x_2, x_3) = \hat{g}(X_1, X_2, X_3) - \hat{\mathbb{E}}[\hat{g}(x_1, x_2, x_3)|X_2 = x_2, X_3 = x_3]$
 - 6: **return** $\hat{f}_0 + \hat{P}'\hat{g}$
-

Algorithm 5 RKHS procedure to estimate $f_0 + P'\hat{g}$

- 1: Estimate $\hat{\mu}_{X|X_2=x_2, X_3=x_3}$
 - 2: Regress $X_3 \rightarrow Y$ to get $x_3 \mapsto \hat{f}_0(x_3)$
 - 3: Take $\tilde{\mathbf{y}} = \mathbf{y} - \hat{f}_0(\mathbf{x}_3)$
 - 4: Take $\hat{\mathbf{g}} = \tilde{\mathbf{y}}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{k}_{\mathbf{x}}$
 - 5: Let $\hat{P}'\hat{\mathbf{g}} = \hat{\mathbf{g}} - \langle \hat{\mu}_{X|X_2=\cdot, X_3=\cdot} \rangle_{\mathcal{H}}$
 - 6: **return** $\hat{f}_0 + \hat{P}'\hat{\mathbf{g}}$
-

Algorithm 6 RKHS procedure to estimate $f_0 + \hat{g}_{P'}$

- 1: Estimate $\hat{\mu}_{X|X_2=x_2, X_3=x_3}$
 - 2: Regress $X_3 \rightarrow Y$ to get $x_3 \mapsto \hat{f}_0(x_3)$
 - 3: Take $\tilde{\mathbf{y}} = \mathbf{y} - \hat{f}_0(\mathbf{x}_3)$
 - 4: Let $\hat{P}'^* \mathbf{k}_x = \mathbf{k}_x - \hat{\mu}_{X|X_2=x_2, X_3=x_3}$
 - 5: Let $\hat{k}_{P'}(x, x') = \langle \hat{P}'^* \mathbf{k}_x, \hat{P}'^* \mathbf{k}_{x'} \rangle_{\mathcal{H}}$
 - 6: Evaluate $\hat{\mathbf{K}}_{P'} = \hat{k}_{P'}(\mathbf{x}, \mathbf{x})$ and $\hat{\mathbf{k}}_{P', \mathbf{x}} = \hat{k}_{P'}(\mathbf{x}, \cdot)$
 - 7: Take $\hat{g}_{P'} = \tilde{\mathbf{y}}^\top (\hat{\mathbf{K}}_{P'} + \lambda \mathbf{I}_n)^{-1} \hat{\mathbf{k}}_{P', \mathbf{x}}$
 - 8: **return** $\hat{f}_0 + \hat{g}_{P'}$
-

E.2. Estimators for a general kernel k

E.2.1. ESTIMATING $\mu_{X|X_2=x_2, X_3=x_3}$

$$\hat{\mu}_{X|X_2=x_2, X_3=x_3} = \mathbf{k}_{\mathbf{x}}^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2, \mathbf{x}_3}(x_2, x_3). \quad (120)$$

E.2.2. ESTIMATING $P'\hat{g}$

$$\hat{P}'\hat{\mathbf{g}} = \tilde{\mathbf{y}}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} (\mathbf{k}_{\mathbf{x}} - \mathbf{K}(\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2, \mathbf{x}_3}). \quad (121)$$

E.3. Estimators when $k = (r + 1) \otimes \ell$

E.3.1. ESTIMATING $\mu_{X|X_2=x_2, X_3=x_3}$

$$\hat{\mu}_{X|X_2=x_2} = \left[\mathbf{r}_{\mathbf{x}_1}^{+\top} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2, \mathbf{x}_3}((x_2, x_3)) \right] \ell_{x_2, x_3}(\cdot) \quad (122)$$

E.3.2. ESTIMATING $P' \hat{g}$

$$\hat{P}' \hat{g} = \tilde{\mathbf{y}}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} [\mathbf{k}_{\mathbf{x}} - \text{Diag}(\ell_{\mathbf{x}_2, \mathbf{x}_3}(\cdot)) \mathbf{R}^+ (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2, \mathbf{x}_3}] \quad (123)$$

E.3.3. ESTIMATING $k_{P'}$

$$\hat{k}_{P'}(x, x') = \ell((x_2, x_3), (x'_2, x'_3)) \quad (124a)$$

$$\times [r^+(x_1, x'_1) \quad (124b)$$

$$- \boldsymbol{\ell}_{\mathbf{x}_2, \mathbf{x}_3}((x_2, x_3))^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \mathbf{r}_{\mathbf{x}_1}^+(x'_1) \quad (124c)$$

$$- \boldsymbol{\ell}_{\mathbf{x}_2, \mathbf{x}_3}((x'_2, x'_3))^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \mathbf{r}_{\mathbf{x}_1}^+(x_1) \quad (124d)$$

$$- \boldsymbol{\ell}_{\mathbf{x}_2, \mathbf{x}_3}((x_2, x_3))^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \mathbf{R}^+ (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \boldsymbol{\ell}_{\mathbf{x}_2, \mathbf{x}_3}((x'_2, x'_3))]. \quad (124e)$$

F. Details on experiments

F.1. Models

➤ **RF** We use the scikit-learn (Pedregosa et al., 2011) `sklearn.ensemble.RandomForestRegressor` implementation which we tune for

- `n_estimators`
- `max_depth`
- `min_samples_split`
- `min_samples_leaf`

using a cross-validated grid search over an independently generated validation set.

➤ **P-RF** Once RF has been fitted as $\hat{f}$, we estimate $\mathbb{E}[\hat{f}(X_1, X_2)|X_2]$ by fitting a linear regression model of X_2 onto $\hat{f}(X_1, X_2)$.

➤ **KRR** We implement our own kernel ridge regression in PyTorch (Paszke et al., 2019). The kernel is taken as

$$k((x_1, x_2), (x'_1, x'_2)) = (\kappa_{\theta_1}(x_1, x'_1) + 1)\kappa_{\theta_2}(x_2, x'_2), \quad (125)$$

where κ_{θ} denotes the Gaussian kernel with lengthscale $\theta > 0$

$$\kappa_{\theta}(u, u') = \exp\left(-\frac{\|u - u'\|_2^2}{\theta}\right). \quad (126)$$

The kernel lengthscales θ_1, θ_2 and the regularisation weight $\lambda > 0$ are tuned using a cross-validated grid search on an independently generated validation set.

➤ **P-KRR** Once KRR has been fitted as $\hat{f} = \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{k}_x$, we estimate the CME and use it to estimate $P\hat{f}(x_1, x_2) = \hat{f}(x_1, x_2) - \langle \hat{f}, \mu_{X|X_2=x_2} \rangle_{\mathcal{H}}$ following

$$\hat{\mu}_{X|X_2=x_2} = \mathbf{k}_x^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{x_2}(x_2) \quad (127)$$

$$\Rightarrow \hat{P} = \text{Id} - \hat{\mu}_{X|X_2=x_2}. \quad (128)$$

$$= \text{Id} - \mathbf{k}_x^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{x_2} \quad (129)$$

$$\Rightarrow \hat{P}\hat{f} = \hat{f} - \hat{f} \mathbf{k}_x^\top (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{x_2} \quad (130)$$

$$= \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{k}_x - \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{K} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{x_2} \quad (131)$$

$$= \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} (\mathbf{k}_x - \mathbf{K} (\mathbf{L} + \gamma \mathbf{I}_n)^{-1} \ell_{x_2}) \quad (132)$$

The kernel on $\mathcal{X}_2$ is taken as $\ell = \kappa_{\theta_2}$. The CME regularisation weight $\gamma > 0$ is tuned using a cross-validated grid search on an independently generated validation set.

➤ **$\mathcal{H}_P$ -KRR** We use the same base kernel as for KRR with again $\ell = \kappa_{\theta_2}$. We implement our estimator of the projected kernel k_P is GPyTorch (Gardner et al., 2018)⁹. The kernel lengthscales and regularisation weights are tuned using a cross-validated grid search on an independently generated validation set.

⁹which can be readily incorporated into GP regression pipelines.

F.2. Simulation example

Data generating process Algorithm 7 outlines the procedure we use to generate a positive definite matrix Σ that encodes independence between X_2 and Y .

Algorithm 7 Procedure to generate Σ

```

1: Input:  $d_1 \geq 1, d_2 \geq 1$ 
2: # Generate a  $4 \times (d_1 + d_2 + 1)$  random matrix
3: for  $i \in \{1, \dots, d_1 + d_2 + 1\}$  do
4:    $M_i \sim \mathcal{N}(0, \mathbf{I}_4)$ 
5:    $M_i \leftarrow M_i / \|M_i\|_2$ 
6: end for
7: # Make  $Y$  column orthogonal to all  $X_2$  columns
8:  $M_Y \leftarrow M_{d_1+d_2+1}$ 
9: for  $i \in \{d_1 + 1, \dots, d_1 + d_2\}$  do
10:   $M_i \leftarrow M_i - (M_i^\top M_Y) M_Y$ 
11: end for
12:  $M \leftarrow [M_1 \mid \dots \mid M_{d_1+d_2} \mid M_Y] \in \mathbb{R}^{4 \times (d_1+d_2+1)}$ 
13:  $\Sigma \leftarrow M^\top M + 0.01 * \mathbf{I}_{d_1+d_2+1}$ 
14: # Normalise variances to 1
15:  $\Lambda \leftarrow \text{Diag}(\Sigma)$ 
16:  $\Sigma \leftarrow \Lambda^{-1/2} \Sigma \Lambda^{-1/2}$ 
    
```

Non-linear mappings The mappings g_1 and g_2 are applied to each component of the input vectors and are given by

$$g_1(u) = u + 0.1 \cos(2\pi u^2) \quad (133)$$

$$g_2(u) = u + 0.1 \sin(2\pi u^2). \quad (134)$$

Statistical significance table

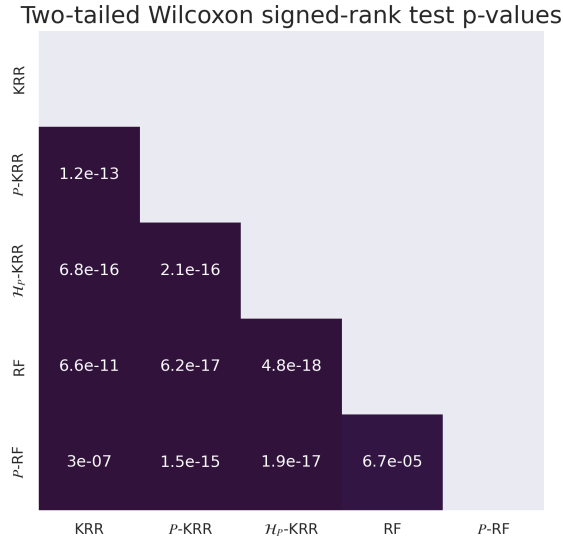


Figure 6. p-values from a two-tailed Wilcoxon signed-rank test between all pairs of methods for the test MSE of the simulation example. The null hypothesis is that scores samples come from the same distribution. We only present the lower triangular matrix of the table for clarity of reading.

F.3. Aerosol radiative forcing

Statistical significance table

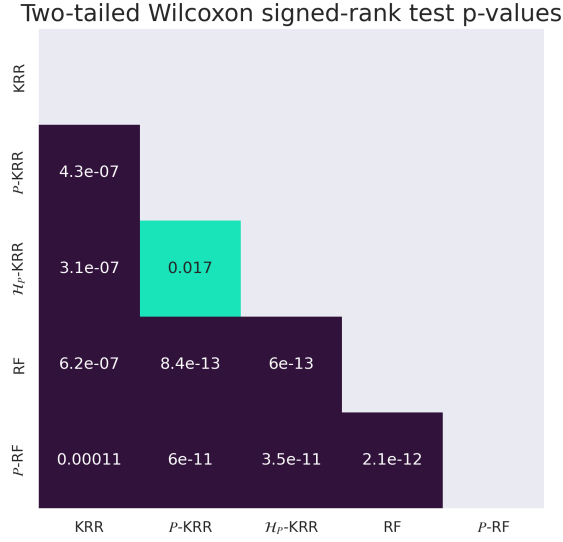


Figure 7. p-values from a two-tailed Wilcoxon signed-rank test between all pairs of methods for the test **MSE** of the aerosol radiative forcing experiment. The null hypothesis is that scores samples come from the same distribution. We only present the lower triangular matrix of the table for clarity of reading.

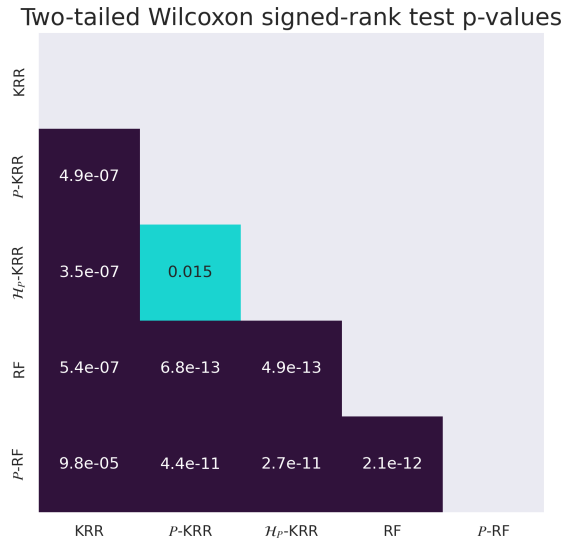


Figure 8. p-values from a two-tailed Wilcoxon signed-rank test between all pairs of methods for the test **SNR** of the aerosol radiative forcing experiment. The null hypothesis is that scores samples come from the same distribution. We only present the lower triangular matrix of the table for clarity of reading.

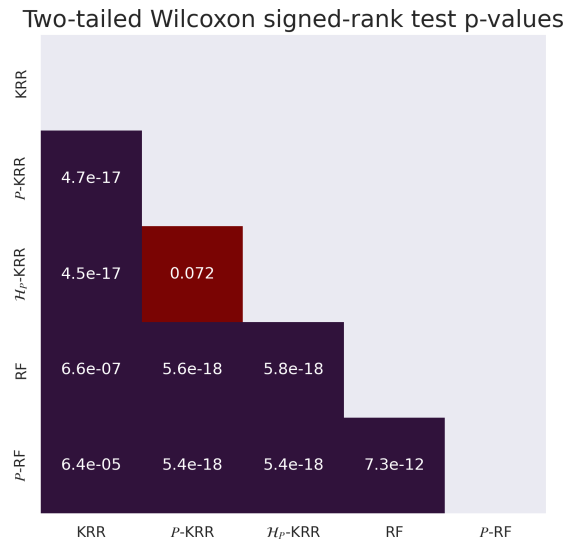


Figure 9. p-values from a two-tailed Wilcoxon signed-rank test between all pairs of methods for the test **correlation** of the aerosol radiative forcing experiment. The null hypothesis is that scores samples come from the same distribution. We only present the lower triangular matrix of the table for clarity of reading.

G. Future direction

G.1. Extension to Gaussian processes

Extension to Gaussian processes The methodology presented can naturally be extended to the Bayesian counterpart of kernel ridge regression, Gaussian processes (GPs) (Rasmussen & Williams, 2005). One can either apply the projection operator $P : L^2(X) \rightarrow L^2(X)$ to the GP prior (or posterior), or use the projected kernel k_P to specify the covariance function¹⁰.

However, such approach raises important questions from a theoretical perspective. If $f \sim \text{GP}(0, k)$, the application of the $L^2(X)$ projection to f will result in a linearly transformed GP $Pf \sim \text{GP}(0, PkP^*)$ (Särkkä, 2011) and its draws will lie in the range of P . In contrast, since draws from a GP almost surely lie outside the RKHS associated with its covariance (Kanagawa et al., 2018), draws from $f \sim \text{GP}(0, k_P)$ will almost surely lie outside $\mathcal{H}_P$. It is therefore unclear whether these draws will lie in the range of the projection and satisfy the desired constraint for f . On the other hand, the posterior mean of the GP will always lie in $\mathcal{H}_P$.

Furthermore, the projection is targeted at improving performance in mean square error. Because this metric is not necessarily adequate to evaluate GPs, it is unclear whether applying the projection would result in a performance improvement on more commonly used metrics for GPs such as maximum likelihood.

¹⁰Our implementation of $\hat{k}_P$ is available in GPyTorch (Gardner et al., 2018) and can be readily incorporated into GP regression pipelines.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	Returning the Favour: When Regression Benefits from Probabilistic Causal Knowledge
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Bouabid, Shahine, Fawkes, Jakes and Sejdinovic, Dino, 2023. Returning the favour: when regression benefits from probabilistic causal knowledge. In <i>International Conference on Machine Learning</i> (pp. 2885-2913). PMLR.

Student Confirmation

Student Name:	Shahine Bouabid		
Contribution to the Paper	I am joint first author of this paper with Jake Fawkes. I came up with the motivation of using knowledge of the causal structure as an inductive bias in regression. Fawkes came up with the idea of looking at collider structures and motivated the collider regression problem. I contributed the formulation of collider regression as a constrained regression problem, the function space projection formulation and proved the theoretical results. Jointly with Fawkes, we contributed the RKHS version of the problem and designed the experiments. I implemented and run the experiments. Jointly with Fawkes, we wrote the paper. Prof. Sejdinovic helped with the RKHS formulation, advised along the way and corrected the paper draft.		
Signature		Date	11/03/24

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof. Dino Sejdinovic		
Supervisor comments		
Signature		Date
		12/03/2024

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 4

FaIRGP: A Bayesian Energy Balance Model for Surface Temperatures Emulation

Bouabid, Shahine, Dino Sejdinovic and Duncan Watson-Parris. “FaIRGP: A Bayesian Energy Balance Model for Surface Temperatures Emulation.” *Journal of Advances in Earth System Modelling*. 2024.

FaIRGP: A Bayesian Energy Balance Model for Surface Temperatures Emulation

Shahine Bouabid¹, Dino Sejdinovic², Duncan Watson-Parris³

¹Department of Statistics, University of Oxford, Oxford, UK

²School of CMS & AIML, University of Adelaide, Adelaide, Australia

³Scripps Institution of Oceanography and Halicioğlu Data Science Institute, University of California, San Diego, US

Key Points:

- We introduce FaIRGP, a Bayesian machine learning emulator for global and local mean surface temperatures that builds upon a physically based simple climate model
- The model improves upon both purely physically-driven and purely data-driven baseline emulators on several metrics across realistic future scenarios.
- The model is fully mathematically tractable, which makes it a convenient and easy-to-use probabilistic tool for the emulation of surface temperatures, but also for downstream applications such as detection and attribution or precipitation emulation.

Corresponding author: Shahine Bouabid, shahine.bouabid@stats.ox.ac.uk

Abstract

Emulators, or reduced complexity climate models, are surrogate Earth system models that produce projections of key climate quantities with minimal computational resources. Using time-series modelling or more advanced machine learning techniques, data-driven emulators have emerged as a promising avenue of research, producing spatially resolved climate responses that are visually indistinguishable from state-of-the-art Earth system models. Yet, their lack of physical interpretability limits their wider adoption. In this work, we introduce FaIRGP, a data-driven emulator that satisfies the physical temperature response equations of an energy balance model. The result is an emulator that *(i)* enjoys the flexibility of statistical machine learning models and can learn from data, and *(ii)* has a robust physical grounding with interpretable parameters that can be used to make inference about the climate system. Further, our Bayesian approach allows a principled and mathematically tractable uncertainty quantification. Our model demonstrates skillful emulation of global mean surface temperature and spatial surface temperatures across realistic future scenarios. Its ability to learn from data allows it to outperform energy balance models, while its robust physical foundation safeguards against the pitfalls of purely data-driven models. We also illustrate how FaIRGP can be used to obtain estimates of top-of-atmosphere radiative forcing and discuss the benefits of its mathematical tractability for applications such as detection and attribution or precipitation emulation. We hope that this work will contribute to widening the adoption of data-driven methods in climate emulation.

Plain Language Summary

Emulators are simplified climate models that can be used to rapidly explore climate scenarios — they can run in less than a minute on an average computer. They are key tools used by the Intergovernmental Panel on Climate Change to explore the diversity of possible future climates. Data-driven emulators use advanced machine learning techniques to produce climate predictions that look very similar to the predictions of complex climate models. However, they are not easy to interpret, and therefore to trust in practice. In this work, we introduce FaIRGP, a data-driven emulator based on physics. The emulator is flexible and can learn from data to improve its predictions, but is also grounded on physical energy balance relationships, which makes it robust and interpretable. The model performs well in predicting future global and local temperatures under realistic future scenarios, outperforming purely physics-driven or purely data-driven models. Further, the probabilistic nature of our model allows for mathematically tractable uncertainty quantification. By gaining trust in such a data-driven yet physically grounded model, we hope the climate science community can benefit more widely from their potential.

1 Introduction

Earth system models (ESMs) (Flato, 2011) are key tools to understand current climate dynamics and climate change responses to greenhouse gas emissions. They constitute an extensive physical simulation of Earth’s atmosphere and ocean fluid dynamics, used for example in the Couple Model Intercomparison Project (Meehl et al., 2007; Taylor et al., 2012; Eyring et al., 2016) to study past and future climate. As such, they offer the most comprehensive view of what future climate could look like. They are also used as an idealized fully controlled environment to study climate dynamics and understand its underlying drivers. In particular, they play a central role in the estimation of key properties of the climate system such as timescales and equilibrium responses to the change in carbon dioxide concentration in the atmosphere (Allen et al., 2009; Collins et al., 2014) and the effect of aerosols (Levy et al., 2013).

Running simulations with an ESM requires an astute understanding of the climate science background, of the numerical schemes used to simulate climate dynamics, and access to an adequate computational infrastructure¹. Therefore, only a limited number of research teams around the world can realistically afford to perform climate simulations. A direct consequence is that a variety of scientific applications relying on future climate projections — such as agricultural studies (Rosenzweig et al., 2014), energy infrastructure models (Turner et al., 2017) or global socio-economic human models (Akhtar et al., 2013) — must settle for publicly available precomputed climate projections that have been selected based on their policy-relevance, and may not be tailored to the application needs. The need for expensive computational resources also serves as an important barrier, making it inaccessible for independent researchers and less well-equipped research teams to run experiments with ESMs. This exacerbates the already unequal representation in high-impact climate science research, where the global north is disproportionately represented (Tandon, 2021).

Furthermore, even when the resource needs are met to run experiments with an ESM, their computational cost remains a critical obstacle. Indeed, the uncertainty over the ESM parameterisation, the climate system internal variability and the emission pathway the world will choose, together span a high-dimensional uncertainty space. Therefore, obtaining a comprehensive coverage of this uncertainty requires running numerous climate simulations, which quickly meets computational cost limitations. As a result, much of the climate variability and potential socio-economic pathways remain in practice unexplored.

Together, these limitations have fostered the emergence of simpler surrogate ESMs which are inexpensive to compute and referred to as *emulators*. Unlike ESMs, emulators do not explicitly model the fluid dynamics of the atmosphere and oceans and focus on a limited number of climate features, such as surface temperatures or precipitations. They can run thousands of years of simulation in less than a minute on an average personal computer (Leach et al., 2021), hence making accessible the emulation of climate projections under configurations unexplored by ESMs.

Simple climate models (SCMs) (Meinshausen et al., 2011, 2020; Millar et al., 2017; Smith et al., 2018; Leach et al., 2021) are an important class of models that can emulate climate projections. They propose a reduced-order representation of the climate system, modelling the carbon cycle, radiative forcing, and temperature response with simplified first order dynamics. A key component of SCMs is their *energy balance model* (EBM) (Rypdal, 2012; Geoffroy et al., 2013; Millar et al., 2017; Fredriksen & Rypdal, 2017; Lovejoy et al., 2021), which describes the changes in global surface temperature through imbalances in Earth’s energy budget (Held et al., 2010). It represents the atmosphere-ocean system as a set of connected boxes forced by radiative flux at the top of the atmosphere. Whilst

¹ As an order of magnitude, running the CESM2 model (Danabasoglu et al., 2020) for a single year ahead takes about 2000 core hours on a supercomputer.

this constitutes a drastic simplification of the climate system, EBMs are robust physically-motivated models, and therefore remain the main tool used to connect the IPCC working physical basis research (Lee et al., 2021) with the adaptation and mitigation efforts (Pörtner et al., 2022; Shukla et al., 2022).

Another important class of emulators are *data-driven emulators*. In contrast to SCMs, they do not explicitly model climate dynamics and rather rely on statistical modelling techniques to emulate key climate variables such as temperature or precipitation (Del Rio Amador & Lovejoy, 2019; Beusch et al., 2020b; Link et al., 2019; Goodwin et al., 2020; Castruccio et al., 2014; Holden & Edwards, 2010). Statistical modelling enjoys greater flexibility and has produced powerful emulators, capable of successfully approximating ESMs’ spatially resolved climate projections with visually indistinguishable outputs. More recently, statistically driven emulators drawing from advances in statistical machine learning have demonstrated a remarkable capacity at regressing global emission profiles onto spatially resolved temperature and precipitation maps (Watson-Parris et al., 2022).

However, both SCMs and data-driven emulators display fundamental limitations. Whilst SCMs provide a robust physical framework to reason about the climate system, they remain a simplistic model which may display a poor fit to ESM’s outputs in certain scenarios (Meinshausen et al., 2011; Geoffroy et al., 2013; Jackson et al., 2022) and can only operate at the global level. Reasoning only in terms of global mean temperatures fails to capture the difference in exposure of different world regions and limits use for scientific applications that require regional projections. On the other hand, data-driven emulators are also limited in their ability to provide a complete, reliable picture of the climate system. Indeed, their lack of robust physical grounding limits their capacity to make inferences about the climate system (Beusch et al., 2020b; Link et al., 2019). Further, the outputs from these emulators are often subject to qualitative evaluation and may not be trustworthy², in part because of limited understanding on how they might behave outside the training data regimes.

In this work, we address these limitations by formulating a hybrid physical/statistical emulator that will both enjoy the robust physical grounding of SCMs, and the flexibility of statistical machine learning methods, thereby combining the advantages of both classes of emulators. We propose a probabilistic emulator for the task of reproducing an ESM temperature response to greenhouse gas and aerosol emissions. Our model builds upon a simple EBM, but draws from the latent force model paradigm (M. Alvarez et al., 2009; M. A. Alvarez et al., 2013) by placing a Gaussian process (GP) prior over the radiative forcing, thereby inducing a stochastic temperature response model. We show that the resulting emulated temperature turns out to also be a GP, with a physically-informed covariance structure that reflects the dynamics of the EBM. In consequence, we can update the EBM with temperature data to learn a posterior distribution over global temperatures, but also over radiative forcing. Further, we demonstrate how our model can be easily extended to emulate spatially-resolved maps of annual surface temperatures.

Experiments demonstrate skilful prediction of global and spatial mean surface temperatures, with improvements over a simple EBM and a simple GP model. We obtain robust predictions even when only training with historical data, or when predicting over scenarios outside the range of emissions of greenhouse gases and aerosols observed during training. Additionally, we show that the model improves over baselines to emulate temperature changes induced by anthropogenic aerosol emissions, and can provide useful estimates of the top-of-atmosphere radiative forcing.

² Whilst statistical explainability methods (Shapley, 1953; Štrumbelj & Kononenko, 2014) may help understand the contributions to predictions, we argue that the lack of a physically grounded model would still harm trust in the predictions.

2 Background

2.1 Energy Balance Models

Energy balance models (EBMs) (Sellers, 1969; Holden & Edwards, 2010) are a key component of SCMs that provide a lower-order representation of the climate system, where the changes in global temperature are explained by the imbalance in the Earth’s energy budget. The most common and established class of EBM are box models. They represent the atmosphere and ocean as a set of vertically stacked boxes, where the uppermost box is exposed to top-of-atmosphere radiative flux.

In a box model EBM, each box has a different heat capacity C_i , heat transfer coefficients κ_i with the adjacent boxes, and its own temperature $T^{(i)}(t)$, as depicted in Figure 1. The uppermost box represents the fast components of the climate system, generally limited to the atmosphere and the land surface, while the following boxes are used to represent slower components of the climate system such as shallow ocean and deep ocean.

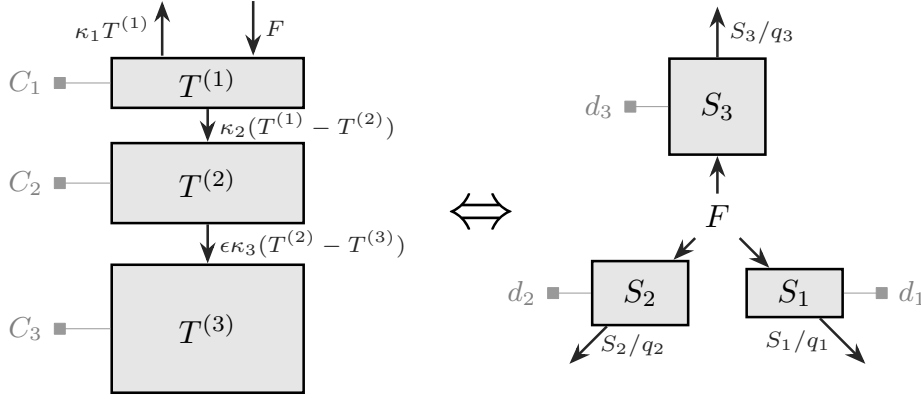


Figure 1: Schematic representation of 3-box energy balance model, the arrows represent the flow of heat. **Left:** Temperature response model representation. Thicker boxes have a larger heat capacity (i.e. $C_1 < C_2 < C_3$) and the arrows represent the flow of heat. $\kappa_1, \kappa_2, \kappa_3$ are the heat transfer coefficients between the boxes and $\epsilon > 0$ denotes the deep ocean heat uptake efficacy. **Right:** Thermal response model representation. Thicker boxes have larger response timescale (i.e. $d_1 < d_2 < d_3$). Adapted from (Figure 1, Cummins et al., 2020).

Let $\mathbf{X}(t) = [T^{(1)}(t) \ \dots \ T^{(k)}(t)]^\top$ be the vector concatenation of the temperatures within the boxes in a model with k boxes. The change in temperature of a k -box EBM is described by the following simple first order linear ordinary differential equation (ODE)

$$\frac{d\mathbf{X}(t)}{dt} = \mathbf{A}\mathbf{X}(t) + \mathbf{b}F(t), \quad (1)$$

where $\mathbf{A}$ is a tridiagonal temperature feedback matrix³ that depends on heat capacities C_i , heat transfer coefficients κ_i , and deep ocean uptake efficacy ϵ , and $\mathbf{b}$ is a radiative forcing feedback vector given by

$$\mathbf{b} = [1/C_1 \ 0 \ \dots \ 0]^\top. \quad (2)$$

$F(t)$ denotes the top of atmosphere effective radiative forcing — radiative forcing for short — i.e. the change in energy flux caused by natural or anthropogenic factors of climate change. It is only applied to the surface box, i.e. the uppermost box.

³ explicit form of the matrix is provided in Appendix B2

Table 1: Notation, description and unit of named parameters.

(a) Deterministic parameters.		
Notation	Unit	Description
$F(t)$	Wm^{-2}	Effective top of atmosphere radiative forcing
$T^{(i)}(t)$	K	Temperature of the i^{th} box
C_i	JK^{-1}	Heat capacity of the i^{th} box
κ_i	$\text{Wm}^{-2}\text{K}^{-1}$	Heat exchange coefficient of the i^{th} box
ε	1	Deep ocean heat uptake efficacy
$T(t)$	K	Global mean surface temperature ($T^{(1)}(t)$)
$S_i(t)$	K	Response of the i^{th} thermal box
d_i	years	Response timescale of i^{th} thermal box
q_i	KW^{-1}m^2	Equilibrium response of i^{th} thermal box
χ	-	Atmospheric agent (e.g. CO_2 , SO_2)
$E(t)$	-	Emission of agents in the atmosphere

(b) Stochastic parameters.		
Notation	Unit	Description
$F(t)$	Wm^{-2}	Effective top of atmosphere radiative forcing
$T(t)$	K	Global mean surface temperature
$S_i(t)$	K	Response of the i^{th} thermal box

Whilst physically motivated, box-models remain an abstract representation of the climate system. Therefore their parameters, such as boxes heat capacities, are not realistic quantities that can be calculated, but rather need to be tuned against data using calibration methods (Meinshausen et al., 2011; Millar et al., 2017) or maximum likelihood strategies (Cummins et al., 2020).

The simplest box models only use $k = 2$ boxes, thereby splitting the climate system into 2 groups: fast and slow components. Whilst reductive, this split has proven to be a robust approximation of the climate system (Held et al., 2010). In fact, 2-box EBM are today the primary tool used for linking the IPCC physical basis research (Lee et al., 2021) with the adaptation and mitigation efforts (Pörtner et al., 2022; Shukla et al., 2022). It is however worth noting recent advocacy for 3-box models (Leach et al., 2021), highlighting the insufficiency of 2-box models to capture the full range of behaviour observed in CMIP6 models (Tsutsui, 2020; Cummins et al., 2020).

2.2 Impulse response formulation

The main box of interest in a box-model is the uppermost box since it describes the global mean surface temperature response $T^{(1)}(t)$ to the radiative forcing $F(t)$. In general, computing $T^{(1)}(t)$ can be simplified by diagonalising the ODE (1). The result is an equivalent impulse response formulation of the temperature response model, depicted on the right in Figure 1, and referred to as the *thermal response ODE* (Millar et al., 2017).

Let us rename the temperature of the first box as $T(t)$ for simplicity. Then, for a k -box model, the thermal response ODE is formally given by

$$\begin{cases} \frac{dS_i(t)}{dt} = \frac{1}{d_i} (q_i F(t) - S_i(t)), & 1 \leq i \leq k \\ T(t) = \sum_{i=1}^k S_i(t), \end{cases} \quad (3)$$

where $S_i(t)$ is the i^{th} thermal response, d_i is the i^{th} response timescale and q_i is the i^{th} equilibrium response. Table 1a provides a brief description of named parameters and units.

Whilst it is harder to interpret the impulse response formulation in terms of physical processes, its primary benefit is that each thermal response ODE can be solved independently, thereby avoiding the intricacies of solving a coupled system of ODE. Further, the timescale and equilibrium parameters d_i and q_i can be expressed in terms of the original boxes heat capacities C_i and heat transfer coefficients κ_i . Detailed derivations can be found in (Geoffroy et al., 2013).

Throughout this work, we will use as a reference SCM FaIRv2.0.0 (Leach et al., 2021) (for Finite amplitude Impulse Response), a recent update of a well established SCM (Millar et al., 2017; Smith et al., 2018), that offers a minimal level of structural complexity. FaIRv2.0.0 is effectively composed of 3 submodels: a gas cycle model, which converts emissions to concentrations, a radiative forcing model, which converts concentrations to radiative forcing, and a temperature response model, which converts radiative forcing into temperatures. The temperature response model of FaIRv2.0.0 exactly corresponds to the impulse response EBM described in (3). We refer to reader to the work of Leach et al. (2021) for a comprehensive presentation of FaIRv2.0.0. In the rest of the paper, we permit ourselves to drop v.2.0.0 and refer to the model as FaIR.

2.3 Gaussian processes

Gaussian processes (GPs) (Rasmussen & Williams, 2005) are a ubiquitous class of Bayesian priors over real-valued functions. They enjoy convenient closed-form expressions, principled uncertainty quantification, and cover a rich class of complex functions. As a result, they have been widely used in various nonlinear and nonparametric regression problems in geosciences (Camps-Valls et al., 2016).

We say that a real-valued stochastic process function $f(t)$ is a GP if any finite collection of its evaluations has a joint multivariate normal distribution. The parameters of this distribution are fully determined by the GP mean function $m(t) := \mathbb{E}[f(t)]$ and its covariance function $k(t, t') := \text{Cov}(f(t), f(t'))$. For example, for two inputs t_1, t_2 we have that

$$\begin{bmatrix} f(t_1) \\ f(t_2) \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} m(t_1) \\ m(t_2) \end{bmatrix}, \begin{bmatrix} k(t_1, t_1) & k(t_1, t_2) \\ k(t_2, t_1) & k(t_2, t_2) \end{bmatrix} \right). \quad (4)$$

Because this extends to any finite collection of inputs $t_1, \dots, t_n$, we say that GPs induce a distribution over the entire function $f(t)$. Figure 2 (a) shows examples of functions sampled from this distribution, referred to as *sample paths*. Formally, since this distribution is fully determined by $m(t)$ and $k(t, t')$, we write

$$f(t) \sim \text{GP}(m, k). \quad (5)$$

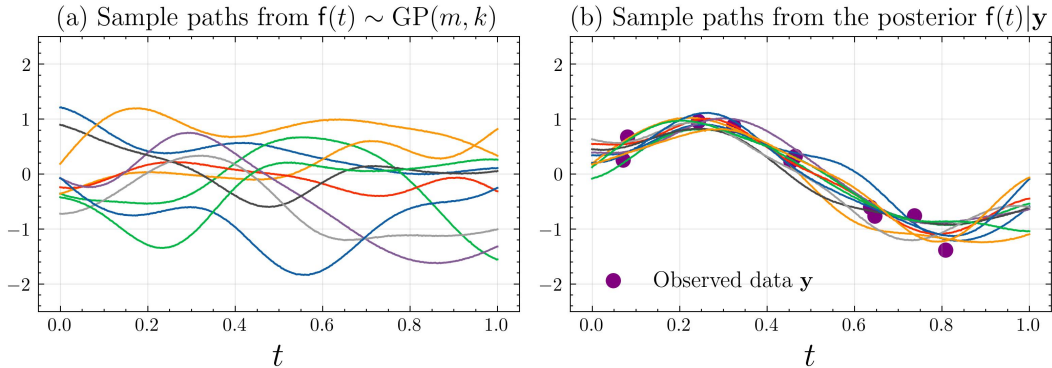


Figure 2: (a): 10 sample paths from $f(t) \sim \text{GP}(m, k)$ with $m(t) = 0$ and $k(t, t') = \sigma_f^2 \exp(-|t - t'|^2 / \ell)$, $\sigma_f^2 = 0.43$, $\ell = 0.17$. (b): 10 sample paths from the posterior $f(t)|\mathbf{y}$ obtained by updating the prior. The data $\mathbf{y}$ is sampled from $y = \sin(2\pi t) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 0.04)$.

When using GPs as a Bayesian prior over a function, the mean and covariance functions are typically user-specified. The mean $m(t)$ can be set as any function specifying the average behaviour of the function. The covariance function $k(t, t')$, commonly called *kernel*, is chosen to specify the correlation between $\mathbf{f}(t)$ and $\mathbf{f}(t')$ by computing a notion of similarity between t and t' . For example, a widely used family of kernels are the Matérn kernels (Stein, 1999), where the one-dimensional Matérn-1/2 kernel is given by

$$k(t, t') = \sigma_f^2 \exp\left(-\frac{|t - t'|}{\ell}\right), \quad (6)$$

with variance parameter σ_f^2 and lengthscale parameter ℓ . When t and t' are similar, i.e. $|t - t'| \ll \ell$, this kernel specifies a strong correlation between $\mathbf{f}(t)$ and $\mathbf{f}(t')$, i.e. $\text{Cov}(\mathbf{f}(t), \mathbf{f}(t'))/\sigma_f^2 \approx 1$. Conversely, if t and t' are dissimilar, i.e. $|t - t'| \gg \ell$, the kernel tends to zero and specifies almost no correlation between $\mathbf{f}(t)$ and $\mathbf{f}(t')$.

Other popular choices of kernel include the squared exponential kernel $k(t, t') = \sigma_f^2 \exp(|t - t'|^2/\ell)$ which specifies a smooth correlation, or the cosine kernel $k(t, t') = \sigma_f^2 \cos(|t - t'|/\ell)$ which specifies a periodic correlation (Rasmussen & Williams, 2005, Chapter 4). More broadly, kernels (and hence GPs) can also be defined over multidimensional inputs, for example by substituting $|t - t'|$ by the Euclidean distance between input vectors. When each dimension has a different lengthscale parameter, the kernel is referred to as an anisotropic, or automatic relevance determination, kernel (Rasmussen & Williams, 2005, Chapter 5.1).

Let $\mathbf{y} = [y_1 \ \dots \ y_n]^\top$ denote independent data samples at inputs $\mathbf{t} = [t_1 \ \dots \ t_n]^\top$ from the noisy data generating process $\mathbf{y} = \mathbf{f}(\mathbf{t}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. It is possible to inform the GP with this data, thereby updating the prior into a posterior. In particular, the posterior distribution gives rise to a *posterior GP*, with a posterior mean function $\bar{m}(t)$ and a posterior covariance function $\bar{k}(t, t')$,

$$\mathbf{f}(t) \mid \mathbf{y} \sim \text{GP}(\bar{m}, \bar{k}). \quad (7)$$

In other words, the posterior GP defines a new distribution over functions, parametrised by $\bar{m}(t)$ and $\bar{k}(t, t')$, which is informed by collected data. Figure 2 (b) shows examples of posterior sample paths, which are now concentrated around the collected data points.

A great advantage of GPs is that the posterior mean and covariance can be analytically computed following

$$\bar{m}(t) = m(t) + k(t, \mathbf{t})(\mathbf{K} + \sigma_\epsilon^2 \mathbf{I}_n)^{-1}(\mathbf{y} - m(\mathbf{t})), \quad (8)$$

$$\bar{k}(t, t') = k(t, t') - k(t, \mathbf{t})(\mathbf{K} + \sigma_\epsilon^2 \mathbf{I}_n)^{-1}k(\mathbf{t}, t), \quad (9)$$

where $\mathbf{I}_n$ denotes the identity matrix of size n , $k(t, \mathbf{t}) = [k(t, t_1) \ \dots \ k(t, t_n)]$, $k(\mathbf{t}, t) = k(t, \mathbf{t})^\top$ and $\mathbf{K} = k(\mathbf{t}, \mathbf{t}) = [k(t_i, t_j)]_{1 \leq i, j \leq n}$ is the covariance matrix.

We provide in Appendix A1 a detailed illustrated walk-through of GP regression, and refer the reader to (Rasmussen & Williams, 2005) for a comprehensive overview of GPs in machine learning.

3 FaIRGP

In this section, we present FaIRGP, a GP emulator for global mean surface temperature that builds upon the thermal response model from FaIR. We begin by motivating a Bayesian treatment of the radiative forcing and formulate a GP prior over the forcing. We show this modification naturally results in a stochastic formulation of the thermal response model, which admits a GP with a physically-informed covariance structure for solution. In addition, we show how this framework can seamlessly account for the climate internal variability. Finally, we provide closed-form expressions for the posterior distributions over temperature and forcing, which can readily be used for emulation.

In what follows, we adopt the following notational conventions: deterministic variables are denoted with serif font (Z), stochastic variables are denoted with sans-serif font ($\mathbf{Z}$) and vector/matrix versions are denoted in bold ($\mathbf{Z}$, $\mathbf{Z}$).

3.1 Epistemic uncertainty over the radiative forcing model

The top-of-atmosphere radiative forcing is the fundamental quantity used to describe Earth’s energy imbalance. The value of this forcing is primarily determined by the emissions of greenhouse gases and aerosols in the atmosphere. In FaIR, the radiative forcing term $F(t)$ drives the temperature response model (1). Having access to a reliable estimate of the forcing, and in particular, to the forcing response to greenhouse gas and aerosol emissions, is therefore critical to produce trustworthy emulated temperatures.

Computing an accurate estimate of the radiative forcing requires modelling how greenhouse gases and aerosols interact with radiations in the atmosphere, accounting for atmospheric adjustments and uncertainty in the measured parameters which may further complicate the calculation. The intricacy of such tasks has led to the development of forcing estimation methods which rely on simplifying assumptions. For example, in FaIR, the forcing model is formulated as a combination of logarithmic, linear and square-root terms of greenhouse gas and aerosol concentrations. Namely, let χ denote a given atmospheric agent (e.g. CO_2 , CH_4 , SO_2), the radiative forcing induced by χ is modeled as

$$F^\chi(t) = \alpha_{\log}^\chi \log \left(\frac{C^\chi(t)}{C_0^\chi} \right) + \alpha_{\text{lin}}^\chi (C^\chi(t) - C_0^\chi) + \alpha_{\text{sqrt}}^\chi \left(\sqrt{C^\chi(t)} - \sqrt{C_0^\chi} \right), \quad (10)$$

where $C^\chi(t)$ denotes the concentration in the atmosphere of agent χ , C_0^χ the agent concentration at pre-industrial period and $\alpha_{\log}^\chi, \alpha_{\text{lin}}^\chi, \alpha_{\text{sqrt}}^\chi$ are scalar sensitivity coefficients. The total radiative forcing is then obtained by combining the contribution of each agent

$$F(t) = \sum_{\chi} F^\chi(t). \quad (11)$$

This choice of forcing model has a physical motivation drawing from model calculations and empirical studies of temperature and historical trajectories of these concentrations (Manabe & Wetherald, 1967; Forster et al., 2007; Myhre et al., 2014), which have provided substantial evidence that the concentration-to-forcing relationship can be reasonably approximated by the combination of terms in (10). However, it is important to recall that such a functional form is primarily based on empirical physics, in contrast with the EBM which has a greater theoretical motivation. In fact, the values of the sensitivity coefficients $\alpha_{\log}^\chi, \alpha_{\text{lin}}^\chi, \alpha_{\text{sqrt}}^\chi$ can vary substantially depending on the data used and the fitting procedure (Leach et al., 2021).

From a statistical machine learning perspective, the concentration-to-forcing relationship is arguably the component of FaIR which is most akin to statistical modelling. Indeed, $\Phi(t) = [\log C^\chi(t) \quad C^\chi(t) \quad \sqrt{C^\chi(t)}]^\top$ can be interpreted as a feature engineering vector, where the features relevance have been identified by physics domain knowledge. The relationship (10) can then be rearranged as postulating a linear model $F^\chi(t) = \boldsymbol{\alpha}^\top \Phi(t) + \alpha_0$, where the coefficients $\boldsymbol{\alpha}$ are fitted against climate model data and the intercept α_0 is determined by the pre-industrial concentration C_0^χ .

Because the radiative forcing is a key driving quantity of the climate system, and therefore of the emulator, we argue it is critical to account for the uncertainty introduced by the FaIR model of radiative forcing. Specifically, this choice of a simplified model of forcing reflects uncertainty caused by the lack of knowledge about the phenomenon we wish to describe, referred to as *epistemic uncertainty* (Hüllermeier & Waegeman, 2021). Drawing from the latent force modelling literature (M. Alvarez et al., 2009), we propose to acknowledge the epistemic uncertainty over the forcing model functional form by treating the radiative forcing as an underspecified latent force. Specifically, following the work of M. A. Alvarez et al. (2013), we propose to account for the epistemic uncertainty over the forcing model through a Bayesian formalism and place a GP prior over the radiative forcing.

3.2 Radiative forcing as a Gaussian process

To specify a GP prior over the radiative forcing, we must first specify a choice of mean function. We propose to use the deterministic forcing function $F(t)$ as the mean function of our GP. This choice ensures that our prior will behave on average like the FaIR forcing model. Second, we must introduce a covariance function that specifies how different forcing levels should correlate as a function of some input variable.

The radiative forcing level is primarily determined by the emission rate of short-lived forcing agents and the cumulative emissions of long-lived forcing agents. Therefore, we anticipate a correlation of forcing levels for similar long-lived and short-lived forcing agent emissions trajectories. Conversely, for dissimilar emissions trajectories, we expect the resulting forcing levels to be fairly different, resulting in a lower correlation. For this reason, we propose to specify our covariance function as a function of long-lived forcing agents' cumulative emissions and short-lived forcing agents' emission rates.

Namely, let $E_r^\chi(t)$ and $E_c^\chi(t)$ be respectively the rate of emission and the cumulative emission of agent χ in the atmosphere at time t . Let $E_c^{\text{LL}}(t) = [E_c^{\chi_1}(t) \dots E_c^{\chi_d}(t)]^\top$ be the vector concatenation of long-lived agents $\{\chi_i\}_{i=1}^p$ cumulative emissions, and let $E_r^{\text{SL}}(t) = [E_r^{\chi_{p+1}}(t) \dots E_r^{\chi_d}(t)]^\top$ be the vector concatenation of short-lived agents $\{\chi_i\}_{i=p+1}^d$ rate of emission. The covariance function we introduce takes the form

$$K(t, t') = \rho(E(t) - E'(t')), \quad (12)$$

where $E(t) = [E_c^{\text{LL}}(t) \ E_r^{\text{SL}}(t)]^\top$ and $E'(t') = [E_c^{\text{LL}}(t') \ E_r^{\text{SL}}(t')]^\top$ are two vectors encapsulating the cumulative emissions of long-lived forcing agents and emission rates of short-lived forcing agents for two different scenarios at times t and t' , and ρ is a user-specified kernel.

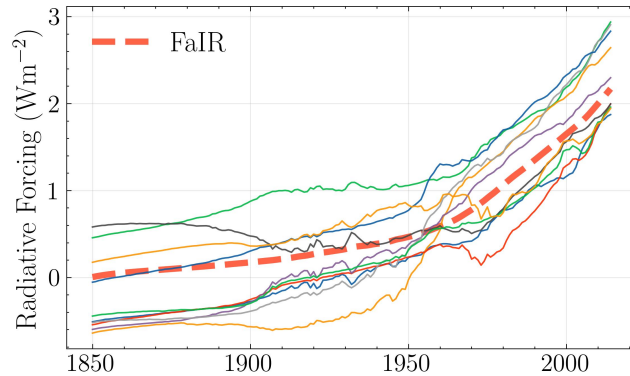


Figure 3: Example of sample paths from the radiative forcing prior $F(t, E(t)) \sim \text{GP}(F, K)$ over the 1850-2014 period. The FaIR forcing response (dashed red) corresponds to the average sample path. The covariance function is taken as a Matérn-3/2 covariance.

With these notations, we can then formally define the prior we specify over the forcing model as

$$F(t, E(t)) \sim \text{GP}(F, K), \quad (13)$$

where we emphasise that F is now a stochastic process by using a sans-serif notation. This Bayesian formulation of the forcing effectively introduces a distribution over the functional forms the forcing response can take, and each sample path drawn from this distribution may describe a different function. Figure 3 shows examples of sample paths from $F(t, E(t))$.

Taking the mean function as the forcing model F from (10) aligns sample paths with the FaIR forcing response on average. However, the probabilistic nature of this prior introduces a relaxation from exactly replicating the FaIR forcing response: sample paths can deviate from the mean, insofar as the forcing trajectories they describe maintain a correlation specified by $K(t, t')$ given their emission trajectories. This degree of freedom introduces a notion of variability over the forcing response, thereby allowing us to account for the epistemic uncertainty over the functional form of the FaIR forcing response.

To concretely understand how the covariance function affects F , recall that for GPs, $\text{Cov}(F(t, E(t)), F(t', E'(t'))) = K(t, t')$. Therefore, if $E(t)$ and $E'(t')$ describe similar emission trajectories, the covariance $K(t, t')$ will take larger values, and specify a strong correlation between $F(t, E(t))$ and $F(t', E'(t'))$. On the other hand, if $E(t)$ and $E'(t')$ describe very different emission trajectories, $K(t, t')$ will take values close to zero, and thus specify a low correlation between $F(t, E(t))$ and $F(t', E'(t'))$. Figure 4 shows an example of a covariance matrix between two scenarios with different CO_2 emissions trajectories.

The choice of kernel ρ is an important choice as it specifies how correlation in forcings should be encoded as a function of emissions. For example, choosing a dirac kernel $\rho(E(t) - E'(t')) = \delta(E(t) - E'(t'))$ makes the forcing levels corresponding to $E(t)$ and $E'(t')$ independent, except if the emission trajectories are identical. Another extreme is to choose $\rho(E(t) - E'(t')) = 1$, which causes the forcing values taken to correlate regardless of emissions. In between these extremes, a widely used class of kernels are the Matérn family (Stein, 1999). They specify an exponentially decaying covariance as a function of the distance between emission trajectories, and most notably allow to control the smoothness of the GP sample path. Appendix A2 provides detailed expressions of the Matérn covariance family and illustrations of their sample paths.

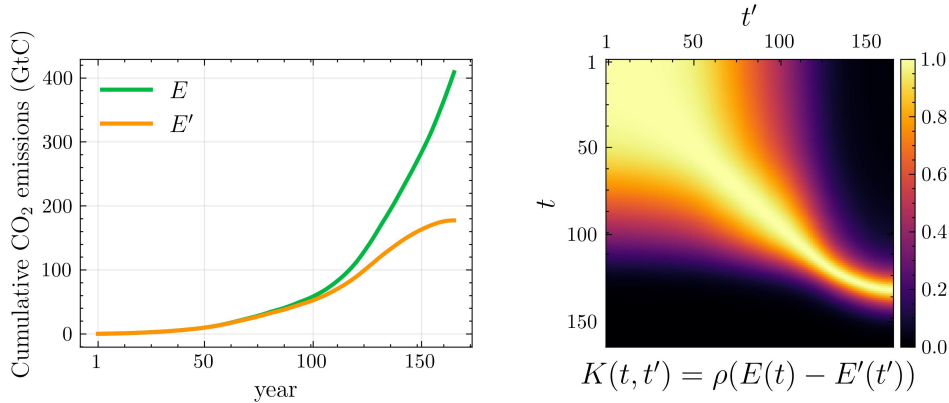


Figure 4: **Left** : Cumulative CO_2 emissions for a scenario with increasing emissions $E(t)$ and a scenario with reduced emissions $E'(t')$. **Right** : Cross-covariance matrix between emission scenarios for a Matérn-3/2 covariance, diagonals gives covariances for $t = t'$ and off-diagonals give covariances for $t \neq t'$. For $t, t' \leq 100$ the emission trajectories are similar, therefore $K(t, t')$ specifies a correlation between $F(t, E(t))$ and $F(t', E'(t'))$. For $t \geq 130$, $E(t)$ becomes dissimilar from the reduced emission scenario, therefore $K(t, t')$ specifies a low correlation. For $t' \geq 150$, $E'(t')$ stabilises to levels similar to $E(t \approx 130)$, therefore $K(t, t')$ specifies a correlation between $F(t', E'(t'))$ and $F(t \approx 130, E(t \approx 130))$.

A popular choice within the Matérn family is the squared exponential kernel⁴, which produces infinitely differentiable GP samples paths. We argue that this may not be faithful to real-world physical processes, which can exhibit abrupt changes. For this reason, we propose in this work to specify ρ as an anisotropic Matérn-3/2 kernel given by

$$C_{3/2}(E(t) - E'(t')) = \sigma_F^2 \left(1 + \sqrt{3 \sum_{i=1}^d \left(\frac{\delta_i}{\ell_i} \right)^2} \right) \exp \left(-\sqrt{3 \sum_{i=1}^d \left(\frac{\delta_i}{\ell_i} \right)^2} \right), \quad (14)$$

where δ_i is the difference between the i^{th} component of $E(t)$ and $E'(t')$, σ_F^2 is a variance parameter, $\ell_1, \dots, \ell_d$ are independent lengthscale parameters associated to each atmospheric agent $\chi_1, \dots, \chi_d$. The Matérn-3/2 covariance yields GP sample paths that are differentiable, thus allowing for smoothness, and yet permitting more abrupt changes compared to a squared exponential kernel.

In the remainder of the paper, we permit ourselves to drop notations and we denote the prior $F(t)$ as a function of time only for the sake of conciseness. However, we emphasise that $F(t)$, along with everything derived from it, fundamentally depends on emissions.

3.3 Thermal response model with GP forcing

We will now show that by combining the GP prior over the radiative forcing and the FaIR thermal response model, we obtain a GP prior over temperatures, which we name FaIRGP. Recall the thermal impulse response model presented in Section 2.2, described by a system of k independent linear first order ODEs

$$\frac{dS_i(t)}{dt} = \frac{1}{d_i} (q_i F(t) - S_i(t)), \quad (15)$$

where the thermal responses $S_i(t)$ are such that $T(t) = \sum_{i=1}^k S_i(t)$ is the global mean surface temperature.

We propose to substitute the deterministic forcing function $F(t)$ in (15) by its Bayesian counterpart $F(t)$. In doing so, we naturally induce a stochastic version of the thermal impulse response model. Namely, the resulting i^{th} stochastic thermal response, which we denote $S_i(t)$, will satisfy a differential equation forced by $F(t) \sim \text{GP}(m, K)$, and given by

$$dS_i(t) = \frac{1}{d_i} (q_i F(t) - S_i(t)) dt. \quad (16)$$

The resolution of this stochastic thermal response differential equation is similar to the resolution of the thermal response ODE. Namely, assuming $S_i(0) = 0$ at pre-industrial time, the solution to (16) takes the canonical form

$$S_i(t) = \frac{q_i}{d_i} \int_0^t F(s) e^{-(t-s)/d_i} ds. \quad (17)$$

However, because $F(t)$ is random, the solution is now a stochastic process. In particular, because $F(t)$ is a GP to which we apply a linear convolution with an exponential function $e^{-(t-s)/d_i}$, we can show that $S_i(t)$ must also be a GP, with its mean and covariance functions shaped by the form of this linear convolution.

⁴ to be exact, the squared exponential kernel corresponds to the limit of the Matérn family for an infinitely smooth kernel.

Namely, the i^{th} thermal response can be characterised as a GP with the following mean and covariance functions

$$\begin{cases} S_i(t) \sim \text{GP}(m_i, k_{ii}) \\ m_i(t) = \frac{q_i}{d_i} \int_0^t F(s) e^{-(t-s)/d_i} ds \\ k_{ii}(t, t') = \left(\frac{q_i}{d_i}\right)^2 \int_0^t \int_0^{t'} K(s, s') e^{-(t-s)/d_i} e^{-(t'-s')/d_i} ds ds'. \end{cases} \quad (18)$$

We observe that the mean function $m_i(t)$ exactly corresponds to the solution of the deterministic thermal response ODE (15). This is consistent with our expectation: because the forcing sample paths from $F(t)$ are specified to behave on average like the FaIR forcing response, the GP thermal response sample paths $S_i(t)$ should behave on average like the FaIR thermal response model. Further, the covariance $k_{ii}(t, t')$ is expressed as a function of the forcing prior covariance K , but also of the parameters of the EBM, d_i and q_i . As such, $k_{ii}(t, t')$ defines a physically-informed covariance structure that propagates the uncertainty over the forcing $F(t)$ — specified by our Bayesian prior — into uncertainty over the thermal response $S_i(t)$.

Under this model, the correlation $k_{ii}(t, t')$ between two stochastic thermal responses $S_i(t)$ and $S_i(t')$ is influenced by how fast the response of the i^{th} thermal box is, and by how much the radiative forcings trajectories leading to $S_i(t)$ and $S_i(t')$ did correlate. Indeed, the covariance in (18), expressed in integral form, specifies a stronger correlation if the past correlation $K(s, s')$ between stochastic forcing trajectories $(F(s))_{s \leq t}$ and $(F(s'))_{s' \leq t'}$ is greater. This aligns with the idea that similar forcing processes $F(s)$ and $F(s')$ should lead to similar thermal responses. Further, a fast thermal box (small d_i) primarily influences the integral with local correlations in forcings just before t and t' , reflecting the short-term impact of fast components of the climate system. Conversely, a slow thermal box (large d_i) accumulates forcing covariance over an extended period before t and t' , reflecting the long-term influence of slow components of the climate system. Figure 5 shows an example of cross-covariance matrices for a stochastic two-box thermal response model.

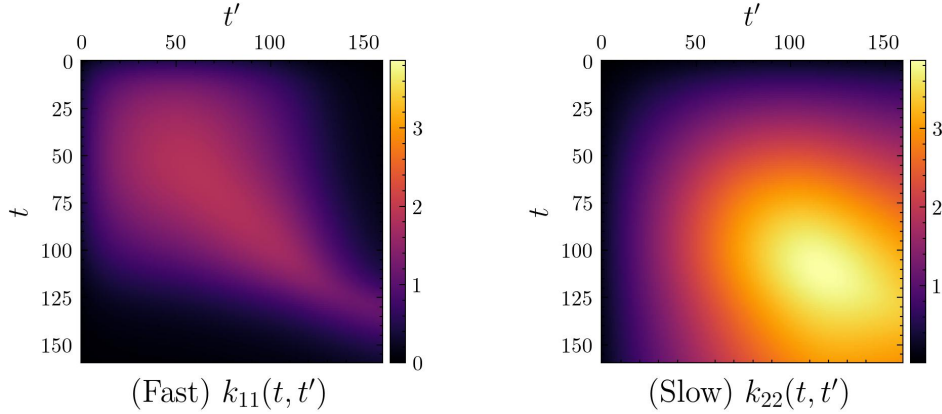


Figure 5: Cross-covariance matrices for the stochastic thermal responses $S_1(t)$ and $S_2(t)$ from a 2-box model under the two emission scenarios from Figure 4. **Left** : Cross-covariance matrix of a fast thermal box with $d_1 = 8$ years and $q_1 = 2 \text{ KW}^{-1}\text{m}^{-2}$. Only the forcings correlations at times right before (t, t') influence $k_{11}(t, t')$, hence the specified correlation is moderate and the form of the thermal box covariance matrix is similar to the form of the forcing covariance matrix $K(t, t')$ in Figure 4. **Right** : Cross-covariance matrix of a slow thermal box with $d_2 = 80$ years and $q_2 = 20 \text{ KW}^{-1}\text{m}^{-2}$. The covariance $k_{22}(t, t')$ integrates forcing correlations over an extended period, which specifies a stronger correlation, and additionally models the memory effect of the climate system: even after the two emission scenarios start to diverge for $t, t' \geq 130$, the cross-covariance remains important.

If we now define the global mean surface temperature as the sum of thermal response GPs $\mathbf{T}(t) = \sum_{i=1}^k \mathbf{S}_i(t)$, then we can show that $\mathbf{T}(t)$ must also be a GP. Namely, let us define the cross-covariance function between thermal responses $\mathbf{S}_i(t)$ and $\mathbf{S}_j(t')$ as

$$k_{ij}(t, t') := \text{Cov}(\mathbf{S}_i(t), \mathbf{S}_j(t')) = \frac{q_i q_j}{d_i d_j} \int_0^t \int_0^{t'} K(s, s') e^{-(t-s)/d_i} e^{-(t'-s')/d_j} ds ds', \quad (19)$$

which describes how the stochastic thermal responses from different boxes should correlate within our model. Then, the global mean surface temperature is a GP specified by the following mean and covariance functions

$$\begin{cases} \mathbf{T}(t) \sim \text{GP}(m_{\mathbf{T}}, k_{\mathbf{T}}), \\ m_{\mathbf{T}}(t) = \sum_{i=1}^k m_i(t), \\ k_{\mathbf{T}}(t, t') = \sum_{i=1}^k \sum_{j=1}^k k_{ij}(t, t'). \end{cases} \quad (20)$$

The mean function $m_{\mathbf{T}}(t)$ exactly corresponds to the FaIR temperature response model from (3). This is again consistent with the expectation that since $\mathbf{F}(t)$ behaves on average like the FaIR forcing model, our GP temperature response $\mathbf{T}(t)$ should behave on average like the FaIR temperature response. The covariance function $k_{\mathbf{T}}(t, t')$, which specifies the correlation between $\mathbf{T}(t)$ and $\mathbf{T}(t')$, includes the covariances of individual thermal boxes $k_{ii}(t, t')$, aligning with the intuition that if thermal responses correlate, so should global mean surface temperatures. However, we note $k_{\mathbf{T}}(t, t')$ also considers correlations between different thermal boxes $k_{ij}(t, t')$ for $i \neq j$. Therefore, if certain components of the climate system exhibit correlated responses across different timescales, this contributes positively to the correlation between $\mathbf{T}(t)$ and $\mathbf{T}(t')$ within our model.

By specifying a GP prior over the radiative forcing, we have obtained a GP prior over the temperature that uses the FaIR thermal response model as its backbone, which we name FaIRGP. Using FaIRGP serves as a principled measure of epistemic uncertainty over the emulator design. In particular, the integral form of its covariance function allows us to account for past trajectories, thereby capturing the climate system memory effect, and producing robust uncertainty estimates. In addition, as we will see in Section 3.5, FaIRGP can go beyond a standard impulse response model by learning from data using standard GP regression techniques. We emphasise that whilst we abuse notations for conciseness, the forcing prior $\mathbf{F}(t)$ is effectively a function of emissions (or cumulative emissions) through its covariance function $\rho(E(t) - E'(t'))$. Therefore, $\mathbf{T}(t)$ is also a function of emissions, and its covariance function can be understood as “ $k_{\mathbf{T}}(E(t), E'(t'))$ ”.

3.4 Accounting for climate internal variability

An important component of the climate system is its internal variability, which integrates the effects of weather phenomena — typically operating on the scale of days — into elements of the climate system, such as the ocean, cryosphere and land vegetation — which rather operate on the scale of months, years or decades (Hasselmann, 1976).

The climate internal variability can classically be modelled in a k -box model by introducing a white noise forcing disturbance over the uppermost box, i.e. the atmosphere and land surface box (Cummins et al., 2020). Formally, let $\mathbf{B}(t)$ be the standard one-dimensional Brownian motion and let $\mathbf{X}(t)$ denote a stochastic version of the k -box model temperatures from (1). The temperature response model with internal variability is given by

$$d\mathbf{X}(t) = \mathbf{A}\mathbf{X}(t) dt + \mathbf{b}\mathbf{F}(t) dt + \sigma \mathbf{b} d\mathbf{B}(t), \quad (21)$$

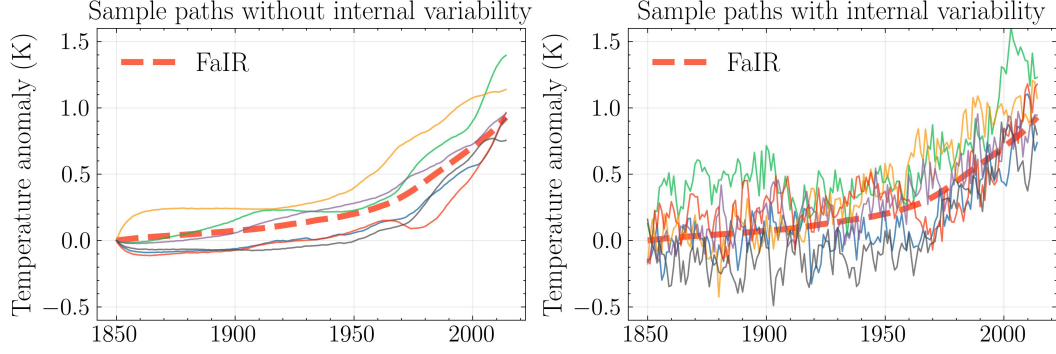


Figure 6: **Left:** Example of sample paths from the FaIRGP prior $T(t) \sim \text{GP}(m_T, k_T)$ without internal variability over the 1850-2014 period. **Right:** Example of sample paths from the FaIRGP prior with internal variability $T(t) \sim \text{GP}(m_T, k_T + \sigma^2 \gamma_T)$ over the same period. The FaIR temperature response (dashed red) corresponds in both cases to the average sample path.

where $\sigma > 0$ is a variance term that controls the amplitude of the white noise, and we recall that $\mathbf{b}$ has zero everywhere but its first entry, therefore the white noise disturbance is only applied to the uppermost box. When the radiative forcing $F(t)$ is deterministic, this is equivalent to adding red noise onto the global mean surface temperature, or equivalently, modelling it as an Ornstein-Uhlenbeck process⁵. The corresponding diagonalised impulse response system is given by

$$dS_i(t) = \frac{1}{d_i} (q_i F(t) - S_i(t)) dt + \sigma \frac{q_i}{d_i} dB(t), \quad (22)$$

where derivations are detailed in Appendix B3. If we now again substitute the deterministic forcing $F(t)$ with the GP forcing $F(t)$, it can be shown that the long time regime solution to this stochastic impulse response system can be expressed as

$$S_i(t) \sim \text{GP}(m_i, k_{ii} + \sigma^2 \gamma_i), \quad (23)$$

where $\gamma_i(t, t')$ is an exponential, or Matérn-1/2, kernel function given by

$$\gamma_i(t, t') = \frac{q_i^2}{2d_i} \exp\left(-\frac{|t - t'|}{d_i}\right). \quad (24)$$

Therefore, FaIRGP can easily model the internal variability of the climate system simply by modifying its covariance structure. We observe that accounting for internal variability essentially corresponds to adding an independent autocorrelated noise process $\xi_i(t) \sim \text{GP}(0, \sigma^2 \gamma_i)$ to the thermal response GP obtained in (18).

Ultimately, we sum the thermal response together to obtain a surface temperature GP $T(t) = \sum_{i=1}^k S_i(t)$, which in the long time regime takes the form

$$\begin{cases} T(t) \sim \text{GP}(m_T, k_T + \sigma^2 \gamma_T) \\ \gamma_T(t, t') = \sum_{i=1}^k \nu_i \gamma_i(t, t'), \end{cases} \quad (25)$$

where ν_i is a dimensionless weight determined by q_i, d_i which accounts for cross-covariances between thermal responses internal variabilities. Its detailed expression can be found in Appendix B3.

⁵ In the literature, it is common to encounter its discrete time analogue, the autoregressive process of order 1 AR(1).

Therefore, we can account for the climate internal variability simply by adding an independent noise process $\xi_{\text{T}}(t) = \sum_{i=1}^k \xi_i(t) \sim \text{GP}(0, \sigma^2 \gamma_{\text{T}})$ to the temperature response GP obtained in (20). Figure 6 shows examples of sample paths from the FaIRGP prior with and without accounting for internal variability.

3.5 Posterior distribution over temperature and radiative forcing

In FaIRGP, the surface temperature response and the radiative forcing model can both be informed by global temperature data. Using standard GP regression techniques (Rasmussen & Williams, 2005), FaIRGP can learn from data how to deviate from its backbone impulse response model to best account for actual temperature simulated by a climate model. The resulting model can then be used to emulate future temperatures, benefiting from both the robustness of its prior and the flexibility of GP regression. As we will demonstrate in Section 5, we can inform FaIRGP with historical observations and climate model data.

Assume we are under a fixed emission scenario e.g. working with historical emissions only. For timesteps $t_1 < \dots < t_n$, suppose we simulate global mean surface temperatures $T_1, \dots, T_n$, and also have access to greenhouse gas and aerosol emissions data $E_1, \dots, E_n \in \mathbb{R}^d$, where d corresponds to the number of atmospheric agents $\chi_1, \dots, \chi_d$. We concatenate this data into $\mathbf{t} = [t_1 \dots t_n]^\top$, $\mathbf{E} = [E_1 \dots E_n]^\top$ and $\mathbf{T} = [T_1 \dots T_n]^\top$.

With notation abuse, the kernel k_{T} allows to specify how the FaIRGP prior correlates between times t_i and t_j by computing the similarity between emission data, i.e. $k_{\text{T}}(t_i, t_j) = k_{\text{T}}(E_i, E_j)$, because the forcing prior kernel ρ is a function of emissions. The internal variability kernel γ_{T} allows us to specify how the additive internal variability noise process correlates between times t_i and t_j by computing $\gamma_{\text{T}}(t_i, t_j)$. By extending this to the entire dataset, we can define the covariance matrices induced by k_{T} and γ_{T} over $\mathbf{E}$ and $\mathbf{t}$ following

$$\mathbf{K} = k_{\text{T}}(\mathbf{E}, \mathbf{E}) = [k_{\text{T}}(E_i, E_j)]_{1 \leq i, j \leq n} \quad (26)$$

$$\mathbf{\Gamma} = \gamma_{\text{T}}(\mathbf{t}, \mathbf{t}) = [\gamma_{\text{T}}(t_i, t_j)]_{1 \leq i, j \leq n}. \quad (27)$$

Intuitively, $\mathbf{K}$ captures how much the values taken by the stochastic global mean temperature response from the FaIRGP prior will correlate between different times. Similarly, $\mathbf{\Gamma}$ captures how much the values taken by the internal variability process ξ_{T} correlate between different time steps. Combining them into $\mathbf{K} + \sigma^2 \mathbf{\Gamma}$, we obtain a covariance matrix that captures the total correlation we have between different times for the values taken by the FaIRGP prior with internal variability.

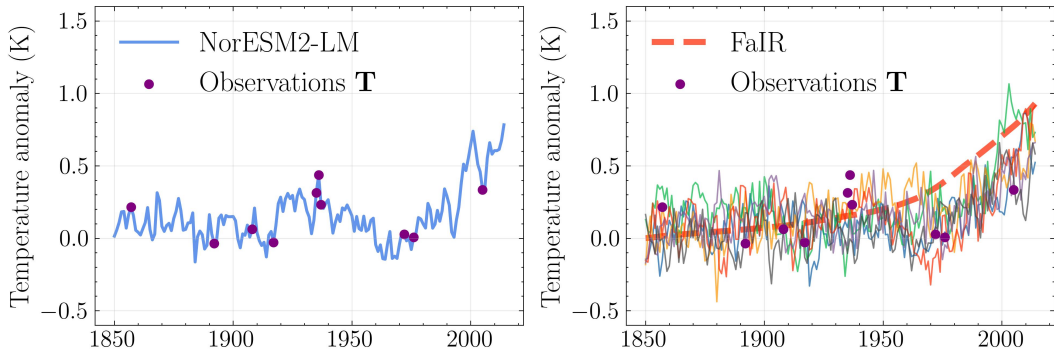


Figure 7: **Left:** Historical global mean surface temperature anomaly simulated with NorESM2-LM from which we randomly extract 10 data points (purple). **Right:** Example of sample paths from the FaIRGP posterior $\text{T}(t)|\mathbf{T}$ conditioned on the data. In contrast with the FaIRGP prior sample paths (see Figure 6), posterior sample paths are more concentrated around data points and deviate from the FaIR temperature response (dashed red).

Because it captures both the effect of internal variability and of the global mean temperature response to changes in emissions, the covariance matrix $\mathbf{K} + \sigma^2 \mathbf{\Gamma}$ is well suited to capture the correlation structure we get from global mean surface temperature data $\mathbf{T}$ simulated with a climate model. Using this new covariance matrix, we can then apply the update rules from (8) and (9) to update our prior over $\mathbf{T}(t)$ with climate model data $\mathbf{T}$, and obtain a posterior GP over global mean surface temperatures given by

$$\mathbf{T}(t) \mid \mathbf{T} \sim \text{GP}(\bar{m}_{\mathbf{T}}, \bar{k}_{\mathbf{T}}), \quad (28)$$

where the posterior mean $\bar{m}_{\mathbf{T}}(t)$ and the posterior covariance $\bar{k}_{\mathbf{T}}(t, t')$ have the following closed-form expressions

$$\underbrace{\bar{m}_{\mathbf{T}}(t)}_{\text{posterior mean}} = \underbrace{m_{\mathbf{T}}(t)}_{\text{prior mean}} + \underbrace{k_{\mathbf{T}}(t, \mathbf{t})(\mathbf{K} + \sigma^2 \mathbf{\Gamma})^{-1}(\mathbf{T} - m_{\mathbf{T}}(\mathbf{t}))}_{\text{posterior correction}} \quad (29)$$

$$\underbrace{\bar{k}_{\mathbf{T}}(t, t')}_{\text{posterior covariance}} = \underbrace{k_{\mathbf{T}}(t, t')}_{\text{prior covariance}} - \underbrace{k_{\mathbf{T}}(t, \mathbf{t})(\mathbf{K} + \sigma^2 \mathbf{\Gamma})^{-1}k_{\mathbf{T}}(\mathbf{t}, t')}_{\text{posterior correction}}. \quad (30)$$

The posterior mean $\bar{m}_{\mathbf{T}}(t)$ explicitly corrects the prior mean $m_{\mathbf{T}}(t)$, which we recall corresponds to the FaIR deterministic temperature response model. The correction introduced by the posterior mean in (29) is a data-driven correction term, informed by the temperature data $\mathbf{T}$, times $\mathbf{t}$, and the emissions $\mathbf{E}$ (which are used to compute the covariance matrix $\mathbf{K}$). Intuitively, the correction term regresses the discrepancy between the simulated temperature data and the FaIR temperature response onto the emission data. Thus, it learns how much to deviate from the FaIR temperature response model to better fit climate model data, given emissions data. Similarly, the posterior covariance $\bar{k}_{\mathbf{T}}(t, t')$ explicitly adjusts the prior covariance $k_{\mathbf{T}}(t, t')$ with a data-driven correction term. Intuitively, this correction term acknowledges the new information we have updated our model with, such that when predicting over emission trajectories similar to the ones from the training data, the variability of the FaIRGP global mean surface temperature response is reduced. Figure 7 shows examples of sample paths from the FaIRGP posterior when updated with simulated data of global mean surface temperature.

This posterior can then be used to emulate temperatures by making predictions at future time steps. A major interest of this formulation is that whilst the emulator hinges on the robustness of the energy balance model for its prior, it can also benefit from the flexibility of GP regression by informing it with data in its posterior, thereby going beyond a simple impulse response model. In addition, the GP approach enjoys full probabilistic tractability, with the analytical expression of the posterior probability distribution provided in Appendix B5. This means we can analytically compute the mean, the 95% credible range and quantiles of the predicted surface temperature anomaly as shown in Figure 8.

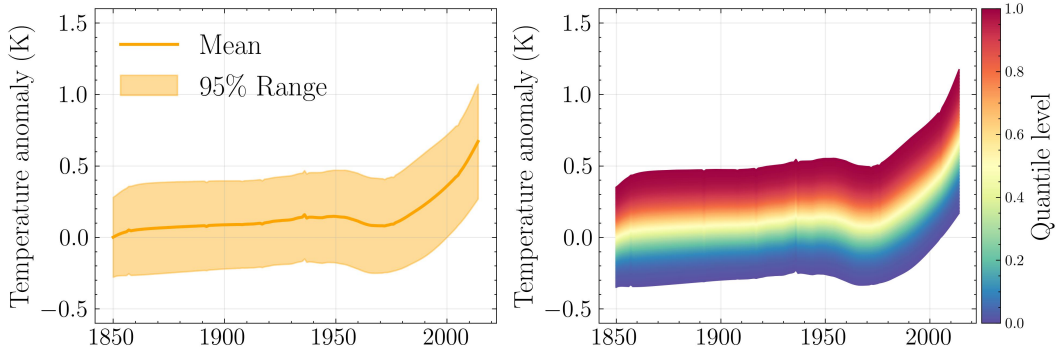


Figure 8: **Left:** FaIRGP posterior mean and 95% credible range. **Right:** FaIRGP posterior quantiles. The posterior distribution $\mathbf{T}(t) \mid \mathbf{T}$ is conditioned on the same 10 random data points used in Figure 7.

Using the same data, we can also update the GP prior over the radiative forcing with temperature data to formulate a posterior distribution over the forcing. Indeed, because the radiative forcing $F(t)$ is also GP, it is jointly Gaussian with the temperature $T(t)$, and thus admits a cross-covariance function that we can derive following

$$k_{FT}(t, t') := \text{Cov}(F(t), T(t')) = \sum_{i=1}^k \frac{q_i}{d_i} \int_0^{t'} K(t, s') e^{-(t'-s')/d_i} ds'. \quad (31)$$

Intuitively, this cross-covariance function plays a mediation role between the temperature and radiative forcing domains: it translates how variations in surface temperatures correlate with variations in radiative forcing under our model. Using this cross-covariance, it is therefore possible to update our prior over $F(t)$ with temperature data $\mathbf{T}$, resulting in a posterior distribution over the radiative forcing given by

$$\begin{cases} F(t) | \mathbf{T} \sim \text{GP}(\bar{m}_F, \bar{k}_F) \\ \bar{m}_F(t) = F(t) + k_{FT}(t, \mathbf{t})(\mathbf{K} + \sigma^2 \mathbf{\Gamma})^{-1}(\mathbf{T} - m_{\mathbf{T}}(\mathbf{t})) \\ \bar{k}_F(t, t') = K(t, t') - k_{FT}(t, \mathbf{t})(\mathbf{K} + \sigma^2 \mathbf{\Gamma})^{-1}k_{FT}(\mathbf{t}, t'). \end{cases} \quad (32)$$

As for temperatures, the posterior mean $\bar{m}_F(t)$ corrects the FaIR forcing model $F(t)$ with a data-informed correction term. The mean correction term again regresses out the discrepancy between $\mathbf{T}$ and the FaIR temperature response onto emissions, but converts this temperature discrepancy signal into a forcing discrepancy signal through the cross-covariance function k_{FT} . The uncertainty quantification given by the posterior covariance $\bar{k}_F(t, t')$ also adjusts the forcing prior covariance $K(t, t')$ to reduce the variability of the stochastic forcing response when predicting over emission trajectories similar to the ones from the training data $\mathbf{E}$. Figure 9 shows examples of sample paths from the forcing posterior when updated with global mean surface temperature data.

Finally, we can verify that solving the thermal response stochastic differential equation for the posterior forcing $F(t) | \mathbf{T}$ yields the posterior temperature $T(t) | \mathbf{T}$ as the solution. Therefore, the forcing posterior and the temperature posterior are consistent with each other. It is important to note that the posterior forcing correction proposed in (32) may correct the forcing prior with sources of temperature discrepancies that are not imputable to the forcing. Indeed, beyond differences between the FaIR and the climate model forcing response, other aspects of the climate system, such as the climate feedback parameter or the ocean inertia, can also contribute to discrepancies between $\mathbf{T}$ and the FaIR temperature response. Because the posterior correction reports all of the temperature response discrepancies onto the forcing, there is a risk of simply compensating errors, which needs cautious examination.

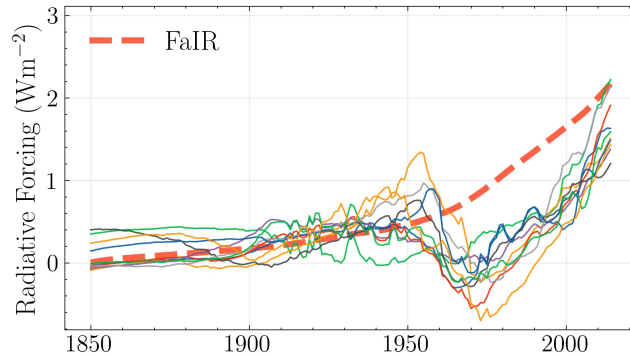


Figure 9: Example of sample paths from the radiative forcing posterior $F(t)|\mathbf{T}$ conditioned on the same 10 random data points used in Figure 7. In contrast with the forcing prior sample paths (see Figure 3), the posterior sample paths are informed by temperature data and deviate from the FaIR forcing response (dashed red).

3.6 Parameters tuning

Access to a closed-form probability density for the prior allows us to tune the model parameters using a maximum likelihood strategy. To illustrate this, consider again the data $\mathbf{t}, \mathbf{E}, \mathbf{T}$ from the previous section. We can derive a closed-form expression for the *marginal log-likelihood* $\log p(\mathbf{T}|\mathbf{E}, \mathbf{t})$ of the data under our FaIRGP prior, where the exact formula can be found in Appendix B5. The marginal log-likelihood represents (the logarithm of) how likely is the simulated data $\mathbf{t}, \mathbf{E}, \mathbf{T}$ to occur under the FaIRGP prior we have specified. Hence, a well-specified prior yields a greater marginal log-likelihood. This motivates a tuning strategy for the model parameters that seeks parameter values that maximise $\log p(\mathbf{T}|\mathbf{E}, \mathbf{t})$.

This can be achieved by using a gradient-based optimisation algorithm to maximise the marginal log-likelihood $\log p(\mathbf{T}|\mathbf{E}, \mathbf{t})$ with respect to the model parameters. Namely, let $\theta = \{\sigma, \sigma_F, \ell_1, \dots, \ell_d\}$ be the internal variability standard deviation and the kernel $K(t, t')$ parameters we wish to tune. We can compute the direction in which a small shift in θ can be taken to increase $\log p(\mathbf{T}|\mathbf{E}, \mathbf{t})$. This direction corresponds to the gradient of the marginal log-likelihood with respect to the model parameters. By iteratively performing this operation, each *gradient step* brings us closer to an optimal choice of model parameters θ that maximises $\log p(\mathbf{T}|\mathbf{E}, \mathbf{t})$. This procedure is known in machine learning as *gradient descent* and illustrated in Algorithm 1. The parameter $\eta > 0$ is typically a small positive number called the learning rate which controls the size of the steps taken in the direction of the gradient at each step. A detailed discussion of its role and of the parameter tuning procedure with gradient descent is provided in Appendix B5.

Algorithm 1 Maximum likelihood tuning of θ with gradient descent

```

1: Input:  $\mathbf{t}, \mathbf{E}, \mathbf{T}, \theta_0, \eta > 0, \text{n\_steps}$ 
2: for  $t \in \{1, \dots, \text{n\_steps}\}$  do
3:   Compute  $\log p(\mathbf{T}|\mathbf{E}, \mathbf{t})$ 
4:   Take  $\theta_t \leftarrow \theta_{t-1} + \eta \nabla_{\theta} \log p(\mathbf{T}|\mathbf{E}, \mathbf{t})$ 
5: end for
6: Return:  $\theta_{\text{n\_steps}}$ 

```

Finally, note that whilst we limit ourselves to tuning the GP and internal variability parameters in this work, we may also want to tune FaIR parameters such as d_i, q_i , the forcing model coefficients $\alpha_{\log}^x, \alpha_{\text{lin}}^x, \alpha_{\text{sqrt}}^x$ or even the temperature response parameters $\mathbf{A}, \mathbf{b}$. This can be achieved using the same procedure by including these parameters in θ .

3.7 Spatial FaIRGP

Whilst informative, the global mean surface temperature fails to capture the difference in exposure of world regions to a changing climate. It is therefore a necessity to invest efforts in obtaining spatially-resolved climate projections. We propose an extension of FaIRGP to emulate spatially-resolved surface temperature maps that grounds itself on a *pattern scaling* prior.

Pattern scaling is a well-established technique to model changes in local surface temperature as a function of changes in global mean surface temperature (Santer et al., 1990; Mitchell et al., 1999; Huntingford & Cox, 2000; Tebaldi & Arblaster, 2014). It consists of a simple scaling of a fixed spatial pattern by global mean temperature changes. Whilst very simple, such approach is supported by findings that regional changes in temperature scale robustly with global temperature (S. I. Seneviratne et al., 2016), and has been successfully used in existing spatial temperature emulation models (Beusch et al., 2020b, 2020a, 2022; Link et al., 2019; Goodwin et al., 2020).

Formally, pattern scaling assumes that for a given spatial location x , the local temperature response $T(x, t)$ is given by

$$T(x, t) = \beta^{(x)}T(t) + \beta_0^{(x)}, \quad (33)$$

with regression coefficient $\beta^{(x)}$ and intercept $\beta_0^{(x)}$. These coefficients are typically obtained by fitting independent local linear regression models of global temperature $T(t)$ onto local temperature $T(x, t)$. If we substitute the deterministic global temperature response $T(t)$ with its GP version $\mathsf{T}(t)$, we obtain a local FaIRGP prior temperature response given by

$$\mathsf{T}(x, t) \sim \text{GP} \left(\beta^{(x)}m_{\mathsf{T}} + \beta_0^{(x)}, (\beta^{(x)})^2k_{\mathsf{T}} \right), \quad (34)$$

which admits the local pattern scaling temperature response for its mean, and a locally scaled covariance.

As for the global response, this local prior can be updated with local temperature data to obtain a posterior temperature response at spatial location x . This gives FaIRGP the flexibility to learn from data how to deviate from a fixed spatial pattern, in order to better account for climate model simulated data.

4 Experimental Setup

In this section, we introduce the dataset used in our emulation experiments, the baseline models we benchmark FaIRGP against, and the evaluation metrics. Code and data to reproduce experiments are publicly available⁶.

4.1 Dataset description

The data is obtained from the ClimateBench v1.0 (Watson-Parris et al., 2022) climate emulation benchmark dataset. ClimateBench v1.0 proposes a curated dataset of local annual mean surface temperature ($\sim 2^\circ$ horizontal resolution), paired with annual emissions for four of the main anthropogenic forcing agents: carbon dioxide (CO_2), methane (CH_4), sulfur dioxide (SO_2) and black carbon (BC). The temperature data is generated from the latest version of the Norwegian Earth System Model (NorESM2) (Seland et al., 2020) as part of the sixth coupled model intercomparison project (CMIP6) (Eyring et al., 2016). The emission data is constructed from the Community Emissions Data System (Hoesly et al., 2018), and corresponds to the input data used to drive the original NorESM2-LM simulations.

The dataset we use includes paired time series of global annual emissions and local annual mean surface temperature for five experiments. We include the CMIP6 *historical* experiment, and four experiments corresponding to different possible shared socio-economic pathways (SSPs) (Riahi et al., 2017) from the ScenarioMIP protocol (O'Neill et al., 2016): *SSP126*, *SSP245*, *SSP370* and *SSP585*. These scenarios are designed to span a range of emissions trajectories corresponding to plausible mitigation scenarios and end-of-century forcing possibilities. Table 2 gives a description of the experiments and the periods covered.

In all experiments, we use as inputs for FaIRGP the global annual cumulative emissions of CO_2 and global annual emissions of CH_4 , SO_2 and BC. In the experiments of Section 5, we focus on emulating global annual mean surface temperature. Therefore, we preprocess the temperature data by computing the global annual mean surface temperatures using an area-weighted mean of the local annual mean surface temperatures data. In Section 6 we focus on emulating local annual mean surface temperatures, thus no preprocessing is required.

⁶<https://github.com/shahineb/FaIRGP>

Table 2: Details of the experiments from ClimateBench v1.0 (Watson-Parris et al., 2022) used in the dataset.

Experiment	Period	Description
<i>historical</i>	1850-2014	A simulation using historical emissions of all forcing agents designed to recreate the historically observed climate
<i>SSP126</i>	2015-2100	A high ambition scenario designed to produce significantly less than 2° warming by 2100
<i>SSP245</i>	2015-2100	A medium forcing future scenario aligned with current trajectories.
<i>SSP370</i>	2015-2100	A medium-high forcing scenario with high emissions of near-term climate forcers such as methane and aerosols
<i>SSP585</i>	2015-2100	A high forcing scenario leading to a large 8.5Wm^{-2} forcing in 2100

Our experiments are effectively conducted on data from a single climate model, and therefore do not reflect the diversity of response of different climate models to a common scenario. However, we argue they still provide a valid assessment of an emulator’s ability to reproduce climate models outputs. Indeed, whilst different climate models exhibit variations in their response to a similar scenario, it has been shown that SCMs are still capable of capturing the diversity of forcing responses spanned (Smith et al., 2018). Therefore, there are reasons to believe that insights drawn from experiments where emulators are tuned on NorESM2 data should carry over to experiments where emulators are tuned on data from different climate models.

4.2 Baseline emulators

To develop intuition on the benefit of combining FaIR with a GP, we propose to compare our model to the temperature anomaly projections obtained with (i) FaIR only and (ii) with a purely data-driven plain GP regression model only. By comparing to the emulation with FaIR, we hope to highlight the benefit that there is to combine the flexibility of a data-driven approach to an impulse response model, and answer the question “*Does introducing a GP in FaIR add value?*”. On the other hand, by comparing FaIRGP with a plain GP regression model, we hope to demonstrate the importance of having a robust physical prior underlying a data-driven approach, and answer the question “*Does a FaIR prior in a GP add value?*”. Both baseline models take as inputs greenhouse gas and aerosol emissions data, and predict annual mean surface temperature anomalies with respect to the preindustrial period.

For FaIR, we use parameter values that have been tuned against NorESM2 simulations (Leach et al., 2021). Therefore, the FaIR model we use in experiments is fully deterministic with fixed parameter values.

The plain GP model is entirely physics-free, and is specified with a zero mean and an anisotropic Matérn-3/2 kernel. This is exactly the same kernel ρ we use in our prior over the forcing in FaIRGP, which takes as inputs the global annual cumulative emissions of CO_2 and global annual emissions of CH_4 , SO_2 and BC. This kernel is to contrast with the physics-informed kernel k_T used for temperatures in FaIRGP. As it is standard practice in GP regression, we assume Gaussian noise with constant variance over the data. The noise variance and the kernel parameters are tuned using marginal likelihood maximisation by gradient descent.

4.3 Evaluation metrics

We use two kinds of metrics to evaluate the predicted probabilistic surface temperatures: deterministic metrics, which compare only the posterior mean prediction to the

ClimateBench temperature data, and probabilistic metrics, which evaluate the entire posterior probability distribution against ClimateBench temperature data. Table 3 provides a brief description of the metrics used.

Table 3: Details on evaluation metrics used in experiments.

	Notation	Description	Best when
<i>Deterministic</i>	RMSE	Root mean square error	close to 0
	MAE	Mean absolute error	close to 0
	Bias	Mean of (prediction – groundtruth)	close to 0
<i>Probabilistic</i>	LL	Log-likelihood	higher
	Calib95	95% calibration score	close to 95%
	CRPS	Continuous ranked probability score	close to 0

The log-likelihood (LL) score evaluates the log probability of the groundtruth temperature data under the predictive posterior probability distribution predicted by our model. Therefore, greater LL means that the predicted distribution is a good fit for the test data. The 95% calibration score (Calib95) computes the percentage of groundtruth temperature data that fall within the 95% credible interval of the predicted posterior distribution. Therefore, if the predicted probability distribution is well calibrated, this score should be close to 95%.

The Continuous Ranked Probability Score (CRPS) is an extension of the RMSE for probability distributions, which measures the discrepancy between the cumulative distribution functions (cdfs) of the predicted posterior distribution and the empirical cdf of the test data. Namely, if Φ denotes the cdf of the predicted posterior distribution and y is a scalar test data point, the CRPS between Φ and y is given by

$$\text{CRPS}(\Phi, y) = \int_{-\infty}^{+\infty} (\Phi(x) - \mathbb{1}_{[y, +\infty)}(x))^2 dx, \quad (35)$$

where $\mathbb{1}_{[y, +\infty)}(x)$ is the indicator function which takes value 0 if $x < y$ and 1 if $x \geq y$.

When evaluating predictions of local annual mean surface temperatures, we compute global mean metrics using an area-weighted mean that accounts for the decreasing grid-cell area toward the poles. Namely, we take

$$\langle \text{Score} \rangle = \frac{1}{N_{\text{lon}} \sum_{i=1}^{N_{\text{lat}}} w_i} \sum_{i=1}^{N_{\text{lat}}} \sum_{j=1}^{N_{\text{lon}}} w_i \text{Score}_{i,j}, \quad (36)$$

where $w_i = \cos(\text{lat}(i))$.

5 Application: global surface temperatures emulation

In this section, we benchmark FaIRGP against baseline models for the task of emulating mean global surface temperatures over SSP scenarios. We first briefly illustrate how the model concretely applies to an example. Then, we evaluate the emulated global temperatures trajectories when the model is trained on historical and SSP data, and when trained on historical data only. Finally, we probe the potential of FaIRGP to infer the top-of-atmosphere radiative forcing.

5.1 Example of global temperature emulation with FaIRGP

We begin with an example, and propose in Figure 10 a concrete illustration of global mean surface temperature emulation for *SSP370* with FaIRGP, that parallels the posterior mean and covariance expressions from (29) and (30).

The prior temperature response over *SSP370* admits as its mean the FaIR response, but with an additional layer of uncertainty quantification that arises from the GP. We then use temperature data from scenarios $\mathcal{D}_{\text{train}} = \{\text{historical}, \text{SSP126}, \text{SSP245}, \text{SSP585}\}$ as training data to learn a posterior correction. The posterior correction allows deviation from the prior — both in mean and variance — to provide a better fit to the training data. Finally, by linearly adding the posterior correction to the prior, we obtain a posterior temperature response over *SSP370*.

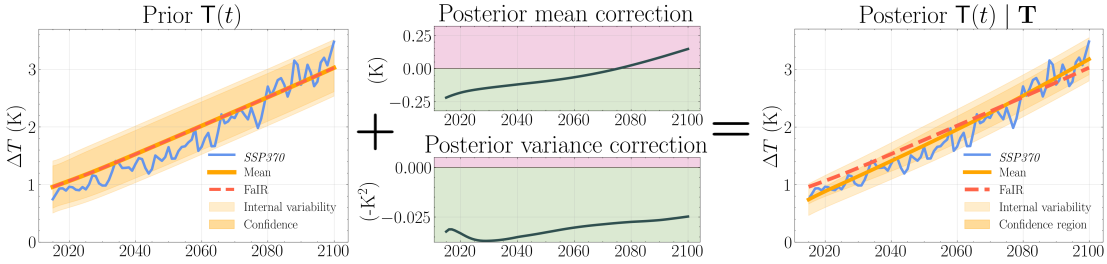


Figure 10: Emulation of NorESM2-LM *SSP370* global mean surface temperature anomaly. **Left:** FaIRGP prior temperature response over *SSP370*. The prior mean response exactly corresponds to the FaIR temperature response, but with prior uncertainty quantification. **Middle:** Posterior correction for the mean and variance that learns from data how to deviate from the prior. The pink area corresponds to deviations toward greater values and the green area corresponds to deviations toward smaller values. **Right:** FaIRGP posterior temperature response when updated with the other scenarios. The posterior response provides a sounder fit to the climate model data, with better calibrated uncertainty. Shading indicates a 5%-95% range.

5.2 Shared socio-economic pathways emulation

In this experiment, we consider the experiment dataset given by $\mathcal{D} = \{\text{historical}, \text{SSP126}, \text{SSP245}, \text{SSP370}, \text{SSP585}\}$. We iteratively remove one SSP experiment from the dataset to construct a training set (e.g. retain *SSP245* to obtain $\mathcal{D}_{\text{train}} = \{\text{historical}, \text{SSP126}, \text{SSP370}, \text{SSP585}\}$) and use it to train FaIRGP and the baseline GP model. We run predictions over the retained SSP experiment and evaluate the emulated global temperatures against NorESM2-LM data. We find that on average, FaIRGP outperforms the baseline models in every evaluation metric, as reported in Table 4. These results suggest that FaIRGP provides a better emulation of the global temperature response to anthropogenic forcing compared to both FaIR and the plain GP model. We include in Appendix C2 a comparison with the GP emulator from ClimateBench (Watson-Parris et al., 2022).

Table 4: Scores of baseline emulators and FaIRGP for the task of emulating SSPs global temperatures over 2015-2100 period; the test experiment is retained from the training data and only used for evaluation; the best emulator scores for each test experiment are highlighted in bold; $\uparrow/\downarrow$ indicates higher/lower is better; we report 1 standard deviation over the mean scores; $\dagger$ indicates our proposed method.

Test experiment	Emulator	RMSE $\downarrow$	MAE $\downarrow$	Bias	LL $\uparrow$	Calib95	CRPS $\downarrow$
<i>SSP126</i>	FaIR	0.17	0.13	0.06	-	-	-
	Plain GP	0.18	0.14	0.08	0.23	1.0	0.10
	FaIRGP $\dagger$	0.15	0.12	0.02	0.48	0.97	0.08
<i>SSP245</i>	FaIR	0.14	0.12	0.06	-	-	-
	Plain GP	0.09	0.07	0.03	0.67	1.0	0.06
	FaIRGP $\dagger$	0.09	0.08	-0.01	0.71	1.0	0.06
<i>SSP370</i>	FaIR	0.22	0.19	0.10	-	-	-
	Plain GP	0.22	0.17	-0.16	0.12	1.0	0.12
	FaIRGP $\dagger$	0.17	0.14	0.06	0.39	0.94	0.10
<i>SSP585</i>	FaIR	0.28	0.22	-0.07	-	-	-
	Plain GP	0.30	0.21	-0.12	0.16	1.0	0.14
	FaIRGP $\dagger$	0.21	0.17	-0.08	0.18	0.88	0.12
<i>Mean</i>	FaIR	0.20 $\pm$ 0.06	0.16 $\pm$ 0.05	0.04 $\pm$ 0.08	-	-	-
	Plain GP	0.20 $\pm$ 0.09	0.15 $\pm$ 0.06	-0.04 $\pm$ 0.11	0.30 $\pm$ 0.25	1.0 $\pm$ 0.0	0.11 $\pm$ 0.03
	FaIRGP $\dagger$	0.16$\pm$0.05	0.12$\pm$0.04	-0.01$\pm$0.06	0.44$\pm$0.22	0.95$\pm$0.05	0.09$\pm$0.02

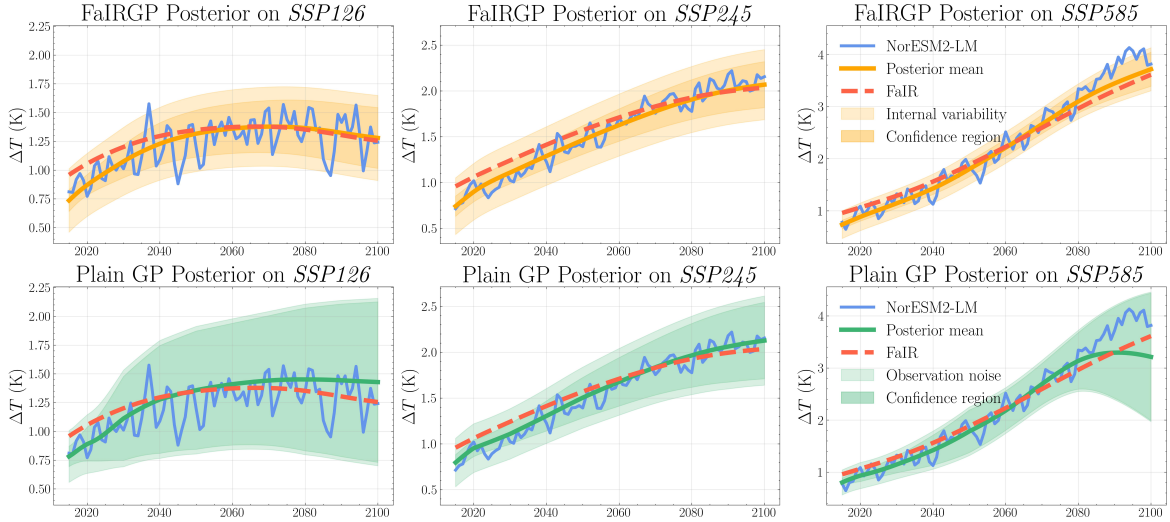


Figure 11: Emulated global mean surface temperature anomaly over 2015-2100 period with our FaIRGP emulator (top) and with a baseline GP emulator (bottom) for a low-forcing future scenario (*SSP126*), a medium forcing future scenario (*SSP245*) and a high forcing future scenario (*SSP585*).

Figure 11 shows that FaIRGP provides a better temperature projection than FaIR in the near future — over the 2015-2050 period — for all scenarios. This is because being informed by data grants FaIRGP the flexibility to deviate from the impulse response prior on which it hinges, and provide predictions that are better aligned with the historical and near future data from the training set. However, when further away from the training data, over the 2080-2100 period, FaIRGP reverts back to the prior behaviour of FaIR. This suggests that FaIRGP is better suited for emulating global temperature response in the near future, and is at least as good as FaIR for longer term emulation.

The plain GP model provides excellent predictions on *SSP245*, and even outperforms FaIRGP by a slight margin on deterministic metrics for this scenario. This is because GPs are notorious for excelling at interpolation tasks, and the *SSP245* scenario is a medium forcing scenario that perfectly lies within the range of the other low and high forcing scenarios used in the training data. However, when the evaluation scenario lies outside of the range of the training data, the plain GP model predictions will simply revert to the prior, and are therefore much less reliable. This is an important drawback of purely data-driven emulation, which by essence are interpolation models, and struggle to extrapolate outside the range of the training set. In Figure 11, this is particularly evident in the plain GP prediction on *SSP585*, where the model struggles at the end of the century due to unprecedented forcing levels that have not been seen in the training data. FaIRGP addresses this shortcoming by using FaIR as its prior, and therefore naturally reverts to an impulse response model when the prediction scenario is too distant from the training data.

Finally, FaIRGP provides a tighter and more robust uncertainty quantification over emulated temperatures compared to the plain GP model, which tends to overestimate the variance. This is reflected in Table 4 by a substantial improvement in LL of FaIRGP against the Plain GP, and suggests that the physics-informed covariance structure of FaIRGP is a sound choice to quantify emulator uncertainty. Further, the model formulated is able to quantify the uncertainty introduced by the internal variability separately from the uncertainty introduced by the emulator design. As shown in Figure 11, the uncertainty due to internal variability dominates in the near future, but the emulator uncertainty becomes predominant toward the end of the century, in accordance with the well-established assessment of Hawkins and Sutton (2009).

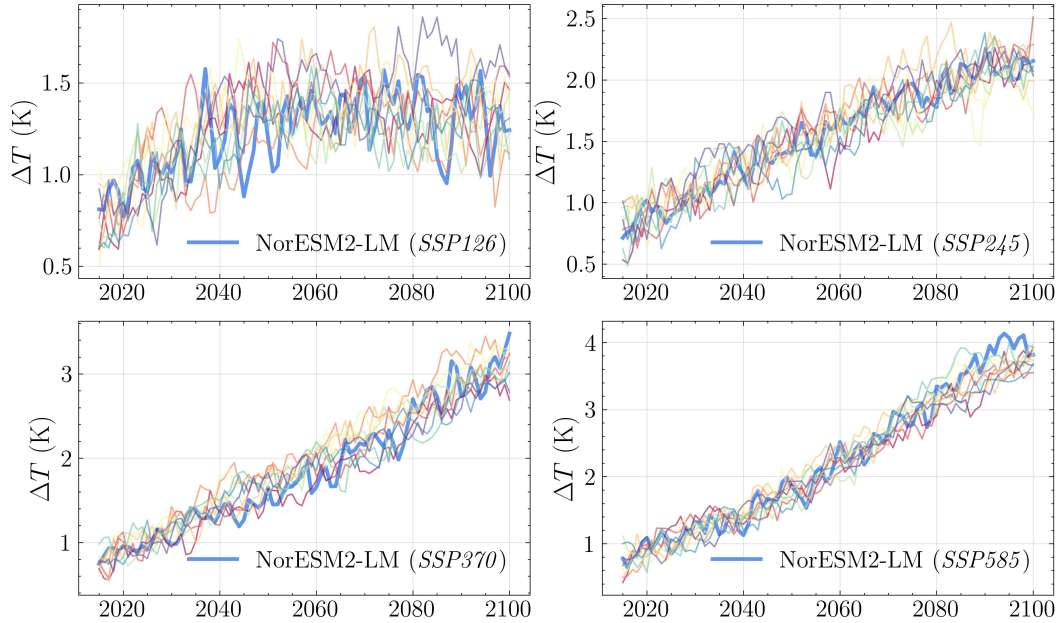


Figure 12: Example of sample paths from the FaIRGP posterior for the emulation of *SSP126* (top-left), *SSP245* (top-right), *SSP370* (bottom-left) and *SSP585* (bottom-right) global mean surface temperature anomaly over the 2015-2100 period.

In Figure 11, the posterior mean lines and the shaded area depict the mean function and the 95% credible region of the predicted posterior distribution over temperatures using FaIRGP. Since this represents a probability distribution, we can generate sample paths from the posterior GP (which will fall within the shaded area with probability 0.95). Figure 12 displays examples of sample paths from the FaIRGP posterior for each scenario. It shows that the smooth shaded areas hide the specific form of the functions described by the

FaIRGP posterior. The sample paths are in fact capable of reproducing the variability of abrupt annual changes in global mean surface temperature anomalies. Further, the sample paths are generally well-concentrated around the NorESM2 simulated temperature anomaly, with the exception of the end of century temperature anomaly under *SSP585* which tends to be underestimated — consistently with the shaded area from Figure 11. Appendix C1 provides additional plots of individual sample paths from the FaIRGP posterior over SSPs.

5.3 Emulation from historical data

In this experiment, we propose to investigate how well can the emulators generalise when they are solely trained on historical data, and no simulated future data⁷. We train the plain GP and FaIRGP using the *historical* experiment only, and evaluate predictions on all SSP scenarios. The results are reported in Table 5. We find that FaIRGP outperforms other models in almost every metric, and displays performance similar to FaIR only in mean bias.

Table 5: Scores of baseline emulators and FaIRGP when models are only trained using the *historical* experiment; scores are evaluated and averaged for the task of emulating *SSP126*, *SSP245*, *SSP370*, *SSP585* global surface temperatures over 2015-2100 period; the best emulator for each metric is highlighted in bold; $\uparrow$ / $\downarrow$ indicates higher/lower is better; $\dagger$ indicates our proposed method.

Emulator	RMSE $\downarrow$	MAE $\downarrow$	Bias	LL $\uparrow$	Calib95	CRPS $\downarrow$
FaIR	0.208	0.163	0.038	-	-	-
Plain GP	1.875	1.690	-1.690	-30.979	0.009	1.558
FaIRGP $\dagger$	0.191	0.140	-0.037	0.314	0.948	0.102

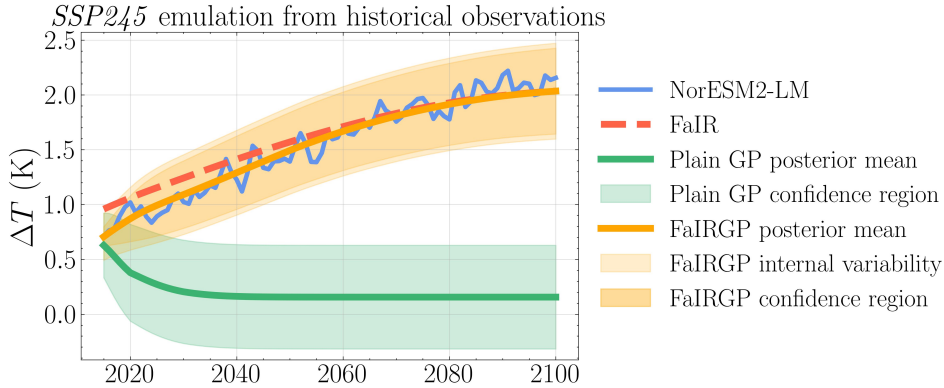


Figure 13: Emulated global mean surface temperature anomaly over 2015-2100 for the *SSP245* scenario with emulators trained over historical temperatures only.

In this context, a purely data-driven method struggles to extrapolate because it cannot marry the new previously unseen emission values with the underlying physical model. Figure 13 shows that the plain GP posterior immediately reverts to its prior after 2015, therefore providing uninformative temperature projections. On the other hand, FaIRGP manages to learn from historical data to provide a better fit to the SSP over the 2015-2050 period, and slowly reverts back to the prior behaviour of FaIR toward the end of the century.

⁷ with the exception of the FaIR parameters which have been calibrated using NorESM2 outputs for future scenarios.

This demonstrates the importance of having a robust underlying physical model for an emulator, and further suggests that FaIRGP is a well suited candidate for this task, and can provide meaningful temperature projections based only on historical data.

5.4 Inferring radiative forcing from temperatures

Whilst FaIRGP is intended to emulate the surface temperature response to changes in emission, its GP nature also allows to make inference of radiative forcing given temperature data as demonstrated in Section 3.5. In this experiment, we propose to assess how global surface temperature anomaly data can be used to inform an estimate of the effective radiative forcing using FaIRGP.

We use the complete dataset of global temperature time series from the *historical* and *SSP126*, *SSP245*, *SSP370*, *SSP585* experiments, which are depicted in Figure 14. We update the GP prior placed over the radiative forcing $F(t)$ with temperature data to obtain a posterior radiative forcing response which incorporates information from temperature data. We emphasise that the posterior over $F(t)$ does not use any forcing data, only simulated temperature data. As for temperature emulation, the posterior radiative forcing response is given in closed-form by (32). We evaluate the posterior historical radiative forcing response against NorESM2-LM historical top-of-atmosphere radiative forcing.

Figure 14 illustrates that the posterior forcing response learns from temperature time series how to deviate from the FaIR forcing trend to better account for the NorESM2 simulated data. This is particularly evident between 1960 and 1980 where the posterior reproduces a decrease in forcing to better account for the global cooling trend in that time period, which FaIR struggles to capture. The results reported in Table 6 show that FaIRGP helps improve over the FaIR forcing in RMSE and Bias. Whilst FaIR performs better at inferring the stable forcing before 1950, informing the inferred forcing with temperature data with FaIRGP helps better account for the forcing variations after 1950.

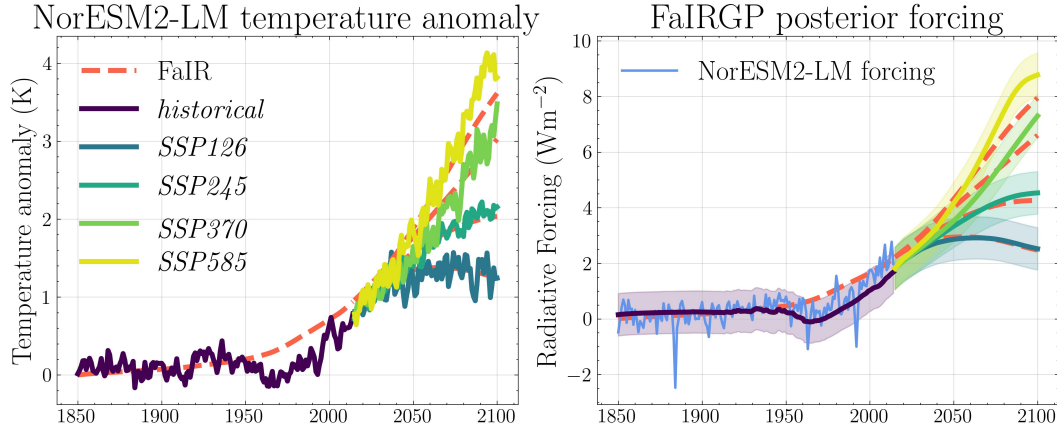


Figure 14: **Left:** Global temperature anomaly responses used to inform the posterior forcing response; **Right:** FaIRGP posterior radiative forcing predicted over experiments; the temperature response and radiative forcing obtained with FaIR are plotted in dashed red lines.

FaIRGP therefore proves to be useful to infer global radiative forcing trends informed by surface temperatures. This is possible because our prior is specified within an energy balance model which explicitly connects the changes in surface temperatures to the changes in radiative forcing. On the contrary, this would not be possible with the plain GP model because it ignores forcing dynamics.

Table 6: Scores of FaIR and FaIRGP at predicting the global radiative forcing over historical periods; scores are evaluated against NorESM2-LM top-of-atmosphere radiative forcing; the best emulator scores for each period are highlighted in bold; $\uparrow/\downarrow$ indicates higher/lower is better; $\dagger$ indicates our proposed method.

Period	Emulator	RMSE $\downarrow$	MAE $\downarrow$	Bias	LL $\uparrow$	Calib95	CRPS $\downarrow$
1850-1950	FaIR	0.396	0.261	-0.008	-	-	-
	FaIRGP $\dagger$	0.421	0.285	0.041	-0.797	0.99	0.25
1950-2014	FaIR	0.622	0.461	0.350	-	-	-
	FaIRGP $\dagger$	0.568	0.466	-0.315	-0.925	0.984	0.328
1850-2014	FaIR	0.496	0.339	0.129	-	-	-
	FaIRGP $\dagger$	0.484	0.357	-0.101	-0.847	0.988	0.281

5.5 FaIR multi-model uncertainty in FaIRGP

In this section, we explore how the uncertainty bounds introduced by the FaIR model compare with those of the FaIRGP uncertainty. FaIR has been widely employed to assess the uncertainty surrounding global temperature projections. The methodology employed by FaIR for generating probabilistic projections is rooted in a constrained ensembling process. This involves specifying prior distributions over model parameters (such as carbon cycle sensitivity, forcing model sensitivity and equilibrium climate sensitivity). These prior distributions are determined from the inter-model variability. Then, a large ensemble of climate projections is drawn using these distributions, and is subsequently constrained based on the likelihood of projections aligning with the simulated present-day level and rate of anthropogenic warming (Leach et al., 2021). The resulting *constrained ensemble* of projections allows to quantify the effect of *FaIR multi-model uncertainty* on temperature projections.

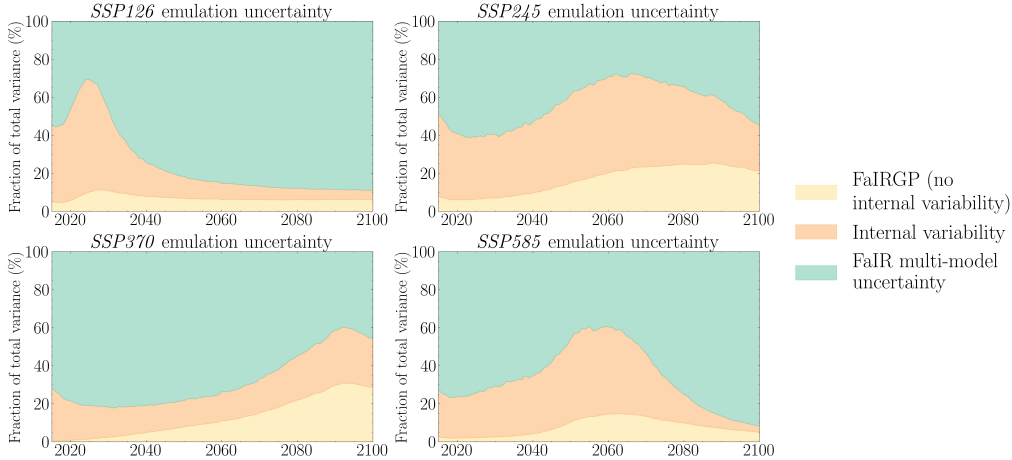


Figure 15: Proportional stacked area charts of the contribution to total variance of the FaIRGP uncertainty without internal variability (yellow), the internal variability modelled in FaIRGP as in Section 3.4 (orange) and the FaIR multi-model uncertainty (green), for the emulation of the global annual mean surface temperature anomaly under *SSP126* (top-left), *SSP245* (top-right), *SSP370* (bottom-left) and *SSP585* (bottom-right).

Following the protocol of Leach et al. (2021), we generate a constrained ensemble of 1000 FaIR surface temperature anomaly projections. For each ensemble member, we then generate a FaIRGP projection. The total variability of this projection can be decomposed into two components: (i) the average variability introduced by FaIRGP across ensemble members and (ii) the variability in the mean FaIRGP response across ensemble members. The first component corresponds to the variability that FaIRGP’s stochasticity introduces

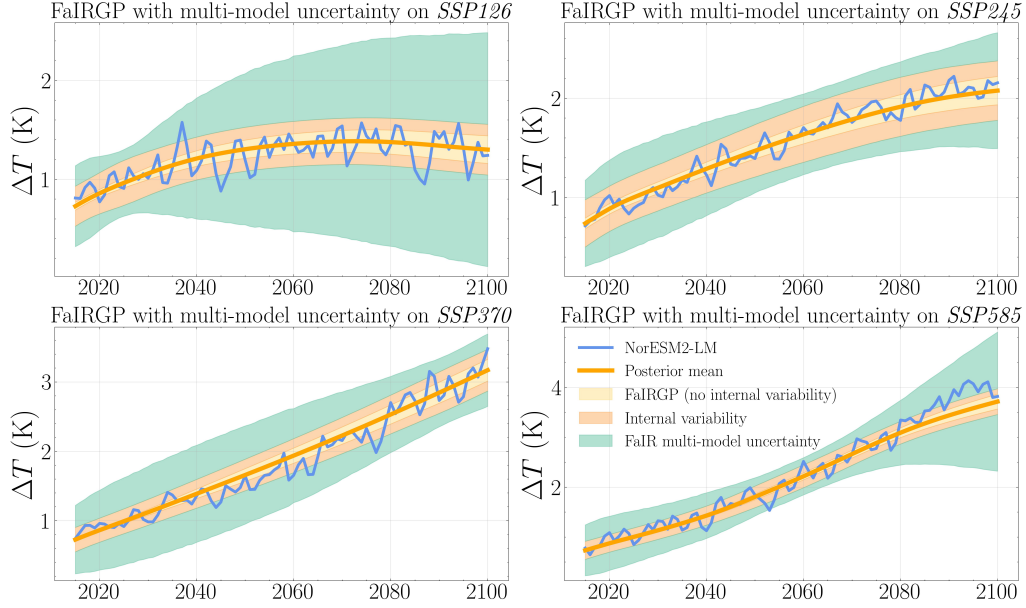


Figure 16: Global annual mean surface temperature anomaly projections over SSPs with contributions to uncertainty from FaIRGP without internal variability (yellow), the internal variability modelled in FaIRGP as in Section 3.4 (orange) and the FaIR multi-model uncertainty (green). Shading indicates 2 standard deviations and is computed by taking the average contribution to the standard deviation of each source of uncertainty

on the deterministic FaIR temperature response, averaged across ensemble member, and can be decomposed into the variability introduced by FaIRGP without internal variability, and the contribution of internal variability alone. For this reason component (i) is viewed as the uncertainty imputable to FaIRGP. The second component corresponds to the variability that sampling across the constrained ensemble introduces in the FaIRGP mean response. For this reason, component (ii) is viewed as the uncertainty imputable to the FaIR multi-model uncertainty. Figure 15 shows how the different sources of uncertainty contribute to the total variance when emulating surface temperature anomalies for SSPs.

A higher FaIR multi-model uncertainty contribution to the total variance (green) indicates significant sensitivity of the FaIRGP mean response to parameter choices within the constrained ensemble. Conversely, a greater FaIRGP uncertainty contribution (orange and yellow) suggests limited sensitivity to parameter choices, with most variability arising from our stochastic treatment of the temperature response. We find that in general, the FaIR multi-model uncertainty contributes the largest fraction of the total variance in projections of future warming. This is particularly evident in 2100 for low/high end forcing scenarios *SSP126* and *SSP585*, since they lie outside of the training data regime used in FaIRGP and the model’s need for extrapolation makes it more sensitive to FaIR’s structural uncertainty.

We find that for intermediate forcing scenarios *SSP245* and *SSP370*, the FaIRGP uncertainty contribution to the total variance remains significant. Because these scenarios lie in between other scenarios from the training data, FaIRGP’s reliance on interpolation between its training data results in lower sensitivity to the multi-model uncertainty, making the idea of an uncertain future warming better constrained. This is supported by Figure 16, depicting how the average contribution of each source of uncertainty to the standard deviation stacks up in predictions. It shows a consistently stable and comparable magnitude of uncertainty introduced by FaIRGP with internal variability across scenarios.

It is important to caveat that the FaIR constrained ensemble samples from inter-model variability, whereas FaIRGP is here only calibrated on NorESM2-LM data. Therefore, one could reasonably expect that a multi-model version of FaIRGP, allowing calibration across multiple climate models, would likely contribute more uncertainty to projections of end-of-century warming.

6 Application: spatial surface temperatures emulation

In this section, we pursue the comparison of FaIRGP to baseline models, but for the task of emulating spatially-resolved temperatures. As for global emulation, we start by briefly illustrating how the model concretely applies. Then, we benchmark it against baseline models for SSP emulation and evaluate the emulation of surface temperatures forced by anthropogenic aerosols only. We conclude by investigating how FaIRGP can help infer spatial top-of-atmosphere radiative forcing maps.

6.1 FaIRGP for spatial temperatures emulation

Figure 17 illustrates the spatial surface temperature anomaly emulation with FaIRGP for test scenario *SSP245*. As in the global case, a prior response based on FaIR is first specified, and then shifted by a data-informed posterior correction map. The posterior correction is learned from training scenarios $\mathcal{D}_{\text{train}} = \{\text{historical}, \text{SSP126}, \text{SSP370}, \text{SSP585}\}$.

The prior spatial response is constructed using a pattern scaling model trained on $\mathcal{D}_{\text{train}}$, and therefore introduces a fixed spatial pattern that is only rescaled by changes in global mean temperature. The posterior correction at location x is obtained by updating the prior over $T(x, t)$ from (34) with local surface temperature data from $\mathcal{D}_{\text{train}}$. The posterior correction maps effectively provide a data-driven way to deviate from this fixed spatial pattern, and better account for the possibly varying spatial temperature patterns of the emulated scenario. Finally, by linearly adding the posterior correction map to the prior, we obtain a posterior spatial temperature response over *SSP245*.

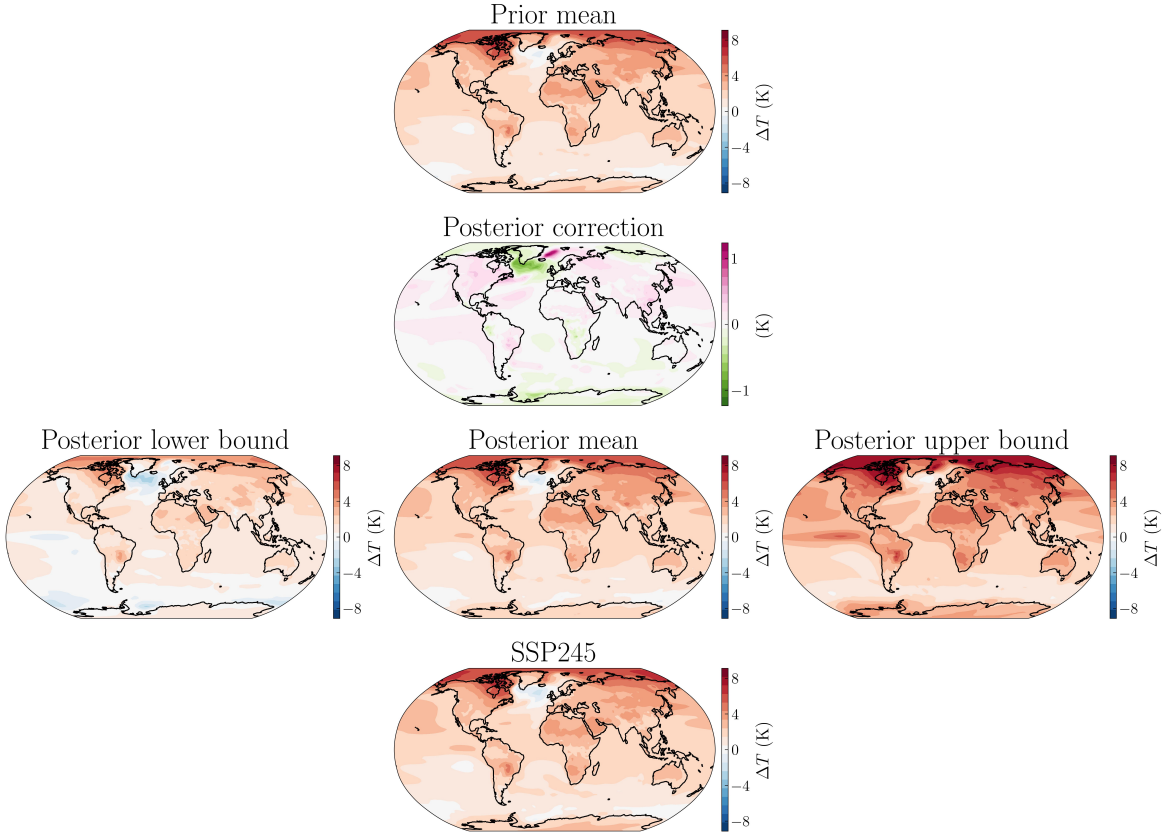


Figure 17: Spatial emulation of NorESM2-LM *SSP245* temperatures. The prior mean (top) corresponds to a pattern scaling model used on the FaIR global temperature response. The posterior correction (middle-top) learns from data how to deviate from the pattern scaling prior to produce a posterior response (middle-bottom) that better emulates the groundtruth surface temperatures (bottom). The posterior lower/upper bounds capture the 95% credible interval. Maps are averaged over the 2080-2100 period.

Table 7: Scores of baseline emulators and FaIRGP for the task of emulating SSPs spatial temperatures over 2080-2100 period; the best emulator for each metric is highlighted in bold; $\uparrow/\downarrow$ indicates higher/lower is better; we report 1 standard deviation; $\dagger$ indicates our proposed method.

Emulator	RMSE $\downarrow$	MAE $\downarrow$	Bias	LL $\uparrow$	Calib95	CRPS $\downarrow$
Pattern scaling	0.637 ± 0.189	0.467 ± 0.139	-0.162 ± 0.194	-	-	-
Plain GP	1.069 ± 0.616	0.761 ± 0.441	-0.362 ± 0.350	-1.625 ± 1.192	0.850 ± 0.141	0.560 ± 0.341
FaIRGP $\dagger$	0.619 ± 0.188	0.451 ± 0.138	-0.113 ± 0.192	-0.853 ± 0.398	0.873 ± 0.093	0.341 ± 0.099

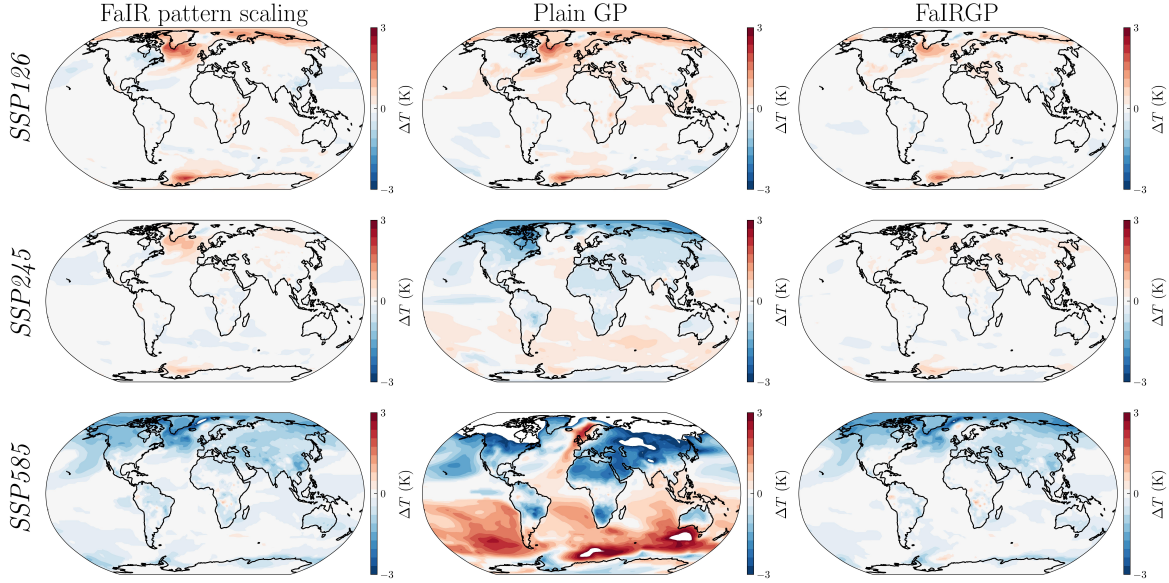


Figure 18: Maps of the mean difference between emulated spatial temperatures and target NorESM2 values for each emulator under a low forcing future scenario (*SSP126*), a medium forcing future scenario (*SSP245*) and a high forcing future scenario (*SSP585*). Red/Blue means the emulator overshoots/undershoots NorESM2 temperatures. Maps are averaged over the 2080-2100 period. Differences insignificant at the $p < 0.05$ level are masked.

6.2 Shared socio-economic pathways emulation

In this experiment, we follow the same procedure as in the global emulation experiment and iteratively train models on the dataset deprived of one SSP scenario, then use the retained scenario as a test scenario for evaluation. We benchmark FaIRGP against a FaIR pattern scaling model and a purely data-driven plain GP emulator. The pattern scaling model is obtained by fitting a linear regression model using the same training data as the other models. The plain GP model is analogous to the baseline GP emulator in Watson-Parris et al. (2022), but differs in two aspects: (i) we adopt a simpler construction for the covariance with an anisotropic Matérn-3/2 kernel, and (ii) our model takes as input global aerosols emissions whereas Watson-Parris et al. (2022) use spatially-resolved aerosols emission maps. Scores are computed over the 2080-2100 period since the start of all SSPs is quite similar. Mean scores are reported in Table 7.

We find that FaIRGP has on average lower error than baseline models. Figure 18 shows that the spatial bias patterns of FaIRGP are similar to the bias patterns obtained with the FaIR pattern scaling model. This indicates that the prior has a strong influence on the

predicted posterior. Nonetheless, we observe that the posterior correction in FaIRGP helps mitigate the spatial inaccuracies of its prior. This is particularly evident for *SSP126* and *SSP245* where the magnitude of the spatial bias in FaIRGP is overall smaller than the one for the FaIR pattern scaling model. Regarding the high forcing scenario *SSP585*, the spatial correction is more subtle. This is due to it becoming an extrapolation task, and as a result, FaIRGP exhibits behaviour closer to its pattern scaling prior.

Figure 18 also shows that, as for global temperature emulation, the plain GP model predicts sound surface temperature maps for low and medium forcing scenarios, but struggles at extrapolating over high forcing scenarios.

6.3 Emulating anthropogenic aerosols forcing

In this experiment, we want to evaluate the emulation of temperature changes induced by anthropogenic aerosol emissions. We use the *hist-aer* experiment from the ClimateBench v1.0 dataset. The *hist-aer* experiment is generated using NorESM2-LM, using only historical anthropogenic aerosol emissions, and setting long-lived greenhouse gas emissions to zero. We emulate surface temperature anomalies over this scenario using emulators trained on all available *historical* and *SSPs* experiments. The FaIR forcing model is modified to include the aerosol-cloud interaction parametrisation from (Leach et al., 2021, Section 2.2.1).

Figure 19 shows the emulated global mean surface temperature anomaly with FaIRGP. Whilst FaIR faces challenges in reproducing the magnitude of the global temperature response to anthropogenic emissions, the posterior correction introduced by FaIRGP allows for a deviation from FaIR, resulting in predicted temperature anomalies that more accurately reflect the NorESM2-LM response. This suggests that FaIRGP can leverage information from the *historical* and *SSPs* experiments to enhance its representation of the temperature response to anthropogenic aerosol emissions.

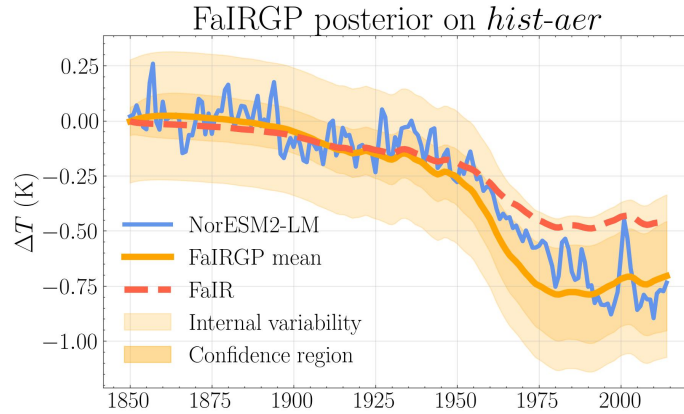


Figure 19: Emulated global mean surface temperature anomaly over 1850-2014 for the *hist-aer* experiment

We now turn to the emulation of local surface temperature anomaly maps for the *hist-aer* experiment. Figure 20 presents a comparison of predictions with the FaIR pattern scaling baseline and FaIRGP over three time periods. Consistently with the global trend from Figure 19, the predicted temperature anomaly magnitudes for 1900-1950 are similar between the pattern scaling model and FaIRGP, and align with the NorESM2-LM response whereas, over periods 1950-1980 and 1980-2014, the magnitude of the FaIRGP spatial response more accurately reflects the magnitude of the NorESM2-LM temperature response.

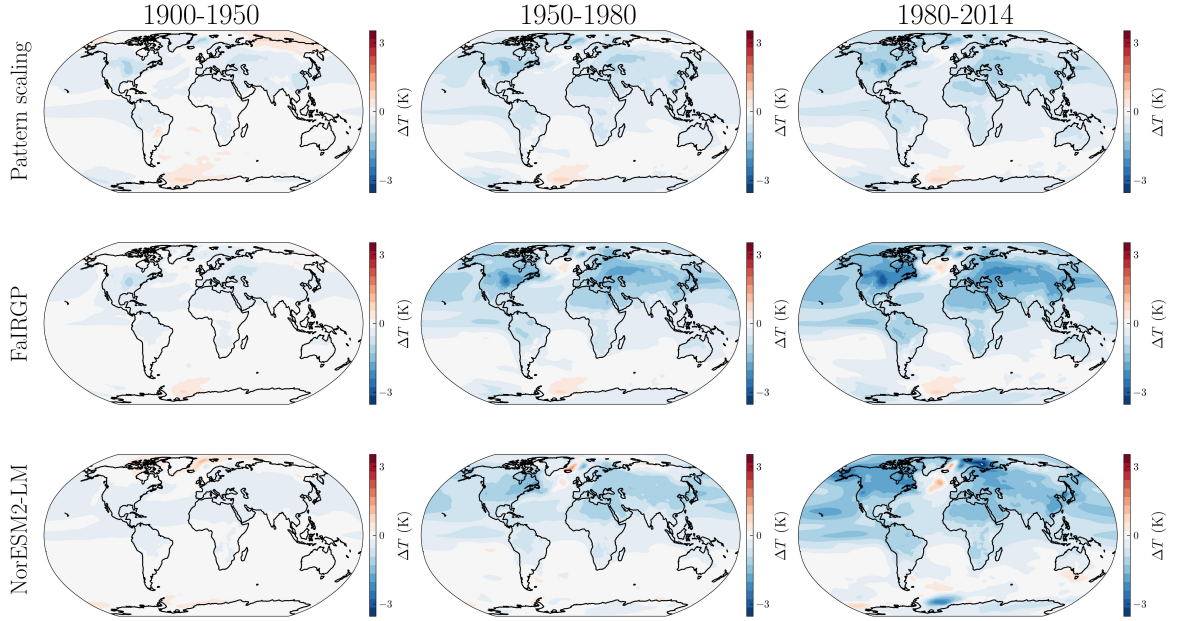


Figure 20: Maps of local surface temperature anomalies for the *hist-aer* experiment averaged over the 1900-1950 period (left), 1950-1980 (left-right center) and 1980-2014 (right). **Top:** Prediction with the FaIR pattern scaling baseline. **Top-Bottom Center:** Prediction with the FaIRGP posterior mean. **Bottom:** Groundtruth NorESM2-LM surface temperature anomaly map.

Regarding predicted spatial patterns, we observe that FaIRGP is able to deviate from its fixed pattern prior to partially reproduce the North Atlantic warming observed in NorESM2-LM after 1950. However, the predicted spatial temperature anomaly pattern with FaIRGP remains strongly influenced by its pattern scaling prior. This is particularly evident in the South Atlantic where FaIRGP reproduces the warming predicted by the pattern scaling model. This task is notoriously difficult for pattern scaling models, which excel at emulating responses to greenhouse gas emissions but struggle under strong aerosol forcing scenarios (May, 2012; Levy et al., 2013). Considering the significant impact of the prior choice on FaIRGP, this fosters advocacy for a prior local response model that goes beyond pattern scaling models. Additionally, using spatially-resolved emissions maps for aerosols, instead of global aerosols emissions, should enhance emulated temperatures given the strong influence spatial patterns of aerosols have over radiative forcing (Williams et al., 2022).

We conclude this experiment with a quantitative evaluation of the predicted local temperature anomaly maps over the 1950-2014 period. This is the period where the temperature response to aerosol emissions is most discernible. Because the anthropogenic aerosols forcing pattern is spatially heterogeneous and mostly localised around industrialised regions (Carslaw & Pringle, 2022), we assess the emulators’ performance over four major industrialised regions: North America, East Asia, South Asia and Europe. The regions’ extents are depicted in Figure 21 and are constructed from world region delineation proposed by (S. Seneviratne et al., 2012). Table 8 shows that, consistently with the prediction maps, the FaIRGP posterior mean spatial response outperforms the FaIR pattern scaling baseline for every region considered. The plain GP model, as anticipated, faces difficulties in emulating local surface temperature anomalies solely based on aerosol emissions inputs. We believe this is primarily because every scenario from its training data includes greenhouse gas emissions. Appendix C4 provides maps emulated with the plain GP model.

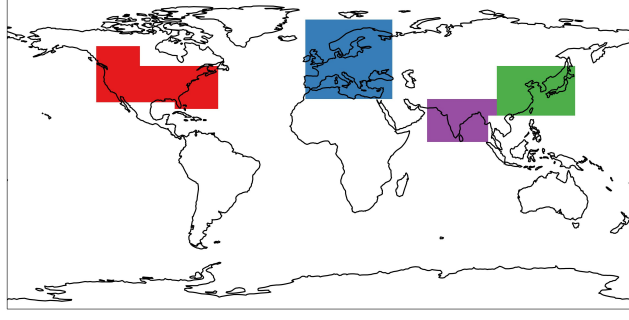


Figure 21: The polygons show the industrialised regions used for the evaluation of the predicted local annual mean surface temperature anomaly under the *hist-aer* experiment: North America (red), Europe (blue), South Asia (purple), East Asia (green).

Table 8: Scores of baseline emulators and FaIRGP for the task of emulating *hist-aer* local annual mean surface temperature anomaly for four industrialised regions over the 1950-2014 period; the best emulator for each metric is highlighted in bold; † indicates our proposed method.

Region	Emulator	RMSE (10^{-1})	MAE (10^{-2})	Bias (10^{-2})
<i>Europe</i>	Pattern scaling	1.29	1.71	1.29
	Plain GP	2.26	3.37	3.07
	FaIRGP [†]	1.05	1.35	-0.313
<i>North America</i>	Pattern scaling	1.10	1.63	1.21
	Plain GP	2.40	3.70	3.65
	FaIRGP [†]	0.998	1.35	-0.795
<i>South Asia</i>	Pattern scaling	0.523	0.594	0.389
	Plain GP	1.26	1.53	1.52
	FaIRGP [†]	0.382	0.404	-0.0100
<i>East Asia</i>	Pattern scaling	1.01	1.32	1.20
	Plain GP	2.22	3.08	3.08
	FaIRGP [†]	0.631	0.772	-0.173

6.4 Inferring radiative forcing from temperatures

As in Section 5.4, we propose to probe whether FaIRGP can be used to infer spatial forcing maps. Since we do not have a spatial forcing model, we simply use a spatially constant forcing prior for $F(x, t)$. We update it with spatially-resolved temperature data from *historical* and SSPs scenarios. Figure 22 compares the obtained posterior mean forcing with historical forcing maps simulated with NorESM2-LM.

The magnitude of the inferred posterior mean forcing with FaIRGP is very conservative, and does not align with the magnitude of the forcing levels observed in the groundtruth NorESM2-LM maps. The sign of the inferred forcing with FaIRGP also disagrees with the groundtruth maps in several regions. This is most evident around 90° latitudes, where FaIRGP infers a positive forcing whilst NorESM2-LM simulations suggest a negative forcing.

We attribute this discrepancy to FaIRGP relying only on local temperature anomalies to infer local radiative forcing. Whilst high latitudes tend to warm faster than around the equator (due among things to surface albedo feedback and atmospheric circulation for equa-

tor to poles), the top-of-atmospheric effective radiative forcing at high latitudes can remain negative (due to higher albedo, larger zenith-angle). This underlines an important point: FaIRGP performs probabilistic inference of local radiative forcing given local temperature anomalies, which alone is not sufficient to account for the complex spatiotemporal dynamics of the climate system. However, there are reason to believe that informing this probabilistic inference with additional covariates, such as surface albedo or net top-of-atmosphere data, could enhance the accuracy of the inferred spatial forcing maps.

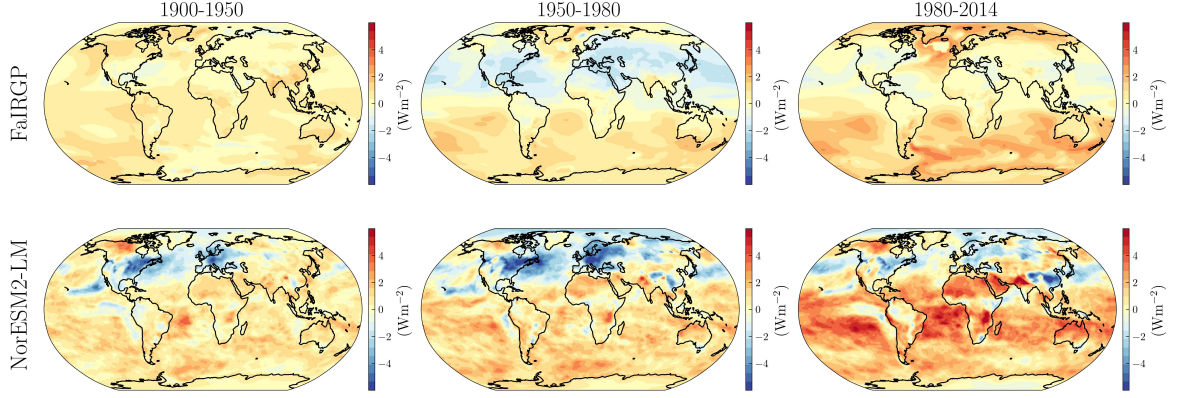


Figure 22: Maps of top-of-atmosphere radiative forcing averaged over 1900-1950, 1950-1980 and 1980-2014 periods of the *historical* scenario. **Top:** Mean posterior radiative forcing inferred with FaIRGP. **Bottom:** NorESM2-LM radiative forcing maps.

Finally, we note that whilst the spatial forcing maps inferred with FaIRGP inadequately capture the groundtruth forcing magnitude and pattern, they still manage to partially reproduce two important features of the climate system: the overall temporal increase in forcing, as well as the mid-1900s hemispheric contrast. In particular, Figure 23 shows that whilst predictions largely underestimate the forcing magnitude, they also display some latitudinal correlation with the NorESM2-LM simulated forcing over the 1940-1980 period. This is encouraging considering that the predicted spatial patterns are solely inferred from temperature patterns.

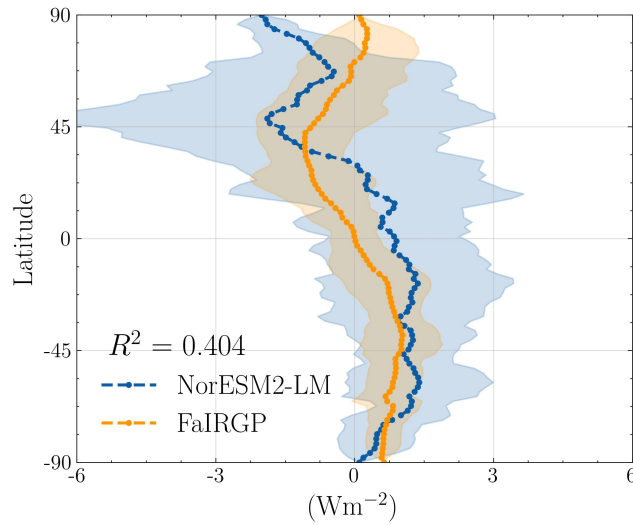


Figure 23: Latitudinal average of FaIRGP inferred radiative forcing and NorESM2-LM radiative forcing over 1940-1980 period. The shaded area reports two standard deviations of longitudinal variability. $R^2 = 0.404$ is computed between latitudinal averages.

7 Discussion

7.1 About the GP approach

7.1.1 *Comparison to placing Bayesian priors over the SCM parameters*

A simple way to introduce model variability in simple climate models is to place Bayesian priors over model parameters, such as the carbon cycle feedback terms or the forcing model coefficients (Leach et al., 2021). Such priors can then be updated with global temperature data to formulate posterior distributions. This specifies a probabilistic climate model emulator calibrated against climate model data. Whilst aligned in spirit with the work proposed in this paper, it is important to highlight the distinctions in objectives and technical scope with our GP-based approach.

Placing Bayesian priors over parameters allows to sample from diverse sources of uncertainty such as uncertainty on the temperature response parameters, forcing sensitivities or carbon cycle sensitivities. Importantly, this allows to explore climate sensitivity and the structural uncertainty of models. In contrast, the framework we propose only addresses the forcing uncertainty and considers fixed values for the SCM parameters. In that sense, FaIRGP, as it is formulated in our work, provides a more restrictive exploration of the uncertainty than methods placing priors over an SCM parameters.

The main advantage of our GP-based approach rather lies in its technical convenience. Chiefly, FaIRGP formulates analytical expressions for the posterior distribution over forcing and temperatures, which have an intuitive interpretation as the sum of a physics-driven prior and a data-driven correction. In contrast, when placing a Bayesian prior over parameters, we do not in general have access to a probability distribution for temperatures. Therefore, probabilistic emulation may need to be sampling-based, and one must resort to more complex Markov-Chain Monte-Carlo (MCMC) techniques to sample from the posterior distribution.

In addition, beyond particular modelling objectives, sampling-based approaches have two main shortcoming from a purely technical perspective: *(i)* a thorough uncertainty quantification requires storing a tremendous amount of scenarios, which can be limited by memory capacity — Beusch et al. (2022) require an ensemble of 9 million emulations; *(ii)* it reduces the statistical representation of uncertainty to summary statistics such as the mean, standard deviation or quantiles. In contrast, having a closed-form expression for the posterior distribution with GPs allows to analytically conduct probabilistic studies using the full probability density, and additionally draw samples, if needed. Finally, having access to the probability density expression allows to evaluate the likelihood of observations. This can critically be used as a maximisation objective to tune the model parameters against observations, but also in the context of Bayesian optimisation routines to find optimal emission trajectories to meet climate goals.

Because of the technical convenience that GP-based approaches display, and the richer structural uncertainty of sampling-based approaches, we argue that rather than opposing them, the two paradigms should be viewed as complementary. Namely, endeavours to formulate a hierarchical Bayesian model that both places priors over the SCM parameters and uses a FaIRGP backbone are promising directions to simultaneously address many sources of uncertainty while retaining some technical advantages of working with GPs.

7.1.2 *Connection with stochastic energy balance models*

Our work is related to the work of Cummins et al. (2020), which formulates a stochastic energy balance model by introducing a white noise variability term in the temperature response, but also in the forcing model. This allows them to account for climate internal variability, and formulate a Kalman filtering strategy to obtain maximum likelihood estimators of the energy balance model parameters. Our work similarly introduces a white noise in the temperature response model to account for climate internal variability, which

results in an additional temporal Matérn-1/2 covariance term in the prior over temperatures. Whilst an extended discussion goes beyond the scope of this work, it can be shown that in the long term regime, these two approaches are equivalent, and that more broadly, Kalman filtering models are in fact equivalent to temporal GPs with Matérn covariance functions (Hartikainen & Särkkä, 2010; Sarkka & Hartikainen, 2012; Sarkka et al., 2013).

Our work differs from Cummins et al. (2020) in that beyond the stochasticity arising from internal variability, we also introduce a GP prior over the radiative forcing. This GP prior introduces stochasticity over the SCM design, which is not only a function of time, but also of emission levels. Because our modelling is not purely temporal (the GP is also a function of emissions), we cannot employ Kalman filtering strategies and instead choose to use GP regression techniques.

7.1.3 Choice of kernel ρ

The choice of covariance function ρ in (12) is an important choice that allows the user to incorporate their domain knowledge into the prior over the radiative forcing. Let u and u' be generic notations for input data (in our work, greenhouse gas and aerosol emissions), the kernel $\rho(u, u')$ specifies how will the prior covary between these two inputs. For example, choosing $\rho(u, u') = \delta(u - u')$ makes the GP independent at any two inputs. On the other hand, choosing $\rho(u, u') = 1$ causes the GP to covary equally between any two inputs.

The Matérn family are a common family of kernel parameterised by a degree ν . They allow to control for the functional regularity of the GP. For $\nu = 1/2$, draws from the GP are continuous functions, for $\nu = 3/2$ they are once differentiable, and in the limit $\nu = \infty$ they become infinitely differentiable⁸. A detailed presentation of the Matérn kernels is provided in Appendix A2.

Additions or multiplications of kernels can be used to construct more elaborate covariance functions that reflect an additive or multiplicative structure in the forcing. Periodic kernels can also be introduced to model seasonality. Going further, more complex choices of kernels include the spectral mixture kernel (A. Wilson & Adams, 2013) which attempts to learn the spectral density of the data, or even kernels parametrised as neural networks (A. G. Wilson et al., 2016; Law et al., 2019).

Whilst kernel selection plays a key role in GP regression, our focus in this work is on the development of the FaIRGP framework rather than refining the kernel itself. The kernel is treated throughout as a modular component, with potential for refinement. As a result, we choose to work with a simple anisotropic Matérn-3/2 kernel throughout. Preliminary findings indicate that the choice of kernel does not significantly degrade the results. Hence, dedicating efforts to constructing more elaborate kernels is likely to yield comparable or better results than our current approach.

7.1.4 Computational efficiency and scalability

We report that to emulate 100 years of surface temperature anomaly with FaIRGP, it takes less than a second for global and spatial emulation on an “average” personal laptop⁹, without requiring any parallelization methods.

Scalability issues are commonly associated with Gaussian Processes (GPs) when the training set grows in size. This is because computing their posterior distribution involves a matrix inversion, which has a cubic computational cost in the number of training samples. Fortunately, unlike neural networks which require large amounts of data (Watson-Parris

⁸ The Matérn- ∞ kernel actually corresponds to the squared exponential kernel.

⁹ 16Go memory

et al., 2022), GPs excel in scenarios with limited data (Rasmussen & Williams, 2005). Consequently, it is possible to develop skilful GP emulators with limited training data.

In cases where using a larger training dataset becomes a necessity, one can still employ linear conjugate gradient methods and parallelisation schemes (Wang et al., 2019) to scale exact GPs to millions of data points. Alternatively, sparse approximation techniques can be used to obtain a scalable estimate of the posterior distribution (Rahimi & Recht, 2007; Titsias, 2009; Matthews et al., 2016).

7.2 Climate modelling considerations

7.2.1 *Beyond FaIR: broader applicability of the method*

While we have chosen to use FaIR as the backbone climate model for our work, the rationale behind the development of FaIRGP is easily transferable to other commonly used simple climate models such as MAGICC (Meinshausen et al., 2011) or OSCAR (Gasser et al., 2017). These models share linear time invariant dynamics, allowing us to incorporate a GP prior into the forcing term of these dynamics. By doing so, we can obtain a GP-based solution that is informed by the model parameters, exactly like in FaIRGP.

The dynamical systems of interest can naturally describe the temperature response to radiative forcing, as is the case in our work. However, we could also imagine extending this to carbon cycle models, where emission levels prescribe the forcing function, and the output of the dynamical system are atmospheric concentrations. This GP framework over dynamical systems is in fact highly general, and is known as latent force modelling. This paradigm was initially introduced by M. Alvarez et al. (2009); M. A. Alvarez et al. (2013) in the context of linear dynamical systems where the forcing function is unknown.

For more complex dynamical systems involving non-linearities, more sophisticated latent force modelling techniques may be applied, such as methodologies based on Volterra series approximation of the dynamical system (Álvarez et al., 2019; Ross et al., 2021), variational approximations of the posterior response (Ward et al., 2020; Moss et al., 2021), with the option of using deep probabilistic modelling of the dynamical system (McDonald & Álvarez, 2021; Baldwin-McDonald & Álvarez, 2023).

7.2.2 *Pixel independence assumption*

The pattern scaling model used in our prior effectively uses independent linear regressions at each location to map changes in global mean temperature onto changes in local temperature. However, this modelling approach challenges our intuition as it overlooks the spatial dependence of temperature fields, despite our expectation that temperatures at nearby locations should covary.

To address this modelling concern, a common solution is to incorporate spatially correlated innovations into the pattern scaling response, which represents the spatial expression of climate internal variability (Beusch et al., 2020b, 2020a, 2022; Castruccio & Stein, 2013; Goodwin et al., 2020). Alternatively, Link et al. (2019) design a procedure based on the Wiener-Khinchin theorem (Champeney & Fuoss, 1973) to emulate a climate variability field with the same variance and spatiotemporal correlation structure as the one in ESMs outputs.

From a statistical modelling standpoint, these approaches introduce such spatially correlated innovations as a multivariate Gaussian variable, their main differences lying in the procedure used to determine its spatial covariance structure. Because of the Gaussian nature of the modelling we propose in this work, a Gaussian spatial correlated innovation can easily be incorporated in the FaIRGP spatial emulation model we propose. However, for the sake of clarity, we chose not to delve into these additional considerations and instead focus on the exposition of the Bayesian energy balance model.

Further, whilst it has been pointed out that, in general, regional changes in temperature scale robustly with global temperature (S. I. Seneviratne et al., 2016), this may not be true under strong mitigation scenarios or under strong aerosol forcing (May, 2012; Levy et al., 2013; Tebaldi & Knutti, 2018; Tebaldi et al., 2020). This fosters advocacy for the development of “spatialised” simple climate models going beyond pattern scaling.

7.2.3 *Opportunities for precipitation emulation*

Going beyond surface temperature emulation, we can explore how FaIRGP can be used to emulate precipitations. One approach is to combine the Gaussian process GP emulator for precipitation described in Watson-Parris et al. (2022) with FaIRGP. By leveraging Gaussian conjugacy relationships, a natural cross-covariance between the two emulators is induced. This enables the exchange of information between precipitation and temperature fields. This is in line with recent advocacy for joint emulation of temperatures and precipitations (Snyder et al., 2019; Schöngart et al., 2022, 2023).

Another option is to use the emulated temperatures from FaIRGP as input for a statistical emulator of precipitation, such as a Gamma regression model (Martinez-Villalobos & Neelin, 2019). Indeed, many climate impacts are routinely assumed to be a function of temperatures. Additionally, the full probability distribution predicted by FaIRGP can be incorporated into the model, allowing for the propagation of epistemic uncertainty associated with FaIRGP.

7.2.4 *Application to detection and attribution*

With FaIRGP, we have access to the analytical expression of the probability density distribution of emulated temperatures. Therefore, we can emulate surface temperatures under historical scenarios, both with and without anthropogenic forcing, and analytically compute the probability of temperature occurrences in each scenario. By comparing these probabilities, we can assess the extent to which human activity has made a certain temperature range more likely, enabling us to conduct attribution studies.

Conducting detection and attribution studies with emulators is not exclusive to FaIRGP, and could in principle be conducted with any emulator as discussed in Watson-Parris et al. (2022). However, the strength of FaIRGP lies in its ability to input temperature ranges directly into a known probability density function, providing a precise probability between 0 and 1 of such temperatures to occur under a given emission scenario.

8 Conclusion and outlooks

Simple climate models (SCMs) are robust physically-motivated emulators of changes in global mean surface temperatures. Gaussian processes (GPs) are powerful Bayesian machine learning models capable of learning complex relationships, and emulate from data how changes in emissions affect changes in surface temperatures. By combining them together, we reconcile these two paradigms of emulator design, which mutually address their respective limitations. We introduce FaIRGP, a Bayesian energy balance model that *(i)* maintains the robustness and interpretability of a simple climate model, *(ii)* gains the flexibility of modern statistical machine learning models with the ability to learn from data, and *(iii)* provides principled uncertainty quantification over the emulator design.

We demonstrate skilful emulation of global mean surface temperatures over realistic emission scenarios. Unlike GPs, FaIRGP has a robust physical grounding which allows it to provide reliable predictions even on out-of-sample scenarios. On the other hand, unlike SCMs, FaIRGP can learn complex non-linear relationships to deviate from an SCM and improve predictions. In particular, FaIRGP better accounts for the temperature response to anthropogenic aerosol emissions. We further show that these findings carry over to the

task of emulating spatially-resolved surface temperature maps. In addition, we find that FaIRGP can also be used to produce estimates of top-of-atmosphere radiative forcing given temperature data.

The full mathematical tractability, with analytical expressions for probability distributions, provides great control over the modelling, and a rich framework to reason about probability distributions over temperatures. This is of great relevance to detection and attribution studies. Further, whilst our work focuses on temperature emulation — which has already been thoroughly studied — we envision FaIRGP as a foundation for the development of robust data-driven emulators for more complex climate variables, such as precipitations. Harnessing the mathematical properties of GPs, we believe that emulating climate impacts using FaIRGP will provide additional control over pure machine learning methods, whilst being able to capture complex non-linear relationships to forcing.

We hope this work will contribute to building trust in data-driven models, and thereby allow the climate science community to benefit more widely from their potential.

Open Research

- Data — The data used to run experiments in this paper is obtained from ClimateBenchv1.0 (Watson-Parris et al., 2022) and available here <https://doi.org/10.5281/zenodo.5196512>.
- Software — The model, evaluation metrics and all code used to run experiments and generate the plots is available here <https://doi.org/10.5281/zenodo.8164335>. All figures were made with Matplotlib version 3.6.2 (Caswell et al., 2020; Hunter, 2007) and Cartopy (Met Office, 2010). Models are implemented with PyTorch version 1.12.1 (Paszke et al., 2019) and GPyTorch version 1.9.0 (Gardner et al., 2018).

Acknowledgments

We would like to thank Maybritt Schillinger for insightful discussions and helpful feedback which has greatly benefited this work. We would also like to thank three anonymous reviewers for their thorough reviews and comments, which contributed significantly to the quality of this work. Shahine Bouabid receives funding from the European Union’s Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 860100.

References

- Akhtar, M. K., Wibe, J., Simonovic, S. P., & MacGee, J. (2013). Integrated assessment model of society-biosphere-climate-economy-energy system. *Environmental Modelling & Software*.
- Allen, M. R., Frame, D. J., Huntingford, C., Jones, C. D., Lowe, J. A., Meinshausen, M., & Meinshausen, N. (2009). Warming caused by cumulative carbon emissions towards the trillionth tonne. *Nature*.
- Alvarez, M., Luengo, D., & Lawrence, N. D. (2009). Latent force models. In *Artificial intelligence and statistics*.
- Alvarez, M. A., Luengo, D., & Lawrence, N. D. (2013). Linear latent force models using gaussian processes. *IEEE transactions on pattern analysis and machine intelligence*.
- Álvarez, M. A., Ward, W., & Guarnizo, C. (2019). Non-linear process convolutions for multi-output gaussian processes. In *The 22nd international conference on artificial intelligence and statistics* (pp. 1969–1977).
- Baldwin-McDonald, T., & Álvarez, M. A. (2023). Deep latent force models: Ode-based process convolutions for bayesian deep learning. *arXiv preprint arXiv:2311.14828*.
- Beusch, L., Gudmundsson, L., & Seneviratne, S. I. (2020a). Crossbreeding cmip6 earth

- system models with an emulator for regionally optimized land temperature projections. *Geophysical Research Letters*.
- Beusch, L., Gudmundsson, L., & Seneviratne, S. I. (2020b). Emulating earth system model temperatures with mesmer: from global mean temperature trajectories to grid-point-level realizations on land. *Earth System Dynamics*.
- Beusch, L., Nicholls, Z., Gudmundsson, L., Hauser, M., Meinshausen, M., & Seneviratne, S. I. (2022). From emission scenarios to spatially resolved projections with a chain of computationally efficient emulators: coupling of magicc (v7. 5.1) and mesmer (v0. 8.3). *Geoscientific Model Development*.
- Camps-Valls, G., Verrelst, J., Munoz-Mari, J., Laparra, V., Mateo-Jiménez, 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.
- Carslaw, K. S., & Pringle, K. (2022). Chapter 4 - Global aerosol properties. In K. S. Carslaw (Ed.), *Aerosols and climate* (p. 101-133). Elsevier. Retrieved from <https://www.sciencedirect.com/science/article/pii/B9780128197660000110> doi: <https://doi.org/10.1016/B978-0-12-819766-0.00011-0>
- Castruccio, S., McInerney, D. J., Stein, M. L., Liu Crouch, F., Jacob, R. L., & Moyer, E. J. (2014). Statistical emulation of climate model projections based on precomputed gcm runs. *Journal of Climate*.
- Castruccio, S., & Stein, M. L. (2013). Global space–time models for climate ensembles. *The Annals of Applied Statistics*.
- Caswell, T. A., Droettboom, M., Lee, A., Hunter, J., Firing, E., Sales De Andrade, E., ... others (2020). matplotlib/matplotlib: REL: v3. 3.1 [Software]. *Zenodo*. Retrieved from <https://github.com/matplotlib/matplotlib>
- Champeney, D. C., & Fuoss, R. M. (1973). Fourier transforms and their physical applications. *Physics Today*.
- Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichet, T., Friedlingstein, P., ... Wehner, M. (2014). Long-term climate change: Projections, commitments and irreversibility pages 1029 to 1076. In *Climate change 2013 – the physical science basis: Working group i contribution to the fifth assessment report of the intergovernmental panel on climate change*.
- Cummins, D. P., Stephenson, D. B., & Stott, P. A. (2020). Optimal estimation of stochastic energy balance model parameters. *Journal of Climate*.
- Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D., DuVivier, A., Edwards, J., ... others (2020). The community earth system model version 2 (cesm2). *Journal of Advances in Modeling Earth Systems*, 12(2), e2019MS001916.
- Del Rio Amador, L., & Lovejoy, S. (2019). Predicting the global temperature with the stochastic seasonal to interannual prediction system (stocsips). *Climate Dynamics*.
- 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*.
- Flato, G. M. (2011). Earth system models: an overview. *WIREs Climate Change*.
- Forster, P., Ramaswamy, V., Artaxo, P., Berntsen, T., Betts, R., Fahey, D. W., ... others (2007). Changes in atmospheric constituents and in radiative forcing. chapter 2. In *Climate change 2007. the physical science basis*.
- Fredriksen, H.-B., & Rypdal, M. (2017). Long-range persistence in global surface temperatures explained by linear multibox energy balance models. *Journal of Climate*.
- Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., & Wilson, A. G. (2018). Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration [Software]. *Advances in neural information processing systems*, 31. Retrieved from <https://github.com/cornellius-gp/gpytorch>
- Gasser, T., Ciais, P., Boucher, O., Quilcaille, Y., Tortora, M., Bopp, L., & Hauglustaine, D. (2017). The compact earth system model oscar v2. 2: description and first results.
- Geoffroy, O., Saint-Martin, D., Olivie, D. J. L., Voldoire, A., Bellon, G., & Tytéca, S.

- (2013). Transient climate response in a two-layer energy-balance model. part i: Analytical solution and parameter calibration using cmip5 aogcm experiments. *Journal of Climate*.
- Goodwin, P., Leduc, M., Partanen, A.-I., Matthews, H. D., & Rogers, A. (2020). A computationally efficient method for probabilistic local warming projections constrained by history matching and pattern scaling, demonstrated by wasp-lgrtc-1.0. *Geoscientific Model Development*.
- Hartikainen, J., & Särkkä, S. (2010). Kalman filtering and smoothing solutions to temporal gaussian process regression models. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*.
- Hasselmann, K. (1976). Stochastic climate models part i. theory. *Tellus*.
- Hawkins, E., & Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, 90(8), 1095–1108.
- Held, I. M., Winton, M., Takahashi, K., Delworth, T., Zeng, F., & Vallis, G. K. (2010). Probing the fast and slow components of global warming by returning abruptly to preindustrial forcing. *Journal of Climate*.
- Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., ... others (2018). Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the community emissions data system (ceds). *Geoscientific Model Development*, 11(1), 369–408.
- Holden, P., & Edwards, N. (2010). Dimensionally reduced emulation of an aogcm for application to integrated assessment modelling. *Geophysical Research Letters*.
- Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment [Software]. *Computing in science & engineering*, 9(03), 90–95. Retrieved from <https://github.com/matplotlib/matplotlib>
- Huntingford, C., & Cox, P. (2000). An analogue model to derive additional climate change scenarios from existing gcm simulations. *Climate dynamics*, 16, 575–586.
- Jackson, L. S., Maycock, A. C., Andrews, T., Fredriksen, H.-B., Smith, C. J., & Forster, P. M. (2022). Errors in simple climate model emulations of past and future global temperature change. *Geophysical Research Letters*.
- Law, H. C., Zhao, P., Chan, L. S., Huang, J., & Sejdinovic, D. (2019). Hyperparameter learning via distributional transfer. *Advances in Neural Information Processing Systems*, 32.
- Leach, N. J., Jenkins, S., Nicholls, Z., Smith, C. J., Lynch, J., Cain, M., ... Allen, M. R. (2021). Fairv2.0.0: a generalized impulse response model for climate uncertainty and future scenario exploration. *Geoscientific Model Development*.
- Lee, J.-Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J. P., ... others (2021). Future global climate: scenario-based projections and near-term information. In *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. 553–672). Cambridge University Press.
- Levy, H., Horowitz, L. W., Schwarzkopf, M. D., Ming, Y., Golaz, J.-C., Naik, V., & Ramaswamy, V. (2013). The roles of aerosol direct and indirect effects in past and future climate change. *Journal of Geophysical Research: Atmospheres*.
- Link, R., Snyder, A., Lynch, C., Hartin, C., Kravitz, B., & Bond-Lamberty, B. (2019). Fldgen v1. 0: an emulator with internal variability and space–time correlation for earth system models. *Geoscientific Model Development*.
- Lovejoy, S., Procyk, R., Hébert, R., & Del Rio Amador, L. (2021). The fractional energy balance equation. *Quarterly Journal of the Royal Meteorological Society*.
- Manabe, S., & Wetherald, R. T. (1967). Thermal equilibrium of the atmosphere with a given distribution of relative humidity.
- Martinez-Villalobos, C., & Neelin, J. D. (2019). Why do precipitation intensities tend to follow gamma distributions? *Journal of the Atmospheric Sciences*.

- Matthews, A. G. d. G., Hensman, J., Turner, R., & Ghahramani, Z. (2016). On Sparse Variational Methods and the Kullback-Leibler Divergence between Stochastic Processes. In *Proceedings of the 19th international conference on artificial intelligence and statistics*.
- May, W. (2012). Assessing the strength of regional changes in near-surface climate associated with a global warming of 2 c. *Climatic change*.
- McDonald, T., & Álvarez, M. (2021). Compositional modeling of nonlinear dynamical systems with ode-based random features. *Advances in Neural Information Processing Systems*, 34, 13809–13819.
- Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F., ... Taylor, K. E. (2007). The wcrp cmip3 multimodel dataset: A new era in climate change research. *Bulletin of the American meteorological society*.
- Meinshausen, M., Nicholls, Z. R., Lewis, J., Gidden, M. J., Vogel, E., Freund, M., ... others (2020). The shared socio-economic pathway (ssp) greenhouse gas concentrations and their extensions to 2500. *Geoscientific Model Development*.
- 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*.
- Met Office, U. (2010). Cartopy: A cartographic python library with a matplotlib interface [Software]. *Exeter, Devon*. Retrieved from <https://github.com/SciTools/cartopy>
- 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*.
- Mitchell, J. F., Johns, T., Eagles, M., Ingram, W., & Davis, R. (1999). Towards the construction of climate change scenarios. *Climatic Change*.
- Moss, J. D., Opolka, F. L., Dumitrescu, B., & Lió, P. (2021). Approximate latent force model inference. *arXiv preprint arXiv:2109.11851*.
- Myhre, G., Shindell, D., Bréon, F.-M., Collins, W., Fuglestad, J., Huang, J., ... Zhang, H. (2014). Anthropogenic and natural radiative forcing. In *Climate change 2013 – the physical science basis: Working group i contribution to the fifth assessment report of the intergovernmental panel on climate change*.
- O’Neill, B. C., Tebaldi, C., Van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., ... others (2016). The scenario model intercomparison project (scenariomip) for cmip6. *Geoscientific Model Development*, 9(9), 3461–3482.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). Pytorch: An imperative style, high-performance deep learning library [Software]. *Advances in neural information processing systems*, 32. Retrieved from <https://github.com/pytorch/pytorch>
- Pörtner, H.-O., Roberts, D. C., Adams, H., Adler, C., Aldunce, P., Ali, E., ... others (2022). *Climate change 2022: Impacts, adaptation and vulnerability*. IPCC Geneva, Switzerland:.
- Rahimi, A., & Recht, B. (2007). Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20.
- Rasmussen, C., & Williams, C. (2005). *Gaussian process for machine learning*. MIT Press, Cambridge, MA, USA.
- Riahi, K., Van Vuuren, D. P., Kriegler, E., Edmonds, J., O’neill, B. C., Fujimori, S., ... others (2017). The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global environmental change*, 42, 153–168.
- Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., ... others (2014). Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. *Proceedings of the national academy of sciences*.
- Ross, M., Smith, M. T., & Álvarez, M. (2021). Learning nonparametric volterra kernels with gaussian processes. *Advances in Neural Information Processing Systems*, 34,

- 24099–24110.
- Rypdal, K. (2012). Global temperature response to radiative forcing: Solar cycle versus volcanic eruptions. *Journal of Geophysical Research: Atmospheres*.
- Santer, B. D., Wigley, T. M., Schlesinger, M. E., & Mitchell, J. F. (1990). Developing climate scenarios from equilibrium gcm results.
- Sarkka, S., & Hartikainen, J. (2012). Infinite-dimensional kalman filtering approach to spatio-temporal gaussian process regression. In *Artificial intelligence and statistics*.
- Sarkka, S., Solin, A., & Hartikainen, J. (2013). Spatiotemporal learning via infinite-dimensional bayesian filtering and smoothing: A look at gaussian process regression through kalman filtering. *IEEE Signal Processing Magazine*.
- Schöngart, S., Lejeune, Q., Gudmundsson, L., Nath, S., Seneviratne, S., & Schleußner, C.-F. (2023). *Extending mesmer-m to jointly emulate earth system model temperature and precipitation realizations* (Tech. Rep.). Copernicus Meetings.
- Schöngart, S., Lejeune, Q., Schleußner, C.-F., Seneviratne, S., Gudmundsson, L., & Beusch, L. (2022). A spatially explicit approach for joint temperature-precipitation emulation of earth system model simulation. In *Egu general assembly conference abstracts* (pp. EGU22–11049).
- Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., ... others (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.
- Sellers, W. D. (1969). A global climatic model based on the energy balance of the earth-atmosphere system. *Journal of Applied Meteorology and Climatology*, 8(3), 392–400.
- Seneviratne, S., Nicholls, N., Easterling, D., Goodess, C., Kanae, S., Kossin, J., ... others (2012). Changes in climate extremes and their impacts on the natural physical environment. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation, A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC)*, 109–230.
- Seneviratne, S. I., Donat, M. G., Pitman, A. J., Knutti, R., & Wilby, R. L. (2016). Allowable co2 emissions based on regional and impact-related climate targets. *Nature*.
- Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*.
- Shukla, P. R., Skea, J., Slade, R., Al Khourdajie, A., Van Diemen, R., McCollum, D., ... others (2022). Climate change 2022: Mitigation of climate change. *Contribution of working group III to the sixth assessment report of the Intergovernmental Panel on Climate Change*, 10, 9781009157926.
- Smith, C. J., Forster, P. M., Allen, M., Leach, N., Millar, R. J., Passerello, G. A., & Regayre, L. A. (2018). Fair v1.3: a simple emissions-based impulse response and carbon cycle model. *Geoscientific Model Development*.
- Snyder, A., Link, R., Dorheim, K., Kravitz, B., Bond-Lamberty, B., & Hartin, C. (2019). Joint emulation of earth system model temperature-precipitation realizations with internal variability and space-time and cross-variable correlation: fldgen v2. 0 software description. *Plos one*, 14(10), e0223542.
- Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.
- Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*.
- Tandon, A. (2021). Analysis: The lack of diversity in climate-science research. *Carbon Brief*.
- Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of cmip5 and the experiment design. *Bulletin of the American Meteorological Society*.
- Tebaldi, C., & Arblaster, J. M. (2014). Pattern scaling: Its strengths and limitations, and an update on the latest model simulations. *Climatic Change*.
- Tebaldi, C., Armbruster, A., Engler, H., & Link, R. (2020). Emulating climate extreme indices. *Environmental Research Letters*, 15(7), 074006.

- Tebaldi, C., & Knutti, R. (2018). Evaluating the accuracy of climate change pattern emulation for low warming targets. *Environmental Research Letters*, 13(5), 055006.
- Titsias, M. K. (2009). Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *Aistats*.
- Tsutsui, J. (2020). Diagnosing transient response to co2 forcing in coupled atmosphere-ocean model experiments using a climate model emulator. *Geophysical Research Letters*, 47(7), e2019GL085844.
- Turner, S. W., Hejazi, M., Kim, S. H., Clarke, L., & Edmonds, J. (2017). Climate impacts on hydropower and consequences for global electricity supply investment needs. *Energy*.
- Wang, K., Pleiss, G., Gardner, J., Tyree, S., Weinberger, K. Q., & Wilson, A. G. (2019). Exact gaussian processes on a million data points. *Advances in neural information processing systems*, 32.
- Ward, W., Ryder, T., Prangle, D., & Alvarez, M. (2020). Black-box inference for non-linear latent force models. In *International conference on artificial intelligence and statistics* (pp. 3088–3098).
- Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls, G., ... others (2022). ClimateBench v1. 0: A Benchmark for Data-Driven Climate Projections [Data]. *Journal of Advances in Modeling Earth Systems*, 14(10), e2021MS002954.
- Williams, A. I., Stier, P., Dagan, G., & Watson-Parris, D. (2022). Strong control of effective radiative forcing by the spatial pattern of absorbing aerosol. *Nature Climate Change*, 12(8), 735–742.
- Wilson, A., & Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th international conference on machine learning*.
- Wilson, A. G., Hu, Z., Salakhutdinov, R., & Xing, E. P. (2016). Deep kernel learning. In *Artificial intelligence and statistics* (pp. 370–378).

Appendix A Complementary material on Gaussian processes

A1 Illustrated walk-through Gaussian process regression

We provide in this section an illustrated walk-through of how a regression task can be achieved using GPs. Begin by considering a GP specified by

$$\begin{cases} f(x) \sim \text{GP}(m, k) \\ m(x) = 0 \\ k(x, x') = \sigma_f^2 \exp\left(-\frac{|x - x'|^2}{\ell}\right). \end{cases} \quad (\text{A1})$$

Further consider a set of $p = 500$ regularly spaced points between 0 and 1, $\mathbf{grid} = [z_1 \dots z_p]$ where

$$z_i = \frac{i}{p}, \quad 1 \leq i \leq p. \quad (\text{A2})$$

Then, by definition of GPs, the evaluation of f on $\mathbf{grid}$ admits a joint multivariate normal distribution. Namely, if we define the covariance matrix

$$\mathbf{K} = \begin{bmatrix} k(z_1, z_1) & \dots & k(z_1, z_p) \\ \vdots & \ddots & \vdots \\ k(z_p, z_1) & \dots & k(z_p, z_p) \end{bmatrix}, \quad (\text{A3})$$

then we have

$$\begin{bmatrix} f(z_1) \\ \vdots \\ f(z_p) \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}). \quad (\text{A4})$$

Therefore, the zero vector $\mathbf{0}$ provides the mean of $f(x)$ at every grid point, and the diagonal elements of the covariance matrix provide the standard deviation of $f(x)$ at every grid point following $\sqrt{k(z_i, z_i)} = \sigma_f$. Figure A1 plots the mean function and the 95% credible interval of $f(x)$ over the $\mathbf{grid}$.

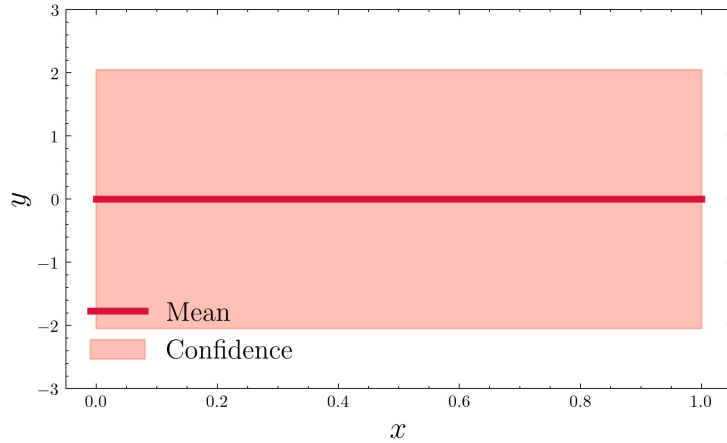


Figure A1: Plot of the $f(x) \sim \text{GP}(0, k)$ prior mean function and 95% confidence region for $\sigma_f^2 = 1$ and $\ell = 0.1$.

Furthermore, we are also able to draw samples from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{K})$ by taking

$$\mathbf{f} = \mathbf{K}^{1/2} \mathbf{u}, \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p). \quad (\text{A5})$$

Such a sample $\mathbf{f} \in \mathbb{R}^p$ corresponds to a sample path from the GP $\mathbf{f}(x)$ evaluated on the grid points. Figure A2 shows the plots of 10 sample paths, from this multivariate normal distribution.

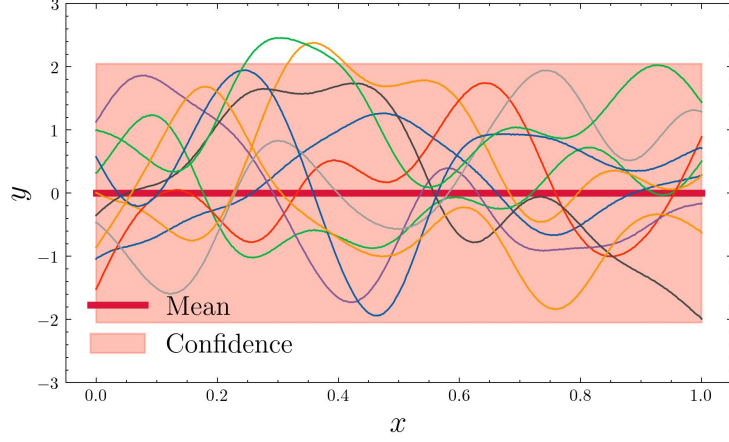


Figure A2: Same as Figure A1, but plotting on top 10 sample paths from $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$. $\sigma_f^2 = 1$ and $\ell = 0.1$.

Consider now the following data generating process

$$y = \sin(2\pi x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 0.04), \quad (\text{A6})$$

from which we simulate $n = 10$ observations $\mathbf{y} = [y_1 \ \dots \ y_n]$ at inputs $\mathbf{x} = [x_1 \ \dots \ x_n]$. Figure A3 shows a plot of the observed data.

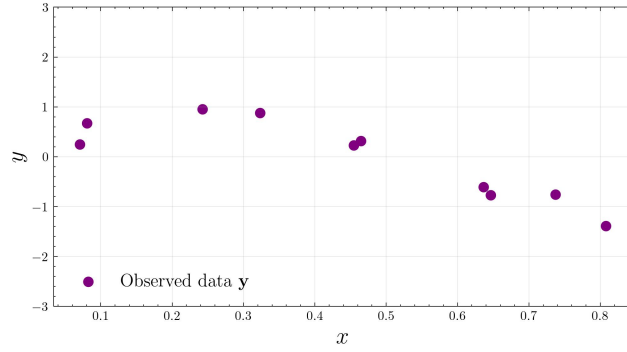


Figure A3: Scatter plot of $(\mathbf{x}, \mathbf{y})$

Suppose that the form of the data generating process is unknown to a user who wishes to use these observations to regress y onto x . One possibility is to use the previously defined GP $\mathbf{f}(x)$ as a Bayesian prior over the space of regression models $x \rightarrow y$ and assume the relationship

$$y_i = \mathbf{f}(x_i) + \varepsilon_i, \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2), \quad (\text{A7})$$

for an unknown noise variance σ_ε^2 .

By definition of GPs, we know that the evaluation of $\mathbf{f}$ on $\mathbf{x}$ will again admit a joint multivariate normal distribution, i.e. by defining the covariance matrix

$$\mathbf{K}_{\mathbf{x}} = \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \dots & k(x_n, x_n) \end{bmatrix} \quad (\text{A8})$$

we get that

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\mathbf{x}}), \quad (\text{A9})$$

and therefore by incorporating the additive noise, it follows that

$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\mathbf{x}} + \sigma_{\varepsilon}^2 \mathbf{I}_n). \quad (\text{A10})$$

Firstly, using this distribution, we can find values of the hyperparameters $\theta = (\sigma_f, \ell, \sigma_{\varepsilon})$ that maximise $\log p(\mathbf{y}|\mathbf{x})$. This corresponds to finding their maximum likelihood estimates under $p(\mathbf{y}|\mathbf{x})$, and can be achieved for example following the gradient descent procedure outlined in the following algorithm. Such a procedure offers a principled and robust way

Algorithm 2 Maximum likelihood calibration of $\theta = (\sigma_f, \ell, \sigma_{\varepsilon})$ with gradient descent

- 1: **Input:** $\mathbf{x}, \mathbf{y}, \theta_0, \eta > 0, N \in \mathbb{N}$
 - 2: **for** $t \in \{1, \dots, N\}$ **do**
 - 3: Compute $\log p(\mathbf{y}|\mathbf{x}) = \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_{\mathbf{x}} + \sigma_{\varepsilon}^2 \mathbf{I}_n)$
 - 4: Take $\theta_t \leftarrow \theta_{t-1} + \eta \nabla_{\theta} \log p(\mathbf{y}|\mathbf{x})$
 - 5: **end for**
 - 6: **Return:** θ_{τ}
-

to calibrate the model hyperparameters against observations. Figure A4 shows a plot of the GP mean and confidence region with sample paths that have been calibrated against observations. Calibrating the prior against $(\mathbf{x}, \mathbf{y})$ yields a smaller variance $\sigma_f^2 = 0.43$, which translates in a tighter confidence region, and a larger lengthscale $\ell = 0.17$, which translates into slower variations of the sample paths.

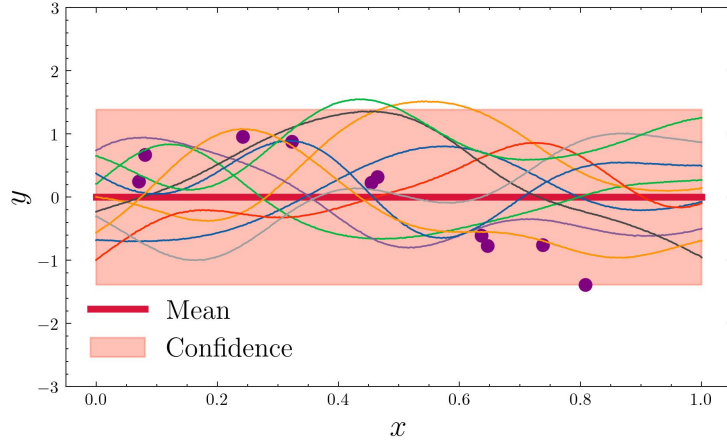


Figure A4: Plot of the $f(x) \sim \text{GP}(0, k)$ prior mean function and 95% confidence region on the `grid` for calibrated hyperparameters $\sigma_f^2 = 0.43$ and $\ell = 0.17$, with 10 sample paths and observed data $(\mathbf{x}, \mathbf{y})$.

Second, we can update the prior $f(x)$ placed over the predictive model into a posterior informed with the observed data $f(x)|\mathbf{y}$. Formally, the mean and covariance of the posterior can be computed analytically following

$$\begin{cases} f(x)|\mathbf{y} \sim \text{GP}(\bar{m}, \bar{k}) \\ \bar{m}(x) = k(x, \mathbf{x})(\mathbf{K}_{\mathbf{x}} + \sigma_{\varepsilon}^2 \mathbf{I}_n)^{-1} \mathbf{y} \\ \bar{k}(x, x') = k(x, x') - k(x, \mathbf{x})(\mathbf{K}_{\mathbf{x}} + \sigma_{\varepsilon}^2 \mathbf{I}_n)^{-1} k(\mathbf{x}, x'), \end{cases} \quad (\text{A11})$$

where $k(x, \mathbf{x}) = [k(x, x_1) \ \dots \ k(x, x_n)]$ and $k(\mathbf{x}, x') = [k(x_1, x') \ \dots \ k(x_n, x')]^\top$.

Since this is still a GP, we can plot it on the `grid` using the exact same procedure as for the prior. Namely, define the posterior mean vector and posterior covariance matrix on `grid`

$$\bar{\mathbf{m}} = \begin{bmatrix} \bar{m}(z_1) \\ \vdots \\ \bar{m}(z_p) \end{bmatrix} \quad \bar{\mathbf{K}} = \begin{bmatrix} \bar{k}(z_1, z_1) & \dots & \bar{k}(z_1, z_p) \\ \vdots & \ddots & \vdots \\ \bar{k}(z_p, z_1) & \dots & \bar{k}(z_p, z_p) \end{bmatrix}. \quad (\text{A12})$$

Then

$$\begin{bmatrix} f(z_1)|\mathbf{y} \\ \vdots \\ f(z_p)|\mathbf{y} \end{bmatrix} \sim \mathcal{N}(\bar{\mathbf{m}}, \bar{\mathbf{K}}). \quad (\text{A13})$$

Therefore, the posterior mean vector $\bar{\mathbf{m}}$ provides the mean of $f(x)|\mathbf{y}$ at every grid point and the diagonal elements of the covariance matrix $\bar{\mathbf{K}}$ provide the standard deviation of $f(x)|\mathbf{y}$ at every grid point (which does not simplify to σ_f this time). Figure A5 illustrates how updating the prior with 1, 2 and 10 observations affects the posterior mean and the posterior credible interval.

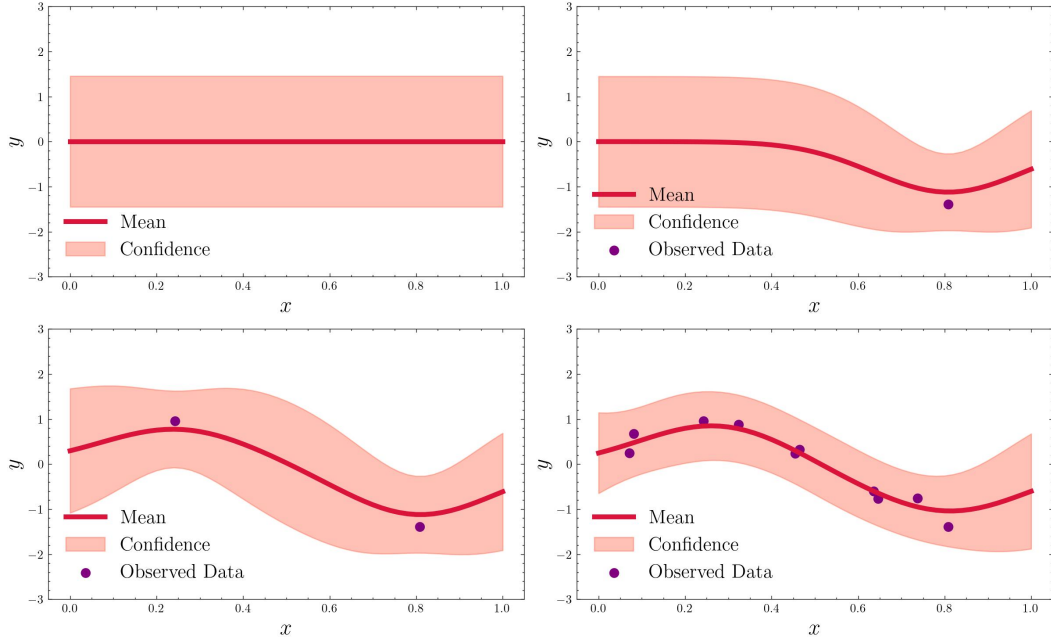


Figure A5: Plots of the posterior mean function and posterior 95% credible interval when updating the prior with 0, (top-left), 1 (top-right), 2 (bottom-left) and 10 (bottom-right) observations.

Furthermore, we are also able to draw samples from the multivariate normal distribution $\mathcal{N}(\bar{\mathbf{m}}, \bar{\mathbf{K}})$ by taking

$$\bar{\mathbf{f}} = \bar{\mathbf{m}} + \bar{\mathbf{K}}^{1/2} \mathbf{u}, \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p). \quad (\text{A14})$$

Such a sample $\bar{\mathbf{f}} \in \mathbb{R}^p$ corresponds to a sample path from the posterior GP $f(x)|\mathbf{y}$ evaluated on the grid points. Figure A6 shows the plots of 10 sample paths from this multivariate normal distribution. Unlike for the prior, the sample paths are now concentrated around the observations and therefore provide reasonable candidates for a predictive model $x \rightarrow y$. In other words, the posterior GP induces a probability distribution over functions that provide a sound fit to observations, and are reasonable predictive models.

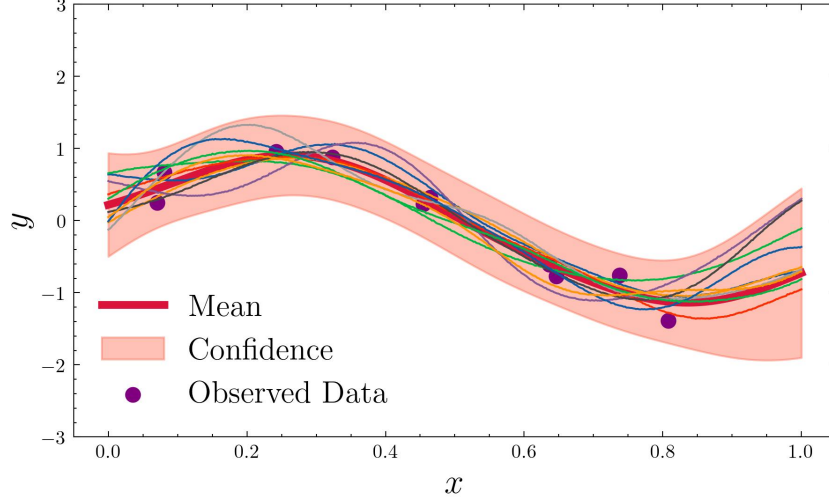


Figure A6: Plots of the posterior mean function and posterior 95% credible interval + plots of 10 sample paths from $\mathbf{f} \sim \mathcal{N}(\bar{\mathbf{m}}, \mathbf{K})$.

A2 Matérn covariance

The Matérn covariances are a class of stationary covariance functions widely used in spatial statistics. The Matérn- ν covariance between two points $x, x' \in \mathbb{R}$ is given by

$$C_\nu(x, x') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|x - x'|}{\ell} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{|x - x'|}{\ell} \right), \quad (\text{A15})$$

where Γ is the gamma function, K_ν is the modified Bessel function, σ^2 is a variance parameter and $\ell > 0$ is a lengthscale hyperparameter.

The covariance function expression considerably simplifies for $\nu = p + 1/2$ where $p \in \mathbb{N}$. For example, for $\nu = 1/2$ ($p = 0$), $\nu = 3/2$ ($p = 1$) and $\nu = 5/2$ ($p = 2$) we have

$$C_{1/2}(x, x') = \sigma^2 \exp \left(-\frac{|x - x'|}{\ell} \right) \quad (\text{A16})$$

$$C_{3/2}(x, x') = \sigma^2 \left(1 + \sqrt{3} \frac{|x - x'|}{\ell} \right) \exp \left(-\sqrt{3} \frac{|x - x'|}{\ell} \right) \quad (\text{A17})$$

$$C_{5/2}(x, x') = \sigma^2 \left(1 + \sqrt{5} \frac{|x - x'|}{\ell} + \frac{5|x - x'|^2}{3\ell^2} \right) \exp \left(-\sqrt{5} \frac{|x - x'|}{\ell} \right) \quad (\text{A18})$$

When using the Matérn covariance as the covariance function of a GP, larger values of ν bestow greater smoothness on the GP sample paths. Namely, when $\nu = 1/2$ sample paths are continuous, when $\nu = 3/2$ sample paths are differentiable and when $\nu = 5/2$ sample paths are twice differentiable. More generally, for a Matérn- $p + 1/2$ covariance function, sample paths from a GP are p times continuously differentiable (with the convention that 0 times means simply continuous). Figure A7 shows sample paths drawn from GPs with Matérn kernels for different values of ν .

In the limit where $\nu \rightarrow \infty$, the Matérn covariance converges to the squared exponential covariance given by

$$C_\infty(x, x') = \sigma^2 \exp \left(-\frac{|x - x'|^2}{2\ell^2} \right). \quad (\text{A19})$$

The sample paths of a GP with a squared exponential covariance function are infinitely differentiable.

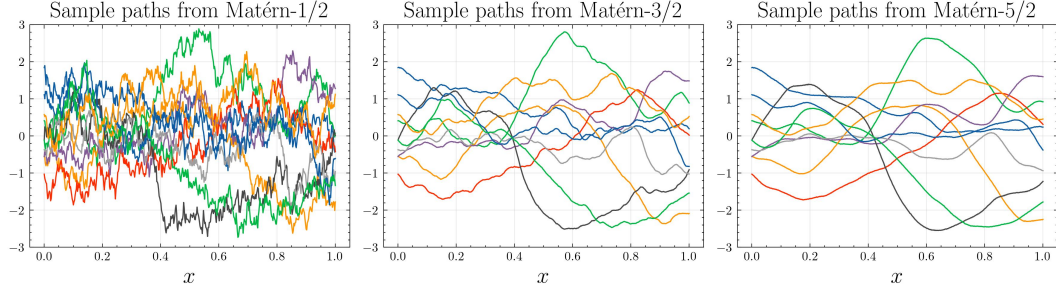


Figure A7: Example of sample paths drawn from a GP with kernel set as a Matérn-1/2, Matérn-3/2 and Matérn-5/2.

When $x, x' \in \mathbb{R}^d$, the distance $|x - x'|$ can be substituted by the norm $\|x - x'\| = \sqrt{\sum_{i=1}^d (x_i - x'_i)^2}$. The covariance is called an anisotropic — or automatic relevance determination — kernel when each dimension has its own independent lengthscale parameter $\ell_i > 0$. For example, the Matérn-1/2 and Matérn-3/2 anisotropic kernels write

$$C_{1/2}(x, x') = \sigma^2 \exp \left(-\sqrt{\sum_{i=1}^d \frac{(x_i - x'_i)^2}{\ell_i^2}} \right) \quad (\text{A20})$$

$$C_{3/2}(x, x') = \sigma^2 \left(1 + \sqrt{3 \sum_{i=1}^d \frac{(x_i - x'_i)^2}{\ell_i^2}} \right) \exp \left(-\sqrt{3 \sum_{i=1}^d \frac{(x_i - x'_i)^2}{\ell_i^2}} \right). \quad (\text{A21})$$

Appendix B Supporting materials for FaIRGP derivation

B1 Useful results

Lemma Appendix B.1. *Let $a, b > 0$ and $u, v \geq 0$. We have*

$$(a + b) \min(u, v) - (au + bv) = \begin{cases} -b|u - v| & \text{if } u \leq v \\ -a|u - v| & \text{if } u \geq v. \end{cases} \quad (\text{B1})$$

Proof. If $u \leq v$,

$$\begin{aligned} (a + b) \min(u, v) - (au + bv) &= (a + b)u - au - bv \\ &= bu - bv \\ &= -b|u - v|. \end{aligned}$$

If $u \geq v$,

$$\begin{aligned} (a + b) \min(u, v) - (au + bv) &= (a + b)v - au - bv \\ &= av - au \\ &= -a|u - v|. \end{aligned}$$

□

B2 Derivation of FaIRGP

Consider a k -box EBM specified by the stochastic temperature response model

$$d\mathbf{X}(t) = \mathbf{A}\mathbf{X}(t) dt + \mathbf{b}F(t) dt, \quad (\text{B2})$$

where $F(t) \sim \text{GP}(F, K)$, a forcing feedback vector $\mathbf{b} = [1/C_1 \ 0 \ \dots \ 0]^\top$, and a forcing feedback matrix given by

$$\mathbf{A} = \begin{bmatrix} -(\kappa_1 + \kappa_2)/C_1 & \kappa_2/C_1 & 0 & \dots & 0 & 0 & 0 \\ \kappa_2/C_2 & -(\kappa_2 + \kappa_3)/C_2 & \kappa_3/C_2 & \dots & 0 & 0 & 0 \\ 0 & \kappa_3/C_3 & -(\kappa_3 + \kappa_4)/C_3 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -(\kappa_{k-2} + \kappa_{k-1})/C_{k-2} & \kappa_{k-1}/C_{k-2} & 0 \\ 0 & 0 & 0 & \dots & \kappa_{k-1}/C_{k-1} & -(\kappa_{k-1} + \kappa_k)/C_{k-1} & \kappa_k/C_{k-1} \\ 0 & 0 & 0 & \dots & 0 & \kappa_k/C_k & -\kappa_k/C_k \end{bmatrix}$$

Denote $\mathbf{A} = \mathbf{\Phi}\mathbf{D}\mathbf{\Phi}^{-1}$ the diagonalisation of the feedback matrix, where $\mathbf{D}$ is a diagonal matrix with diagonal elements

$$\mathbf{D} = \text{Diag}(D_1, \dots, D_k). \quad (\text{B3})$$

Then we have

$$d\mathbf{X}(t) = \mathbf{A}\mathbf{X}(t) dt + \mathbf{b}F(t) dt \quad (\text{B4})$$

$$\Rightarrow \mathbf{\Phi}^{-1} d\mathbf{X}(t) = \mathbf{D}\mathbf{\Phi}^{-1}\mathbf{X}(t) dt + \mathbf{\Phi}^{-1}\mathbf{b}F(t) dt \quad (\text{B5})$$

$$\Rightarrow d[\mathbf{\Phi}^{-1}\mathbf{X}(t)] = \mathbf{D}[\mathbf{\Phi}^{-1}\mathbf{X}(t)] dt + [\mathbf{\Phi}^{-1}\mathbf{b}] F(t) dt. \quad (\text{B6})$$

Therefore, by using the notations from (Millar et al., 2017)

$$D_i = -1/d_i \quad (\text{B7})$$

$$\mathbf{\Phi}^{-1}\mathbf{X}(t) = [\mathbf{S}_1(t) \ \dots \ \mathbf{S}_k(t)]^\top \quad (\text{B8})$$

$$\mathbf{\Phi}^{-1}\mathbf{b} = [q_1/d_1 \ \dots \ q_k/d_k]^\top, \quad (\text{B9})$$

the diagonalised system can be written as an impulse response system, where for $i \in \{1, \dots, k\}$ we have

$$\boxed{dS_i(t) = -\frac{1}{d_i} S_i(t) dt + \frac{q_i}{d_i} F(t) dt = \frac{1}{d_i} (q_i F(t) - S_i(t)) dt.} \quad (\text{B10})$$

Consider now the thermal response $S_i(t)$ of the i^{th} stochastic differential equation (SDE). This thermal response is given by

$$S_i(t) = \frac{q_i}{d_i} \int_0^t F(s) e^{-(t-s)/d_i} ds, \quad (\text{B11})$$

where we recall that $F(t) \sim \text{GP}(F, K)$. Because GPs are closed under linear transformations, $S_i(t)$ must also be GPs for all $i \in \{1, \dots, k\}$. Their mean functions can be computed following

$$m_i(t) := \mathbb{E}[S_i(t)] = \mathbb{E} \left[\frac{q_i}{d_i} \int_0^t F(s) e^{-(t-s)/d_i} ds \right] \quad (\text{B12})$$

$$= \frac{q_i}{d_i} \int_0^t \mathbb{E}[F(s)] e^{-(t-s)/d_i} ds \quad (\text{B13})$$

$$= \frac{q_i}{d_i} \int_0^t F(s) e^{-(t-s)/d_i} ds. \quad (\text{B14})$$

Their cross-covariance functions can be computed following

$$k_{ij}(t, t') := \text{Cov}(S_i(t), S_j(t')) \quad (\text{B15})$$

$$= \text{Cov} \left(\frac{q_i}{d_i} \int_0^t F(s) e^{-(t-s)/d_i} ds, \frac{q_j}{d_j} \int_0^{t'} F(s') e^{-(t'-s')/d_j} ds' \right) \quad (\text{B16})$$

$$= \frac{q_i q_j}{d_i d_j} \int_0^t \int_0^{t'} \text{Cov}(F(s), F(s')) e^{-(t-s)/d_i} e^{-(t'-s')/d_j} ds ds' \quad (\text{B17})$$

$$= \frac{q_i q_j}{d_i d_j} \int_0^t \int_0^{t'} K(s, s') e^{-(t-s)/d_i} e^{-(t'-s')/d_j} ds ds'. \quad (\text{B18})$$

$$(\text{B19})$$

So we have $S_i(t) \sim \text{GP}(m_i, k_{ii})$ for any $i \in \{1, \dots, k\}$. Finally, if we define $T(t) = \sum_{i=1}^k S_i(t)$, since the sum of Gaussians is still a Gaussian, we know that $T(t)$ must also be a GP. And we can compute its mean function following

$$m_T(t) := \mathbb{E}[T(t)] = \mathbb{E} \left[\sum_{i=1}^k S_i(t) \right] = \sum_{i=1}^k \mathbb{E}[S_i(t)] = \sum_{i=1}^k m_i(t), \quad (\text{B20})$$

and its covariance function following

$$k_T(t, t') := \text{Cov}(T(t), T(t')) = \text{Cov} \left(\sum_{i=1}^k S_i(t), \sum_{j=1}^k S_j(t') \right) \quad (\text{B21})$$

$$= \sum_{i=1}^k \sum_{j=1}^k \text{Cov}(S_i(t), S_j(t')) \quad (\text{B22})$$

$$= \sum_{i=1}^k \sum_{j=1}^k k_{ij}(t, t'). \quad (\text{B23})$$

We conclude that

$$\boxed{\begin{cases} \mathbf{T}(t) \sim \text{GP}(m_{\mathbf{T}}, k_{\mathbf{T}}) \\ m_{\mathbf{T}}(t) = \sum_{i=1}^k m_i(t) \\ k_{\mathbf{T}}(t, t') = \sum_{i=1}^k \sum_{j=1}^k k_{ij}(t, t'). \end{cases}} \quad (\text{B24})$$

B3 Accounting for climate internal variability in FaIRGP

Consider again the same problem, but introducing an additional white noise term in the temperature response SDE

$$d\mathbf{X}(t) = \mathbf{A}\mathbf{X}(t) dt + \mathbf{b}F(t) dt + \sigma \mathbf{b} dB(t), \quad (\text{B25})$$

where $B(t)$ denotes the standard Brownian motion.

Following the same derivation steps we get

$$d\mathbf{X}(t) = \mathbf{A}\mathbf{X}(t) dt + \mathbf{b}F(t) dt + \sigma \mathbf{b} dB(t) \quad (\text{B26})$$

$$\Rightarrow \Phi^{-1} d\mathbf{X}(t) = \mathbf{D}\Phi^{-1}\mathbf{X}(t) dt + \Phi^{-1}\mathbf{b}F(t) dt + \Phi^{-1}\mathbf{c} dB(t) \quad (\text{B27})$$

$$\Rightarrow d[\Phi^{-1}\mathbf{X}(t)] = \mathbf{D}[\Phi^{-1}\mathbf{X}(t)] dt + [\Phi^{-1}\mathbf{b}] F(t) dt + \sigma [\Phi^{-1}\mathbf{b}] dB(t). \quad (\text{B28})$$

which gives an impulse response form

$$\boxed{dS_i(t) = \frac{1}{d_i}(q_i F(t) - S_i(t)) dt + \sigma \frac{q_i}{d_i} dB(t), \quad \forall i \in \{1, \dots, k\}. \quad (\text{B29})$$

The solution to the i^{th} SDE is now given by

$$S_i(t) = \underbrace{\frac{q_i}{d_i} \int_0^t F(s) e^{-(t-s)/d_i} ds}_{S_i^\circ(t)} + \sigma \underbrace{\frac{q_i}{d_i} \int_0^t e^{-(t-s)/d_i} dB(s)}_{\xi_i(t)} \quad (\text{B30})$$

$$= S_i^\circ(t) + \sigma \xi_i(t). \quad (\text{B31})$$

$S_i^\circ(t)$ corresponds to the GP we have obtained in the previous derivation without white noise, i.e. $S_i^\circ(t) \sim \text{GP}(m_i, k_{ii})$. $\xi_i(t)$ is also a GP, with mean zero, and cross-covariance

function given by

$$\begin{aligned}
\text{Cov}(\xi_i(t), \xi_j(t')) &= \mathbb{E}[(\xi_i(t) - \mathbb{E}[\xi_i(t)])(\xi_j(t') - \mathbb{E}[\xi_j(t')])] \\
&= \mathbb{E}[\xi_i(t)\xi_j(t')] \\
&= \mathbb{E}\left[\left(\frac{q_i}{d_i} \int_0^t e^{-(t-s)/d_i} d\mathbf{B}(s)\right) \left(\frac{q_j}{d_j} \int_0^{t'} e^{-(t'-s')/d_j} d\mathbf{B}(s')\right)\right] \\
&= \frac{q_i q_j}{d_i d_j} e^{-t/d_i - t'/d_j} \mathbb{E}\left[\int_0^t e^{s/d_i} d\mathbf{B}(s) \int_0^{t'} e^{s'/d_j} d\mathbf{B}(s')\right] \\
&= \frac{q_i q_j}{d_i d_j} e^{-t/d_i - t'/d_j} \int_0^{\min(t, t')} e^{(1/d_i + 1/d_j)s} ds \\
&= \frac{q_i q_j}{d_i d_j} e^{-t/d_i - t'/d_j} \frac{d_i d_j}{d_i + d_j} \left(e^{\frac{d_i + d_j}{d_i d_j} \min(t, t')} - 1\right) \\
&= \frac{q_i q_j}{d_i + d_j} e^{-(d_j t + d_i t')/d_i d_j} \left(e^{\frac{d_i + d_j}{d_i d_j} \min(t, t')} - 1\right) \\
&= \begin{cases} \frac{q_i q_j}{d_i + d_j} \left(e^{-|t-t'|/d_i} - e^{-(d_j t + d_i t')/d_i d_j}\right) & \text{if } t \leq t' \\ \frac{q_i q_j}{d_i + d_j} \left(e^{-|t-t'|/d_j} - e^{-(d_j t + d_i t')/d_i d_j}\right) & \text{if } t \geq t' \end{cases} \quad (\text{Lemma Appendix B.1}) \\
&\sim \begin{cases} \frac{q_i q_j}{d_i + d_j} e^{-|t-t'|/d_i} & \text{if } t \leq t' \\ \frac{q_i q_j}{d_i + d_j} e^{-|t-t'|/d_j} & \text{if } t \geq t' \end{cases} \quad \text{when } t \gg d_i \text{ or } t' \gg d_j.
\end{aligned}$$

Therefore, if we define

$$\gamma_{ij}(t, t') = \frac{q_i q_j}{d_i + d_j} \exp\left(\frac{-|t - t'|}{d_i \mathbb{1}_{\{t \leq t'\}} + d_j \mathbb{1}_{\{t > t'\}}}\right), \quad (\text{B32})$$

which simplifies when $i = j$ to

$$\gamma_i(t, t') = \gamma_{ii}(t, t') = \frac{q_i^2}{2d_i} \exp\left(-\frac{|t - t'|}{d_i}\right), \quad (\text{B33})$$

we obtain that in the long time regime, we can approximate $\xi_i(t) \sim \text{GP}(0, \gamma_i)$. And because $S_i^\circ(t)$ and $\xi_i(t)$ are independent processes, we obtain that

$$\boxed{S_i(t) \sim \text{GP}(m_i, k_{ii} + \sigma^2 \gamma_i)}. \quad (\text{B34})$$

Finally, if we take again $\mathbb{T}(t) = \sum_{i=1}^k S_i(t)$, then $\mathbb{T}(t)$ must be a GP. Its mean function is given by

$$m_{\mathbb{T}}(t) := \mathbb{E}[\mathbb{T}(t)] = \mathbb{E}\left[\sum_{i=1}^k S_i(t)\right] = \sum_{i=1}^k \mathbb{E}[S_i(t)] = \sum_{i=1}^k m_i(t), \quad (\text{B35})$$

and its covariance function is given by

$$k_{\mathbf{T}}(t, t') := \text{Cov}(\mathbf{T}(t), \mathbf{T}(t')) \quad (\text{B36})$$

$$= \text{Cov} \left(\sum_{i=1}^k S_i(t), \sum_{j=1}^k S_j(t') \right) \quad (\text{B37})$$

$$= \sum_{i,j=1}^k \text{Cov}(S_i(t), S_j(t')) \quad (\text{B38})$$

$$= \sum_{i,j=1}^k \text{Cov}(S_i^{\circ}(t) + \sigma \eta_i(t), S_j^{\circ}(t') + \sigma \eta_j(t')) \quad (\text{B39})$$

$$= \sum_{i,j=1}^k \text{Cov}(S_i^{\circ}(t), S_j^{\circ}(t')) + \sigma^2 \text{Cov}(\eta_i(t), \eta_j(t')) \quad (S_i^{\circ}, S_j^{\circ} \perp\!\!\!\perp \xi_i, \xi_j) \quad (\text{B40})$$

$$= \sum_{i,j=1}^k k_{ij}(t, t') + \sigma^2 \sum_{i,j=1}^k \gamma_{ij}(t, t'). \quad (\text{B41})$$

However, if $t \leq t'$ we have

$$\sum_{i,j=1}^k \gamma_{ij}(t, t') = \sum_{i,j=1}^k \frac{q_i q_j}{d_i + d_j} \exp\left(-\frac{|t - t'|}{d_i}\right) \quad (\text{B42})$$

$$= \sum_{i=1}^k \left(\sum_{j=1}^k \frac{q_i q_j}{d_i + d_j} \right) \exp\left(-\frac{|t - t'|}{d_i}\right) \quad (\text{B43})$$

$$= \sum_{i=1}^k \underbrace{\frac{2d_i}{q_i^2} \left(\sum_{j=1}^k \frac{q_i q_j}{d_i + d_j} \right)}_{\nu_i} \frac{q_i^2}{2d_i} \exp\left(-\frac{|t - t'|}{d_i}\right) \quad (\text{B44})$$

$$= \sum_{i=1}^k \nu_i \gamma_i(t, t'). \quad (\text{B45})$$

When $t \geq t'$ we can refactor terms similarly into a sum over j . By symmetry of the indices we conclude that

$$\left\{ \begin{array}{l} \mathbf{T}(t) \sim \text{GP}(m_{\mathbf{T}}, k_{\mathbf{T}}) \\ m_{\mathbf{T}}(t) = \sum_{i=1}^k m_i(t) \\ k_{\mathbf{T}}(t, t') = \sum_{i,j=1}^k k_{ij}(t, t') + \sigma^2 \sum_{i=1}^k \nu_i \gamma_i(t, t') \\ \nu_i = \sum_{j=1}^k \frac{2d_i q_j}{q_i(d_i + d_j)} \end{array} \right. \quad (\text{B46})$$

B4 Accounting for unforced fluctuations in top-of-atmosphere radiative flux

In addition to the variability arising from unforced transfers of energy modelled in the last section, Cummins et al. (2020) propose to also account for unforced fluctuations in the top-of-atmosphere radiative flux. Let us denote by $\tilde{\mathbf{F}}(t)$ the radiative forcing model proposed by Cummins et al. (2020), to distinguish it from the FaIRGP prior $\mathbf{F}(t)$. They propose to

model $\tilde{F}(t)$ as an Ornstein-Uhlenbeck process,

$$d\tilde{F}(t) = \theta(F(t) - \tilde{F}(t)) dt + \varsigma d\tilde{B}(t) \quad (\text{B47})$$

where $\theta > 0$ is the autocorrelation parameter, $\varsigma > 0$ is the white noise variance, $\tilde{B}(t)$ is another standard Brownian motion (independent from $B(t)$ in the temperature response SDE), and $F(t)$ is a deterministic forcing component.

Assuming that $F(0) = 0$, we know that the solution of this SDE is given by

$$\tilde{F}(t) = \underbrace{\int_0^t \theta F(s) e^{-\theta(t-s)} ds}_{\tilde{F}(t)} + \varsigma \underbrace{\int_0^t e^{-\theta(s-t)} d\tilde{B}(s)}_{\eta(t)} \quad (\text{B48})$$

$$= \tilde{F}(t) + \varsigma \eta(t). \quad (\text{B49})$$

$\tilde{F}(t)$ is a deterministic quantity, whereas $\eta(t)$ is as before a GP with mean zero, and a covariance function γ_θ which can be approximated in the long time regime by

$$\gamma_\theta(t, t') = \frac{1}{2\theta} \exp(-\theta|t - t'|). \quad (\text{B50})$$

Therefore, we get that in the long term regime

$$\tilde{F}(t) \sim \text{GP}(\tilde{F}, \varsigma \gamma_\theta). \quad (\text{B51})$$

This can be combined with the FaIRGP prior by substituting the deterministic forcing $F(t)$ by the FaIRGP forcing prior $F(t) \sim \text{GP}(F, K)$. Namely, if we define

$$\tilde{K}(t, t') = \int_0^t \int_0^{t'} \theta^2 K(s, s') e^{-\theta(t-s)} e^{-\theta(t'-s')} ds ds', \quad (\text{B52})$$

then we obtain a new FaIRGP prior over the forcing

$$\tilde{F}(t) \sim \text{GP}(\tilde{F}, \tilde{K} + \varsigma \gamma_\theta), \quad (\text{B53})$$

which accounts for the unforced fluctuations in the top-of-atmosphere radiative flux like in (Cummins et al., 2020).

B5 Analytical expression for FaIRGP probability distribution

A key benefit of FaIRGP is its complete mathematical tractability. This tractability in particular includes access to the analytical expression of: (i) the marginal probability distribution over the training data, which can be used as an objective to tune the model hyperparameters and (ii) of the probability distribution predicted over emulated temperatures, which can be useful for downstream applications of emulation.

In what follows, we assume access to a training set $\mathcal{D} = \{\mathbf{t}, \mathbf{E}, \mathbf{T}\}$ of size n , where we follow the notational conventions from Section 3.5.

B51 Analytical expression of the marginal log-likelihood

Using the prior mean and covariance of the FaIRGP prior over $\mathbf{T}(t)$, we define a prior mean vector and a prior covariance matrix over $\mathbf{t}, \mathbf{E}$ given by

$$\mathbf{m} = m_{\mathbf{T}}(\mathbf{t}) = \begin{bmatrix} m_{\mathbf{T}}(t_1) \\ \vdots \\ m_{\mathbf{T}}(t_n) \end{bmatrix} \quad (\text{B54})$$

$$\mathbf{K} = k_{\mathbf{T}}(\mathbf{E}, \mathbf{E}) = [k_{\mathbf{T}}(E_i, E_j)]_{1 \leq i, j \leq n}. \quad (\text{B55})$$

Further, consider the temperatures internal variability covariance matrix defined by

$$\mathbf{\Gamma} = \gamma_{\mathbf{T}}(\mathbf{t}, \mathbf{t}) = [\gamma_{\mathbf{T}}(t_i, t_j)]_{1 \leq i, j \leq n}. \quad (\text{B56})$$

Then the prior distribution over temperatures is exactly given by the multivariate normal distribution $\mathcal{N}(\mathbf{m}, \mathbf{K} + \sigma^2 \mathbf{\Gamma})$. Let us denote $\mathbf{K}_{\sigma^2} = \mathbf{K} + \sigma^2 \mathbf{\Gamma}$ for conciseness. Then, we can exactly evaluate the marginal probability of $\mathbf{T} \in \mathbb{R}^n$ under the prior distribution following

$$p(\mathbf{T}|\mathbf{E}, \mathbf{t}) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{K}_{\sigma^2})}} \exp\left(-\frac{1}{2}(\mathbf{T} - \mathbf{m})^\top \mathbf{K}_{\sigma^2}^{-1}(\mathbf{T} - \mathbf{m})\right), \quad (\text{B57})$$

which can be used as a maximisation objective to tune model hyperparameters such that the observed temperatures $\mathbf{T}$ have the greatest possible likelihood under the prior. In practice, we prefer working with the marginal log-likelihood for computational stability. It is given in closed-form by

$$\log p(\mathbf{T}|\mathbf{E}, \mathbf{t}) = -\frac{1}{2} \left\{ n \log(2\pi) + \log \det(\mathbf{K}_{\sigma^2}) + (\mathbf{T} - \mathbf{m})^\top \mathbf{K}_{\sigma^2}^{-1} (\mathbf{T} - \mathbf{m}) \right\}. \quad (\text{B58})$$

Let θ denote a set of model parameters we wish to tune against simulated data $\mathbf{t}, \mathbf{E}, \mathbf{T}$. For example, we can choose the internal variability variance and GP kernel parameters $\theta = \{\sigma, \sigma_{\text{F}}, \ell_1, \dots, \ell_d\}$, FaIR temperature response parameters $\theta = \{d_1, \dots, d_k, q_1, \dots, q_k\}$, the radiative forcing model parameters $\theta = \{\alpha_{\log}^X, \alpha_{\text{lin}}^X, \alpha_{\text{sqr}}^X\}$, or any combination of these options. Let θ_0 be the initial values we arbitrarily choose for these parameters, and $\eta > 0$ is a typically small positive number called the *learning rate*. Then the gradient descent algorithm outlined below proposes a principled procedure to tune θ such that it maximises the marginal log-likelihood $\log p(\mathbf{T}|\mathbf{E}, \mathbf{t})$.

Algorithm 3 Maximum likelihood calibration of θ with gradient descent

- 1: **Input:** $\mathbf{t}, \mathbf{E}, \mathbf{T}, \theta_0, \eta > 0, N \in \mathbb{N}$
 - 2: **for** $t \in \{1, \dots, N\}$ **do**
 - 3: Compute $\log p(\mathbf{T}|\mathbf{E}, \mathbf{t}) = \log \mathcal{N}(\mathbf{T}|\mathbf{m}, \mathbf{K} + \sigma^2 \mathbf{\Gamma})$
 - 4: Take $\theta_t \leftarrow \theta_{t-1} + \eta \nabla_{\theta} \log p(\mathbf{T}|\mathbf{E}, \mathbf{t})$
 - 5: **end for**
 - 6: **Return:** θ_N
-

The learning rate $\eta > 0$ controls the size of the steps taken in the direction of the gradient. A higher learning rate results in larger steps, while a lower learning rate leads to smaller steps. If the learning rate is too high, the algorithm may overshoot the maximum and fail to converge. On the other hand, if the learning rate is too low, the algorithm may take a long time to converge.

In the presented gradient descent algorithm, the algorithm stops after the maximum number of iterations N has been reached. More sophisticated stopping criteria can be devised, such as stopping if the absolute change in the marginal log-likelihood falls below a predefined threshold.

Finally, it may be that the size n of the dataset $\mathcal{D} = \{\mathbf{t}, \mathbf{E}, \mathbf{T}\}$ is so large that the evaluation of $\log p(\mathbf{T}|\mathbf{E}, \mathbf{t})$ is in practice computationally prohibitive. In this case, a popular variant of gradient descent called *mini-batch stochastic gradient descent* can be employed. At each iteration, instead of computing the marginal log-likelihood over the entire dataset, we compute it over a randomly selected subset of the training data, called a mini-batch. This mini-batch typically has a size $m \ll n$. Because only a subset of the data is used

for each update, the updates are noisier, introducing more variability in the optimisation process. It is however more computationally efficient, especially for large datasets, as it processes only a small subset of the data in each iteration.

B52 Analytical expression of the posterior distribution

Suppose that we want to emulate the global temperature response for a different emission scenario where we have access to greenhouse gas and aerosol emission data $E_1^*, \dots, E_m^*$ at times $t_1^* < \dots < t_m^*$. We concatenate them into

$$\mathbf{t}^* = \begin{bmatrix} t_1^* \\ \vdots \\ t_m^* \end{bmatrix} \in \mathbb{R}^m, \quad \mathbf{E}^* = \begin{bmatrix} E_1^* \\ \vdots \\ E_m^* \end{bmatrix} \in \mathbb{R}^{m \times d}, \quad (\text{B59})$$

where $d \geq 0$ is the number of emission agents $\chi_1, \dots, \chi_d$ we observe emissions from.

Using the posterior mean and covariance functions when $\mathbf{T}(t)$ is updated with the training set $\mathcal{D}$, we define a posterior mean vector and posterior covariance matrix over $\mathbf{t}^*, \mathbf{E}^*$ given by

$$\bar{\mathbf{m}}^* = \bar{m}_{\mathbf{T}}(\mathbf{t}^*) \quad (\text{B60})$$

$$\bar{\mathbf{K}}^* = \bar{k}_{\mathbf{T}}(\mathbf{E}^*, \mathbf{E}^*). \quad (\text{B61})$$

Further, consider the emulated temperatures internal variability covariance matrix defined by

$$\mathbf{\Gamma}^* = \gamma_{\mathbf{T}}(\mathbf{t}^*, \mathbf{t}^*). \quad (\text{B62})$$

Then, the posterior distribution over emulated temperatures is exactly given by the multivariate normal distribution $\mathcal{N}(\bar{\mathbf{m}}^*, \bar{\mathbf{K}}^* + \sigma^2 \mathbf{\Gamma}^*)$. Let us denote $\bar{\mathbf{K}}_{\sigma^2}^* = \bar{\mathbf{K}}^* + \sigma^2 \mathbf{\Gamma}^*$ for conciseness. This means that for a given temperature vector $\mathbf{T}^* \in \mathbb{R}^m$, we can exactly evaluate the probability of $\mathbf{T}^*$ under the predicted distribution following

$$p(\mathbf{T}^* | \mathcal{D}, \mathbf{E}^*, \mathbf{t}^*) = \frac{1}{\sqrt{(2\pi)^m \det(\bar{\mathbf{K}}_{\sigma^2}^*)}} \exp \left(-\frac{1}{2} (\mathbf{T}^* - \bar{\mathbf{m}}^*)^\top (\bar{\mathbf{K}}_{\sigma^2}^*)^{-1} (\mathbf{T}^* - \bar{\mathbf{m}}^*) \right). \quad (\text{B63})$$

The vector $\mathbf{T}^*$ may be a retained test scenario, in which case computing $p(\mathbf{T}^* | \mathcal{D}, \mathbf{E}^*, \mathbf{t}^*)$ provides an evaluation of the emulated posterior $\mathcal{N}(\bar{\mathbf{m}}^*, \bar{\mathbf{K}}_{\sigma^2}^*)$. $\mathbf{T}^*$ may also simply correspond to temperature for which we would like to assess the probability under emission scenario $\{\mathbf{t}^*, \mathbf{E}^*\}$. A natural application is in detection attribution studies. Indeed, we can take $\mathbf{T}^*$ to be observed historical temperatures, and $\{\mathbf{t}^*, \mathbf{E}^*\}$ to be a counterfactual historical scenario without anthropogenic forcing. By evaluating $p(\mathbf{T}^* | \mathcal{D}, \mathbf{E}^*, \mathbf{t}^*)$ we would be able to assess the probability of historical temperature observations to occur in a scenario without anthropogenic forcing.

The posterior distribution can also be used to draw a sample $\mathbf{V}$ if needed following

$$\mathbf{V} = (\bar{\mathbf{K}}_{\sigma^2}^*)^{1/2} \mathbf{U} + \bar{\mathbf{m}}^*, \quad \mathbf{U} \sim \mathcal{N}(0, \mathbf{I}_m). \quad (\text{B64})$$

Finally, whilst the above is formulated for global mean surface temperatures, we can also derive analytical forms for the posterior distribution over the radiative forcing and for spatially-resolved emulation.

Appendix C Complementary experimental results

C1 FaIRGP posterior sample paths over SSPs

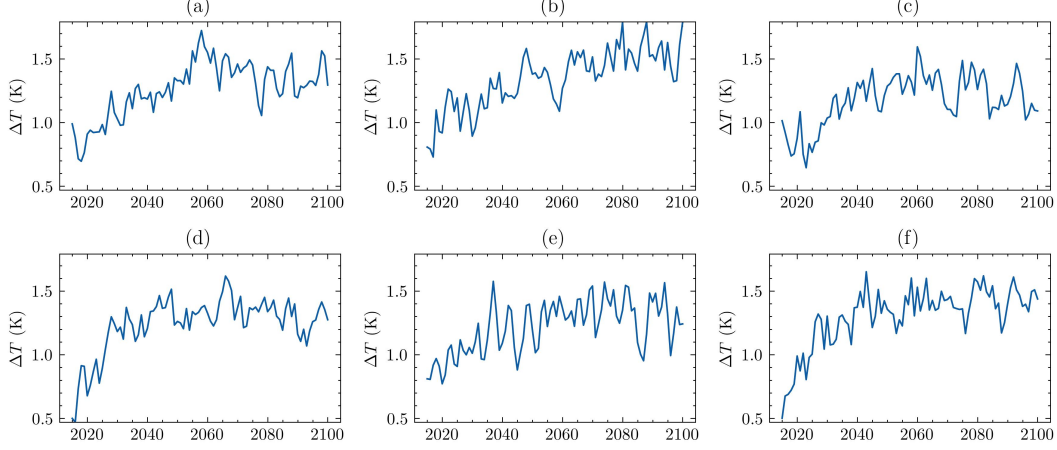


Figure C1: Example of 5 sample paths from the predicted posterior distribution over global annual mean surface temperature anomaly under *SSP126* with FaIRGP + NorESM2-LM simulated global annual mean surface temperature anomaly. The posterior is conditioned on experiments $\mathcal{D}_{\text{train}} = \{\text{historical}, \text{SSP245}, \text{SSP370}, \text{SSP585}\}$. The position of the NorESM2-LM simulated data is given at the end of the section.

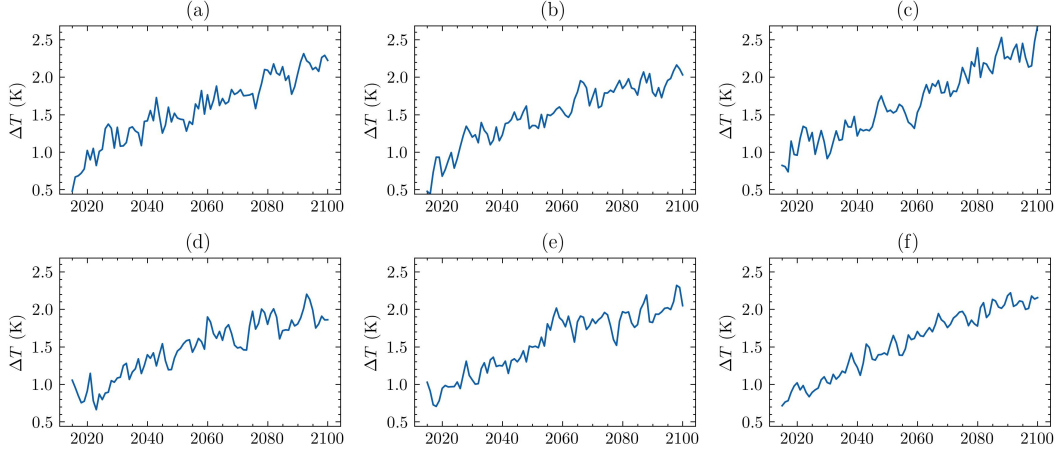


Figure C2: Example of 5 sample paths from the predicted posterior distribution over global annual mean surface temperature anomaly under *SSP245* with FaIRGP + NorESM2-LM simulated global annual mean surface temperature anomaly. The posterior is conditioned on experiments $\mathcal{D}_{\text{train}} = \{\text{historical}, \text{SSP126}, \text{SSP370}, \text{SSP585}\}$. The position of the NorESM2-LM simulated data is given at the end of the section.

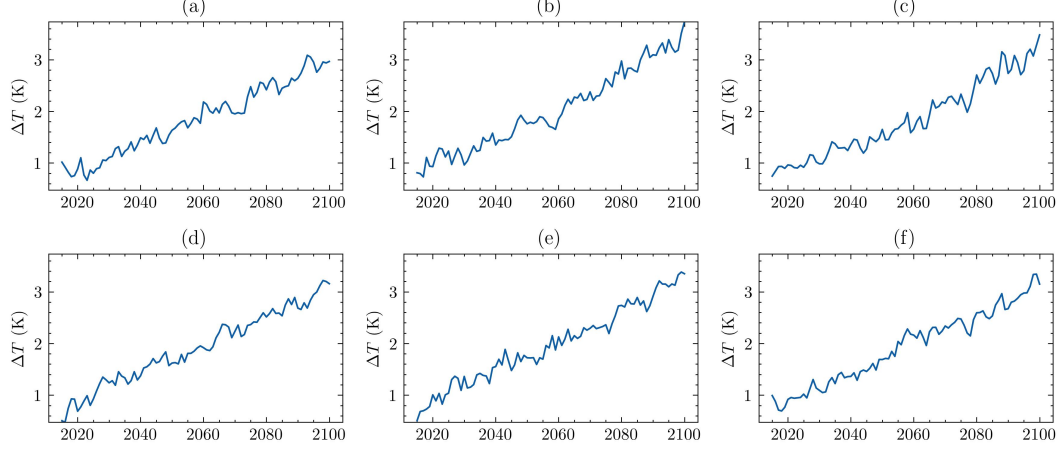


Figure C3: Example of 5 sample paths from the predicted posterior distribution over global annual mean surface temperature anomaly under *SSP370* with FaIRGP + NorESM2-LM simulated global annual mean surface temperature anomaly. The posterior is conditioned on experiments $\mathcal{D}_{\text{train}} = \{\text{historical}, \text{SSP126}, \text{SSP245}, \text{SSP585}\}$. The position of the NorESM2-LM simulated data is given at the end of the section.

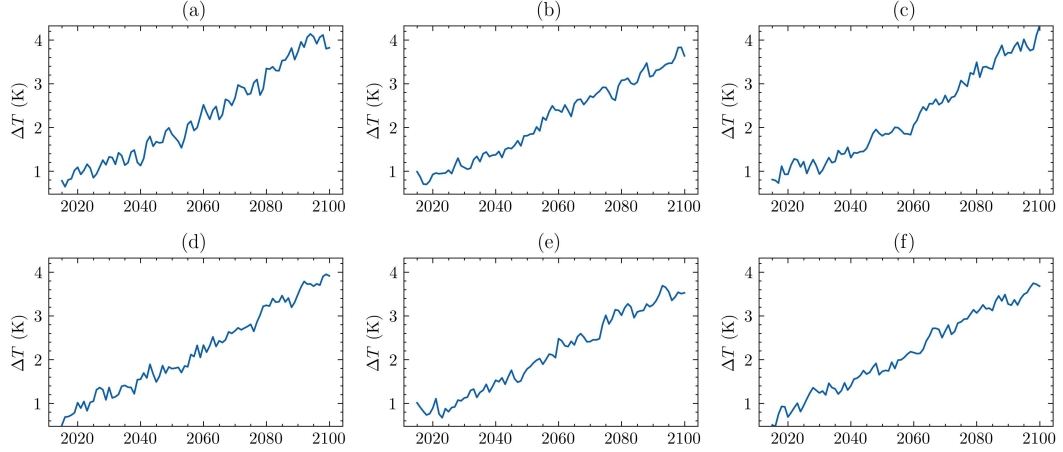


Figure C4: Example of 5 sample paths from the predicted posterior distribution over global annual mean surface temperature anomaly under *SSP585* with FaIRGP + NorESM2-LM simulated global annual mean surface temperature anomaly. The posterior is conditioned on experiments $\mathcal{D}_{\text{train}} = \{\text{historical}, \text{SSP126}, \text{SSP245}, \text{SSP370}\}$. The position of the NorESM2-LM simulated data is given at the end of the section.

Positions of NorESM2-LM simulated data : (e) in Figure C1, (f) in Figure C2, (f) in Figure C3 and (a) in Figure C4.

C2 Comparison of plain GP baseline with ClimateBench GP emulator from Watson-Parris et al. (2022)

C21 Modelling differences

The plain GP baseline we use is analogous to the GP emulator from ClimateBench (Watson-Parris et al., 2022), but differs in two aspects.

First, we use in our plain GP baseline an anisotropic covariance structure, denoted as $\rho(E, E') = C_{3/2}(E, E')$, where $C_{3/2}$ denotes the Matérn-3/2 covariance (see Appendix A2).

This kernel introduces a different lengthscale parameter ℓ_χ for each atmospheric agent χ , which is tuned through maximizing the marginal loglikelihood. It is explicitly given by

$$\rho(E, E') = C_{3/2}(E, E') = \left(1 + \sqrt{3 \sum_\chi \frac{(E^\chi - E'^\chi)^2}{\ell_\chi^2}}\right) \exp\left(-\sqrt{3 \sum_\chi \frac{(E^\chi - E'^\chi)^2}{\ell_\chi^2}}\right), \quad (\text{C1})$$

where E^χ denotes global emission level for agent $\chi \in \{\text{CH}_4, \text{SO}_2, \text{BC}\}$, while for $\chi = \text{CO}_2$, it denotes global cumulative emission levels.

In contrast, Watson-Parris et al. (2022) adopt an additive kernel structure given by

$$\rho(E, E') = \sum_\chi \sigma_\chi C_{3/2}(E^\chi, E'^\chi) = \sum_\chi \sigma_\chi \left(1 + \sqrt{3} \frac{|E^\chi - E'^\chi|}{\ell_\chi}\right) \exp\left(-\sqrt{3} \frac{|E^\chi - E'^\chi|}{\ell_\chi}\right). \quad (\text{C2})$$

This additive structure introduces additional complexity through the variance terms σ_χ , which modulate the of each atmospheric agent to the covariance, and need to be tuned alongside the lengthscales ℓ_χ . From a functional perspective, choosing an additive kernel is equivalent to representing the temperature response as the sum of independent GPs, where each GP models the response to a single atmospheric agent. Further details on the covariance function be found in (Watson-Parris et al., 2022, Appendix A2).

Second, in our plain GP baseline E^{SO_2} and E^{BC} correspond to global emission levels, whereas Watson-Parris et al. (2022) used the 5 principal components of spatial emission maps for SO_2 and BC . We anticipate that spatially-resolved inputs for aerosol emissions should improve the model’s ability to predict spatially-resolved surface temperature features.

C22 Predictive performance comparison

To compare the predictive performance against our plain GP and FaIRGP, we evaluate the ClimateBench GP for the emulation of global and spatial mean surface temperature anomaly and report scores in Table C1. All emulators are trained on the same training data: *historical*, *SSP126*, *SSP370*, *SSP585*. Predictive performance is evaluated for the emulation of *SSP245* since this is the test data used in (Watson-Parris et al., 2022).

Overall, FaIRGP demonstrates improved scores compared to the ClimateBench baseline GP across various metrics, both in terms of global and spatial emulation. In terms of global emulation, the predictive performance of the Plain GP appears to be better than that of the ClimateBench GP. However, when it comes to spatial surface temperature emulation, the ClimateBench GP outperforms the Plain GP. This improvement can likely be attributed to the utilization of spatially-resolved aerosol information as input in the ClimateBench GP.

Table C1: Scores of ClimateBench GP, our Plain GP baseline and FaIRGP for the task of emulating global and spatial mean surface temperatures on *SSP245*; scores are computed over 2015-2100 period for global emulation and over 2080-2100 period for spatial emulation; the best emulator scores are highlighted in bold; $\uparrow/\downarrow$ indicates higher/lower is better.

	Emulator	RMSE $\downarrow$	MAE $\downarrow$	Bias	LL $\uparrow$	Calib95	CRPS $\downarrow$
<i>Global</i>	ClimateBench GP	0.12	0.10	0.01	-0.45	1.0	0.15
	Plain GP	0.09	0.07	0.03	0.67	1.0	0.06
	FaIRGP	0.09	0.08	-0.01	0.71	1.0	0.06
<i>Spatial</i>	ClimateBench GP	0.43	0.32	-0.13	-0.69	0.98	0.25
	Plain GP	0.56	0.40	-0.19	-0.70	0.97	0.28
	FaIRGP	0.36	0.26	-0.05	-0.34	0.99	0.20

C3 Tuned model parameters values

Training data	σ^2	σ_F^2	ℓ_{CO_2}	ℓ_{CH_4}	ℓ_{SO_2}	ℓ_{BC}
All SSPs and historical	0.754	0.244	1.60	3.79	2.98	3.797865
All but SSP126	0.679	0.271	1.57	3.90	2.91	3.86
All but SSP245	0.845	0.276	1.52	3.61	2.86	3.89
All but SSP370	0.709	0.241	1.87	3.87	2.91	3.40
All but SSP585	0.647	0.192	1.58	3.22	2.41	3.18
Historical only	0.385	0.263	3.33	2.57	1.02	3.48

Table C2: Values obtained for internal variability variance σ^2 , forcing kernel variance σ_F^2 and forcing kernel lengthscales $\ell_{CO_2}, \ell_{CH_4}, \ell_{SO_2}, \ell_{BC}$ after tuning using the gradient descent maximum likelihood procedure described in Section 3.6.

C4 Emulating anthropogenic aerosols forcing

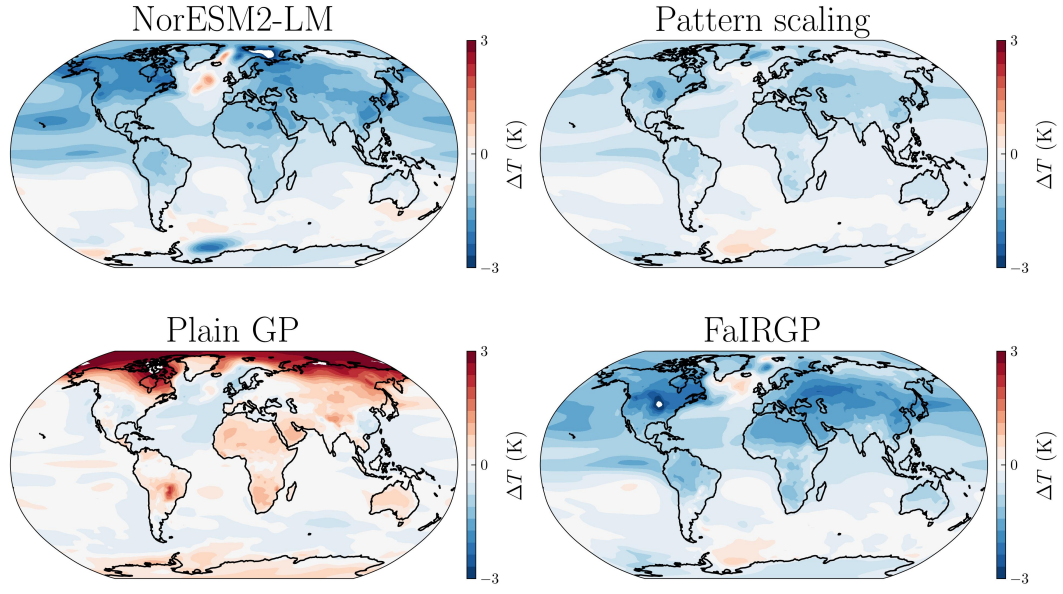


Figure C5: Maps of surface temperatures averaged for the *hist-aer* experiment averaged over 1980-2014 period. **Top-Left:** Groundtruth NorESM2-LM surface temperatures map. **Top-Right:** Prediction with the FaIR pattern scaling baseline. **Bottom-Left:** Prediction with Plain GP posterior mean. **Bottom-Right:** Prediction with the FaIRGP posterior mean.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	FaIRGP: A Bayesian Energy Balance Model for Surface Temperatures Emulation
Publication Status	☐ Published ☒ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Bouabid, Shahine, Sejdinovic, Dino and Watson-Parris, Duncan, 2023. FaIRGP: A Bayesian Energy Balance Model for Surface Temperatures Emulation (to appear in Journal of Advances in Modelling Earth Systems)

Student Confirmation

Student Name:	Shahine Bouabid		
Contribution to the Paper	I am first author of this paper. I came up with the idea of combining an energy balance model with Gaussian process modeling to formulate a hybrid physics-data driven surface temperature emulator. I wrote up the precise model formulation, the derivation of the Bayesian thermal response stochastic differential system, how to account for climate internal variability and how to extend this model to emulate surface temperatures spatially. Prof. Watson-Parris curated the dataset used in the experiments. I designed, implemented and conducted the experiments. I wrote the paper. Prof. Watson-Parris and Prof. Sejdinovic advised along the way and corrected the paper draft.		
Signature		Date	11/03/24

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof. Dino Sejdinovic		
Supervisor comments		
Signature		Date
		12/03/2024

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 5

Aerosol Optical Depth Disaggregation: Toward Global Aerosol Vertical Profile

Bouabid, Shahine, Duncan Watson-Parris, Sofija Stefanović, Athanasios Nenes, Dino Sejdinovic. “Aerosol optical depth disaggregation: toward global aerosol vertical profiles”. *Environmental Data Science*. 2024.

RESEARCH ARTICLE

AODisaggregation: toward global aerosol vertical profiles

Shahine Bouabid^{1*} , Duncan Watson-Parris² , Sofija Stefanović³, Athanasios Nenes⁴  and Dino Sejdinovic⁵ 

¹Department of Statistics, University of Oxford, Oxford, UK

²Scripps Institution of Oceanography and Halicioğlu Data Science Institute, University of California, San Diego, US

³Department of Earth Sciences, University of Cambridge, Cambridge, UK

⁴Laboratory of Atmospheric Processes and their Impacts, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

⁵School of Computer and Mathematical Sciences & AIML, University of Adelaide, Adelaide, Australia

*Corresponding author. Email: shahine.bouabid@stats.ox.ac.uk

Received: 07 May 2022; **Revised:** 23 May 2023; **Accepted:** 10 April 2024

Keywords: AOD; aerosol extinction profiles; vertical disaggregation; Gaussian processes; ECHAM-HAM

Abstract

Aerosol-cloud interactions constitute the largest source of uncertainty in assessments of the anthropogenic climate change. This uncertainty arises in part from the difficulty in measuring the vertical distributions of aerosols, and only sporadic vertically resolved observations are available. We often have to settle for less informative *vertically aggregated* proxies such as aerosol optical depth (AOD). In this work, we develop a framework for the vertical disaggregation of AOD into extinction profiles, i.e. the measure of light extinction throughout an atmospheric column, using readily available vertically resolved meteorological predictors such as temperature, pressure or relative humidity. Using Bayesian nonparametric modelling, we devise a simple Gaussian process prior over aerosol vertical profiles and update it with AOD observations to infer a distribution over vertical extinction profiles. To validate our approach, we use ECHAM-HAM aerosol-climate model data which offers self-consistent simulations of meteorological covariates, AOD and extinction profiles. Our results show that, while very simple, our model is able to reconstruct realistic extinction profiles with well-calibrated uncertainty, outperforming by an order of magnitude the idealised baseline which is typically used in satellite AOD retrieval algorithms. In particular, the model demonstrates a faithful reconstruction of extinction patterns arising from aerosol water uptake in the boundary layer. Observations however suggest that other extinction patterns, due to aerosol mass concentration, particle size and radiative properties, might be more challenging to capture and require additional vertically resolved predictors.

Impact Statement

Aerosol-cloud interactions (ACIs) represent the largest uncertainty in assessments of global warming, with uncertainty bounds that could offset global warming or double its effects. This uncertainty arises in part from the inability to observe aerosol amounts at an appropriate resolution. Instead aerosol optical depth (AOD) — observed globally through satellite products — is commonly used as a two-dimensional proxy for aerosols. Yet, obtaining a global estimate of aerosol vertical distribution in the atmosphere would be more informative and help better constrain ACIs uncertainty. Here, we show that using AOD and readily available vertically resolved meteorological predictors (temperature, pressure, relative humidity, updraft velocity), a simple probabilistic model can already yield realistic vertical extinction profiles together with appropriate uncertainty quantification. This highlights that such simple modelling can benefit satellite products, leading to more accurate priors over aerosol vertical profiles.

1. Introduction

Aerosols are microscopic particles ($< 5 \mu m$) suspended in the atmosphere. They can come from natural sources (e.g. dust, sea salt) or be emitted by human activity (e.g. black carbon).

They influence the Earth's energy budget with a negative radiative forcing that counteracts the global warming from anthropogenic greenhouse gases emissions. A fraction of this negative forcing stems from aerosols' direct scattering of incoming solar radiation [33]: this is the *direct effect*. A larger fraction of this forcing is due to their modulation of radiative properties of clouds: this is the *indirect effect*. By acting as cloud condensation nuclei (CCN), additional aerosols can drive up the cloud droplet number while driving down the mean cloud droplet size. The resulting clouds are brighter, larger and last longer [2, 52]. They hence reflect more solar radiation and cool the Earth.

Unlike the direct effect which can, in principle, be well constrained [54], the magnitude of the forcing induced by the indirect effect is difficult to estimate. There are two reasons for this: (i) the physical processes underpinning aerosol-cloud interactions (ACIs) are not yet fully understood, which hinders the estimation of present day forcing; (ii) the present day forcing must be compared to the forcing at pre-industrial state, which is also particularly challenging to quantify [7]. In fact, observational and model-based studies of ACIs still disagree on the magnitude of this forcing. As a result, ACIs contribute the largest uncertainty in present day global warming [31].

To better estimate present day forcing we require accurate, global measurements of CCN concentrations to assess radiative properties of clouds [2, 52]. Unfortunately, measuring CCN concentrations can only be achieved in-situ, and while field campaigns have already undertaken to collect detailed CCN observations, these measurements are spatio-temporally sparse and provide insufficient constraint on global distribution of aerosols [3, 43].

For lack of better observations, the Aerosol Optical Depth (AOD) has been widely adopted as a first order proxy of CCN concentration in an atmospheric column [3, 11, 37]. The AOD is a measure of the extinction of solar radiation through an atmospheric column. It is denoted by τ and defined at a given wavelength, time, latitude and longitude by

$$\tau = \int_0^H b_{\text{ext}}(h) dh, \quad (1)$$

where b_{ext} is the extinction coefficient¹ and the integral is taken over the height H of an atmospheric column. The AOD is appealing because it is routinely observed on a global scale by satellite products [40] which, unlike in situ observations, offer long term global records.

However, the AOD is a column-integrated quantity and does not provide information on the vertical distribution of aerosols. This is limiting as their vertical distribution strongly influences both the magnitude and even the sign of the forcing induced by the indirect effect. For example, both modelling [47] and observational studies [38] find AOD inadequate for assessing ACIs over vast subtropical ocean areas, which play a key role in determining the radiation balance of the Earth. Yet, in both studies, the vertically resolved aerosol extinction coefficient b_{ext} shows significantly higher correlation with CCN concentrations. Stier [47] also highlights the importance of determining aerosol vertical distributions to provide stronger constraints on CCN at specific altitudes. In particular, the AOD fails to describe near-surface properties such as the concentration of aerosols in the boundary layer.

In this work, we propose to probe whether AOD observations can be used to constrain a global prior over aerosol vertical distributions. Formally, given an AOD observation τ , we want to reconstruct

¹the sum of contribution from particle-light scattering plus absorption of light by particles

the corresponding extinction coefficient profile b_{ext} . This amounts to the task of reconstructing three-dimensional (3D) profiles using height-integrated two-dimensional (2D) observations and quantities that are easier to obtain in 3D, such as temperature and relative humidity.

Motivated by the study of cloud vertical structures, this task has been framed in the past as fully-supervised learning [25], i.e. assuming observations of groundtruth vertical profiles were available. Collecting vertically resolved observations of aerosols optical properties is also possible, using lidar-based remote sensing instruments [55] or groundbased sun-photometers [16]. While valuable, these observations are however limited by their low spatiotemporal coverage and prone to corruption (e.g. low signal-to-noise ratio, clear-sky requirement). Compiling high quality observational data of aerosol vertical profiles at large scale is thus challenging, making fully-supervised learning approaches inadequate.

Instead, we propose to draw from spatial disaggregation methodologies that only require observations at the aggregated level. Spatial disaggregation is the task of inferring subgrid details given coarse resolution spatial observations. Postulating an underlying fine grained spatial field that aggregates into coarse observations, this problem can be framed as weakly supervised learning [61] with aggregated targets. While existing works [26, 49, 53, 58, 60] have only considered aggregation processes happening on a 2D field, this rationale can be extended to disaggregate quantities along a third dimension — height. Since τ corresponds to the vertical integration of b_{ext} , we propose to frame the reconstruction of aerosol vertical profiles as the vertical disaggregation of AOD observations.

Using Gaussian processes (GPs) [39], we design a Bayesian model that maps vertically resolved meteorological variables (e.g. pressure, relative humidity) to a probabilistic estimate of the extinction coefficient that integrates into the AOD. The model formulation is simple and makes assumptions explicit, hence granting control and interpretability over predictions while offering built-in uncertainty quantification.

In order to be able to fully validate the proposed methodology, we use ECHAM-HAM global aerosol-climate model simulation data [45, 46, 59]. While our prime motivation is to reconstruct b_{ext} from satellite observations of AOD, the intricacies of combining measurements from different instruments makes it challenging to validate any proposed methodology. On the other hand, ECHAM-HAM is a self-consistent climate model that offers readily available aerosol vertical profiles, and is better suited for model development. We demonstrate our model is able to reconstruct natural patterns that arise in aerosol vertical distribution, in particular in the boundary layer. We show that very simple and readily available meteorological predictors suffice to obtain a good estimation of the extinction coefficient.

The aims of this study are as follows:

- Formulate a probabilistic vertical disaggregation methodology for the task of reconstructing aerosol vertical profiles using readily available vertically resolved covariates.
- Validate the proposed methodology using climate model data — where access to groundtruth extinction coefficient profiles enables quantitative evaluation.

Section 2 outlines the design of the vertical disaggregation model and describes the inference procedure as well as the hyperparameter selection. Section 3 describes the dataset and experimental setup used to validate the model. Section 4 presents the experimental results and Section 5 discusses, while introducing avenues of future research.

2. Model Design

In this section, we first outline the design of the prior distribution we place on the extinction coefficient profile using GPs. We then describe the observation model that connects our prior to the observations of the AOD. Finally, we present how the posterior inference over the extinction profiles is conducted.

2.1. Design of the Prior

2.1.1. An idealised vertical prior

In passive satellite sensors, AOD retrieval algorithms are used to produce an estimate of the AOD from measurements of top-of-atmosphere reflectance. They are typically based on a forward model of top-of-atmosphere reflectance [57]. The parameters required to run this forward model are obtained using a look-up-table algorithm, which admits as indices aerosol properties such as: the aerosol type, the AOD and the form of the vertical profile. When the satellite sensor captures a measure of reflectance, the look-up-table can be used to solve the corresponding inverse problem and estimate the AOD, provided the aerosol properties have been predetermined. Consequently, to address the inverse problem, AOD retrieval algorithms need to assume a form for aerosol vertical profile.

The form of the vertical profile is typically idealised, using either a Gaussian profile, or in the simplest case an exponential profile [28, 57]. The exponential profile takes the form $b_{\text{ext}}(h) \propto e^{-h/L}$, where L is a fixed height scale parameter that is typically taken as the top altitude of the boundary layer (2 km) — although different algorithms may adopt varying values.

Whilst idealised, these profiles capture a key element of aerosol vertical distribution: most CCN lies at low altitude in the boundary layer. This constitutes a form of domain knowledge, which we argue is important to encode within the design of a prior over aerosol vertical profiles. Thus, we propose to use an idealised exponential profile as a structural component of the prior placed over the extinction profile.

2.1.2. Weighting the ideal profile

Clouds' local meteorology influences clouds properties, and hence the sign and magnitude of the effective radiative forcing due to ACIs [1, 9, 12, 42] — this source of heterogeneity is at the heart of the uncertainty over how clouds impact climate.

The impact of local meteorology on ACIs can be characterized by the following set of environmental confounders [21, 51]: temperature T , pressure P , relative humidity RH and updraft velocity ω . Standing as a proxy for aerosols, the AOD is also impacted by its surrounding meteorology [10, 17]. These meteorological variables should thus also modulate the extinction coefficient.

A great advantage of these simple meteorological variables is that they (or their proxies) are readily available on multiple pressure levels in reanalysis data. Reanalysis data, such as ERA5 [35], combines observations and model simulations through physical laws to provide the most accurate representation of past and present climate and meteorology. Hence, T , P , RH and ω can be reliably used as vertically resolved predictors. Finally, note that b_{ext} is also a spatiotemporal field which exhibits smoothness across spatial and temporal dimensions. Such regularity should be included in the modelling.

Let $x|h$ denote a d -dimensional vector resulting from the concatenation of spatiotemporal and meteorological variables for a given altitude h . For example, one can consider

$$x = (t, \text{lat}, \text{lon}, T, P, \text{RH}, \omega), \quad (2)$$

where lat and lon respectively denote latitude and longitude. In the remainder of this paper, we will denote $\mathcal{X} \subseteq \mathbb{R}^d$ the space in which x takes values.

We propose to model the extinction coefficient b_{ext} by weighting an idealised exponential vertical profile with a positive weight function $w : \mathcal{X} \rightarrow (0, +\infty)$. Namely, the chosen prior for the extinction coefficient profile is denoted φ and takes the simple form

$$\varphi(x|h) = w(x|h)e^{-h/L}. \quad (3)$$

This weight function is meant to capture finer details of variability in the extinction profile, putting more mass in regions where meteorological predictors suggest aerosol loading is likely to be higher.

2.1.3. Probabilistic modelling of the weighting function

Whilst the extinction coefficient should indeed be modulated by $x|h$, we expect this relationship to be non-trivial and highly non-linear [17]. For this reason, we propose to learn the weighting function w using non-linear (e.g. kernel-based) statistical machine learning methodologies. Furthermore, we want the prior to reflect our lack of knowledge about the relationship between $x|h$ and $b_{\text{ext}}(h)$. To account for this epistemic uncertainty, we propose a Bayesian formulation of the weighting function.

GPs [39] are a ubiquitous class of expressive Bayesian priors over real-valued functions. They have been widely used in various nonlinear and nonparametric regression problems in geosciences [5]. A GP is fully determined by its mean function and its covariance function. The covariance function — called kernel — is typically user-specified as a positive definite bivariate function on the input data.

We place a GP prior over the weight function. To ensure the weights are strictly positive, we further warp the GP with a positive transform $\psi : \mathbb{R} \rightarrow (0, +\infty)$. Formally, let $m : \mathcal{X} \rightarrow \mathbb{R}$ be a mean function and $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ denote a positive definite kernel, we model the weighting function as

$$w(x|h) = \psi(f(x|h)) \quad \text{where } f \sim \text{GP}(m, k). \quad (4)$$

The Bayesian prior defined by f represents a latent variable that maps through the positive link function ψ onto the weight function. A simple choice for ψ is the exponential function, which as we will see in Section 3 is a mathematically convenient choice.

Whilst $\psi \circ f$ describes an expressive probability distribution over nonlinear positive functions, it remains interpretable. The choice of kernel specifies how covariance is encoded, i.e. $\text{Cov}(f(x), f(x')) = k(x, x')$ and controls the functional class the GP belongs to. For example, the Matérn class of covariance functions offers control over the functional smoothness of the GP [39, 44].

2.2. Design of the Observation Models

In this section, we establish a connection between the extinction coefficient prior $\varphi(x|h)$ (constructed in Section 2.1) and the observed values of extinction coefficient and AOD. To simplify notations, we focus on a single air column with a height of H , where we have observations of the extinction coefficient $b_{\text{ext}}(h)$ and the AOD τ .

Ideally, we would like to achieve a perfect match between $b_{\text{ext}}(h)$ and $\varphi(x|h)$, as well as between τ and $\int_0^H \varphi(x|h)$. Unfortunately this is unrealistic because observations are prone to measurement error, which must critically be accounted for. To address this, we formulate probabilistic observation models for b_{ext} and τ in what follows.

2.2.1. Relevance of the log-normal distribution

The log-normal distribution has been reported to provide a good fit to the AOD in studies focusing on locations in North America and Europe [36]. This distribution is particularly suitable because the AOD is strictly positive and exhibits a right-skewed distribution concentrated around small values (≈ 0.14).

The log-normal distribution is a right-skewed continuous probability distribution with support $(0, +\infty)$. It is specified by two parameters: a location parameter $\mu \in \mathbb{R}$, and a scale parameter $\sigma > 0$. In particular, if $Z \sim \mathcal{N}(0, 1)$ is a standard normal random variable, then $e^{\mu + \sigma Z}$ follows a log-normal distribution with location μ and scale σ , which we denote $\mathcal{LN}(\mu, \sigma)$.

It is often preferable to express the log-normal distribution as a function of its mean rather than its location, as it is more interpretable and offers a convenient parametrisation for including higher-order model parameters. Formally, let $\eta = \mathbb{E}[\mathcal{LN}(\mu, \sigma)]$ denote the expectation of the distribution and suppose the scale σ is known. The η can be expressed using μ and σ as $\eta = e^{\mu + \sigma^2/2} \Leftrightarrow \mu = \log \eta - \frac{\sigma^2}{2}$,

which yields the mean-reparametrised formulation of the log-normal distribution

$$\mathcal{LN}(\mu, \sigma) = \mathcal{LN}\left(\log \eta - \frac{\sigma^2}{2}, \sigma\right). \quad (5)$$

We collect AERONET sun-photometers AOD measurements from 1315 stations between 1993 and 2021 [16], and show in Figure 1 that the empirical marginal distribution of AERONET observations can indeed be appropriately fitted using a log-normal density. Whilst Figure 1 only depicts a marginal distribution for the AOD, we draw motivation from this to formulate a log-normal observation model for b_{ext} and τ conditional on $\varphi(x|h)$.

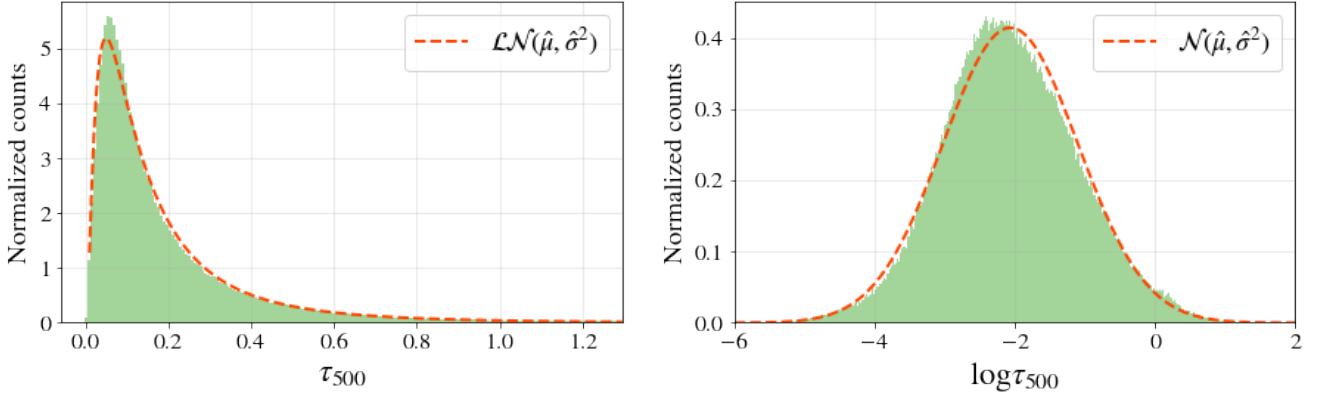


Figure 1. Left: Log-normal density (red) fitted with maximum likelihood estimates to the empirical distribution of AERONET AOD at 500nm (τ_{500}) from 1315 stations between 1993 and 2021 (green). **Right:** logspace plot of the left panel. It demonstrates a sound fit for the normal distribution in the logspace, albeit with a slight right-skew; $\hat{\mu} = -2.06$, $\hat{\sigma} = 0.96$.

2.2.2. An observation model for b_{ext}

Whilst we do not usually observe $b_{\text{ext}}(h)$ it is still useful to specify an observation model for the extinction coefficient. Indeed, at inference stage, $\varphi(x|h)$ might fail to capture observational noise which will inevitably hinder variance calibration for the predicted extinction coefficient profile distribution. Using an observation model over b_{ext} introduces an additional degree of freedom that will help calibrate the variance of observations.

We propose a simple log-normal observation model with mean $\varphi(x|h)$ for the extinction coefficient,

$$b_{\text{ext}}(h) \mid \varphi(x|h) \sim \mathcal{LN}\left(\log \varphi(x|h) - \frac{\sigma_{\text{ext}}^2}{2}, \sigma_{\text{ext}}\right), \quad (6)$$

where $\sigma_{\text{ext}} > 0$ is a scale parameter. This observation model conserves the mean² from our prior $\varphi(x|h)$ and simply includes observational noise through the scale parameter σ_{ext} . When extinction coefficient observations are available, σ_{ext} can be calibrated following the procedure described in Section 3.3.3.

We emphasise that tuning σ_{ext} requires only a small number of extinction coefficient observations. This is in contrast to the AOD observations, which constitute the principal source of data we observe in the context of this work.

² $\mathbb{E}[b_{\text{ext}}(h)] = \mathbb{E}[\mathbb{E}[b_{\text{ext}}(h) \mid \varphi(x|h)]] = \mathbb{E}\left[\exp\left(\log \varphi(x|h) - \frac{\sigma_{\text{ext}}^2}{2} + \frac{\sigma_{\text{ext}}^2}{2}\right)\right] = \mathbb{E}[\varphi(x|h)]$

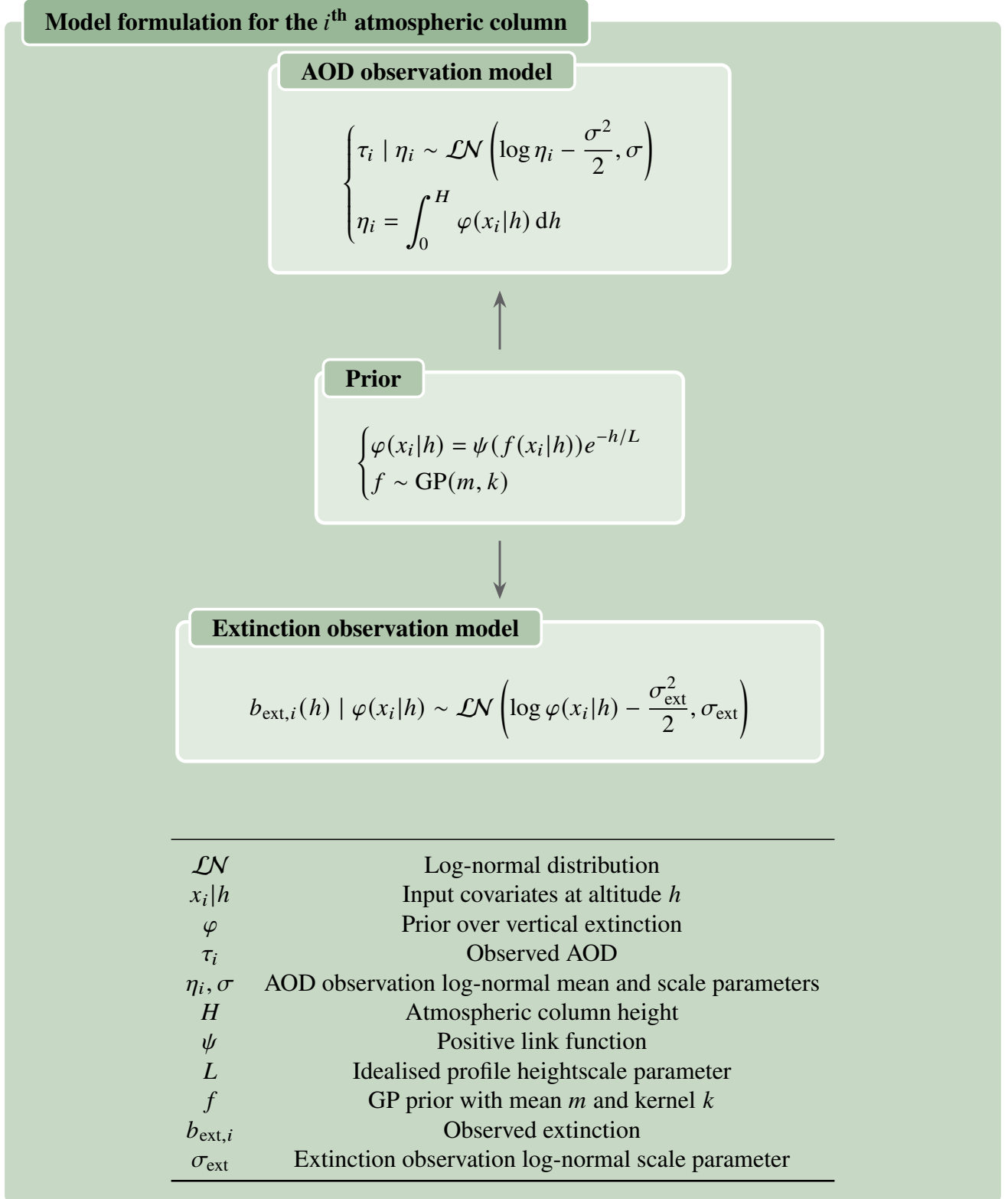


Figure 2. Observation models and prior formulation for the i^{th} atmospheric column. The model follows a hierarchical Bayesian structure represented by the arrows. We first specify the prior, and then express the observation models conditionally on the prior.

2.2.3. An observation model for τ

AOD observations constitute the primary source of information on aerosols optical properties in the context of this work since it is collected routinely at global scale by satellite products. Satellite retrievals are however prone to distortion, that may arise from observational noise [40] or assumptions made in the AOD retrieval algorithm [34, 56]. Therefore, as for the extinction coefficient, we propose to model these distortions using a log-normal observation model for the AOD conditional on $\varphi(x|h)$.

Formally, let $\eta = \int_0^H \varphi(x|h) dh$ be the column integration of our prior, then we propose a mean-parametrised observation model for the AOD conditional on η , given by

$$\tau | \eta \sim \mathcal{LN}\left(\log \eta - \frac{\sigma^2}{2}, \sigma\right) \quad (7)$$

$$\eta = \int_0^H \varphi(x|h) dh, \quad (8)$$

with a scale parameter $\sigma > 0$ which can be tuned against AOD observations jointly with other model parameters following the procedure described in Section 2.3.

When working with multiple AOD observations $\tau_1, \dots, \tau_n$, the scale parameter σ will be assumed shared among atmospheric columns. The mean parameter η (or location parameter μ for the canonical parametrisation) will however be column-specific.

2.3. Tuning and Posterior Inference

2.3.1. Finite-sample problem formulation

We now make things concrete and assume we observe the AOD for n columns, which we stack into the vector $\boldsymbol{\tau} = [\tau_1 \dots \tau_n]^\top \in \mathbb{R}^n$. For the i^{th} column, we also observe m_i vertically resolved meteorological covariates $x_i^{(1)}, \dots, x_i^{(m_i)}$ and their respective altitudes $h_i^{(1)} < \dots < h_i^{(m_i)}$, such that $x_i^{(j)} \sim p(x|h_i^{(j)})$. We concatenate these observations into a dataset $\mathcal{D} = \left\{ \left(x_i^{(j)}, h_i^{(j)} \right)_{j=1}^{m_i}, \tau_i \right\}_{i=1}^n$ and denote $M = \sum_{i=1}^n m_i$ the total number of vertically resolved samples. The model description for the i^{th} column is summarized in Figure 2.

Our objective is to use $\mathcal{D}$ to learn the mapping $\varphi(x|h)$. Within the Bayesian framework, this objective is two-fold:

- (A) Update the prior placed on φ with the observations from $\mathcal{D}$. This corresponds to computing the posterior distribution of φ given $\boldsymbol{\tau}$.
- (B) Tune the model hyperparameters by maximising the marginal log-likelihood $\log p(\boldsymbol{\tau})$. The hyperparameters include the log-normal scale parameter σ and any parameter from the GP mean m and kernel k .

Regarding point (A), since φ results from a transformation of GP f , we will rather focus on the posterior distribution of the GP directly for convenience. If $f \in \mathbb{R}$ denotes a realization of f at any input $x|h$, the posterior distribution of $f(x|h)$ given observations is denoted $p(f|\boldsymbol{\tau})$. Having access to the posterior $p(f|\boldsymbol{\tau})$ allows computation of the predictive mean and variance of $\varphi(x|h)$ following

$$\mathbb{E}[\varphi(x|h)|\boldsymbol{\tau}] = \int_{\mathbb{R}} \psi(f) e^{-h/L} p(f|\boldsymbol{\tau}) df \quad (9)$$

$$\text{Var}(\varphi(x|h)|\boldsymbol{\tau}) = \mathbb{E}[\varphi(x|h)^2|\boldsymbol{\tau}] - \mathbb{E}[\varphi(x|h)|\boldsymbol{\tau}]^2. \quad (10)$$

The above can be estimated with Monte Carlo by drawing samples from the posterior $p(f|\boldsymbol{\tau})$. In Section 3.3.1 we will obtain a simpler closed-form solution for the particular choice $\psi = \exp$.

Regarding point (B), the marginal log-likelihood $\log p(\boldsymbol{\tau})$ is unfortunately intractable. Indeed, let $\mathbf{x} = [x_1^{(1)} \dots x_n^{(m_n)}]^\top \in \mathcal{X}^M$ denote the concatenation of all input entries from the dataset and let $\mathbf{f} \in \mathbb{R}^M$ denote a realisation of $f(\mathbf{x})$. Assuming independence of AOD observations $\tau_1, \dots, \tau_n$ conditionally on the GP realisation over atmospheric columns $\mathbf{f}$, the marginal likelihood $p(\boldsymbol{\tau})$ is expressed in terms of the observation model and prior distributions following

$$p(\boldsymbol{\tau}) = \int_{\mathbb{R}^M} p(\boldsymbol{\tau}|\mathbf{f})p(\mathbf{f}) d\mathbf{f} \quad \text{with} \quad p(\boldsymbol{\tau}|\mathbf{f}) = \prod_{i=1}^n p(\tau_i|\mathbf{f}_i), \quad (11)$$

where $\mathbf{f}_i = [\mathbf{f}_i^{(1)} \dots \mathbf{f}_i^{(m_i)}]^\top \in \mathbb{R}^{m_i}$ corresponds to the GP realization over the i^{th} column only. Because the observation model is log-normal, this integral is however intractable. To circumvent this intractability, we propose to tackle objective (B) by maximising a proxy of the marginal log-likelihood presented in Section 2.3.3.

2.3.2. A sparse variational approximation of the posterior

We start by addressing objective (A). The predictive posterior distribution of interest is given by

$$p(\mathbf{f}|\boldsymbol{\tau}) = \frac{p(\boldsymbol{\tau}|\mathbf{f})p(\mathbf{f})}{\int_{\mathbb{R}} p(\boldsymbol{\tau}|\mathbf{f})p(\mathbf{f}) d\mathbf{f}}. \quad (12)$$

The integral denominator is not available in closed-form, making the posterior intractable. We propose to use a variational approximation scheme [24, 32, 50] to substitute this intractable inference problem with a tractable optimization problem. In addition, we make the variational approximation sparse [50] such that the model can scale to large amounts of data.

Let $\mathbf{z} = [z_1 \dots z_p]^\top \in \mathcal{X}^p$ be a set of $p \ll M$ inducing locations over the space of inputs. Their evaluation by the GP follows a multivariate normal distribution $f(\mathbf{z}) \sim \mathcal{N}(0, \mathbf{K}_{\mathbf{zz}})$, where $\mathbf{K}_{\mathbf{zz}} = k(\mathbf{z}, \mathbf{z})$. We denote $\mathbf{u} = f(\mathbf{z}) \in \mathbb{R}^p$ and refer to this vector as *inducing variables*.

A p -dimensional parametric distribution is set over these inducing variables. We choose this distribution as a multivariate normal defined by $q(\mathbf{u}) := \mathcal{N}(\mathbf{u}|\boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}})$. $\boldsymbol{\mu}_{\mathbf{z}} \in \mathbb{R}^p$, $\boldsymbol{\Sigma}_{\mathbf{z}} \in \mathbb{R}^{p \times p}$ are called the *variational parameters* and need to be tuned such that $q(\mathbf{u})$ best approximates the true posterior $p(\mathbf{u}|\boldsymbol{\tau})$.

Once this is achieved, we take as an approximation to $p(\mathbf{f}|\boldsymbol{\tau})$ the variational posterior defined by $q(\mathbf{f}) := \int_{\mathbb{R}^p} p(\mathbf{f}|\mathbf{u})p(\mathbf{u}) d\mathbf{u}$, which is given in closed-form by

$$q(\mathbf{f}) = \mathcal{N}(\mathbf{f} | \bar{\boldsymbol{\mu}}_{x|h}, \bar{\boldsymbol{\Sigma}}_{x|h}) \quad (13)$$

$$\bar{\boldsymbol{\mu}}_{x|h} = k(x|h, \mathbf{w})\mathbf{K}_{\mathbf{zz}}^{-1}\boldsymbol{\mu}_{\mathbf{z}} \quad (14)$$

$$\bar{\boldsymbol{\Sigma}}_{x|h} = k(x|h, x|h) - k(x|h, \mathbf{z}) \left(\mathbf{K}_{\mathbf{zz}}^{-1} - \mathbf{K}_{\mathbf{zz}}^{-1}\boldsymbol{\Sigma}_{\mathbf{z}}\mathbf{K}_{\mathbf{zz}}^{-1} \right) k(\mathbf{z}, x|h). \quad (15)$$

Naturally, (13) can be extended to describe a variational posterior over multiple GP entries. Namely, let $\mathbf{x}^* \in \mathcal{X}^D$ denote a vector of input entries. If $\mathbf{f}^* \in \mathbb{R}^D$ denotes a realization of $f(\mathbf{x}^*)$, then the associated variational posterior is given by

$$q(\mathbf{f}^*) = \mathcal{N}(\mathbf{f}^* | \bar{\boldsymbol{\mu}}_{\mathbf{x}^*}, \bar{\boldsymbol{\Sigma}}_{\mathbf{x}^*}) \quad (16)$$

$$\bar{\boldsymbol{\mu}}_{\mathbf{x}^*} = k(\mathbf{x}^*, \mathbf{w})\mathbf{K}_{\mathbf{zz}}^{-1}\boldsymbol{\mu}_{\mathbf{z}} \in \mathbb{R}^D \quad (17)$$

$$\bar{\boldsymbol{\Sigma}}_{\mathbf{x}^*} = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{z}) \left(\mathbf{K}_{\mathbf{zz}}^{-1} - \mathbf{K}_{\mathbf{zz}}^{-1}\boldsymbol{\Sigma}_{\mathbf{z}}\mathbf{K}_{\mathbf{zz}}^{-1} \right) k(\mathbf{z}, \mathbf{x}^*) \in \mathbb{R}^{D \times D}. \quad (18)$$

The sparse nature of this approach becomes apparent in (17) and (18). Indeed, regardless of the number of samples we wish to evaluate the variational posterior over, we only need to invert a $p \times p$ matrix, incurring a $O(p^3)$ computational cost.

2.3.3. Learning the variational parameters μ_z, Σ_z

As mentioned above, the variational parameters μ_z, Σ_z need to be tuned such that $q(\mathbf{u})$ best approximates the posterior $p(\mathbf{u}|\tau)$, which is intractable. This problem is casted as the maximisation of an objective called the evidence lower-bound (ELBO), given by

$$\text{ELBO}(q) = \mathbb{E}_{q(\mathbf{f})} [\log p(\tau|\mathbf{f})] - \text{KL}(q(\mathbf{u})||p(\mathbf{u})). \quad (19)$$

The ELBO is a lower-bound to the marginal log-likelihood $\log p(\tau) \geq \text{ELBO}(q)$. It can thus be used as a proxy of $\log p(\tau)$ to also tune the model hyperparameters and fulfill objective (B).

The second term in (19) is the Kullback-Leibler divergence between two multivariate normal distributions. It admits a closed-form expression and can be computed. The first term, on the other hand, is an expected log-likelihood under the variational posterior which cannot be analytically computed. It can be decomposed into column-wise terms

$$\mathbb{E}_{q(\mathbf{f})} [\log p(\tau|\mathbf{f})] = \sum_{i=1}^n \mathbb{E}_{q(\mathbf{f}_i)} \left[\log \mathcal{LN} \left(\tau_i \left| \log \eta_i - \frac{\sigma^2}{2}, \sigma \right. \right) \right]. \quad (20)$$

To estimate (20), we must first evaluate the mean parameter $\eta_i = \int_0^H \psi(f(x_i|h)) e^{-h/L} dh$ with the finite number of GP evaluations $\mathbf{f}_i$ we have access to. We propose to use the trapezoidal integration scheme given by

$$\hat{\eta}_i = \sum_{j=1}^{m_i-1} \frac{\psi(\mathbf{f}_i^{(j+1)}) e^{-h_i^{(j+1)}/L} - \psi(\mathbf{f}_i^{(j)}) e^{-h_i^{(j)}/L}}{2} (h_i^{(j+1)} - h_i^{(j)}). \quad (21)$$

While we choose the trapezoidal rule for simplicity, we note that alternative finite-sample integration schemes can be chosen here in accordance with the needs.

Second, because of the log-normal observation model, the expected log-likelihood remains intractable. To estimate it, while allowing backpropagation through the variational parameters, we use a reparametrization trick [20]. Namely, we sample $\epsilon_i \sim \mathcal{N}(0, \mathbf{I}_{m_i})$ and compute $\mathbf{f}_i = \bar{\mu}_i + \bar{\Sigma}_i^{1/2} \epsilon_i$, where $\bar{\mu}_i, \bar{\Sigma}_i$ are the variational posterior parameters for the i^{th} column and are obtained by application of (17) and (18) over the predictors of the i^{th} column. The resulting GP sample $\mathbf{f}_i$ is then used to estimate the mean parameter $\hat{\eta}_i$ following (21) and we can approximate the expected log-likelihood with its one-sample estimate

$$\mathbb{E}_{q(\mathbf{f}_i)} \left[\log \mathcal{LN} \left(\tau_i \left| \log \eta_i - \frac{\sigma^2}{2}, \sigma \right. \right) \right] \approx \log \mathcal{LN} \left(\tau_i \left| \log \hat{\eta}_i - \frac{\sigma^2}{2}, \sigma \right. \right). \quad (22)$$

This method allows estimation of the ELBO objective, which in turn can be maximised with respect to the variational parameters μ_z, Σ_z using a stochastic gradient approach for example. As mentioned above, the model hyperparameters can also be tuned jointly with this objective, hence fulfilling objective (B). These include the log-normal scale σ or the kernel k hyperparameters (e.g. variances and lengthscales), with an option to parametrise these kernels using feature maps given by deep neural networks [23]. As it is standard in sparse variational GPs [50], we also learn the inducing locations $\mathbf{z}$.

3. Experimental Setup

Our motivating application is to reconstruct aerosol vertical profiles from remote-sensing AOD observations. However, to validate the proposed methodology, we also need to observe extinction coefficient profiles. While the latter can be collected from vertically resolved remote sensing instruments [16, 55], the collocation of measurements from different devices raises non-trivial questions regarding the self-consistency of the collocated dataset — and thus the validation procedure. In contrast, climate models are self-consistent and offer readily available simulations of gridded AOD and extinction coefficient profiles. We hence propose to evaluate our model using observations from a global climate model, the ECHAM-HAM global aerosol-climate model simulation [45, 46, 59].

The objective of this experiment is to apply our model to the vertical disaggregation of the AOD simulated with ECHAM-HAM, and compare our prediction to the extinction coefficient simulated with ECHAM-HAM. In this section, we first describe the ECHAM-HAM dataset used in the experiments. We then outline the model setup and the inference procedure. Finally, we present the evaluation metrics used to quantify the quality of the reconstructed vertical profiles.

3.1. Dataset

3.1.1. The ECHAM-HAM dataset

The aerosol-climate model ECHAM-HAM is a self-consistent global climate model of aerosol radiative properties and CCN which demonstrates excellent agreement with AOD measurements from ground-based sun-photometers and satellite retrievals [45]. It computes the evolution of log-normal aerosol mass and number modes — for species sulfates, black carbon, organic carbon, sea salt and mineral dust — by taking into account physical and chemical particle processes. The simulation used includes aerosol optical properties, aerosol tracers and meteorological variables at a $1.8^\circ \times 1.8^\circ$ horizontal resolution on a Gaussian grid and over 31 levels of vertical resolution and for 8 regularly spaced time steps over a day (06/06/2008). The resulting dataset counts 147 456 atmospheric columns used as training points. Table 1 provides its detailed dimensions.

The dataset captures two important aspects of aerosol properties variability: (i) because the dataset is global, it captures the spatial variability of aerosols, which is arguably the greatest contribution to the overall variability of aerosol properties. Indeed, the distribution of particles in the troposphere is highly heterogeneous due to the combination of spatially diverse sources and the relatively short lifetime of particles and gases [6]; (ii) the dataset covers a full day, which provides a good coverage of the diurnal cycle of aerosols.

However, a single day of data may fail to account for the seasonal variations of aerosol properties. The seasonal variability of aerosols is strongly regionally dependent, where variations in aerosol concentration can exceed a factor of 10 in regions prone to episodic emissions (e.g. wildfires, human activity), but vary only by 10% over a year in regions with persistent atmospheric circulation such as the Pacific subtropics [6]. Therefore, we expect that the demonstrated predictive performance of our model over this dataset should carry over to more stable regions for different seasons, but it remains uncertain from experiments on this dataset whether the model may generalise to the regional temporal variability induced by episodic events.

3.1.2. Choice of variables

ECHAM-HAM offers simulations of aerosol properties at particular wavelengths that matches the wavelengths at which operate specific instruments for evaluation. We use the ECHAM-HAM AOD at 550nm as the response variable to vertically disaggregate and the ECHAM-HAM extinction coefficient at 533nm as the groundtruth variable to evaluate our predictions against. The AOD and the extinction coefficients are simulated in ECHAM-HAM at slightly different wavelengths in order to match

Table 1. Gridded variables from ECHAM-HAM simulation data. The grid includes 8 time steps (t), 96 latitude levels (lat), 192 longitude levels (lon) and 31 vertical pressure levels (lev) — which is a proxy for h . Our objective is to vertically disaggregate the response τ using the vertically resolved predictors (T, P, RH, ω) and spatiotemporal columns locations (t, lat, lon).

	Name	Notation	Dimensions
<i>Predictors</i>	Temperature	T	(t, lat, lon, lev)
	Pressure	P	(t, lat, lon, lev)
	Relative humidity	RH	(t, lat, lon, lev)
	Vertical velocity	ω	(t, lat, lon, lev)
<i>Response</i>	AOD 550nm	τ	(t, lat, lon)
<i>Groundtruth</i>	Extinction coefficient 533nm	b_{ext}	(t, lat, lon, lev)

the wavelength at which operate the respective sensors (550nm for the MODIS AOD product [40] and 532nm for the CALIOP extinction [55]).

In order to make the model easily applicable to any satellite AOD retrieval, we select vertically-resolved predictors among standard meteorological variables, that could in practice be easily obtained from reanalysis data: T , P , RH and ω . By limiting ourselves to these variables, which are comprehensively covered by reanalysis products, we favour a global applicability of the model to AOD observations from most times and locations.

In what follows, the input variable writes $x = (t, lat, lon, T, P, RH, \omega)$. We standardize meteorological predictors and transform them using a rank-based inverse normal mapping to achieve approximately normal distributions.

3.2. Model Setup and Tuning

3.2.1. Choice of idealised profile heightscale L

A simple choice for the idealised profile heightscale is the boundary layer upper altitude, typically taken at 2 km. This is the value assumed by MISR, VIIRS and MODIS C6_DT AOD retrieval algorithms [18, 22, 27]. In our experiment, we also choose to set $L = 2$ km.

3.2.2. Choice of mean and covariance structure

We set the GP with a zero mean function $m = 0$ and kernel taken as the sum of a spatiotemporal and a meteorological contribution to the covariance

$$k(x|h, x'|h') = \gamma_1 k_{\text{ST}}(\{t, lat, lon\}, \{t', lat', lon'\}) + \gamma_2 k_{\text{meteo}}(\{T, P, \omega, RH\}|h, \{T', P', \omega', RH'\}|h'), \quad (23)$$

with respective variances γ_1 and γ_2 . We propose to parametrise the kernels using the Matérn family, which has been widely used to work with spatiotemporal data [44]. We denote C_ν the Matérn- ν covariance function with automatic relevance determination (i.e. independent lengthscales for each entry). This choice of Matérn order ν guarantees a desired level of regularity for the resulting GP. Details on the Matérn covariance functions are provided in Appendix A.

The spatiotemporal kernel is taken as the product kernel

$$k_{\text{ST}}(\{t, lat, lon\}, \{t', lat', lon'\}) = C_{3/2}(|t - t'|) C_{3/2} \left(\text{Hav} \left(\begin{bmatrix} lat \\ lon \end{bmatrix}, \begin{bmatrix} lat' \\ lon' \end{bmatrix} \right) \right), \quad (24)$$

where Hav denotes the Haversine distance — or great-circle distance — that determines the distance between two points on a sphere given their latitudes and longitudes. Each Matérn covariance admits a lengthscale, respectively denoted ℓ_t and $\ell_{lat,lon}$. Note that $C_{3/2}$ may not be positive definite over Hav for

certain lengthscales [4, 13], thus we set an upper bound on $\ell_{\text{lat},\text{lon}}$ to ensure positive-definiteness. The Matérn order $\nu = 3/2$ ensures the GP is continuously differentiable with respect to time and space.

The spatiotemporal kernel (24) makes the predicted extinction smooth across time and space. Its product structure ensures that for distant times (t, t') or distance locations $((\text{lat}, \text{lon}), (\text{lat}', \text{lon}'))$, the spatiotemporal covariance vanishes to zero. Hence, the GP covariance of predictions distant in space or time will disregard spatiotemporal predictors. This prevents overfitting over spatiotemporal predictors.

The meteorological kernel is taken as an automatic relevance determination covariance

$$k_{\text{meteo}}(\{T, P, \omega, \text{RH}\}|h, \{T', P', \omega', \text{RH}'\}|h') = C_{1/2} \left(\begin{bmatrix} |T - T'| \\ |P - P'| \\ |\text{RH} - \text{RH}'| \\ |\omega - \omega'| \end{bmatrix} \right)^3, \quad (25)$$

where each dimension admits its own lengthscale parameter, respectively denoted $\ell_T, \ell_P, \ell_{\text{RH}}$ and ℓ_ω . The Matérn order $\nu = 1/2$ ensures that the GP is continuous with respect to meteorological predictors.

The kernel on meteorological predictors (25) allows to GP to learn how meteorological predictors should modulate extinction. It introduces covariance between predicted extinctions if meteorological predictors are close, even if the spatiotemporal covariance is zero (because of the additive structure in (23)). In practice, we find that setting a unit variance for this kernel $\gamma_2 = 1$ helps prevent vanishing kernel signal when tuning the hyperparameters.

We include in Appendix B a comparison of this choice of covariance structure with a more conventional tensor form covariance and evaluates their respective predictive performance.

3.2.3. Sparse variational GP setup and tuning

We use $p = 200$ inducing locations $\mathbf{z}$ as an arbitrary choice (fewer or more inducing locations can be used following needs). They are initialized by randomly drawing samples at boundary layer altitude from $\mathbf{x}$. The variational distribution $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}})$ is initialized with $\boldsymbol{\mu}_{\mathbf{z}} = 0$ and $\boldsymbol{\Sigma}_{\mathbf{z}} = \mathbf{I}_p$. We denote $\Theta = \{\boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}}, \mathbf{z}, \sigma, \gamma_1, \gamma_2, \ell_t, \ell_{\text{lat},\text{lon}}, \ell_T, \ell_P, \ell_{\text{RH}}, \ell_\omega\}$ the joint set of variational and model hyperparameters. We maximise the ELBO objective (19) with respect to Θ using Adam optimiser [19].

Since the dataset includes $8 \times 96 \times 192 = 147\,456$ atmospheric columns, evaluating the ELBO for all columns at each optimisation step is computationally prohibitive. Instead, we use a stochastic gradient approach and sample random mini-batches of columns over which we compute an unbiased noisy estimate of the ELBO gradient. Whilst compromising statistical efficiency, the high variance estimates of gradients have regularising properties that mitigates overfitting, and this approach is guaranteed to converge to a local optima for Θ [15]. To meet memory limitations, we set a mini-batch size of 64, and therefore choose a small learning rates of 0.01 to avoid taking large steps based on few samples. When possible, a larger mini-batch size couple with a larger learning rate should allow to reach a better local optimum faster [15].

3.3. Inference Procedure

3.3.1. Exponential link

We choose an exponential link function $\psi = \exp$ to warp the GP. This particular choice is mathematically convenient as it makes the prior $\varphi(x|h)$ a log-normal random variable given by

$$\varphi(x|h) = e^{f(x|h) - h/L} \sim \mathcal{LN}\left(m(x|h) - h/L, \sqrt{k(x|h, x|h)}\right). \quad (26)$$

³we permit ourselves to drop the conditioning on h and h' for conciseness

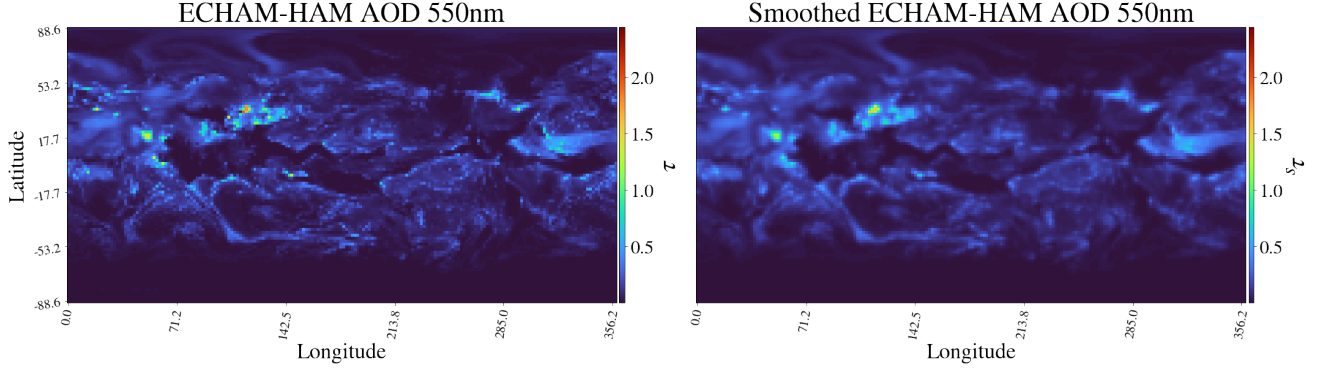


Figure 3. *Left: ECHAM-HAM 550nm AOD. Right: Spatially smoothed ECHAM-HAM 550nm AOD.*

A direct consequence is that we know the analytical expressions of its moments and quantiles, which alleviates the need for estimation procedures. This becomes of particular interest when making prediction, since the variational posterior that approximates $\varphi(x|h)|\tau$ is also a log-normal distribution given by $\mathcal{LN}(\bar{\mu}_{x|h} - h/L, \bar{\Sigma}_{x|h}^{1/2})$ following notations from Section 2.3.2.

Therefore, we obtain approximations of the posterior mean, variance and quantiles in closed-form following

$$\mathbb{E}[\varphi(x|h)|\tau] \approx \exp\left(\bar{\mu}_{x|h} - \frac{h}{L} + \frac{1}{2}\bar{\Sigma}_{x|h}\right) \quad (27)$$

$$\text{Var}(\varphi(x|h)|\tau) \approx \left(e^{\bar{\Sigma}_{x|h}} - 1\right) \exp\left(2\bar{\mu}_{x|h} - \frac{2h}{L} + \bar{\Sigma}_{x|h}\right) \quad (28)$$

$$Q_{\alpha}(\varphi(x|h)|\tau) \approx \exp\left(\bar{\mu}_{x|h} - \frac{h}{L} + \sqrt{2\bar{\Sigma}_{x|h}} \text{erf}^{-1}(2\alpha - 1)\right), \quad (29)$$

where Q_{α} denotes the α -quantile function and erf denotes the error function. This allows for an efficient computation of the predicted mean extinction profile and confidence regions.

3.3.2. Rescaling the predicted profiles

The posterior distribution $p(\mathbf{f}|\tau)$ updates the GP with the information that the mapping φ should (in expectation) vertically integrate to the observed AOD τ . While instructive, this is a weak constraint on the prior and does not guarantee that the prediction will effectively integrate to the AOD of the atmospheric column.

To get a stronger enforcement of column-integrated values and ensure predictions effectively integrate to the AOD in expectation, we propose to rescale the model against observed AOD. Because observations $\tau_1, \dots, \tau_n$ are corrupted with high-frequency observational noise, we introduce a spatially smoothed version of the AOD observations that filters out this noise. By applying a spatial Gaussian smoothing filter to the ECHAM-HAM AOD, we obtain spatially smoothed versions of the observations which we denote ${}^s\tau_1, \dots, {}^s\tau_n$. A comparison of the fields is displayed in Figure 3.

At inference time, we then substitute the predicted posterior of extinction with its rescaled version given by

$${}^s\varphi(x_i|h)|\tau = \frac{{}^s\tau_i}{\int_0^H \mathbb{E}[\varphi(x_i|h)|\tau] dh} \varphi(x_i|h)|\tau \quad (30)$$

for the i^{th} column. This ensures that, in expectation, ${}^s\varphi(x_i|h)|\tau$ effectively integrates along height to ${}^s\tau$. When choosing $\psi = \exp$, performing this rescale is equivalent to shifting the location of the posterior in (26) by $\log {}^s\tau_i - \log \int_0^H \mathbb{E}[\varphi(x_i|h)|\tau] dh$. The integral is approximated with a trapezoidal scheme.

3.3.3. From ${}^s\varphi(x|h)|\tau$ to $b_{\text{ext}}(h)|\tau$

The final stage of inference is to estimate the posterior distribution of b_{ext} following the observation model (6). Because the chosen observation model conserves the mean, the posterior mean of $b_{\text{ext}}(h)$ is the same as the posterior mean of ${}^s\varphi(x|h)$. Higher-order moments however cannot be obtained analytically and will require estimation by sampling from $b_{\text{ext}}(h)|\tau$. This can be achieved by sampling from ${}^s\varphi(x|h)|\tau$ and then plugging the samples in the observation model (6) to sample from $b_{\text{ext}}(h)|{}^s\varphi(x|h)$.

The observation model scale parameter σ_{ext} needs to be calibrated against observational noise. We set σ_{ext} to the value that minimises the integrated calibration index (ICI) given by

$$\text{ICI} = \int_0^1 |N_{\sigma_{\text{ext}}}(\alpha) - (1 - \alpha)| d\alpha, \quad (31)$$

where $N_{\sigma_{\text{ext}}}(\alpha)$ is the percentage of ECHAM-HAM extinction coefficient observations that fall within the $(1 - \alpha)$ credible interval of the distribution of $b_{\text{ext}}|\tau$. The ICI is evaluated against $n_{\text{calib}} = 200$ random atmospheric columns from the dataset. These columns form a separate calibration set which is solely used to calibrate σ_{ext} . This separate calibration set is negligible in size in comparison with the 147 456 atmospheric columns available in the simulation.

3.4. Evaluation procedure

We compare the predicted extinction coefficient profile $b_{\text{ext}}|\tau$ to the extinction coefficient from ECHAM-HAM simulations. For clarity, we refer to our proposed methodology as *VAExtGP* (for Variational Aerosol Extinction GP) in the presentation of experimental results.

3.4.1. Baseline models

We use as a comparative baseline an idealised exponential profile, similar to the vertical profiles postulated in AOD retrieval algorithms. For a fair comparison, we rescale the baseline profiles such that their integrated values match the AOD, following the procedure described in Section 3.3.2. We complement this baseline with a log-normal observation model for the extinction coefficient and calibrate its scale parameter against a separate calibration following the procedure outlined in Section 3.3.3. This enables probabilistic evaluation of the extinction coefficient profiles predicted with the idealised baseline.

In addition, to examine the contribution to predictions of our modelling choices, we consider the following ablated versions of *VAExtGP*:

- *GP only* : we remove the idealised exponential vertical contribution to the prior. The prior is specified as the exponential of a GP only, i.e. $\varphi(x|h) = \exp(f(x|h))$ with $f \sim \text{GP}(m, k)$.
- *ST only* : we only use spatiotemporal covariates as input variables for the model, i.e. $x = (t, \text{lat}, \text{lon})$. The kernel is set to $k = k_{\text{ST}}$.
- *Meteo only* : we only use meteorological covariates as input variables for the model, i.e. $x = (T, P, \text{RH}, \omega)$. The kernel is set to $k = k_{\text{meteo}}$.

3.4.2. Evaluation metrics

We use two categories of metrics to quantify the quality of the predicted vertical profiles: deterministic metrics, which compare only the posterior mean prediction to the groundtruth ECHAM-HAM extinction coefficient profiles, and probabilistic metrics which evaluate the entire posterior probability distribution against the extinction coefficients from ECHAM-HAM simulation and thus also assess the quality of the uncertainty quantification in the posteriors. The metrics used are detailed in Table 2.

Table 2. Evaluation metrics; Deterministic metrics compare the predicted posterior mean $\mathbb{E}[b_{\text{ext}}|\tau]$ to the ECHAM-HAM extinction coefficient; Probabilistic metrics evaluate the complete predicted posterior probability distribution of $b_{\text{ext}}|\tau$ against ECHAM-HAM extinction coefficient.

	Metric	Description	Best when
<i>Deterministic</i>	RMSE	Root mean square error	close to 0
	MAE	Mean absolute error	close to 0
	Corr	Pearson correlation	close to 1
	Bias	Mean bias	close to 0
	Bias98	Bias in the 98th percentile	close to 0
<i>Probabilistic</i>	ELBO	Evidence lower-bound of groundtruth b_{ext}	higher
	Calib95	95% calibration score, i.e. $N_{\sigma_{\text{ext}}}(\alpha = 0.05)$	close to 95%
	ICI	Integrated calibration index — see (31)	close to 0

Table 3. Scores of the idealised exponential baseline and VAExtGP for the task of predicting ECHAM-HAM extinction profiles ; “Entire column” means scores are computed for every altitude levels ; “Boundary layer” means scores are computed for altitude levels of the boundary level only (< 2 km) ; the best scores for each metric and region are highlighted in bold ; we report 1 standard deviation.

Region	Method	RMSE (10^{-5})	MAE (10^{-6})	Corr	Bias (10^{-6})	Bias98 (10^{-5})	ELBO ⁴	Calib95 (%)	ICI (10^{-2})
<i>Entire column</i>	<i>Idealised</i>	4.10	6.65	0.51	-2.40	-4.08	13.1	96.0	5.05
	VAExtGP	3.21 ± 0.04	5.29 ± 0.06	0.73 ± 0.01	-0.005 ± 0.161	-0.482 ± 0.212	13.0 ± 0.1	93.5 ± 0.1	6.60 ± 0.45
<i>Boundary layer</i>	<i>Idealised</i>	7.55	16.8	0.54	-12.9	-11.7	10.2	93.5	19.1
	VAExtGP	5.91 ± 0.07	15.3 ± 0.1	0.72 ± 0.06	-0.404 ± 0.71	-4.53 ± 0.44	10.7 ± 0.1	98.7 ± 0.1	6.26 ± 0.34

For each metric, we compute scores for pixels along the entire atmospheric columns and for pixels lying within the boundary layer only (< 2 km). Scores are averaged across all pixels.

4. Results

4.1. Comparison to an idealised exponential baseline

Table 3 shows that the posterior mean profile predicted with the proposed method outperforms the idealised exponential baseline across deterministic metrics. We note that evaluating over the entire column consistently yields better scores compared to evaluating only over the boundary layer. This is to be expected as the extinction coefficient outside the boundary layer tends to vanish to zero and most of the variability occurs within the boundary layer.

The improvement in RMSE and MAE against the idealised baseline is slightly more significant when computed within the boundary layer rather than over the entire column. This supports the idea that VAExtGP improves upon an idealised exponential profile at predicting extinction since the boundary layer concentrates most extinction variability. The improvement is greatest in Bias and Bias98, where the proposed model consistently improves over the idealised baseline by one or multiple orders of magnitude. In particular, the improvement in Bias98 suggests that VAExtGP better captures extreme extinction values.

When considering the evaluation over the entire column, the ELBO and ICI scores are comparable for both methods. However, within the boundary layer, VAExtGP outperforms the baseline, by a substantial margin in terms of ICI. This indicates that, within the boundary layer, the predicted posterior

⁴for *Idealised*, we report the marginal log-likelihood of ECHAM-HAM extinction under the predicted profiles instead of the ELBO. This an upper bound to the ELBO, which could possibly be lower.

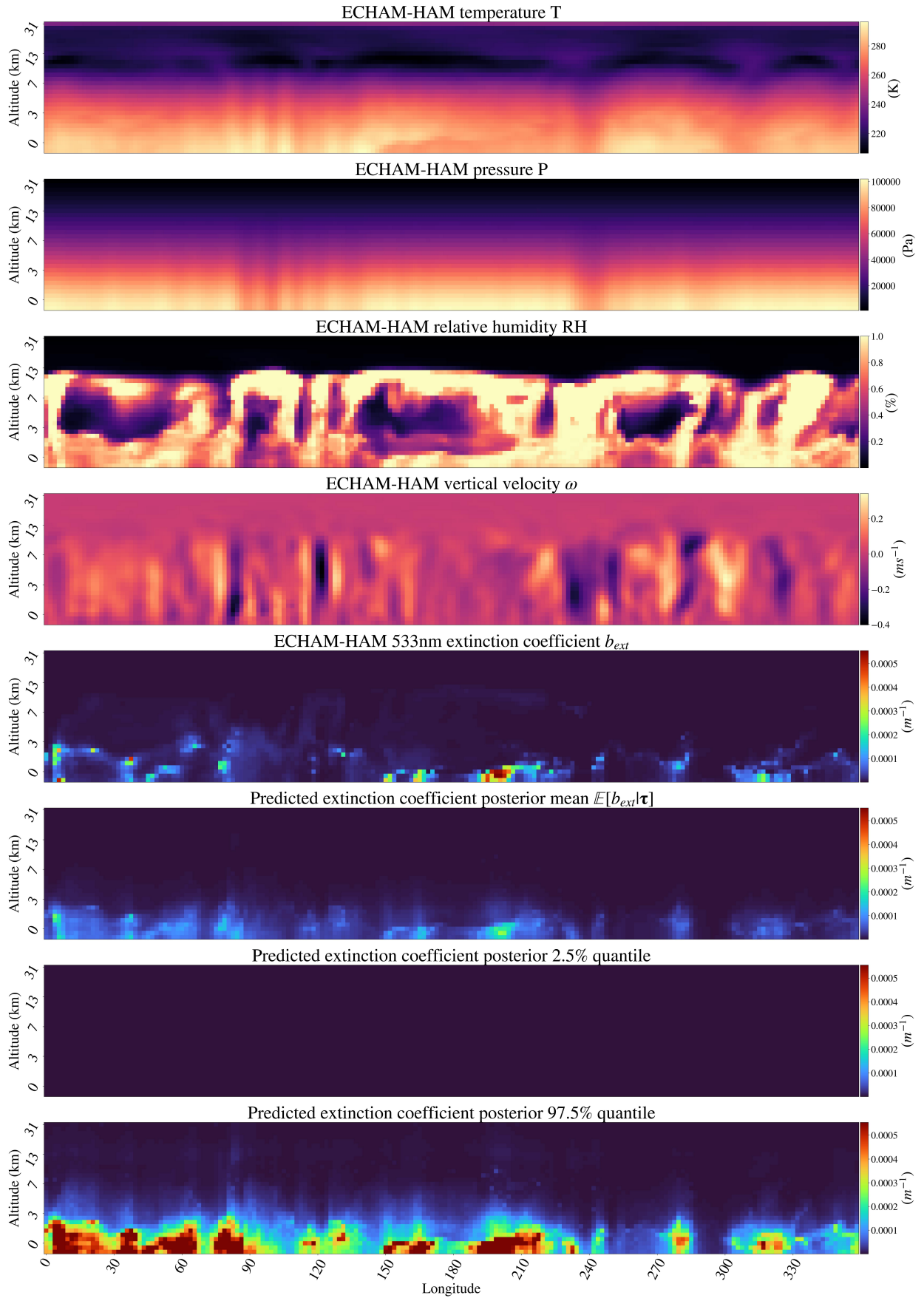


Figure 4. Vertical slices at latitude 51.29° of meteorological predictors (T, P, RH, ω), groundtruth extinction coefficient, predicted extinction coefficient posterior mean, 2.5% and 97.5% quantiles of the predicted extinction coefficient posterior distribution.

probability distribution of *VAExtGP* better captures the ECHAM-HAM extinction variability. The 95% calibration scores are similar for both methods, with a slight advantage for the idealised baseline.

4.2. Reconstructed vertical profiles

Figure 4 displays slices for a fixed latitude of the vertically resolved predictors used, the ECHAM-HAM extinction coefficient and the predicted extinction coefficient with *VAExtGP*. For comparison, the prediction at the same latitude with the idealised exponential baseline and additional predictions with *VAExtGP* for different latitudes are provided in Appendix C.

We observe that our predicted mean profile is able to reconstruct extinction patterns that are visually very similar to the groundtruth extinction coefficient. In comparison with the profiled predicted with the idealised exponential baseline in Figure 9, the extinction profiles predicted with our method look much more realistic. This is encouraging considering that the only aerosol optical property used is the AOD. We also observe that extreme extinction coefficient values appear to be well captured within the 95% confidence region of the predicted posterior distribution.

Since aerosol water uptake is related to relative humidity, we observe a good capacity to recover patterns corresponding to extinction due to aerosol swelling at low altitudes in the boundary layer. This is particularly visible over the Pacific (longitudes $144^\circ - 228^\circ$) and the Atlantic (longitudes $305^\circ - 350^\circ$), where the predicted patterns of extinction tend to best reproduce the ECHAM-HAM extinction patterns.

However, the predicted mean profile faces greater difficulties to accurately reproduce the extinction patterns for longitudes $0^\circ - 100^\circ$. This is particularly noticeable in Figure 4 around longitude 71.4° nearby Astana in Kazakhstan, and around longitude 17.3° in the Silesia region in Poland. The geographical locations suggest these are extinction patterns induced by human activity, which cannot be captured solely by meteorological variables. We conjecture these patterns are imputable to aerosols mass concentration, particles size and radiative properties. These properties, unlike extinction due to swelling, are more complex and cannot be fully characterized by relative humidity, temperature, pressure, and updraft velocity. It is worth highlighting that, whilst difficult to predict, these extinctions pattern appear nonetheless to be well captured within the 95% confidence region of the predicted posterior distribution.

In general, the posterior mean predictions tend to be smoother than the groundtruth extinction coefficient. The smoothing effect is a regularizing property of the GP prior, resulting in an overestimation of the extinction coefficient in certain regions with low extinction. For example, around longitude 200° in the Pacific, the ECHAM-HAM extinction in Figure 4 displays an extinction pocket with a sharp limit around altitude 1 km , above which extinction is virtually absent. In contrast, our predicted posterior mean is more diffuse and tends spread out lightly above 1 km .

The density plots in Figure 5 support this observation, as a large mass of extinction coefficient around 10^{-7} m^{-1} tends to be overestimated by an order of magnitude. However, this trend is significantly reduced when focusing on the boundary layer, where most of the mass lies around the axis $y = x$. This suggests that most of the overestimation occurs above the boundary layer for low extinction ($< 10^{-6}\text{ m}^{-1}$). Within the boundary layer, however, the mean predictions are reasonably aligned with the groundtruth for high extinction coefficient values ($> 10^{-5}\text{ m}^{-1}$). Although low extinction coefficient values are less dominant in the boundary layer, they also tend to be overestimated. We attribute this to the smoothing effect of the GP prior, which is prone to overestimation between extinction pockets.

4.3. Ablation study

We examine the contribution to predictions of our modelling choices by comparing the scores obtained with *VAExtGP* to the scores obtained with the 3 ablated models described in Section 3.4.1. Figure 6 shows the relative changes in scores with the ablated models, defined as

$$\% \text{Relative change} = \frac{\text{Score } VAExtGP - \text{Score ablated model}}{\text{Score } VAExtGP} \times 100, \quad (32)$$

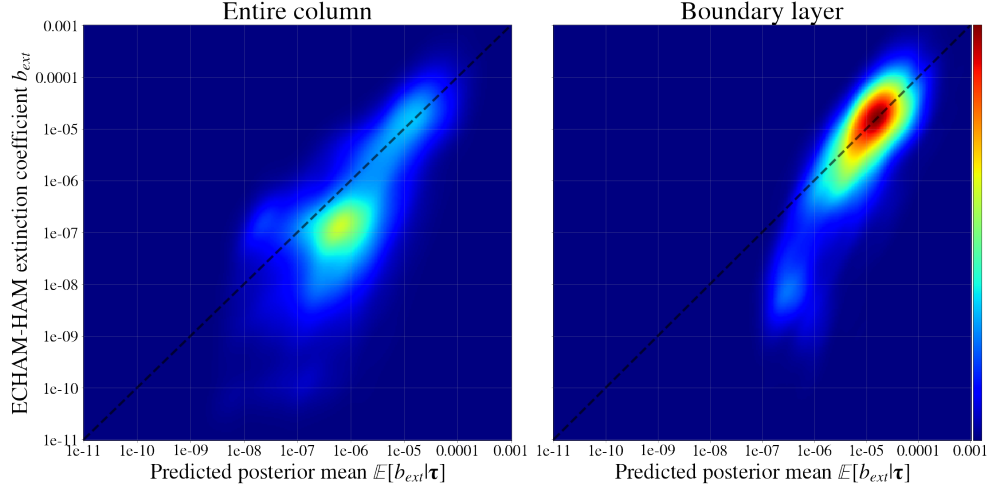


Figure 5. Density plots of groundtruth extinction coefficient values against predicted posterior mean extinction coefficient; **Left:** plotted for the entire column; **Right:** plotted for the boundary layer only; density plots are computed on a random subset on a random subset of 1000 samples drawn for the entire column (left) and in the boundary layer (right).

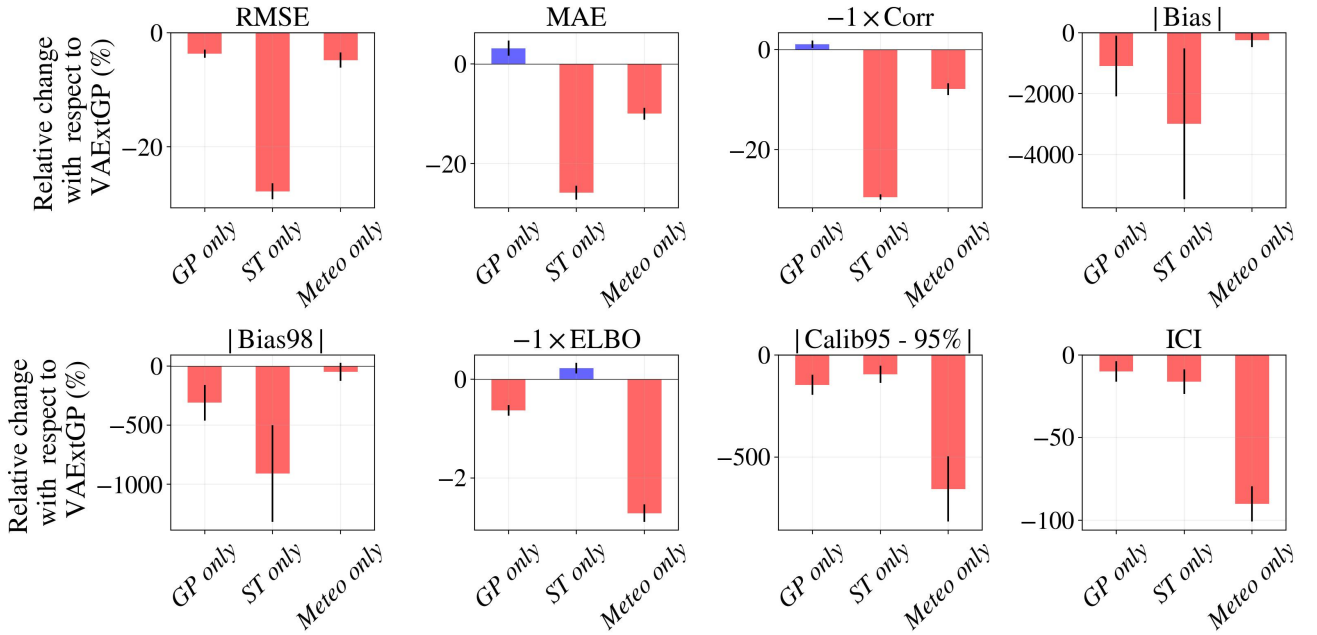


Figure 6. Relative changes in scores with respect to VAExtGP for the 3 ablated models : “GP only” (no idealised exponential term $e^{-h/L}$ in the prior), “ST only” (only spatiotemporal covariates in the input variable) and “Meteo only” (only meteorological covariates in the input variable). Red/Blue indicates that the performance is on average worse/better for the ablated model. We report 1 standard deviation..

where a positive value (blue) corresponds to an improvement for the ablated model, and a negative value (red) corresponds to a decrease in performance for the ablated model. Transformations of Corr, Bias, Bias98, ELBO and Calib95 are taken such that smaller values correspond to better performance, and are detailed in Appendix C.

Overall, we find that excluding a component from the model leads to a decrease in performance across most metrics. Specifically, when using only a GP without the idealized exponential term in the prior (*GP only*), only spatiotemporal covariates (*ST only*), or only meteorological covariates (*Meteo only*), the model’s performance is adversely affected.

For deterministic metrics, the ablation of meteorological covariates (*ST only*) has a particularly pronounced impact, with drastic drop in Bias and Bias98. This highlights the key role of learning the mapping from meteorological variables to extinction profiles. For probabilistic metrics, we observe that the ablation of spatiotemporal covariates (*Meteo only*) leads to the largest decrease in scores, indicating that incorporating spatiotemporal information plays a significant role in calibrating the predicted posterior distribution. Furthermore, we note that whilst the performance of the *GP only* model is generally inferior to that of *VAExt*, emphasizing the importance of encoding prior knowledge about aerosol vertical distribution through the idealized exponential profile, it is also the ablated model that least degrades performance. This highlights the remarkable adaptivity of GPs, and supports their use to learn non-trivial relationship in data.

In conclusion, these results demonstrate the significance of incorporating both spatiotemporal and meteorological covariates, as well as the idealised exponential profile, in our proposed model.

4.4. Feature importance analysis

In this section, we investigate the role predictors in *VAExtGP* play in the prediction of the extinction profiles. We begin by examining the model hyperparameters to gain understanding of how do predictors overall contribute to the covariance structure. We then ask the following questions:

- How do meteorological predictors modulate the predicted extinction?
- Do temperature and pressure provide valuable meteorological information for prediction or are they just proxies for height?

To address these questions, we propose a more detailed feature importance analysis that elucidates the contribution of predictors to individual predictions: Shapley values (SVs) [41]. Originally introduced as a game-theoretic concept, SVs have been widely adopted in machine learning to design local feature importance explanation models [14, 29, 30, 48]. With SVs, we can estimate for each individual prediction, how much each predictor contributed, and hence explain predictions. In our study, we use the KernelSHAP SV-based explanation model [29].

4.4.1. Global analysis with model hyperparameters

Table 4 reports the hyperparameters values after tuning them to maximise the ELBO objective as described in Section 2.3.3. These values provide a global description of the importance of predictors contribution to the covariance structure. γ_1 and γ_2 respectively correspond to the variances of the spatiotemporal and meteorological kernels from (23). A larger variance hyperparameter indicates a greater contribution to the covariance. The lengthscales $\ell_{\text{lat,lon}}$, ℓ_T , ℓ_P , ℓ_{RH} and ℓ_ω correspond to the lengthscales of each standardised predictor in the spatiotemporal kernel (24) and (25). A smaller lengthscale parameter means that the covariance is more sensitive to small variations in the predictors.

Table 4. Hyperparameter values for *VAExtGP* tuned for standardised input variables following procedure described in Section 2.3.3 ; lengthscales are unitless and apply to standardised predictors ; hyperparameters are tuned using 10 different initialisation seeds ; we report 1 standard deviation ; † indicates the value is fixed in the model and not tuned against data.

γ_1	γ_2	ℓ_t	$\ell_{\text{lat,lon}}$	ℓ_T	ℓ_P	ℓ_{RH}	ℓ_ω
0.26±0.02	1.0†	7.6±0.2	0.85±0.06	2.5±0.1	3.4±0.1	3.4±0.1	6.6±0.1

The value of γ_1 indicates that meteorological predictors globally contribute more to the covariance than spatiotemporal predictors. The large value of the temporal lengthscale ℓ_t (greater than 5) suggests a minimal sensitivity of the model to time. This is certainly because the dataset covers a single day

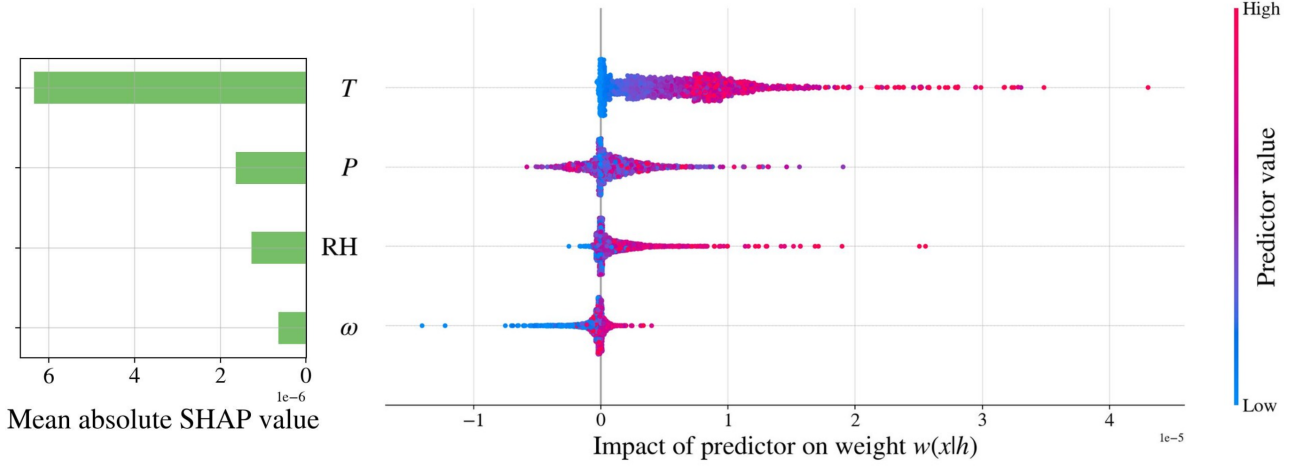


Figure 7. Left: Mean absolute Shapley values of meteorological predictors in the boundary layer; **Right:** Beeswarm plot of Shapley values of individual predictions in the boundary layer for each meteorological predictor; red/blue dots indicate a sample where the predictor value lies on the upper/lower end of its distribution; Shapley values are computed on a random subset of 2000 samples in the boundary layer.

and therefore exhibits limited temporal variability. In contrast, the small value of the spatial lengthscale $\ell_{\text{lat, lon}}$ indicates a high sensitivity to spatial location and suggests that spatial correlation occurs within a restricted spatial range. This likely arises from the spatial heterogeneity of aerosol concentration, resulting in localized, short-range spatial patterns.

Regarding meteorological predictors, the lengthscales indicate that — in general — predictions are more sensitive to variations in temperatures and less sensitive to variations in updraft velocity. In fact, the large value of the updraft velocity lengthscale ℓ_{ω} indicates, in general, a minimal correlation decay over the range of inputs.

4.4.2. How do meteorological predictors modulate the predicted extinction?

To gain a refined understanding of how meteorological predictors affect predictions, we conduct a SV analysis of their contribution to individual predictions. Since the meteorological predictors only modulate predictions through the weighting function, we only study the impact meteorological predictors have on the weight $w(x|h)$ predicted for the posterior mean response. We also focus on samples within the boundary layer since this is where most of the variability happens. Figure 7 displays the mean absolute SV obtained for each meteorological predictor and the contribution each predictor has on a randomly selected set of individual predictions.

We observe that in absolute mean, temperature is the factor that most influences predictions of the weighting function in the boundary layer. This is sensible as temperature is a smooth spatial field that gradually decreases with altitude, making T characteristic of the altitude level. Since the altitude typically correlates with extinction values, T becomes an informative proxy of extinction. This is corroborated by the plot of contributions to individual predictions: greater temperature (red) — lower altitudes — contribute to a greater weight $w(x|h)$ and hence greater extinction prediction while lower temperature (blue) — higher altitudes — yield lower impact on the weight and hence lower extinction prediction. It is interesting to notice that while pressure is also a proxy of altitude just like temperature, its mean absolute contribution is lower and the individual contribution plot does not display a clear identifiable trend. This suggests that the information conveyed by pressure might be redundant with the information conveyed by temperature, and could possibly be discarded from the model without loss of predictive performance.

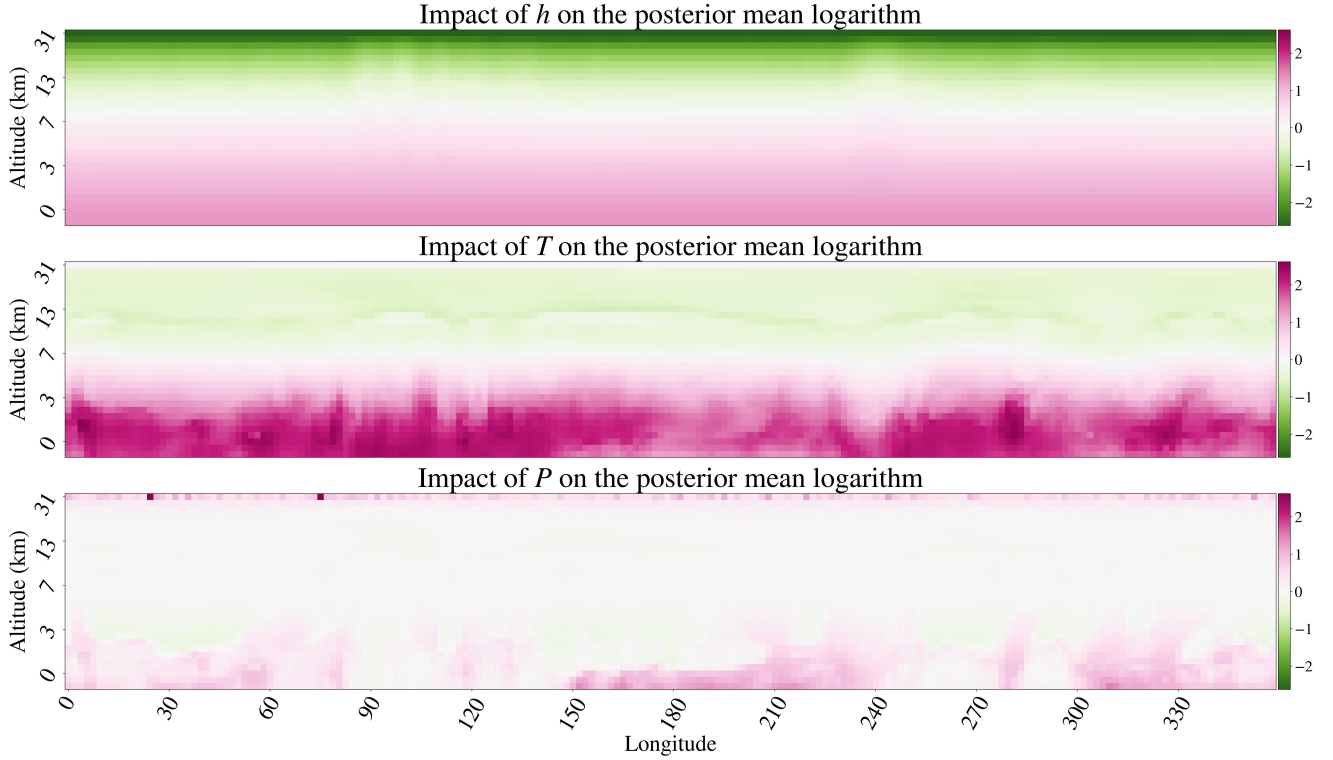


Figure 8. Vertical slices at latitude 51.29° (same as Figure 4) of the Shapley values for height h , temperature T and pressure P ; the Shapley values are computed for predicting the logarithm of the posterior mean, i.e. $\log \mathbb{E}[\varphi(x|h)|\tau]$; green/pink indicates a sample where the predictor has contributed to a lower/higher predicted extinction.

Whilst the mean absolute contribution of relative humidity seems marginal compared to temperature, we notice in individual contributions that when RH contributes to a greater weight $w(x|h)$ (points lying on the right), it consistently corresponds to locations with high RH. This supports the observation that our model captures extinction due to aerosol water uptake. Similarly, the mean absolute contribution of ω is relatively small, but a clear trend appears in the individual contributions: a positive updraft velocity consistently corresponds to a positive contribution of ω to the weight function, whereas a negative updraft velocity corresponds to a negative contribution to the weight function. This is consistent with intuition as strong updraft velocities lead to humidification of aerosols, and an increase in water content in aerosols increases in turn the extinction due to aerosol water uptake. Conversely, a negative updraft velocity should lead to lower humidification of aerosols, and therefore a decrease in extinction due to aerosols water uptake.

4.4.3. Temperature and pressure: proxies for height?

Temperature and pressure carry significant information about height, and may therefore inform *VAExtGP* primarily as a proxy for height. Indeed, the vertical slices of T and P in Figure 4 exhibit an approximately linear behavior with height. However, for a fixed height, temperature and pressure also exhibit variations along longitude, and may therefore provide additional information about meteorological conditions, beyond serving as a proxy for height. To investigate this, we conduct a SV analysis of the contribution of height, temperature and pressure to the predicted logarithm of the posterior mean $\log \mathbb{E}[\varphi(x|h)|\tau]$. Figure 8 shows the reported SVs for the same vertical slice as Figure 4.

The contribution of h to predicted extinction is clearly visible, where greater altitude contributes to a lower predicted extinction and lower altitude contributes to a greater extinction. This corresponds to the domain knowledge we encode within the prior by using an idealised exponential component. The contribution of T displays a similar mode, in agreement with the analysis from Section 4.4.2: at high

Table 5. Percentages of relative change in scores in with respect to VAExtGP when removing P and (T, P) from the input meteorological predictors. Red/Blue indicates that the performance is worse/better for the ablated model with statistical significance at $p < 0.05$ level. We report 1 standard deviation.

Remove	RMSE	MAE	$-1 \times \text{Corr}$	$ \text{Bias} $	$ \text{Bias98} $	$-1 \times \text{ELBO}$	$ \text{Calib95} - 95\% $	ICI
P	+2.7 $\pm$ 1.1	-0.77 $\pm$ 1.4	+1.5 $\pm$ 0.93	-52 $\pm$ 1.8 $\cdot 10^2$	+64 $\pm$ 27	+0.39 $\pm$ 0.13	+46 $\pm$ 18	+18 $\pm$ 7.7
T, P	-25 $\pm$ 2	-22 $\pm$ 1.6	-23 $\pm$ 0.02	-2.7 $\cdot 10^3 \pm 2.4 \cdot 10^3$	-8.9 $\cdot 10^2 \pm 4.2 \cdot 10^2$	+8.0 $\cdot 10^{-2} \pm 1.6 \cdot 10^{-1}$	-23 $\pm$ 42	-7.9 $\pm 1.4 \cdot 10^1$

altitude, lower temperatures contribute to lower predicted extinction, whereas at low altitudes, higher temperatures contribute to greater predicted extinction. The contribution of P is in general marginal in comparison, which aligns with the conjecture that P carries redundant information and may be a confounding feature which can be discarded.

In the stratosphere (altitude $> 10\text{km}$), the contribution of T becomes marginal in comparison with the contribution of h . This suggests that the contribution of the exponential profile dominates at high altitudes. In the troposphere (altitude $< 10\text{km}$), the contribution of T dominates, indicating that temperature plays a key role in predictions at lower altitudes. This is particularly visible in the boundary layer (altitude $< 2 - 3\text{km}$). Importantly, the temperature SVs display spatial heterogeneity within the boundary layer, where the contribution at a fixed altitude varies across longitude. This is different from the height SVs, and suggests that T informs the model with valuable meteorological information, beyond being a proxy for height.

To evaluate the contribution of T and P to prediction performance, we conduct ablation experiments where we train and evaluate VAExtGP by (i) removing P and (ii) removing both T, P from the input meteorological predictors. Table 5 shows the percentages of relative change in score in comparison with the entire VAExtGP model. We follow the same conventions than in Section 4.3 and take transformations of Corr, Bias, Bias98, ELBO and Calib95 such that smaler values correspond to better performance.

We find that: (i) predictive performance is overall improved by removing P . This supports the conjecture that P is a confounding factor, which carries redundant information for the prediction of extinction, and thereby is a source of noise in predictions. (ii) removing both T and P overall harms predictive performance. This supports the observation that, whilst strongly correlated with height, temperature and pressure provide additional predictive utility.

5. Conclusion and Outlooks

In this work, we introduce a GP-based methodology to vertically disaggregate the AOD using simple vertically resolved meteorological predictors such as temperature, pressure or relative humidity. Our approach emphasises uncertainty quantification using a Bayesian formalism. A successful application of our methodology to the vertical disaggregation of ECHAM-HAM simulated AOD is demonstrated. Our model outperforms an idealised baseline and displays capacity to recover realistic extinction patterns, in particular for extinction patterns arising from aerosol swelling in the boundary layer.

The simplicity of the explicit modelling assumptions grants better control and interpretability over the model and makes the choice of analysis strategy less subjective. While such simplicity can never account for the complex phenomena underpinning ACIs, this is balanced by a particular emphasis on quantification of epistemic uncertainty through a principled Bayesian formalism.

Naturally, the modelling assumptions can be adapted to reflect different modelling choices. For example, where more relevant vertically resolved quantities can be obtained, they can seamlessly be incorporated into the predictors. The idealised exponential component of the prior can be replaced by an idealised Gaussian profile to reflect the assumptions made by different AOD retrieval algorithms [28, 57]. The kernel design is also flexible and allows the user to specify a covariance structure that best

accounts for pairwise dependencies of the predictors. Similarly, the positive link function, integration scheme or number of inducing locations can be modified to fit specific needs.

Experiments demonstrate the proposed model is able to realistically reproduce vertical structures of extinction. In particular — for the chosen set of meteorological predictors — we are able to reliably predict extinction due to water vapour in the boundary layer. Because of the smoothness bestowed by the GP prior, the model tends to overestimate low extinction above the boundary layer and between extinction packets in the boundary layer. It is conjectured that the remaining unexplained extinctions patterns can be attributed to aerosols mass concentration, particles size and radiative properties, which are more challenging to model and would require additional vertically resolved covariates. This constitutes an important area for future work.

Regarding methodology, several extensions remain to be explored. For example, while we only consider aerosol extinction at a single wavelength, the AOD is in general retrieved for multiple wavelengths ranging from 470 nm to 870 nm [40]. By making the GP a function of the wavelength $f(\lambda, x|h)$, it should be possible to leverage multiple AOD observations by introducing a notion of functional smoothness across the electromagnetic spectrum. Another exciting direction would be to allow the use of AOD observations which are not matched with vertically resolved predictors. This can be achieved by adapting the work of Chau et al. [8] to our model, using a globally observed 2D field that would mediate the learning between unmatched AOD and vertically resolved predictors. Such addition would be particularly welcome as it would allow prediction of aerosol vertical profiles even at locations where AOD is not observed but vertically resolved predictors are.

The proposed method can also be useful to diagnose the extinction profiles produced by Earth system models (ESMs). For example, by training and validating our model on ESMs, one can produce a predictive model that learns ESM extinction characteristics. Predictions from this model can then be evaluated against real extinction observations, and thereby provide a quantitative indicator of whether ESM extinction characteristics match the characteristics of observed extinction.

Finally, a different exciting use-case of the proposed methodology can be considered. If working with climate model data only, then one can choose to use any vertically resolved aerosol tracers as predictors. These include mass and number concentrations which are simulated for several aerosol modes (nucleation, aiten, coarse and accumulation) and species (dust, sea salt, sulfate, black and organic carbon). Leveraging the interpretability of the GP covariance function, the kernel hyperparameters can then indicate which mode/specie did contribute to predictions, hence providing a tool to better understand factors that modulate the optical properties of aerosols.

In future work we intend to apply our model to AOD arising from satellite observations and meteorological predictors from reanalysis data. Since observations of groundtruth extinction coefficients are not available in 2D satellite products, we intend to collocate observations from MODIS 2D AOD products [40] with CALIOP vertical lidar measurements [55] to validate the model.

Acknowledgments. We would like to thank two anonymous reviewers for their thorough reviews and comments, which contributed significantly to the quality of this work.

Funding Statement. SB, DWP, AN, DS receive funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 860100. DWP acknowledges funding from NERC project NE/S005390/1 (ACRUISE). SS receives funding from the Engineering and Physical Sciences Research Council (EPSRC) via the UK Research and Innovation (UKRI) Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks (AI4ER, EP/S022961/1).

Competing Interests. None

Data Availability Statement. The data and code that support the findings of this study are openly available at <https://doi.org/10.5281/zenodo.11028725>.

Ethical Standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Author Contributions. Conceptualization: SB, SS. Data curation: SB, DWP. Investigation: SB. Methodology: SB, DS, DWP. Software: SB. Visualization: SB. Writing – original draft: SB. All authors approved the final submitted draft.

Supplementary Material. Supplementary material has been provided with the submission.

Nomenclature

Table 6. *Acronyms*

ACIs	Aerosols-cloud interactions
AOD	Aerosol optical depth
CALIOP	Cloud-Aerosol LIdar with Orthogonal Polarization
CCN	Cloud condensation nuclei
ECHAM-HAM	ECMWF Hamburg-Hamburg Aerosol Model
ECMWF	European Centre for Medium-Range Weather Forecasts
ELBO	Evidence lower-bound
ICI	Integrated calibration index
GP	Gaussian process
MISR	Multi-angle Imaging Spectroradiometer
MODIS	Moderate Resolution Imaging Spectroradiometer
SV	Shapley value
VIIRS	Visible Infrared Imaging Radiometer Suite

Table 7. *Greek letters*

α	Truncated credible interval size
γ_1	Spatiotemporal kernel variance hyperparameter
γ_2	Meteorological predictors kernel variance hyperparameter
η	Generic notation for the mean parameter of a log-normal distribution
η_i	Mean parameter of the log-normal observation model for the i^{th} column
$\hat{\eta}_i$	Approximated mean parameter of the log-normal observation model for the i^{th} column
λ	Generic notation for wavelength
μ	Generic notation for the location parameter of a log-normal distribution
μ_z	Mean vector of the variational distribution
$\bar{\mu}(\cdot)$	Mean vector of the posterior variational distribution evaluated at $(\cdot)$
ν	Matérn covariance function order
σ	Scale parameter of the AOD log-normal observation model
σ_{ext}	Scale parameter of the extinction coefficient log-normal observation model
Σ_z	Covariance matrix of the variational distribution
$\bar{\Sigma}(\cdot)$	Covariance of the posterior variational distribution evaluated at $(\cdot)$
τ	Generic notation for the observed AOD
τ_i	Observed AOD for the i^{th} column
$^s\tau_i$	Spatially smoothed AOD observation for the i^{th} column
τ	Vector of observed AOD for the n columns
φ	Prior over extinction coefficient profile
$^s\varphi$	Rescaled prior over extinction coefficient profile
ψ	Positive link function warping the GP

Table 8. Latin letters

b_{ext}	Extinction coefficient
d	Dimensionality of the vertically resolved covariates vector
$\mathcal{D}$	Dataset
f	Gaussian process
$\mathbf{f}$	Generic notation for a realization of the GP at vertically resolved covariates $x h$
$\mathbf{f}_i$	Realization of the GP over all vertically resolved inputs in the dataset
$\mathbf{f}_i$	Realization of the GP over all vertically resolved inputs from the i^{th} column
h	Generic notation for height
$h_i^{(j)}$	Altitude of the j^{th} vertically resolved covariates vector for the i^{th} column
H	Generic notation for the total height of an atmospheric column
$\mathbf{I}_p$	Identity matrix of size p
k	GP covariance function or kernel
$\mathbf{K}_{\mathbf{zz}}$	Covariance matrix evaluated at inducing locations
$\ell_{(\cdot)}$	Kernel lengthscale hyperparameter for variable $(\cdot)$
L	idealised exponential profile lengthscale
lat	Generic notation for latitude
lon	Generic notation for longitude
m	GP mean function
m_i	Number of observed altitude levels for the i^{th} column
M	Total number of vertically resolved observations across the dataset
n	Number of columns observed in the dataset
$N_{\sigma_{\text{ext}}}(\alpha)$	Percentage of groundtruth observation falling within the $1 - \alpha$ credible interval
p	Number of inducing locations
P	Generic notation for pressure
q	Variational distribution
S	Generic notation for supersaturation
t	Generic notation for time
T	Generic notation for temperature
$\mathbf{u}$	Inducing variables of the variational distribution
w	Weighting function for the idealised exponential profile
$\mathbf{z}$	Inducing locations of the variational distribution
x	Generic notation for the vertically resolved covariates vector
$x_i^{(j)}$	Observed vertically resolved covariates at altitude $h_i^{(j)}$ for the i^{th} column
$\mathbf{x}$	Vector of all vertically resolved covariates from the dataset
$\mathcal{X}$	Space in which the vertically resolved covariates vector take values

References

1. AS Ackerman, MP Kirkpatrick, DE Stevens, and OB Toon. The impact of humidity above stratiform clouds on indirect aerosol climate forcing. *Nature*, 2004.
2. Bruce A. Albrecht. Aerosols, Cloud Microphysics, and Fractional Cloudiness. *Science*, 1989.
3. M. O. Andreae. Correlation between cloud condensation nuclei concentration and aerosol optical thickness in remote and polluted regions. *Atmospheric Chemistry and Physics*, 2009.
4. Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, et al. Matérn gaussian processes on riemannian manifolds. *Advances in Neural Information Processing Systems*, 33:12426–12437, 2020.
5. Gustau Camps-Valls, Jochem Verrelst, Jordi Muñoz-Marí, Valero Laparra, Fernando Mateo-Jimenez, and Jose L. Gómez-Dans. A Survey on Gaussian Processes for Earth-Observation Data Analysis: A Comprehensive Investigation. *IEEE Geoscience and Remote Sensing Magazine*, 2016.
6. Ken S. Carslaw and Kirsty Pringle. Chapter 4 - Global aerosol properties. In Ken S. Carslaw, editor, *Aerosols and Climate*, pages 101–133. Elsevier, 2022. ISBN 978-0-12-819766-0. doi: <https://doi.org/10.1016/B978-0-12-819766-0.00011-0>. URL <https://www.sciencedirect.com/science/article/pii/B9780128197660000110>.

7. KS Carslaw, LA Lee, CL Reddington, KJ Pringle, A Rap, PM Forster, GW Mann, DV Spracklen, MT Woodhouse, LA Regayre, et al. Large contribution of natural aerosols to uncertainty in indirect forcing. *Nature*, 2013.
8. Siu Lun Chau, Shahine Bouabid, and Dino Sejdinovic. Deconditional Downscaling with Gaussian Processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
9. Yi-Chun Chen, Matthew W. Christensen, Graeme L. Stephens, and John H. Seinfeld. Satellite-based estimate of global aerosol-cloud radiative forcing by marine warm clouds. *Nature Geoscience*, 2014.
10. M. W. Christensen, D. Neubauer, C. A. Poulsen, G. E. Thomas, G. R. McGarragh, A. C. Povey, S. R. Proud, and R. G. Grainger. Unveiling aerosol–cloud interactions – Part 1: Cloud contamination in satellite products enhances the aerosol indirect forcing estimate. *Atmospheric Chemistry and Physics*, 2017.
11. Antony Clarke and Vladimir Kapustin. Hemispheric Aerosol Vertical Profiles: Anthropogenic Impacts on Optical Depth and Cloud Nuclei. *Science*, 2010.
12. A. Douglas and T. L’Ecuyer. Quantifying cloud adjustments and the radiative forcing due to aerosol–cloud interactions in satellite observations of warm marine clouds. *Atmospheric Chemistry and Physics*, 2020.
13. Aasa Feragen and Søren Hauberg. Open problem: Kernel methods on manifolds and metric spaces. what is the probability of a positive definite geodesic exponential kernel? In *Conference on learning theory*, pages 1647–1650. PMLR, 2016.
14. Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*. PMLR, 2019.
15. Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.
16. B.N. Holben, T.F. Eck, I. Slutsker, D. Tanré, J.P. Buis, A. Setzer, E. Vermote, J.A. Reagan, Y.J. Kaufman, T. Nakajima, F. Lavenu, I. Jankowiak, and A. Smirnov. AERONET: A Federated Instrument Network and Data Archive for Aerosol Characterization. *Remote Sensing of Environment*, 1998.
17. Andrew Jesson, Alyson Douglas, Peter Manshausen, Nicolai Meinshausen, Philip Stier, Yarin Gal, and Uri Shalit. Scalable Sensitivity and Uncertainty Analysis for Causal-Effect Estimates of Continuous-Valued Interventions. *arXiv preprint arXiv:2204.10022*, 2022.
18. Ralph A. Kahn and Barbara J. Gaitley. An analysis of global aerosol type as retrieved by MISR. *Journal of Geophysical Research (Atmospheres)*, 2015.
19. Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015.
20. Durk P Kingma, Tim Salimans, and Max Welling. Variational Dropout and the Local Reparameterization Trick. In *Advances in Neural Information Processing Systems*, 2015.
21. Hilding Köhler. The nucleus in and the growth of hygroscopic droplets. *Trans. Faraday Soc.*, 1936.
22. I. Laszlo and H.Q. Liu. EPS Aerosol Optical Depth (AOD) Algorithm Theoretical Basis Document. *NOAA NESDIS Center for Satellite Application and Research*, 2016.
23. Ho Chung Law, Peilin Zhao, Leung Sing Chan, Junzhou Huang, and Dino Sejdinovic. Hyperparameter Learning via Distributional Transfer. In *Advances in Neural Information Processing Systems*, 2019.
24. Felix Leibfried, Vincent Dutordoir, ST John, and Nicolas Durrande. A tutorial on sparse Gaussian processes and variational inference. *arXiv preprint arXiv:2012.13962*, 2020.
25. Jussi Leinonen, Alexandre Guillaume, and Tianle Yuan. Reconstruction of Cloud Vertical Structure With a Generative Adversarial Network. *Geophysical Research Letters*, 2019.
26. Ho Chung Leon Law, Dino Sejdinovic, Ewan Cameron, Tim C.D. Lucas, Seth Flaxman, Katherine Battle, and Kenji Fukumizu. Variational learning on aggregate outputs with Gaussian processes. In *Advances in Neural Information Processing Systems*, 2018.
27. Robert C. Levy, Lorraine A. Remer, Shana Mattoo, Eric F. Vermote, and Yoram J. Kaufman. Second-generation operational algorithm: Retrieval of aerosol properties over land from inversion of Moderate Resolution Imaging Spectroradiometer spectral reflectance. *Journal of Geophysical Research: Atmospheres*, 2007.
28. Chong Li, Jing Li, Oleg Dubovik, Zhao-Cheng Zeng, and Yuk L. Yung. Impact of Aerosol Vertical Distribution on Aerosol Optical Depth Retrieval from Passive Satellite Sensors. *Remote Sensing*, 2020.
29. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.
30. Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.
31. Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (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*. Cambridge University Press, 2021.
32. Alexander G. de G. Matthews, James Hensman, Richard Turner, and Zoubin Ghahramani. On Sparse Variational Methods and the Kullback-Leibler Divergence between Stochastic Processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016.
33. Robert A. McCormick and John H. Ludwig. Climate Modification by Atmospheric Aerosols. *Science*, 1967.

34. T. Mielonen, R. C. Levy, V. Aaltonen, M. Komppula, G. de Leeuw, J. Huttunen, H. Lihavainen, P. Kolmonen, K. E. J. Lehtinen, and A. Arola. Evaluating the assumptions of surface reflectance and aerosol type selection within the MODIS aerosol retrieval over land: the problem of dust type selection. *Atmospheric Measurement Techniques*, 2011.
35. J Muñoz-Sabater, E Dutra, A Agustí-Panareda, C Albergel, G Arduini, G Balsamo, S Boussetta, M Choulga, S Harrigan, H Hersbach, et al. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications, 2021.
36. O'Neill N. T., A. Ignatov, B. N. Holben, and T. F. Eck. The lognormal distribution as a reference for reporting aerosol optical depth statistics; Empirical tests using multi-year, multi-site AERONET Sunphotometer data. *Geophysical Research Letters*, 2000.
37. Teruyuki Nakajima, Akiko Higurashi, Kazuaki Kawamoto, and Joyce E. Penner. A possible correlation between satellite-derived cloud and aerosol microphysical parameters. *Geophysical Research Letters*, 2001.
38. D. Painemal, F.-L. Chang, R. Ferrare, S. Burton, Z. Li, W. L. Smith Jr, P. Minnis, Y. Feng, and M. Clayton. Reducing uncertainties in satellite estimates of aerosol–cloud interactions over the subtropical ocean by integrating vertically resolved aerosol observations. *Atmospheric chemistry and physics*, 2020.
39. C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, USA, 2005.
40. L. A. Remer, Y. J. Kaufman, D. Tanré, S. Mattoo, D. A. Chu, J. V. Martins, R.-R. Li, C. Ichoku, R. C. Levy, R. G. Kleidman, T. F. Eck, E. Vermote, and B. N. Holben. The MODIS Aerosol Algorithm, Products, and Validation. *Journal of the Atmospheric Sciences*, 2005.
41. Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 1953.
42. Jennifer D. Small, Patrick Y. Chuang, Graham Feingold, and Hongli Jiang. Can aerosol decrease cloud lifetime? *Geophysical Research Letters*, 2009.
43. D. V. Spracklen, K. S. Carslaw, U. Pöschl, A. Rap, and P. M. Forster. Global cloud condensation nuclei influenced by carbonaceous combustion aerosol. *Atmospheric Chemistry and Physics*, 2011.
44. Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999.
45. P. Stier, J. Feichter, S. Kinne, S. Kloster, E. Vignati, J. Wilson, L. Ganzeveld, I. Tegen, M. Werner, Y. Balkanski, M. Schulz, O. Boucher, A. Minikin, and A. Petzold. The aerosol-climate model ECHAM5-HAM. *Atmospheric Chemistry and Physics*, 2005.
46. P. Stier, J. H. Seinfeld, S. Kinne, and O. Boucher. Aerosol absorption and radiative forcing. *Atmospheric Chemistry and Physics*, 2007.
47. Philip Stier. Limitations of passive remote sensing to constrain global cloud condensation nuclei. *Atmospheric Chemistry and Physics*, 2016.
48. Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 2014.
49. Yusuke Tanaka, Toshiyuki Tanaka, Tomoharu Iwata, Takeshi Kurashima, Maya Okawa, Yasunori Akagi, and Hiroyuki Toda. Spatially aggregated Gaussian processes with multivariate areal outputs. *Advances in Neural Information Processing Systems*, 2019.
50. Michalis K. Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *AISTATS*, 2009.
51. S. Twomey. Pollution and the planetary albedo. *Atmospheric Environment*, 1974.
52. S. Twomey. The Influence of Pollution on the Shortwave Albedo of Clouds. *Journal of Atmospheric Sciences*, 1977.
53. Arto Klam Ville Tanskanen , Krista Longi. Non-Linearities in Gaussian Processes with Integral Observations. *IEEE international Workshop on Machine Learning for Signal*, 2020.
54. D. Watson-Parris, N. Bellouin, L. T. Deaconu, N. A. J. Schutgens, M. Yoshioka, L. A. Regayre, K. J. Pringle, J. S. Johnson, C. J. Smith, K. S. Carslaw, and P. Stier. Constraining Uncertainty in Aerosol Direct Forcing. *Geophysical Research Letters*, 2020.
55. D. M. Winker, J. L. Tackett, B. J. Getzewich, Z. Liu, M. A. Vaughan, and R. R. Rogers. The global 3-D distribution of tropospheric aerosols as characterized by CALIOP. *Atmospheric Chemistry and Physics*, 2013.
56. Yerong Wu, Martin De Graaf, and Massimo Menenti. The Sensitivity of AOD Retrieval to Aerosol Type and Vertical Distribution over Land with MODIS Data. *Remote Sensing*, 2016.
57. Yerong Wu, Martin de Graaf, and Massimo Menenti. The impact of aerosol vertical distribution on aerosol optical depth retrieval using calipso and modis data: Case study over dust and smoke regions. *Journal of Geophysical Research: Atmospheres*, 122(16):8801–8815, 2017.
58. Fariba Yousefi, Michael Thomas Smith, and Mauricio A. Álvarez. Multi-task learning for aggregated data using Gaussian processes. In *Advances in Neural Information Processing Systems*, 2019.
59. K. Zhang, D. O'Donnell, J. Kazil, P. Stier, S. Kinne, U. Lohmann, S. Ferrachat, B. Croft, J. Quaas, H. Wan, S. Rast, and J. Feichter. The global aerosol-climate model ECHAM-HAM, version 2: sensitivity to improvements in process representations. *Atmospheric Chemistry and Physics*, 2012.
60. Yivan Zhang, Nontawat Charoenphakdee, Zhenguo Wu, and Masashi Sugiyama. Learning from Aggregate Observations. In *Advances in Neural Information Processing Systems*, 2020.
61. Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 2017.

A. Matérn covariance

The Matérn covariances are a class of stationary covariance functions widely used in spatial statistics. The Matérn- ν covariance with unit variance between two points $x, x' \in \mathbb{R}$ is given by

$$C_\nu(|x - x'|) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|x - x'|}{\ell} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{|x - x'|}{\ell} \right), \quad (33)$$

where Γ is the gamma function, K_ν is the modified Bessel function and $\ell > 0$ is a lengthscale hyperparameter.

The covariance function expression considerably simplifies for $\nu = p + 1/2$ where $p \in \mathbb{N}$. For example, for $\nu = 1/2$ ($p = 0$) and $\nu = 3/2$ ($p = 1$) we have

$$C_{1/2}(|x - x'|) = \exp \left(-\frac{|x - x'|}{\ell} \right) \quad (34)$$

$$C_{3/2}(|x - x'|) = \left(1 + \sqrt{3} \frac{|x - x'|}{\ell} \right) \exp \left(-\sqrt{3} \frac{|x - x'|}{\ell} \right) \quad (35)$$

When $x, x' \in \mathbb{R}^d$, the distance $|x - x'|$ can be substituted by the norm $\|x - x'\| = \sqrt{\sum_{i=1}^d (x_i - x'_i)^2}$. The covariance is called an *automatic relevance determination* (ARD) kernel when each dimension has its own independent lengthscale parameter $\ell_i > 0$. For example, the Matérn-1/2 and Matérn-3/2 ARD kernel write

$$\|x - x'\|_\ell = \sqrt{\sum_{i=1}^d \frac{(x_i - x'_i)^2}{\ell_i}} \quad (36)$$

$$C_{1/2}(\|x - x'\|_\ell) = \exp(-\|x - x'\|_\ell) \quad (37)$$

$$C_{3/2}(\|x - x'\|_\ell) = \left(1 + \sqrt{3} \|x - x'\|_\ell \right) \exp \left(-\sqrt{3} \|x - x'\|_\ell \right) \quad (38)$$

When a Matérn- $p + 1/2$ covariance function is used as a kernel for a GP, draws from the GP are p times continuously differentiable (with convention that 0 times means simply continuous).

B. Comparison with tensor product covariance structure

A standard choice in design of covariance structures is to assume a tensor product form of the covariance with independent correlation over each input dimension, i.e. $k(x, x') = \prod_i k(x_i, x'_i)$. We compare our choice of covariance structure from Section 3.2.2 to the product covariance function given by

$$k(x|h, x'|h') = \gamma_2 C_{3/2}(|t - t'|) C_{3/2} \left(\text{Hav} \left(\begin{bmatrix} \text{lat} \\ \text{lon} \end{bmatrix}, \begin{bmatrix} \text{lat}' \\ \text{lon}' \end{bmatrix} \right) \right) \quad (39)$$

$$\times C_{1/2}(|T - T'|) C_{1/2}(|P - P'|) C_{1/2}(|\text{RH} - \text{RH}'|) C_{1/2}(|\omega - \omega'|), \quad (40)$$

where we maintain the same Matérn covariances for a the sake of comparison. This product structures introduces a global variance parameter γ_2 and independent lengthscale parameters $\ell_t, \ell_{\text{lat,lon}}, \ell_T, \ell_P, \ell_{\text{RH}}, \ell_\omega$ for each input variable, resulting in 7 hyperparameters that need to be tuned.

In contrast, the covariance structure we propose in Section 3.2.2 introduces two variance parameters γ_1, γ_2 and an equal number of lengthscale parameters for each input variables (through the ARD Matérn kernel on meteorological predictors), resulting in 8 hyperparameters that need to be tuned. However, because we fix $\gamma_2 = 1$ in our model, the number of hyperparameters that effectively require tuning is 7.

Table 9. Hyperparameter values for our choice of covariance and a tensor product covariance structure. Models are tuned for standardised input variables following procedure described in Section 2.3.3; hyperparameters are tuned using 10 different initialisation seeds; we report 1 standard deviation; † indicates the value of the parameter is fixed in the model and not tuned against data.

Covariance	γ_1	γ_2	ℓ_t	$\ell_{\text{lat,lon}}$	ℓ_T	ℓ_P	ℓ_{RH}	ℓ_ω
Ours	0.26±0.02	1.0†	7.6±0.2	0.85±0.06	2.5±0.1	3.4±0.1	3.4±0.1	6.6±0.1
Product	-	0.92±0.1	8.8±0.1	1.2±0.1	5.6±0.1	6.2±0.1	5.8±0.1	8.3±0.1

Table 9 reports the hyperparameters values for both the product covariance and our covariance, after tuning them to maximise the ELBO objective as described in Section 2.3.3. We find that the product covariance structure leads to larger lengthscale parameters for the meteorological predictors. This means that, in general, predictions with the product covariance structure should have less sensitivity to changes in meteorological predictors.

Using the tuned models, we then perform inference and predict the ECHAM-HAM extinction coefficients over the dataset with both models following the procedure outlined in Section 2.3. Table 10 reports the ECHAM-HAM extinction prediction scores for the product covariance and our covariance. It shows the proposed covariance provides substantial improvement in deterministic metrics compared to the tensor product covariance. Regarding probabilistic scores, the ELBO and Calib95 are less sensitive to the choice of covariance structure. We find that the ICI is improved over the entire column with a product covariance structure, but degraded within the boundary layer. These results suggest that separating the contribution of spatiotemporal and meteorological predictors in the covariance benefits extinction predictive performance.

Table 10. Scores of the VAExtGP with our choice of covariance and a product covariance structure for the task of predicting ECHAM-HAM extinction profiles ; “Entire column” means scores are computed and averaged for every altitude levels ; “Boundary layer” means scores are computed and averaged for altitude levels of the boundary level only (< 2 km) ; the best scores for each metric and region are highlighted in bold ; we report 1 standard deviation.

<i>Region</i>	<i>Covariance</i>	RMSE (10^{-5})	MAE (10^{-6})	Corr	Bias (10^{-6})	Bias98 (10^{-5})
<i>Entire column</i>	<i>Product</i>	3.65 $\pm$ 0.03	5.92 $\pm$ 0.09	0.61 $\pm$ 0.01	-0.54 $\pm$ 0.01	-1.15 $\pm$ 0.01
	<i>Ours</i>	3.21 $\pm$ 0.04	5.29 $\pm$ 0.06	0.73 $\pm$ 0.01	-0.005 $\pm$ 0.161	-0.482 $\pm$ 0.212
<i>Boundary layer</i>	<i>Product</i>	6.73 $\pm$ 0.05	16.5 $\pm$ 0.1	0.59 $\pm$ 0.01	-3.46 $\pm$ 0.22	-5.64 $\pm$ 0.28
	<i>Ours</i>	5.91 $\pm$ 0.07	15.3 $\pm$ 0.1	0.72 $\pm$ 0.06	-0.404 $\pm$ 0.71	-4.53 $\pm$ 0.44

<i>Region</i>	<i>Covariance</i>	ELBO	Calib95 (%)	ICI (10^{-2})
<i>Entire column</i>	<i>Product</i>	13.0 $\pm$ 0.1	93.9 $\pm$ 0.1	4.45 $\pm$ 0.15
	<i>Ours</i>	13.0 $\pm$ 0.1	93.5 $\pm$ 0.1	6.60 $\pm$ 0.45
<i>Boundary layer</i>	<i>Product</i>	10.6 $\pm$ 0.1	98.3 $\pm$ 0.1	7.16 $\pm$ 0.20
	<i>Ours</i>	10.7 $\pm$ 0.1	98.7 $\pm$ 0.1	6.26 $\pm$ 0.34

C. Additional experimental details

C.1. Ablation study: details on relative change computation

To evaluate the relative change in scores of the ablated model against *VAExtGP*, we compute

$$\% \text{Relative change} = \frac{\text{Score } VAExtGP - \text{Score ablated model}}{\text{Score } VAExtGP} \times 100, \quad (41)$$

where a positive percentage indicates improvement with respect to *VAExtGP* and a negative percentage indicates deterioration with respect to *VAExtGP*. However, this is only a useful indicator of change in performance across scores if all scores are improved when smaller. This is true for RMSE, MAE and ICI, but not for the other metrics used. Therefore, we take the following transformations reported in Table 11 to ensure that smaller values correspond to an improvement in score.

Table 11. Evaluation metrics and transformed evaluation metrics used to compute the relative change in scores for ablation experiments.

Metric	Best when	Transformed metric	Best when
RMSE	lower	RMSE	lower
MAE	lower	MAE	lower
Corr	higher	$-1 \times \text{Corr}$	lower
Bias	close to 0	$ \text{Bias} $	lower
Bias98	close to 0	$ \text{Bias98} $	lower
ELBO	higher	$-1 \times \text{ELBO}$	lower
Calib95	close to 95%	$ \text{Calib95} - 95\% $	lower
ICI	lower	ICI	lower

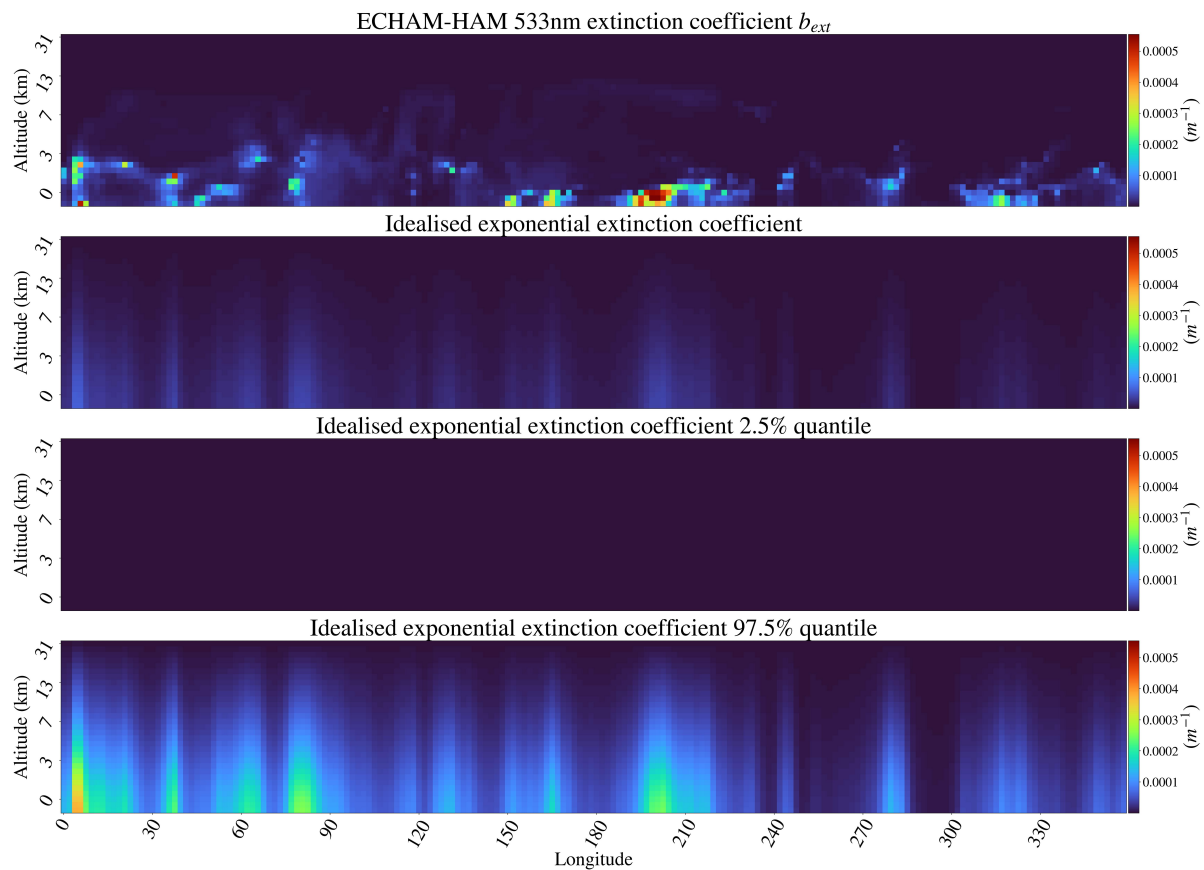
C.2. Additional sliced plots

Figure 9. Vertical slices at latitude 51.29° of groundtruth extinction coefficient, idealised exponential extinction coefficient, 2.5% and 97.5% quantiles of the idealised exponential extinction distribution.

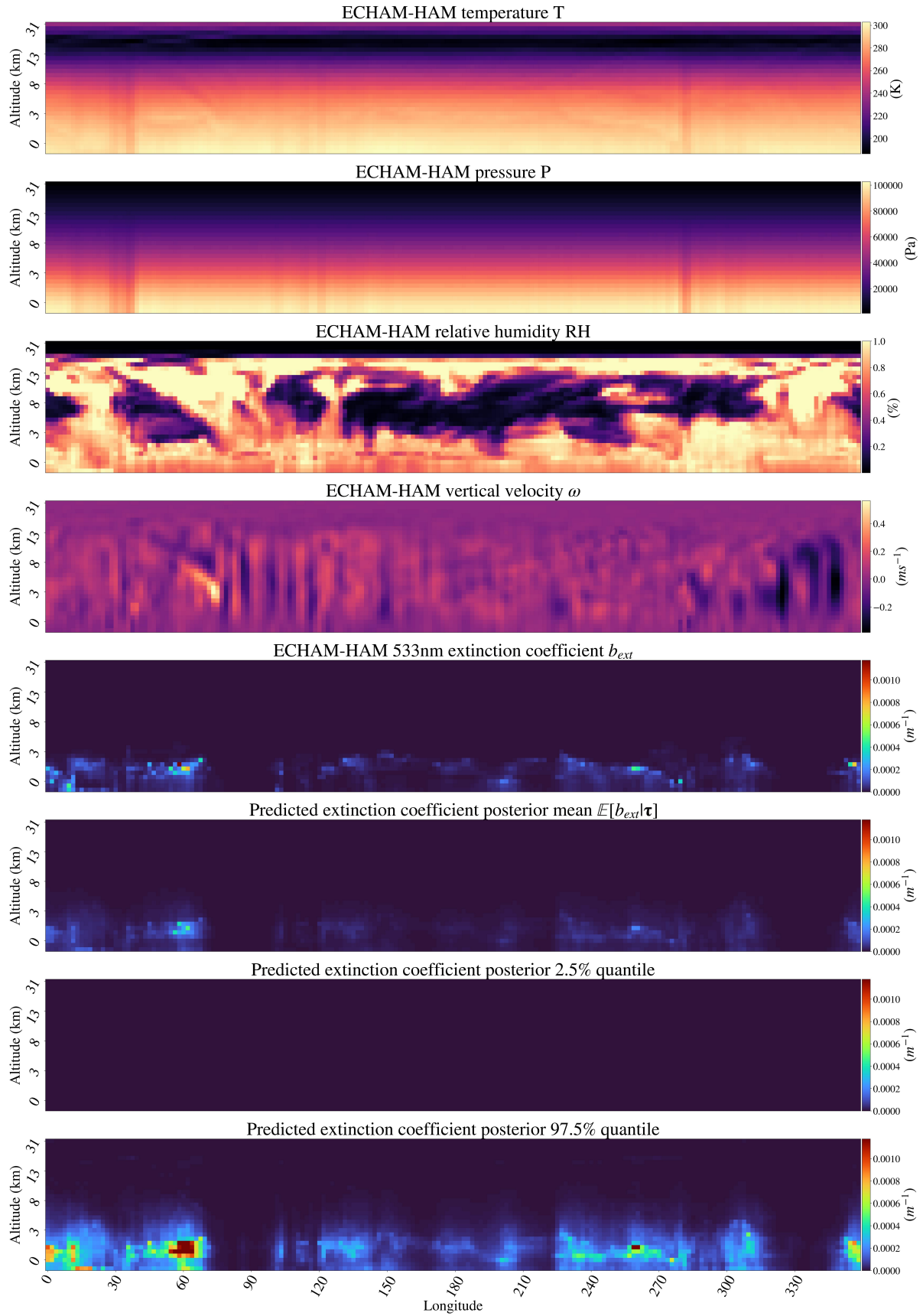


Figure 10. Vertical slices at latitude -0.93° of meteorological predictors (T, P, RH, ω), groundtruth extinction coefficient, predicted extinction coefficient posterior mean, 2.5% and 97.5% quantiles of the predicted extinction coefficient posterior distribution.

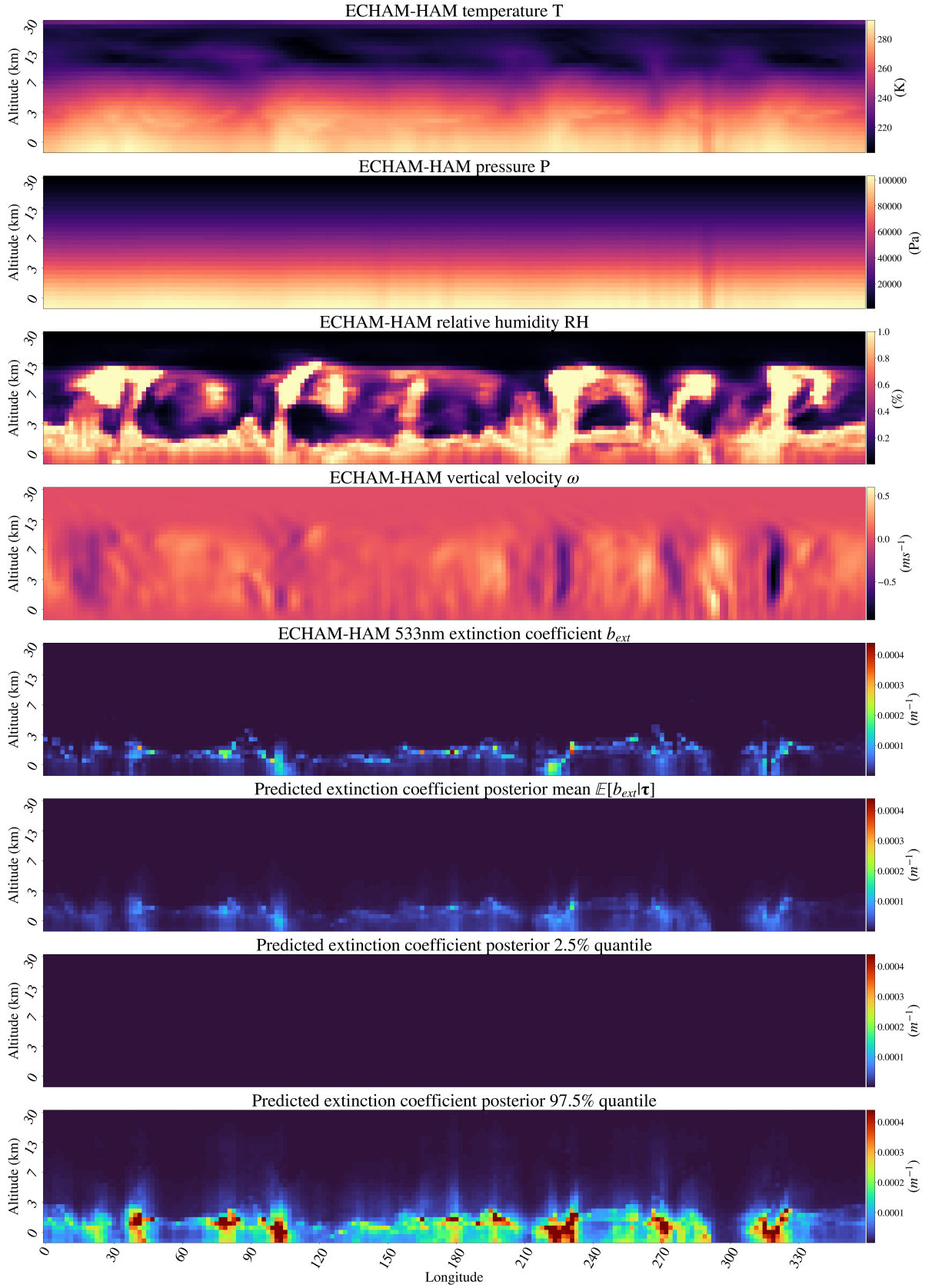


Figure 11. Vertical slices at latitude -38.2° of meteorological predictors (T, P, RH, ω), groundtruth extinction coefficient, predicted extinction coefficient posterior mean, 2.5% and 97.5% quantiles of the predicted extinction coefficient posterior distribution.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	AODisaggregation: toward global aerosol vertical profiles
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Bouabid, Shahine, Watson-Parris, Duncan, Stefanović, Sofija, Nenes, Athanasios and Sejdinovic, Dino, 2022. AODisaggregation: toward global aerosol vertical profiles.

Student Confirmation

Student Name:	Shahine Bouabid		
Contribution to the Paper	I am the first autor of this paper. The idea to use satellite data and vertical meteorological variables to reconstruct aerosol vertical profiles was conceptualized jointly with Sofija Stefanović. I came up with a Bayesian model based on Gaussian processes to frame this problem as a vertical disaggregation problem and proposed an approximate variational formulation of the posterior distribution on aerosol vertical extinction. Prof. Sejdinovic provided guidance on how to simplify the model formulation with log-normal distributions. Prof. Watson-Parris contributed the dataset we use in experiments and guidance on the choice of meteorological predictors. I designed, implemented and conducted the experiments and the feature importance analysis. I wrote the paper. Prof. Sejdinovic and Prof. Watson-Parris corrected the draft.		
Signature		Date	11/03/24

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof. Dino Sejdinovic			
Supervisor comments			
Signature		Date	12/03/2024

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 6

Conclusion and Outlooks

In this final chapter, we summarise the results presented in the thesis, discuss limitations, and outline directions for future research.

6.1 Summary

In this thesis, we demonstrate the role of statistical machine learning in helping address modelling and observational challenges in climate science through methodological and applied contributions. An important emphasis is placed on the integration of domain knowledge from climate science directly in the modelling, the algorithmic transparency of the model formulated, and the use of probabilistic modelling, which propagates into uncertainty quantification on predictions. These are guiding principles we advocate in the design of machine learning models in climate science. In particular, we focus on the use of kernel-based learners, which are transformed to incorporate domain knowledge, across different climate science challenges.

In Chapter 3, we contribute methodological work to incorporate a particular form of domain knowledge into a regression problem: knowledge of the causal structure underlying the data generating process. When interested in discriminative modelling of $Y|X$, knowledge of the graphical structure allows one to perform feature selection and discard from X any covariates which are uninformative for $Y|X$. We show that once all uninformative covariates have been discarded, collider structures are the only remaining structures that provide a useful inductive bias for regression. We formulate *collider regression* as a constrained regression problem that imposes this inductive bias on the regression problem, and prove that incorporating this knowledge can only benefit the model generalisation. Further, we prove this generalisation gain is strictly positive in expectation for kernel ridge regression, and grows with limited training data and complexity of the relationship between vertices of the collider. We

demonstrate empirically the improvement in test error in a synthetic experiment, and for the regression of aerosol radiative forcing using data simulated with a simple climate model.

In Chapter 4, we demonstrate how GPs can be used to represent the radiative forcing component in FaIR, a simple climate model. This offers a principled framework to acknowledge the uncertainty in the radiative forcing model, and derive a Bayesian prior over the temperature response, which we call FaIRGP. This prior can be analytically conditioned in climate model data, thereby allowing a more flexible emulation of surface temperatures beyond what a simple climate model alone can achieve. The GP treatment offers a transparent and fully mathematically tractable framework for probabilistic modelling which can seamlessly account for climate internal variability. Meanwhile, incorporating an energy balance model in the prior grants FaIRGP robustness and physical interpretability. We demonstrate empirically that FaIRGP outperforms both purely physically-driven and purely data-driven models for emulating realistic future scenarios, notably better predicting the temperature response to anthropogenic aerosol emissions.

In Chapter 5, we consider the problem of predicting aerosol vertical distribution globally. We propose a Bayesian hierarchical model using GPs to infer the vertical extinction of aerosols, based on aerosol optical depth — a column-integrated quantity — and vertically-resolved meteorological variables that modulate aerosol extinction, such as temperature, relative humidity or updraft wind velocity. The model we formulate incorporates heuristic domain knowledge about aerosol vertical distribution from satellite retrieval algorithms, and acknowledges structural uncertainties through our Bayesian formalism. This in turn allows for uncertainty quantification in predictions. We use a variational approximation to the posterior to address the intractability of direct posterior inference, and demonstrate the model can predict realistic extinction profiles with sound uncertainty quantification using climate model data.

6.2 Limitations

After presenting our contributed methods in the earlier chapters, we would like to discuss some limitations to our work.

Applicability and guarantees from collider regression The collider regression framework proposed offers a way to translate knowledge of the presence of a collider

into an inductive bias for regression. However, this framework relies on several strong assumptions.

First, a key limitation is that we require full knowledge of the causal structure. The causal structure can in some scenarios be provided by expert knowledge, with examples involving climate data in our work or in genetics (Day et al., 2016), and assuming access to the causal structure is common in many areas of causal inference, with examples in optimal adjustment sets (Runge, 2021) and causal Bayesian optimisation (Aglietti et al., 2020). Nonetheless, assuming to know the causal graph remains in general a strong assumption for practitioners, one for which we should however remain hopeful that further progress in causal discovery can help. Additionally, the issue of hidden and potentially confounding variables, which can affect observed outcomes, poses another critical challenge. Approaches that enable experts to define a level of hidden confounding offer potential pathways to tackle this issue (Jesson et al., 2022).

Second, the theoretical guarantees we provide for collider regression are formulated in terms of the exact conditional expectation operator E , or rather its complement $P = \text{Id} - E$. This is however a theoretical object to which we do not in practice have access. Rather, the projection P needs to be estimated from data, using for example one of the algorithms we present. Whilst this does not diminish the theoretical results contributed in our work, it limits the practical applicability of the guarantees they offer.

Spatiotemporal and uncertainty modelling in FaIRGP FaIRGP, as it currently stands, only models the global mean annual temperature response to changes in emissions, building directly upon the scope of FaIR. This narrow focus masks the disparities in exposure and vulnerability to climate change of different regions at different times of the year, and therefore limits the relevance of the emulator to study climate impacts: it is not only a luxury but a necessity to invest efforts in obtaining climate projection with finer spatial and temporal granularity. This necessity for finer granularity in climate data applies broadly, not just to climate emulation, and we tackle this for example in Chapter 5 where we aim to add a vertical dimension to aerosol extinction observations.

In FaIRGP, this is something we partially address with the introduction of the pattern scaling model. However, this approach has known limitations as it breaks under certain scenario (May, 2012; Levy et al., 2013; Tebaldi and Knutti, 2018; Tebaldi et al., 2020). It also raises modelling concerns since it overlooks the spatial dependence

of temperature fields, despite our expectation that temperatures at nearby locations should covary.

Moreover, whilst FaIRGP allows modelling the uncertainty over the radiative forcing response, this is alone insufficient to acknowledge other important dimensions of uncertainty, such as uncertainty on the climate sensitivity or the carbon cycle (Leach et al., 2021). Rigorously accounting for the contribution of each source of uncertainty is key to obtaining a complete picture of what an uncertain warming at the end of the century may look like, and this is something our GP modelling approach alone cannot achieve.

Transparency and limited scope in AODisaggregation In our aerosol vertical extinction inference model, posterior intractability requires us to use an approximate posterior inference method. We propose a stochastic variational approximation, substituting the posterior inference problem with an optimisation problem. Whilst tractable, and scalable to large datasets, this approach also harms the algorithmic transparency of our model. Indeed, this approximation is known to converge to a local optimum (Hoffman et al., 2013), but no guarantee is provided on which optimum will that be, and whether this will be a global optimum. Further, there exists evidence that variational approximations tend to underestimate the variance of the true posterior (Blei et al., 2017), which also challenges the robustness of the probabilistic modelling.

An arguably more important limitation, which lies on the side of observational limitations, is that globally accessible vertically-resolved covariates are primarily meteorological variables, such as temperature, relative humidity or updraft. Meteorology mainly provides insights into aerosol extinction driven by water uptake, but does not inform on aerosol characteristics crucial for extinction, such as mass concentration, particle size, and radiative properties. Acquiring global data on these additional extinction characteristics poses a significant challenge. Hence, whilst the model we propose marks a significant improvement over simplified vertical models by capturing extinction due to water uptake, it falls short of addressing extinction from these critical aerosol properties. This represents a substantial limitation, despite our model’s demonstrated capability to infer realistic vertical structures.

6.3 Outlooks

We would like to outline some possible future directions for the work presented in this thesis.

Quantifying the impact of domain knowledge on generalisation Our work with collider regression in Chapter 3 demonstrates that incorporating domain knowledge about colliders can indeed offer generalisation benefits in regression tasks. Yet, our experiments, particularly with aerosol radiative forcing, suggest the magnitude of these benefits can vary, and sometimes yield only marginal improvements. Conversely, our experience with FaIRGP in Chapter 3 illustrates a notable improvement in model performance by the integration of domain knowledge in the form of the FaIR backbone. This raises the question of whether all forms of domain knowledge are equal in terms of the improvement in generalisation they provide. Studying this question from the angle of learning theory could allow us to determine specific aspects of the domain knowledge encoded that significantly improve generalisation in the sense of a lower excess risk. For instance, it seems plausible that imposing more restrictive constraints may narrow down the hypothesis space, enabling the model to generalise better with fewer training data.

Climate emulators for fast attribution studies An exciting application of climate emulators, such as FaIRGP, lies in their potential to perform fast attribution studies (Watson-Parris et al., 2022b). For example, with FaIRGP, we can emulate surface temperatures under historical scenarios, both with and without anthropogenic forcing, and analytically compute the probability of temperature occurrences in each scenario. By comparing these probabilities, we can assess the extent to which human activity has made a certain temperature range more likely, enabling us to conduct attribution studies. Whilst this may have limited relevance for the mean surface temperature, the development of climate emulators for extremes (Quilcaille et al., 2022, 2023) may soon allow for an efficient alternative methodology to perform attribution of extreme events.

Improving aerosol extinction inference with unmatched observations The methodology presented in our aerosol extinction inference work relies on a synthetic setup where aerosol optical depth (AOD) data is spatially aligned with meteorological predictors. Although feasible, for example by collocating satellite and reanalysis data, a more innovative approach could bypass the need for collocation of measurement

from different instruments. Leveraging the fact that both instruments have an overlap in their observation domains, methods using mediating fields or two-stage regressions (Chau et al., 2021) could allow us to infer aerosol vertical extinction without strict spatial and temporal matching of AOD measurements and meteorological predictors. Such an approach better fits the nature of remote sensing observations which are captured along the satellite’s orbit, and accommodates regions where AOD data may be compromised, for example by cloud cover, thus not wasting valuable meteorological information.

Outlook on machine learning for climate science We would like to conclude this thesis with a perspective on the contributions of machine learning to climate science. The already growing literature at the intersection of both fields indicates a great enthusiasm for what machine learning can bring to climate science research. This is most evident from the recent launch of specialised journals like *Environmental Data Science*, *Artificial Intelligence for the Earth Systems*, *Journal of Geophysical Research: Machine Learning and Computation*, dedicated sessions at both machine learning and climate science gatherings, or the growing number of submissions at intersectional conferences like Climate Informatics. Looking ahead, we anticipate the emergence of the first climate projections that include machine learning components, with promising directions like deep learning to represent subgrid processes (Rasp et al., 2018), neural GCMs (Kochkov et al., 2023) or ACE (Watt-Meyer et al., 2023), and advancements in our ability to better constrain the aerosol-cloud interaction forcing using data fusion strategies.

However, for the advances promised by machine learning to withstand the scientific process and lead to real world impact, a consensus from the climate science community should emerge and rigorously examine their trustworthiness. In this thesis, we propose guidelines for developing machine learning models in climate science that could facilitate such discussions by offering transparency, trust, and nuanced conclusions. The contributed chapters further demonstrate how to design machine learning models that adhere to these guidelines across different problems in climate science. We hope this contribution will guide machine learning practitioners eager to make a difference in climate science towards creating useful and reliable solutions for the climate science community.

Bibliography

Simon L Lewis and Mark A Maslin. Defining the anthropocene. *Nature*, 519(7542): 171–180, 2015. 1

Jason W Moore. The Capitalocene, Part I: on the nature and origins of our ecological crisis. *The Journal of peasant studies*, 44(3):594–630, 2017. 1

Françoise Vergès. *Racial Capitalocene. Futures of Black radicalism*. Verso Books, 2017. 1

Veronika Eyring, NP Gillett, Krishna Achutarao, Rondrotiana Barimalala, Marcelo Barreiro Parrillo, Nicolas Bellouin, Christophe Cassou, Paul Durack, Yu Kosaka, Shayne McGregor, et al. Human influence on the climate system. in climate change 2021: the physical science basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. *IPCC Sixth Assessment Report*, 2021a. 1

World Meteorological Organization. WMO confirms that 2023 smashes global temperature record [Press release]. 2024. Published: 2024-01-12; Accessed: 2024-03-18. 1

Copernicus. Copernicus: In 2024, the world experienced the warmest January on record [Press release]. 2024. Published: 2024-02-08; Accessed: 2024-03-18. 2

Hans O Pörtner, Debra C Roberts, Helen Adams, Carolina Adler, Paulina Aldunce, Elham Ali, Rawshan Ara Begum, Richard Betts, Rachel Bezner Kerr, Robbert Biesbroek, et al. Climate change 2022: Impacts, adaptation and vulnerability. Technical report, IPCC Sixth Assessment Report, 2022. 2

Sonia I Seneviratne, Xuebin Zhang, Muhammad Adnan, Wafae Badi, Claudine Dereczynski, Alejandro Di Luca, Subimal Ghosh, I Iskander, James Kossin, Sophie Lewis, et al. Weather and climate extreme events in a changing climate (Chapter 11). *IPCC Sixth Assessment Report*, 2021. 2

- Joern Birkmann, Emma Liwenga, Rajiv Pandey, Emily Boyd, Riyanti Djalante, François Gemenne, Walter Leal Filho, Patricia Pinho, Lindsay Stringer, and David Wrathall. Poverty, livelihoods and sustainable development. *IPCC Sixth Assessment Report*, 2022. 2
- Geoffrey K Vallis. *Atmospheric and oceanic fluid dynamics*. Cambridge University Press, 2017. 2
- John H Seinfeld and Spyros N Pandis. *Atmospheric chemistry and physics: from air pollution to climate change*. John Wiley & Sons, 2016. 2
- Dennis L Hartmann. *Global physical climatology*, volume 103. Newnes, 2015. 2
- Norman A Phillips. The general circulation of the atmosphere: A numerical experiment. *Quarterly Journal of the Royal Meteorological Society*, 82(352):123–164, 1956. 2
- Edward Lorenz. The nature and theory of the general circulation of the atmosphere. *World meteorological organization*, 161, 1967. 2
- Marko Scholze, J. Icarus Allen, William J. Collins, Sarah E. Cornell, Chris Huntingford, Manoj M. Joshi, Jason A. Lowe, Robin S. Smith, and Oliver Wild. *Earth system models*. Cambridge University Press, 2012. 2
- Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 2016. 2, 3
- Deliang Chen, Maisa Rojas, Bjørn H Samset, Kim Cobb, Aida Diongue Niang, Paul Edwards, Seita Emori, Sergio Henrique Faria, Edward Hawkins, Pandora Hope, et al. Framing, Context, and Methods (Chapter 1). *IPCC Sixth Assessment Report*, 2021. 2
- V Balaji, Fleur Couvreur, Julie Deshayes, Jacques Gautrais, Frédéric Hourdin, and Catherine Rio. Are general circulation models obsolete? *Proceedings of the National Academy of Sciences*, 119(47):e2202075119, 2022. 2
- Mario C Acosta, Sergi Palomas, Stella V Paronuzzi Ticco, Jean-Claude Andre, Joachim Biercamp, Pierre-Antoine Bretonniere, Reinhard Budich, Miguel Castrillo, Arnaud Caubel, Francisco Doblas-Reyes, et al. The computational and energy cost of

- simulation and storage for climate science: lessons from CMIP6. *Geoscientific Model Development Discussions*, 2023:1–21, 2023. 3
- Gokhan Danabasoglu, J-F Lamarque, J Bacmeister, DA Bailey, AK DuVivier, Jim Edwards, LK Emmons, John Fasullo, R Garcia, Andrew Gettelman, et al. The community earth system model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001916, 2020. 3
- Ayesha Tandon. Analysis: The lack of diversity in climate-science research. *Carbon Brief*, 2021. 3
- Ed Hawkins and Rowan Sutton. The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, 90(8):1095–1108, 2009. 3
- Lea Beusch, Lukas Gudmundsson, and Sonia I Seneviratne. Emulating Earth system model temperatures with MESMER: from global mean temperature trajectories to grid-point-level realizations on land. *Earth System Dynamics*, 2020. 3, 5
- Veronika Eyring, Vimal Mishra, Gary P Griffith, Lei Chen, Trevor Keenan, Merritt R Turetsky, Sally Brown, Frank Jotzo, Frances C Moore, and Sander Van der Linden. Reflections and projections on a decade of climate science. *Nature Climate Change*, 11(4):279–285, 2021b. 3
- Tim Palmer. Climate forecasting: Build high-resolution global climate models. *Nature*, 515(7527):338–339, 2014. 3
- Christian Jakob. Accelerating progress in global atmospheric model development through improved parameterizations: Challenges, opportunities, and strategies. *Bulletin of the American Meteorological Society*, 91(7):869–876, 2010. 3
- Steven C Sherwood, Sandrine Bony, and Jean-Louis Dufresne. Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature*, 505(7481):37–42, 2014. 3
- Peter M Caldwell, Mark D Zelinka, Karl E Taylor, and Kate Marvel. Quantifying the sources of intermodel spread in equilibrium climate sensitivity. *Journal of Climate*, 29(2):513–524, 2016. 3

- CR Terai, MS Pritchard, P Blossey, and CS Bretherton. The impact of resolving sub-kilometer processes on aerosol-cloud interactions of low-level clouds in global model simulations. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002274, 2020. 3
- Yan Ma, Haiping Wu, Lizhe Wang, Bormin Huang, Rajiv Ranjan, Albert Zomaya, and Wei Jie. Remote sensing big data computing: Challenges and opportunities. *Future Generation Computer Systems*, 51:47–60, 2015. 4
- Athos Agapiou. Remote sensing heritage in a petabyte-scale: Satellite data and heritage Earth Engine© applications. *International Journal of Digital Earth*, 10(1):85–102, 2017. 4
- Charlie J Tomlinson, Lee Chapman, John E Thornes, and Christopher Baker. Remote sensing land surface temperature for meteorology and climatology: A review. *Meteorological Applications*, 18(3):296–306, 2011. 4
- Marco Tedesco. Remote sensing and the cryosphere. *Remote Sensing of the Cryosphere*, pages 1–16, 2015. 4
- Jun Yang, Peng Gong, Rong Fu, Minghua Zhang, Jingming Chen, Shunlin Liang, Bing Xu, Jiancheng Shi, and Robert Dickinson. The role of satellite remote sensing in climate change studies. *Nature climate change*, 3(10):875–883, 2013. 4
- Daniel P Grosvenor, Odran Sourdeval, Paquita Zuidema, Andrew Ackerman, Mikhail D Alexandrov, Ralf Bennartz, Reinout Boers, Brian Cairns, J Christine Chiu, Matthew Christensen, et al. Remote sensing of droplet number concentration in warm clouds: A review of the current state of knowledge and perspectives. *Reviews of Geophysics*, 56(2):409–453, 2018. 4
- S. Twomey. Pollution and the planetary albedo. *Atmospheric Environment*, 8(12):1251–1256, January 1974. doi: 10.1016/0004-6981(74)90004-3. 4
- Bruce A. Albrecht. Aerosols, Cloud Microphysics, and Fractional Cloudiness. *Science*, 1989. 4
- Nicolas Bellouin, Johannes Quaas, Edward Gryspeerdt, Stefan Kinne, Philip Stier, Duncan Watson-Parris, Olivier Boucher, Ken S Carslaw, Matt Christensen, A-L Daniau, et al. Bounding global aerosol radiative forcing of climate change. *Reviews of Geophysics*, 58(1):e2019RG000660, 2020. 4

- Piers Forster, Trude Storelvmo, Kyle Armour, William Collins, Jean-Louis Dufresne, David Frame, Dan Lunt, Thorsten Mauritsen, Matthew Palmer, Masahiro Watanabe, et al. The earth’s energy budget, climate feedbacks, and climate sensitivity. *IPCC Sixth Assessment Report*, 2021. 4
- Teruyuki Nakajima, Akiko Higurashi, Kazuaki Kawamoto, and Joyce E. Penner. A possible correlation between satellite-derived cloud and aerosol microphysical parameters. *Geophysical Research Letters*, 2001. 4
- Philip Stier. Limitations of passive remote sensing to constrain global cloud condensation nuclei. *Atmospheric Chemistry and Physics*, 2016. 4
- D. Painemal, F.-L. Chang, R. Ferrare, S. Burton, Z. Li, W. L. Smith Jr, P. Minnis, Y. Feng, and M. Clayton. Reducing uncertainties in satellite estimates of aerosol–cloud interactions over the subtropical ocean by integrating vertically resolved aerosol observations. *Atmospheric chemistry and physics*, 2020. 4
- Gustau Camps-Valls, Jochem Verrelst, Jordi Muñoz-Marí, Valero Laparra, Fernando Mateo-Jimenez, and Jose L. Gómez-Dans. A Survey on Gaussian Processes for Earth-Observation Data Analysis: A Comprehensive Investigation. *IEEE Geoscience and Remote Sensing Magazine*, 2016. 4
- Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and fnm Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019. 4, 10
- Christopher Irrgang, Niklas Boers, Maike Sonnewald, Elizabeth A Barnes, Christopher Kadow, Joanna Staneva, and Jan Saynisch-Wagner. Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nature Machine Intelligence*, 3(8):667–674, 2021. 4
- Tom Beucler, Imme Ebert-Uphoff, Stephan Rasp, Michael Pritchard, and Pierre Gentine. Machine learning for clouds and climate. *Clouds and their climatic impacts: Radiation, circulation, and precipitation*, pages 325–345, 2023. 4, 10
- Duncan Watson-Parris, Yuhan Rao, Dirk Olivieé, Øyvind Seland, Peer Nowack, Gustau Camps-Valls, Philip Stier, Shahine Bouabid, Maura Dewey, Emilie Fons, et al. ClimateBench v1. 0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 14(10):e2021MS002954, 2022a. 5, 11

- Catherine O de Burgh-Day and Tennessee Leeuwenburg. Machine learning for numerical weather and climate modelling: a review. *Geoscientific Model Development*, 16(22):6433–6477, 2023. 5
- Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022. 5
- Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical fourier neural operators: Learning stable dynamics on the sphere. In *International conference on machine learning*, pages 2806–2823. PMLR, 2023. 5
- Suyash Bire, Björn Lütjens, Kamyar Azizzadenesheli, Anima Anandkumar, and Christopher N Hill. Ocean emulation with fourier neural operators: Double gyre. *Authorea Preprints*, 2023. 5
- Stephan Rasp, Michael S Pritchard, and Pierre Gentine. Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115(39):9684–9689, 2018. 5, 171
- Janni Yuval and Paul A O’Gorman. Stable machine-learning parameterization of sub-grid processes for climate modeling at a range of resolutions. *Nature communications*, 11(1):3295, 2020. 5
- Pierre Gentine, Veronika Eyring, and Tom Beucler. Deep learning for the parametrization of subgrid processes in climate models. *Deep learning for the Earth sciences: A comprehensive approach to remote sensing, climate science, and geosciences*, pages 307–314, 2021. 5
- Oliver Watt-Meyer, Noah D Brenowitz, Spencer K Clark, Brian Henn, Anna Kwa, Jeremy McGibbon, W Andre Perkins, Lucas Harris, and Christopher S Bretherton. Neural network parameterization of subgrid-scale physics from a realistic geography global storm-resolving simulation. *Journal of Advances in Modeling Earth Systems*, 16(2):e2023MS003668, 2024. 5

- Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, James Lottes, Stephan Rasp, Peter Düben, Milan Klöwer, et al. Neural general circulation models. *arXiv preprint arXiv:2311.07222*, 2023. 5, 171
- Laure Zanna and Thomas Bolton. Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, 47(17):e2020GL088376, 2020. 5
- Gustau Camps-Valls, Andreas Gerhardus, Urmi Ninad, Gherardo Varando, Georg Martius, Emili Balaguer-Ballester, Ricardo Vinuesa, Emiliano Diaz, Laure Zanna, and Jakob Runge. Discovering causal relations and equations from data. *Physics Reports*, 1044:1–68, 2023. 5
- Arthur Grundner, Tom Beucler, Pierre Gentine, and Veronika Eyring. Data-driven equation discovery of a cloud cover parameterization. *Journal of Advances in Modeling Earth Systems*, 16(3):e2023MS003763, 2024. 5
- Edward Gryspeerdt, Tom Goren, Odran Sourdeval, Johannes Quaas, Johannes Mülmenstädt, Sudhakar Dipu, Claudia Unglaub, Andrew Gettelman, and Matthew Christensen. Constraining the aerosol influence on cloud liquid water path. *Atmospheric Chemistry and Physics*, 19(8):5331–5347, 2019. 5
- Daniel T McCoy, Paul Field, Hamish Gordon, Gregory S Elsaesser, and Daniel P Grosvenor. Untangling causality in midlatitude aerosol–cloud adjustments. *Atmospheric Chemistry and Physics*, 20(7):4085–4103, 2020. 5
- Emilie Fons, Jakob Runge, David Neubauer, and Ulrike Lohmann. Stratocumulus adjustments to aerosol perturbations disentangled with a causal approach. *npj Climate and Atmospheric Science*, 6(1):130, 2023. 5
- Marlene Kretschmer, Dim Coumou, Jonathan F Donges, and Jakob Runge. Using causal effect networks to analyze different Arctic drivers of midlatitude winter circulation. *Journal of climate*, 29(11):4069–4081, 2016. 5
- Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019a. 5
- Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in Earth system sciences. *Nature communications*, 10(1):2553, 2019b. 5

- Philine Bommer, Marlene Kretschmer, Anna Hedström, Dilyara Bareeva, and Marina M-C Höhne. Finding the right XAI method—A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science. *arXiv preprint arXiv:2303.00652*, 2023. 6
- Kai Jeggle, David Neubauer, Gustau Camps-Valls, and Ulrike Lohmann. Understanding cirrus clouds using explainable machine learning. *Environmental Data Science*, 2:e19, 2023. 6
- Fredrik Gustafsson. *Statistical sensor fusion*. Studentlitteratur, 2010. 6
- Marina Cocchi. *Data fusion methodology and applications*. Elsevier, 2019. 6
- Tong Meng, Xuyang Jing, Zheng Yan, and Witold Pedrycz. A survey on machine learning for data fusion. *Information Fusion*, 57:115–129, 2020. 6
- Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, and Jocelyn Chanussot. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112:102926, 2022. 6
- Feng Gao, Thomas Hilker, Xiaolin Zhu, Martha Anderson, Jeffrey Masek, Peijuan Wang, and Yun Yang. Fusing Landsat and MODIS data for vegetation monitoring. *IEEE Geoscience and Remote Sensing Magazine*, 3(3):47–60, 2015. 6
- Michel Deudon, Alfredo Kalaitzis, Israel Goytom, Md Rifat Arefin, Zhichao Lin, Kris Sankaran, Vincent Michalski, Samira E Kahou, Julien Cornebise, and Yoshua Bengio. Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery. *arXiv preprint arXiv:2002.06460*, 2020. 6
- D James Fulton, Ben J Clarke, and Gabriele C Hegerl. Bias correcting climate model simulations using unpaired image-to-image translation networks. *Artificial Intelligence for the Earth Systems*, 2(2):e220031, 2023. 6
- Antti Arola, Antti Lipponen, Pekka Kolmonen, Timo H Virtanen, Nicolas Bellouin, Daniel P Grosvenor, Edward Gryspeerdt, Johannes Quaas, and Harri Kokkola. Aerosol effects on clouds are concealed by natural cloud heterogeneity and satellite retrieval errors. *Nature Communications*, 13(1):7357, 2022. 6
- Matthew Hutson. Artificial intelligence faces reproducibility crisis, 2018. 6

- Zachary C Lipton and Jacob Steinhardt. Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17(1):45–77, 2019. 6
- Marlene Kretschmer, Samantha V Adams, Alberto Arribas, Rachel Prudden, Niall Robinson, Elena Saggioro, and Theodore G Shepherd. Quantifying causal pathways of teleconnections. *Bulletin of the American Meteorological Society*, pages 1–34, 2021. 8
- Chris K Folland, Jeff Knight, Hans W Linderholm, David Fereday, Sarah Ineson, and James W Hurrell. The summer North Atlantic Oscillation: past, present, and future. *Journal of Climate*, 22(5):1082–1103, 2009. 8
- Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018. 8
- Kevin P Murphy. *Probabilistic machine learning: Advanced topics*. MIT press, 2023. 9
- Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. Integrating physics-based modeling with machine learning: A survey. *arXiv preprint arXiv:2003.04919*, 1(1):1–34, 2020. 10
- Siu Lun Chau, Shahine Bouabid, and Dino Sejdinovic. Deconditional downscaling with Gaussian processes. *Advances in Neural Information Processing Systems*, 34: 17813–17825, 2021. 11, 171
- Anurag Singh, Siu Lun Chau, Shahine Bouabid, and Krikamol Muandet. Domain generalisation via imprecise learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 45544–45570. PMLR, 21–27 Jul 2024. 11
- Shahine Bouabid, Jake Fawkes, and Dino Sejdinovic. Returning the favour: when regression benefits from probabilistic causal knowledge. In *International Conference on Machine Learning*, pages 2885–2913. PMLR, 2023. 11
- Shahine Bouabid, Dino Sejdinovic, and Duncan Watson-Parris. Fairgp: A bayesian energy balance model for surface temperatures emulation. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS003926, 2024a. 12

- Shahine Bouabid, Duncan Watson-Parris, Sofija Stefanović, Athanasios Nenes, and Dino Sejdinovic. Aerosol optical depth disaggregation: toward global aerosol vertical profiles. *Environmental Data Science*, 3:e16, 2024b. 12
- Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998. 16
- Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004. 16
- Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *International conference on algorithmic learning theory*, pages 13–31. Springer, 2007. 16
- Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017. 16
- Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006. 16
- Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The annals of statistics*, pages 2263–2291, 2013. 16
- Siu Lun Chau. *Towards trustworthy machine learning with kernels*. PhD thesis, University of Oxford, 2023. 16
- Jean-Francois Ton. *Causal reasoning and meta learning using kernel mean embeddings*. PhD thesis, University of Oxford, 2022. 16
- Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. *Advances in Neural Information Processing Systems*, 32, 2019. 16
- Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002. 16, 22

- Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008a. 16
- Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011. 16
- Vern I Paulsen and Mrinal Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*, volume 152. Cambridge university press, 2016. 18
- Laurent Schwartz. Sous-espaces hilbertiens d’espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *Journal d’analyse mathématique*, 13:115–256, 1964. 19
- Julien Mairal and Jean-Philippe Vert. Machine learning with kernel methods. *Lecture Notes, January*, 10:52, 2018. 19, 24
- Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950. 20
- Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999. 21
- David Duvenaud, James Lloyd, Roger Grosse, Joshua Tenenbaum, and Ghahramani Zoubin. Structure discovery in nonparametric regression through compositional kernel search. In *International Conference on Machine Learning*, pages 1166–1174. PMLR, 2013. 22
- Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 2011. 22, 23
- Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 2008b. 23
- Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018. 25
- C Rasmussen and C Williams. *Gaussian Processes for Machine Learning*, 2005. 25, 27
- Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016. 29

- Bryn Elesedy. Provably strict generalisation benefit for invariance in kernel methods. *Advances in Neural Information Processing Systems*, 34:17273–17283, 2021. 29
- Simo Särkkä. Linear operators and stochastic partial differential equations in Gaussian process regression. In *Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14–17, 2011, Proceedings, Part II 21*, pages 151–158. Springer, 2011. 30
- Christian Agrell. Gaussian processes with linear operator inequality constraints. *Journal of Machine Learning Research*, 20(135):1–36, 2019. 30
- Felix R Day, Po-Ru Loh, Robert A Scott, Ken K Ong, and John RB Perry. A robust example of collider bias in a genetic association study. *The American Journal of Human Genetics*, 98(2):392–393, 2016. 168
- Jakob Runge. Necessary and sufficient graphical conditions for optimal adjustment sets in causal graphical models with hidden variables. *Advances in Neural Information Processing Systems*, 34:15762–15773, 2021. 168
- Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3155–3164. PMLR, 2020. 168
- Andrew Jesson, Alyson Douglas, Peter Manshausen, Maëlys Solal, Nicolai Meinshausen, Philip Stier, Yarin Gal, and Uri Shalit. Scalable sensitivity and uncertainty analyses for causal-effect estimates of continuous-valued interventions. *Advances in Neural Information Processing Systems*, 35:13892–13907, 2022. 168
- Wilhelm May. Assessing the strength of regional changes in near-surface climate associated with a global warming of 2 C. *Climatic change*, 2012. 168
- Hiram Levy, Larry W Horowitz, M Daniel Schwarzkopf, Yi Ming, Jean-Christophe Golaz, Vaishali Naik, and V Ramaswamy. The roles of aerosol direct and indirect effects in past and future climate change. *Journal of Geophysical Research: Atmospheres*, 2013. 168
- Claudia Tebaldi and Reto Knutti. Evaluating the accuracy of climate change pattern emulation for low warming targets. *Environmental Research Letters*, 13(5):055006, 2018. 168

- C Tebaldi, A Armbruster, HP Engler, and R Link. Emulating climate extreme indices. *Environmental Research Letters*, 15(7):074006, 2020. 168
- N. J. Leach, S. Jenkins, Z. Nicholls, C. J. Smith, J. Lynch, M. Cain, T. Walsh, B. Wu, J. Tsutsui, and M. R. Allen. FaIRv2.0.0: a generalized impulse response model for climate uncertainty and future scenario exploration. *Geoscientific Model Development*, 2021. doi: 10.5194/gmd-14-3007-2021. 169
- Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013. 169
- David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017. 169
- Duncan Watson-Parris, Yuhan Rao, Dirk Olivié, Øyvind Seland, Peer Nowack, Gustau Camps-Valls, Philip Stier, Shahine Bouabid, Maura Dewey, Emilie Fons, et al. ClimateBench v1. 0: A Benchmark for Data-Driven Climate Projections. *Journal of Advances in Modeling Earth Systems*, 14(10):e2021MS002954, 2022b. 170
- Yann Quilcaille, Lukas Gudmundsson, Lea Beusch, Mathias Hauser, and Sonia I Seneviratne. Showcasing MESMER-X: Spatially Resolved Emulation of Annual Maximum Temperatures of Earth System Models. *Geophysical Research Letters*, 49(17):e2022GL099012, 2022. 170
- Yann Quilcaille, Lukas Gudmundsson, and Sonia I Seneviratne. Extending MESMER-X: A spatially resolved Earth system model emulator for fire weather and soil moisture. *Earth System Dynamics*, 14(6):1333–1362, 2023. 170
- Oliver Watt-Meyer, Gideon Dresdner, Jeremy McGibbon, Spencer K Clark, Brian Henn, James Duncan, Noah D Brenowitz, Karthik Kashinath, Michael S Pritchard, Boris Bonev, et al. ACE: A fast, skillful learned global atmospheric model for climate prediction. *arXiv preprint arXiv:2310.02074*, 2023. 171