Machine Learning Approaches for Estimating the Causal Effects of Aerosols on Clouds

Candidate no. 1060314

Word count: 16493

Submitted in partial completion of the

Master of Science

August 2022

Acknowledgements

This endeavour would not have been possible without my supervisors Dr Yarin Gal, Dr Alyson Douglas and Andrew Jesson. I am beyond grateful to them for offering me this opportunity to work with them at OATML and AOPP, and would like to thank them for their great guidance throughout the project. It was a true pleasure to work with them, especially when we were able to meet in person, either in Oxford or London.

I am also grateful to my supervisor Dr Bartek Klin, my mentor Dr Piotr Mirowski, and my advisor Dr Quentin Miller who supported and encouraged me throughout my entire year at Oxford. Our discussions were invaluable, helped me navigate the challenges I faced, and inspired me to grow.

Lastly, I would like to thank my family and friends for their unwavering support and encouragement.

Abstract

Aerosol cloud interactions (ACI) include various effects that result from aerosols entering a cloud, acting as cloud condensation nuclei (CCN) and affecting cloud properties. In general, an increase in aerosol concentration results in smaller droplet sizes which leads to larger, brighter, longer-lasting clouds that reflect more sunlight and contribute to cooling the earth. The strength of the effect is however heterogeneous over meteorological regimes, making ACI the most uncertain driver of radiative forcing due to human activities, and the largest source of uncertainty in our current climate models. In our work, we estimate ACI from observational data through the potential outcomes approach to causal inference. Based on [26], we use machine learning approaches to estimate plausible ranges for the causal effects of aerosols on clouds. Specifically, using the proposed method and models, we look at satellite data from different regions, resolutions, and timescales to study how different levels of confounding affect uncertainty bounds. To a larger extent, our work contributes to understanding the climatological impacts of human emissions on cloud properties. We highlight the importance of uncertainty and assumptions to correctly assess interventions that aim to reduce global warming like geoengineering.

Contents

List of Figures

List of Tables

List of Abbreviations

N_d . . . Cloud droplet number.

r_e . . . . Mean cloud droplet size.

τ . . . . Cloud optical depth.

ω500 . . Vertical motion at 500 millibar.

ACI . . . Aerosol-cloud interactions.

ANN . . Artificial neural network.

AOD . . Aerosol optical depth.

APO . . Average potential outcome.

ARI . . Aerosol-radiation interactions.

CAPO . Conditional average potential outcome.

CCN . . Cloud condensation nucleus.

CF . . . Cloud fraction.

CMSM . Continuous treatment-effect marginal sensitivity model.

CWP . . Cloud water path.

EIS . . . Effective inversion strength.

GMM . Gaussian mixture model.

i.i.d. . . Independent and identically distributed (random variables).

LTS . . . Lower tropospheric stability.

MAB . . Multi-head attention block.

MLP . . Multi-layer perceptron.

ResNet . Residual network.

RHx . . Relative humidity at x millibar.

SST . . . Sea surface temperature.

1 Introduction

1.1 Motivation

Climate change is one of the major challenges of our time. The impacts are global and unprecedented, requiring immediate and drastic actions to limit the number of irreversible changes. Scientists use climate models to understand future projections due to climate change and attribute shifts to anthropogenic or natural sources. These models also allow to test different carbon emissions scenarios and help decision makers find appropriate policies to reduce global warming. However, climate model predictions come with uncertainties that arise from being unable to explicitly model small-scale interactions, such as aerosol-cloud interactions (ACI) [7, 32].

Aerosol is a suspension system of fine liquid or solid particles usually non-uniformly distributed in a gas (commonly air). There are different types of aerosols coming from either natural or anthropogenic sources. Examples of aerosols from natural sources include dust, sand, volcanic ash, and sea salt. Examples of aerosols from anthropogenic sources include particulate air pollutants, smoke, and sprayed pesticides. The overall effect of aerosols on climate is cooling, directly or indirectly. By reflecting incoming solar radiation, and thus reducing the amount of sunlight reaching the surface of the planet, aerosols can have a direct cooling effect. Aerosols also serve as cloud condensation nuclei (CCN) allowing clouds to form and altering

their properties like the size of their droplets, their precipitation efficacy, or their lifetime [52]. This causes larger, brighter, longer-lasting clouds, contributing to aerosols' indirect cooling effect on climate. These "aerosol-cloud interactions" are the focus of our work.

Different types of aerosols have different effects on clouds. Tthe magnitude and sign of these effects can vary under different environmental conditions [17]. For instance, aerosols can act to either decrease or increase the size of water droplets in a cloud [51]. To understand ACI, it is therefore crucial to identify sources of heterogeneity.

Observing ACI using satellite, as we do, leads to a confounding problem because aerosols are impossible to measure directly. The present work relies on a proxy, aerosol optical depth (AOD), which is affected by nearby clouds and the local environment [12]. Further uncertainty arises from working with finite data. Overall, from a causal point of view, the environment acts as a confounder on ACI, which is itself a heterogeneous effect due to its modulation by the environment. This motivates our investigation of the effects of aerosols on cloud properties, whilst accounting for different environmental factors. We are especially interested in estimating the uncertainty bounds of the treatment effect of aerosols on cloud properties for different levels of confounding.

Estimating the uncertainty related to ACI is crucial for decision-making and scientific understanding. It is thought that the cooling effect of ACI counteracts global warming, but the uncertainties on the sign and magnitude of this effect are very large and justify contradicting hypotheses. Current climate models fail to emulate ACI and have uncertainty bounds that could offset global warming completely or double the effects of rising carbon dioxide [7]. It is therefore of utmost importance that we improve our understanding of these interactions to reduce their underlying uncertainties.

1.2 Objectives

Our work uses machine learning approaches to estimate plausible ranges for the causal effects of aerosols on clouds and derive uncertainty bounds. We report ranges rather than point estimates because of the uncertainties arising from unobserved confounding and working with finite data. We base our research on [26], which we refer to as "Overcast". There, the authors propose a method and models to estimate continuous treatment effects. They develop a statistical uncertainty-aware sensitivity model, the continuous treatment-effect marginal sensitivity model (CMSM), based on the potential outcomes approach to causal inference, and relying on a feed-forward neural network and a transformer. They evaluate their methods on a synthetic dataset and on real-world data from satellite observations to estimate ACI.

The objective of this project is to further investigate ACI and their uncertainties using the Overcast models. From a causal point of view, we aim to understand how unmeasured confounding can change treatment-effect estimates. This can be summarised as follows. First, we evaluate the method and models proposed in Overcast by implementing baselines. Second, we look into geographical dependencies of aerosols by studying two geographical regions (the South-East Pacific and the South Atlantic) and two different levels of data resolution. Third, we perform an uncertainty-aware causal sensitivity analysis to study how unmodelled confounding variables can influence the range of plausible treatment effects for a given dataset.

1.3 Contributions

Our contributions are threefold and relate to the objectives outlined previously.

First, we find that whilst the predictor for cloud properties proposed by Overcast is weak, it agrees with off-the-shelf regression models. Furthermore, we identify that the Overcast transformer performs better than the feed-forward neural network in terms of predictive power and estimates of treatment effects but has larger ignorance regions.

Second, we observe that Overcast models emulate ACI better in the Atlantic than in the Pacific with low-resolution data, but worse with higher spatio-temporal resolution data in the Pacific region. This work allows us to investigate how varying levels of hidden confounding arising from geographical dependencies generate different plausible ranges of treatment effects.

Third, we study unmeasured confounding across different datasets and perform an uncertainty-aware sensitivity analysis. We find that omitting covariates reduces uncertainty, as expected, but also yields less accurate dose-response curves. This work is an extension to the original Overcast article in that we propose a methodology for setting important parameters introduced in the paper, such as Λ which defines the possible spread of outcomes. Our study also shows the importance of accounting for violations of causal assumptions to derive realistic uncertainty bounds.

To a larger extent, our work contributes to highlighting the importance of uncertainty when studying climate projection models to take appropriate measures.

1.4 Outline

This text contains eight chapters. The current chapter, Chapter 1 presents an introduction to this project, highlighting our motivations, our objectives and our contributions, and outlines the thesis structure. Chapter 2 provides background on the topics of aerosol-cloud interactions, causal inference, and relevant topics in machine learning. Chapter 3 formalises the problem setting. Chapter 4 presents our methods, describing the datasets used, the Overcast models, and the experiments performed. Chapters 5 through 7 present and analyse our results. Chapter 5 refers to off-the-shelf regression baselines and compares the performance of the two Overcast models. Chapter 6 investigates geographical dependencies by presenting our work on datasets from different geographical regions and resolutions. Chapter 7 exposes our uncertainty-aware sensitivity analysis where we further investigate unobserved confounding. The text ends with our conclusions in Chapter 8, summarising our research, considering its implications and limitations, and reflecting on potential future works.

2
Background

In this chapter, the background necessary for understanding our work is provided. Section 2.1 describes aerosols, clouds, and their interactions (ACI), as well as their importance for climate modelling. Section 2.2 concerns causal inference, explaining causality and the potential outcomes framework using formal statistical terms. Section 2.3 explains core concepts of machine learning, specifically, the basics of regression, artificial neural networks, attention mechanisms and transformer.

2.1 Aerosols, clouds and their interactions

2.1.1 Aerosols

Aerosol is a suspension system of fine liquid or solid particles usually non-uniformly distributed in a gas (usually air). Aerosols come from various sources, with approximately 90% of the total aerosol mass from natural origins, and the remaining 10% from anthropogenic sources [8]. Examples of aerosols from natural sources include dust, sand, volcanic ash, and sea salt. Examples of aerosols from anthropogenic sources include particulate air pollutants, smoke, and sprayed pesticides. The key aerosol groups are sulphates, organic carbon, black carbon, nitrates, mineral dust, and sea salt. Since aerosols often clump together to form complex mixtures, it is

difficult to trace back their origins. They are therefore described based on their shape, size, chemical composition, and other properties.

Aerosols can be measured by satellite and aircraft over both water and land, and by ground-based instruments. Due to the complexity of their composition, multiple properties are required for a full characterisation, making them difficult to observe. However, light measurements taken by radiometers serve us as good indicators of aerosol levels. The single most comprehensive variable for the remote assessment of the aerosol load in the atmosphere is the aerosol optical depth (AOD), also known as the aerosol optical thickness. It is a one-dimensional measure of the amount of light that aerosols scatter and absorb in the atmosphere, with low values of AOD indicating a clear atmosphere with quite high visibility, and higher values representing higher concentrations of aerosols.

Aerosols impact climate by typically working in opposition to greenhouse gases, exerting an overall cooling influence on the Earth both directly through their interactions with radiation (ARI) and indirectly through their interactions with clouds (ACI). These interactions are represented in Figure 2.2, with ARI on the left, and ACI in the centre and on the right. Aerosol-radiation interactions (ARI) stem from direct scattering or absorption of solar and terrestrial radiations by aerosols [52]. The precise effect on light depends on aerosols' properties such as their composition, the colour of their particles, and environmental conditions. For instance, pure sulphates and nitrates reflect nearly all radiation they encounter, leading to cooling, whereas black carbon absorbs radiation and therefore warms the atmosphere. Overall, the effect of black carbon is overwhelmed by the effect of sulphates and nitrates, we thus say that aerosols have an overall cooling influence. The indirect cooling effect of aerosols stems from aerosol-cloud interactions (ACI) and is the focus of our project. We describe these further in the following subsections.

2.1.2 Clouds

A cloud is a visible mass of water droplets or ice crystals floating in the sky. There are diverse types of clouds, categorised based on their location in the sky and

their shape, which are partly determined by atmospheric conditions like pressure, temperature, and winds.

Clouds form when water vapour condenses to liquid form in the sky. For clouds to form, the air must not only be cool enough for water to condense but also contain enough aerosols. Aerosols provide the non-gaseous surface required for water to transition into the liquid state, creating a water droplet. We say that aerosol particles acting as cloud seeds are activated as cloud condensation nuclei (CCN). The activation itself depends on environmental properties like the level of supersaturation within a cloud, and aerosol properties like size, shape, and hygroscopicity, which is the capacity of a particle to attract moisture from the air. Precipitation occurs when cloud droplets become larger because gravity causes these droplets to fall through the air faster. The process of cloud droplet formation is called cloud micro-physics.

To study cloud micro- and macro-physics, multiple measurements are used, as represented in Figure 2.1 and summarised in Table 2.1. The diagrams in the top row of this figure represent high values for each variable whereas the bottom row represents low values. The cloud droplet concentration N_d is an indicator of the total number of droplets present in any given volume of air. The mean cloud droplet radius r_e measures the size of cloud droplets. The cloud optical thickness, or cloud optical depth, τ measures the amount of light that a cloud prevents from passing through it. It is linked to the brightness of a cloud. The cloud water path CWP is the total amount of liquid water droplets in the atmosphere above a unit surface area on the earth. The cloud fraction CF is the portion of each pixel of the sky that is covered by clouds. It is an indicator of cloud coverage. The properties are ordered according to the chain of causation both in the enumeration above, the table and the figure. Namely, the number of cloud droplets affects their size which affects a cloud's optical thickness as well as the amount of water in the atmosphere and the cloud coverage in the sky.

Cloud property	Notation	Description
Cloud droplet number	N_d	Number of water droplets in a given volume of air
Cloud droplet radius	r_e	Mean radius of cloud droplets
Cloud optical depth	τ	Amount of light that a cloud prevents from passing through it
Cloud water path	CWP	Amount of liquid water droplets in the atmosphere above a unit surface area on the earth
Cloud fraction	CF	Portion of each pixel of the sky that is covered by clouds

Table 2.1: Cloud properties: notations and descriptions

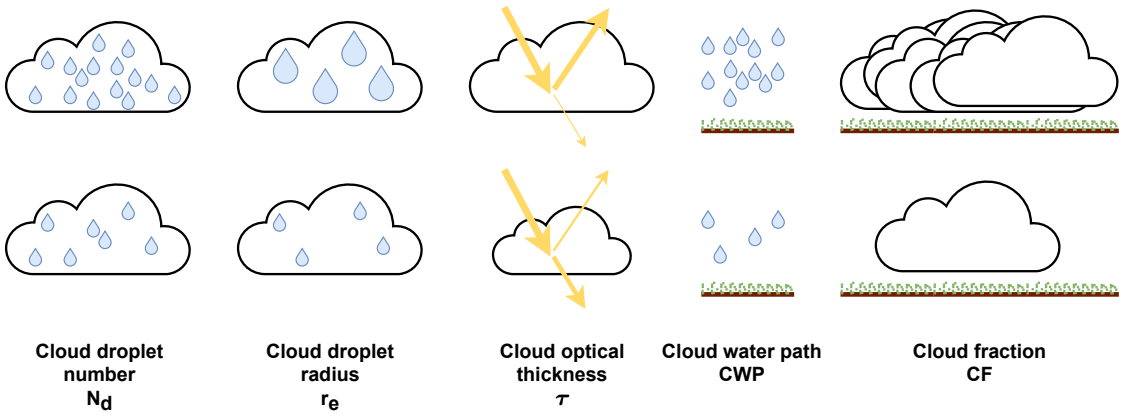


Figure 2.1: Cloud properties. Ordered according to the chain of causation: cloud droplet concentration N_d, cloud droplet radius r_e, cloud optical thickness τ, cloud water path CWP, and cloud fraction CF. The top row represents larger values of these variables. The bottom row represents lower values of these variables.

2.1.3 Aerosol-cloud interactions

Aerosols influence clouds' micro-physical and macro-physical properties in multiple ways, englobed in the term aerosol-cloud interactions (ACI). In this subsection, two effects of aerosols on cloud micro-physics are described: the Twomey effect on cloud reflectivity, and the Albrecht effect on cloud lifetime and coverage. The caveats of these theories and the consequent difficulty to model ACI are then developed.

The Twomey and Albrecht effects have similar underlying mechanisms resulting from an increase in aerosols. First, aerosols are emitted and enter a cloud where they activate as CCN. Provided that the amount of liquid water in the cloud

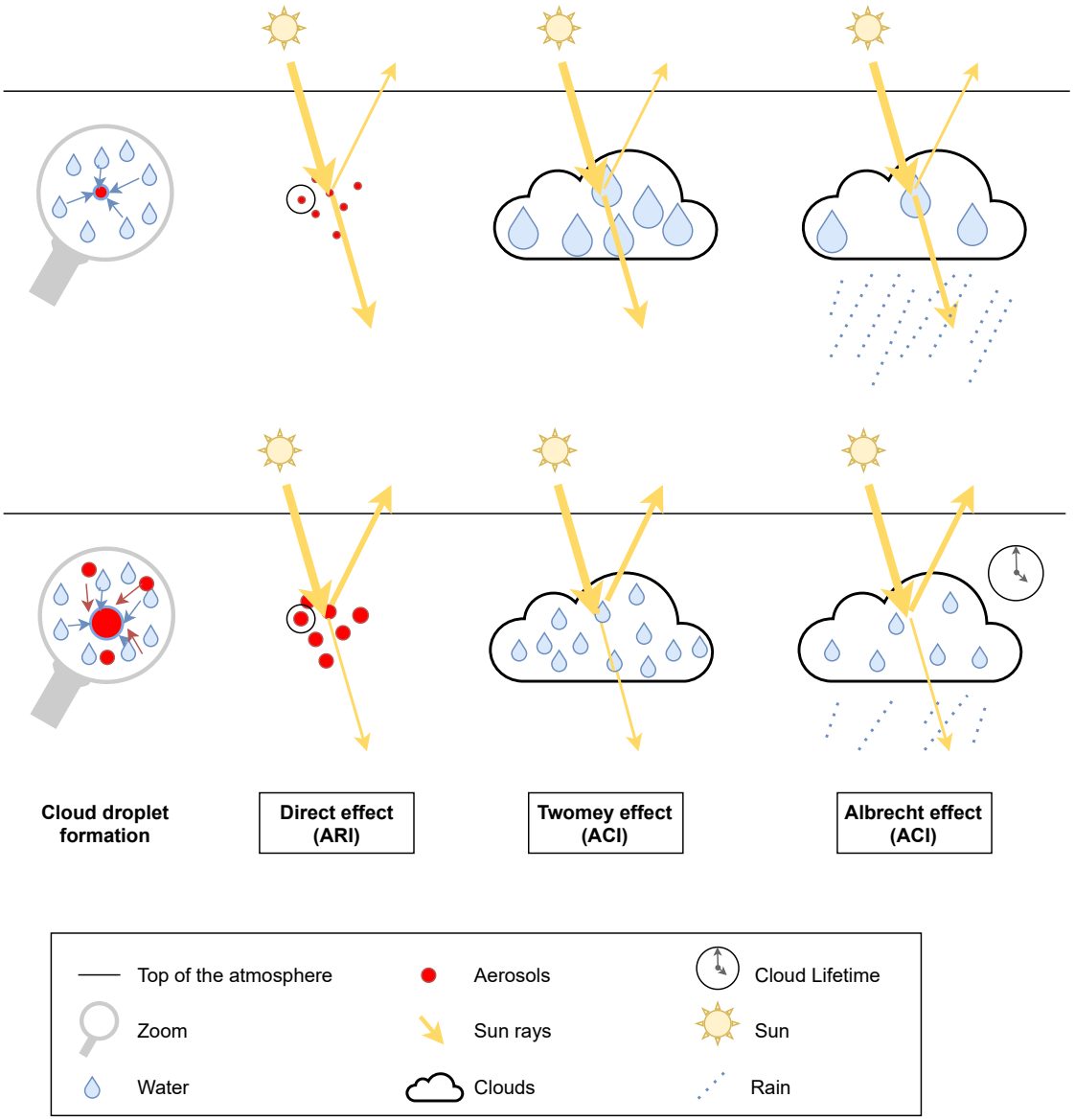


Figure 2.2: Aerosols' cooling effect through interactions with radiation and clouds (ARI and ACI). The top row represents higher levels of aerosols. The bottom row represents lower levels of aerosols. Left: direct cooling effect of aerosols (ARI). Middle and Right: indirect cooling effect of aerosols (ACI): Twomey effect on cloud reflectivity, and Albrecht effect on cloud lifetime.

remains constant, the same amount of liquid water is divided between more CCN which leads to an increase in the number of droplets N_d but a decrease in their size r_e. Both of these effects are represented in Figure 2.2, with the top row showing lower levels of aerosols and the bottom row representing higher levels of aerosols.

Twomey effect: cloud reflectivity The Twomey effect asserts that an increase in aerosols leads to an increase in cloud reflectivity [53]. The increase in the number of cloud droplets N_d described previously results in more light being scattered. This leads to brighter and more reflective clouds. In terms of measurements, this can be observed by an increase in τ.

Albrecht effect: cloud lifetime and coverage The Albrecht effect states that an increase in aerosols leads to an increase in cloud lifetime and cloud coverage [1]. The decrease in cloud droplet size r_e leads to droplets not being large enough to precipitate, and thus reduced precipitation efficiency. Hence, the cloud liquid water stays suspended for longer. The clouds then last longer or spread out more thus increasing cloud coverage CF.

Caveats All clouds are affected by aerosols, but the effects depend on the type of aerosols, the type of clouds and the environmental conditions [17]. There are therefore caveats to these two effects. For example, if the aerosol is black carbon, it absorbs the light thus making the cloud less reflective in contradiction with the Twomey effect [9]. Aerosols can also absorb solar radiation directly thus changing the clouds' environment and leading to their evaporation and a decrease in cloud coverage, contradicting the Albrecht effect. These examples give us a taste of the fact that ACI have high spatio-temporal variability and have a non-linear dependence on meteorological drivers like temperature, and winds.

Modelling ACI ACI are the most uncertain driver of radiative forcing due to human activities [7, 32]. It is difficult to understand and model ACI, especially on a global scale. This is because aerosols and aerosol types are unevenly distributed

around the globe, travel thanks to winds and have a limited lifetime. Moreover, since diverse types of aerosols and environmental conditions have different effects on clouds in terms of magnitude and sign, their impacts are mostly regional. For instance, sand and black carbon are found in distinct geographical regions and affect clouds differently. In technical terms, it is said that the effect of aerosols on cloud properties is heterogeneous and confounded by environmental conditions. The difficulty to measure aerosols and aerosol-clouds interactions, the inability to perform experiments, and the lack of baselines for the pre-industrial era all hinder research and contribute to these uncertainties.

2.1.4 Importance of ACI for climate modelling

Climate models are computer simulations of the Earth's climate system which are used to understand and predict its behaviour. Models differ in their method, they can be physical or statistical, and the scale of the processes they model. Physical models rely on mathematical equations of physical, chemical and biological processes whilst statistical models make use of data to uncover physical relations.

Modelling and experimentation increase our understanding of complex processes involved in climate change as a function of anthropogenic and natural changes that are affecting our climate. Climate models are essential to test different carbon emissions scenarios and help decision makers find appropriate policies to reduce global warming. It is therefore of utmost importance to improve accuracy and reduce uncertainties of these models to act appropriately to reduce climate change.

Although models are tested by running simulations of past events and improved using real-world observations, there remain uncertainties and approximations. Most current climate models, called general circulation models, do not have a high enough resolution to capture cloud processes. ACI thus remain the largest source of model uncertainty [7, 32]. Overall, it is estimated that the cooling effect of aerosols overcomes their warming effect and that aerosols counteract the effects of greenhouse gases. To give an order of magnitude, current climate models fail to emulate ACI to the extent that they have uncertainty bounds that could offset global

warming completely or double the effects of rising carbon dioxide [7]. Understanding how human emissions affect the ability of clouds to cool our planet and reducing uncertainties is crucial to accurately assess climate intervention methods, particularly geoengineering methods involving cloud seeding. This highly motivates research on ACI and encourages scientists to investigate the effects of the environmental surroundings of clouds and aerosols on cloud properties.

2.2 Causal inference

In this section, background on causality and its core principles are given. We start by defining and motivating causal inference, then illustrate our point with Pearl's ladder of causation, and finally introduce the potential outcomes framework. We use results and definitions from [34, 37–40].

2.2.1 Definition and motivations

Causal inference is the process of drawing conclusions about cause and effect. It consists in analysing the response of an effect variable, also called outcome, when one of the causes of this variable, also called treatment, is changed. Formally, we are interested in estimating the effect of a treatment (aerosols), denoted by the random variable $\mathrm{T} \in \mathcal{T}$, on outcomes of interest (cloud properties), denoted by the random variable $\mathrm{Y} \in \mathcal{Y}$, for a unit i described by covariates (environment) represented by the random variable $\mathbf{X} \in \mathcal{X}$.

The main problem motivating causal inference can be summarised in the phrase "association is not causation", more widely known as "correlation does not imply causation". This amounts to saying that observing statistical association between two variables Y and T does not allow to conclude on the effect of T on Y. For example, we can observe that as ice cream sales increase, the rate of drowning deaths increases. This observation is however insufficient to conclude about the causal relationship between ice cream consumption and drowning, and we must account for the season, temperature and exposure to water-based activities, which are all confounding effects. This example highlights that statistical association is a mixture

of causal association and confounding association. Causal inference aims to uncover causal patterns from mere association, notably by controlling for confounders.

It is a critically important task in a variety of fields such as climate science, healthcare, and economics. We may be interested in studying the effect of environmental policies on emissions and climate change, or the effect of a treatment on a disease. Overall, causal inference is not merely a tool but a paradigm to answer a specific type of question.

2.2.2 Ladder of causation

In [40], Pearl describes three steps of learning about causality, represented by what he calls the Ladder of Causation, represented in Figure 2.3. At the first rung, we uncover association by seeing, and answering the question "What if I see ...". It entails observing regularities or patterns in data. The second rung is that of intervention, where we learn by doing and address the question "What if I do ...". We predict the effects of deliberate actions and infer causal relationships. The third rung is about imagining and constructing a theory explaining why actions have specific effects and reason about the absence of these actions. It sits in the realm of counterfactuals where questions are of the form "What if I had done ...".

Standard statistical methods and machine learning enable us to access the first rung of this ladder. Scientific experimentation elevates us to the second rung, provided we rigorously follow the scientific method. This rung and the third one can also be attained through causal knowledge and further assumptions.

Let us now reason about the example of aerosols and clouds, and address the question: "Does an increase in aerosol lead to a decrease in cloud droplet size?". The naive way to answer this question consists in comparing conditional expectations of the size of cloud droplets given different aerosol levels. Unfortunately, due to confounding effects like meteorological conditions, this reasoning only allows us to infer statistical association and not causation. The most natural way to reason causally would be to climb to the second rung of the ladder and perform experiments. Unfortunately, experimenting is impractical in this case, and in general

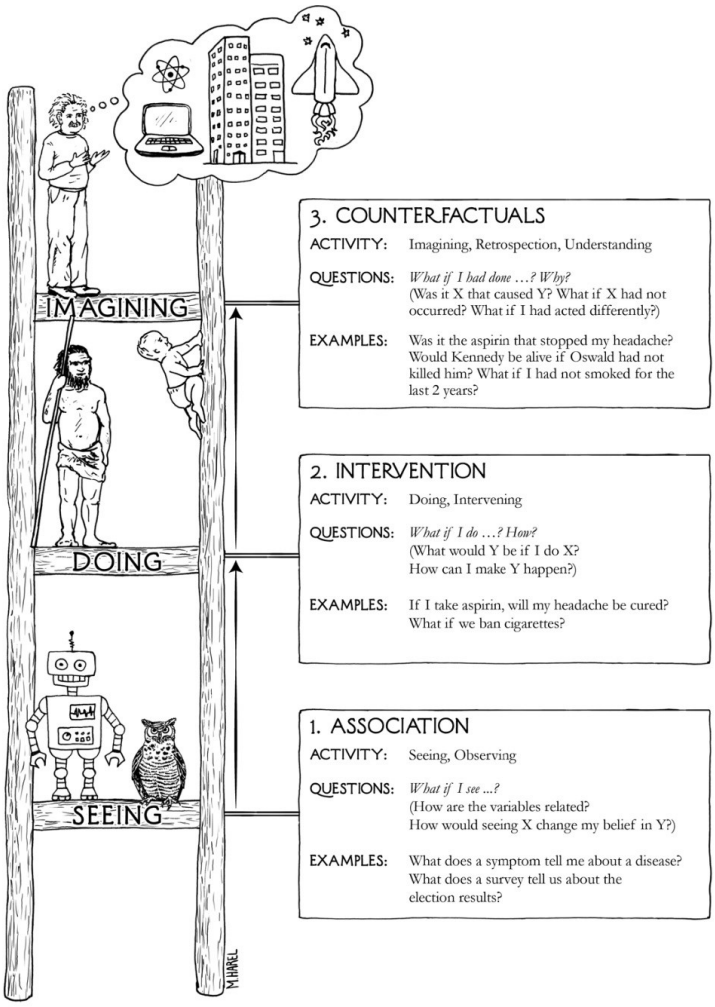


Figure 2.3: "**The Ladder of Causation**, with representative organisms at each level. Most animals, as well as present-day learning machines, are on the first rung, learning from association. Tool users, such as early humans, are on the second rung if they act by planning and not merely by imitation. We can also use experiments to learn the effects of interventions, and presumably, this is how babies acquire much of their causal knowledge. Counterfactual learners, on the top rung, can imagine worlds that do not exist and infer reasons for observed phenomena. (*Source*: Drawing by Maayan Harel.)" [40]

often expensive or unethical [43]. Some scientists turn to natural experiments, where they study natural perturbations such as ship tracks, or hemispheric differences and consider differences in the northern and the southern hemispheres as a proxy to estimate post-industrial and pre-industrial aerosols respectively [11, 21, 47]. It remains however unclear how representative these natural experiments are in the grander scheme of studying the global response of clouds to aerosols. Climate scientists should therefore climb to the last rung of the ladder, modelling with

both experimental and observational data, and using further assumptions to reason in the world of counterfactuals.

2.2.3 Potential outcomes framework

We describe one of the frameworks to study the causal question "What if I had done ...", the potential outcomes framework, also known as the Neyman-Rubin causal model [43, 44, 48, 49].

Notation Recall that we are interested in estimating the effect of a treatment, denoted by the random variable $\mathrm{T} \in \mathcal{T}$, on outcomes of interest, denoted by the random variable $\mathrm{Y} \in \mathcal{Y}$, for a unit i described by covariates represented by the random variable $\mathbf{X} \in \mathcal{X}$. In what follows, we use upper-case letters to denote random variables and lower-case letters to denote values that these random variables take on. We assume that we have observational data $\mathcal{D}_n$ consisting of n realisations of the random variables so that $\mathcal{D}_n = \{(\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i)\}_{i=1}^n$. We call a potential outcome denoted by Y_t what the outcome would be if the treatment were t. It is distinct from the observed outcome Y because not all potential outcomes are observed. We assume that the tuples $(\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i)$ are independent and identically distributed (i.i.d.) samples from the joint distribution $P(\mathbf{X}, \mathrm{T}, \mathrm{Y}_\mathrm{T})$ where $\mathrm{Y}_\mathrm{T} = \{\mathrm{Y}_\mathrm{t} \mid \mathrm{t} \in \mathcal{T}\}$.

Treatment effect and identifiability A causal estimand of interest is the individual treatment effect. The best way to evaluate this quantity would be to observe all potential outcomes for a given individual, but this is impossible and known as the fundamental problem of causal inference [25, 43]. One can however study other causal estimands, like the conditional average potential outcome (CAPO) and the average potential outcome (APO):

$$\text{CAPO} = \mu(\mathbf{x}, \mathrm{t}) \coloneqq \mathbb{E}\left[\mathrm{Y}_\mathrm{t} \mid \mathbf{X} = \mathbf{x}\right] \quad \text{and} \quad \text{APO} = \mu(\mathrm{t}) \coloneqq \mathbb{E}\left[\mu(\mathbf{X}, \mathrm{t})\right]. \tag{2.1}$$

Unfortunately, with data from the observational distribution $P(\mathbf{X}, \mathrm{T}, \mathrm{Y}_\mathrm{T})$, one can only compute the following estimates:

$$\tilde{\mu}(\mathbf{x}, \mathrm{t}) = \mathbb{E}\left[\mathrm{Y} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}\right] \quad \text{and} \quad \tilde{\mu}(\mathrm{t}) = \mathbb{E}\left[\tilde{\mu}(\mathbf{X}, \mathrm{t})\right]. \tag{2.2}$$

Naturally, once can wonder how the APO $\mu(\mathrm{t})$ and the CAPO $\mu(\mathbf{x}, \mathrm{t})$ relate to the observational estimates $\tilde{\mu}(\mathrm{t})$ and $\tilde{\mu}(\mathbf{x}, \mathrm{t})$. Formally, this is known as identifying a causal effect, that is, reducing a causal expression to a purely statistical expression. Identifying the APO and the CAPO from observational data requires additional assumptions which we describe before formally proving identifiability.

Unconfoundedness The first assumption is ignorability and states that units are randomly assigned their treatment. It guarantees that the treatment groups are comparable. It is written:

$$\mathrm{Y}_{\mathrm{T}} \perp\!\!\!\perp \mathrm{T}. \tag{2.3}$$

In practice, this assumption is unrealistic as there are likely to be confounding variables, especially with observational data. Confounding variables are variables that have an impact on the results of a statistical test but are not the variables that causal inference is studying. For example, meteorological conditions (like winds and pressure) affect both aerosols and cloud properties and therefore confound the influence of aerosols on cloud properties. We therefore adjust our assumption by controlling for confounding variables by conditioning. This assumption is called unconfoundedness, or conditional ignorability, and is written:

$$\mathrm{Y}_{\mathrm{T}} \perp\!\!\!\perp \mathrm{T} \mid \mathbf{X}. \tag{2.4}$$

It implies that $\mathbb{E}\left[\mathrm{Y}_{\mathrm{t}} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x}\right] = \mathbb{E}\left[\mathrm{Y}_{\mathrm{t}} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}\right]$ for any $\mathrm{t}, \mathrm{t}' \in \mathcal{T}$.

Positivity Intuitively, it seems like we can fit as many covariates into $\mathbf{X}$ as possible to ensure unconfoundedness. Unfortunately, doing so can be detrimental to the positivity assumption. This assumption states that all subgroups of the data with different covariates have a non-zero probability of receiving any dose of treatment. Formally, for any $\mathbf{x} \in \mathbf{X}$ such that $\mathbb{P}\left[\mathbf{X} = \mathbf{x}\right] > 0$, we must have

$$\mathbb{P}\left[\mathrm{T} = \mathrm{t} \mid \mathbf{X} = \mathbf{x}\right] > 0 \quad \forall \mathrm{t} \in \mathcal{T}. \tag{2.5}$$

There is therefore a trade-off between positivity and unconfoundedness due to the curse of dimensionality and working with finite data, as with large $\mathbf{X}$ and continuous treatment, it is unlikely that we observe all treatment levels for each $\mathbf{x} \in \mathbf{X}$.

No interference Furthermore, we need the no interference assumption which states that a single unit's outcome is not affected by other units' treatment.

Consistency Finally, we assume consistency which states that a unit's observed outcome Y given treatment t is identical to their potential outcome, written:

$$\mathrm{Y} = \sum_{\mathrm{t} \in \mathcal{T}} \mathrm{Y}_{\mathrm{t}} \cdot \mathbb{1}[\mathrm{T} = \mathrm{t}], \tag{2.6}$$

where $\mathbb{1}$ denotes the indicator function so that $\mathbb{1}[\mathrm{T} = \mathrm{t}]$ is 1 if $\mathrm{T} = \mathrm{t}$ otherwise 0.

Proof of identifiability Let us now assume no interference and prove that the CAPO and the APO are identifiable from the observational distribution $P(\mathbf{X}, \mathrm{T}, \mathrm{Y}_{\mathrm{T}})$ by adapting a proof from [34]:

$$\begin{aligned}
\mu(\mathrm{t}) &= \mathbb{E}\left[\mu(\mathbf{X}, \mathrm{t})\right] && \text{by definition} \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathrm{Y}_{\mathrm{t}} \mid \mathbf{X}\right]\right] && \text{by definition} \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathrm{Y}_{\mathrm{t}} \mid \mathrm{T} = \mathrm{t}, \mathbf{X}\right]\right] && \text{by unconfoundedness (2.4) and positivity (2.5)} \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathrm{Y} \mid \mathrm{T} = \mathrm{t}, \mathbf{X}\right]\right] && \text{by consistency (2.6)} \\
&= \mathbb{E}\left[\tilde{\mu}(\mathbf{X}, \mathrm{t})\right] && \text{by definition} \\
&= \tilde{\mu}(\mathrm{t}). && \text{by definition}
\end{aligned}$$

In summary, with the assumptions enumerated above, we can infer causal relationships from observational data. In our case, using satellite observations theoretically allows us to infer the effect of aerosols on clouds. In practice, the present thesis explores limitations of this reasoning, especially when some of the assumptions are violated.

2.3 Machine learning

Machine learning aims to uncover complex relationships and patterns in datasets. It builds methods that "learn" from data to perform specific tasks. Algorithms are assessed on their capacity to generalise to unseen data rather than just memorising training data.

Formally, the aim is to learn a function $f_{\boldsymbol{\theta}}$ from an input space $\mathcal{X}$ to an output space $\mathcal{Y}$. Methods are designed to learn this function automatically from data, adjusting the parameters $\boldsymbol{\theta}$. It can be viewed as an optimisation problem, searching for appropriate values of $\boldsymbol{\theta}$ to optimise a loss function $\mathcal{L}(\boldsymbol{\theta})$. For this, we make use of a training dataset $\mathcal{D}$. In supervised learning, $\mathcal{D}$ consists of input output pairs $(\mathbf{x}, \mathbf{y}) \in (\mathcal{X}, \mathcal{Y})$, whereas in unsupervised learning $\mathcal{D}$ only consists of input data $\mathbf{x} \in \mathcal{X}$.

2.3.1 Regression methods

Ordinary least squares linear regression Assume the dataset is labelled, $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, and the aim is to predict real-valued numbers, $\mathcal{Y} = \mathbb{R}$ from D-dimensional input, $\mathcal{X} = \mathbb{R}^D$. If we assume a linear relationship between inputs and output, the data can be modelled using linear regression. A prediction for a single data point $(\mathbf{x}, y) \in (\mathbb{R}^D, \mathbb{R})$ is then of the form:

$$\hat{\bar{y}} = f_{\boldsymbol{\theta}}(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_D x_D = w_0 + \sum_{i=1}^{D} w_i x_i,$$

where $\bar{y} = \mathbb{E}[y \mid \mathbf{x}]$ and $\mathbf{w}$ is one of the parameters included in $\boldsymbol{\theta}$. For ease of notation, a dimension $x_0 = 1$ is added to the input data so that

$$\hat{\bar{y}} = f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}.$$

The least squares estimate looks for $\mathbf{w}$ that minimises the following:

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \hat{\bar{y}}_i)^2.$$

Through basic calculus, we obtain a closed-form optimal solution

$$\mathbf{w}^\star = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{y},$$

where $\mathbf{X}$ is the matrix of all input points (called the input matrix), and $\mathbf{y}$ is the vector of all outputs (called the output vector).

Regularisation Overfitting is a problem that arises when the model is too complex, fits exactly the training data, and cannot generalise to unseen data. To prevent overfitting, we can use regularisation techniques. The idea is to add a penalty or regularisation term to the loss function. This term is a function of the model's parameters $\boldsymbol{\theta}$. For example, the Ridge regression penalises large weights and takes the following form:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{D} w_i^2,$$

where $\lambda > 0$ is a hyperparameter that controls the strength of the regularisation. There are other forms of regularisation, such as Lasso, which are not discussed here.

Gradient descent Depending on the form of the loss function, its minimum does not necessarily have a closed-form solution. In such case, we can use optimisation techniques such as gradient descent, or variants, to approximate $\boldsymbol{\theta}^\star$ [45]. The idea is to iteratively update the parameters $\boldsymbol{\theta}$ by taking a step in the direction of the gradient. If we denote the gradient of the loss function as $\nabla\mathcal{L}(\boldsymbol{\theta})$, then the update is given by:

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} - \eta\nabla\mathcal{L}(\boldsymbol{\theta}),$$

where $\eta > 0$ is the learning rate, a hyperparameter that controls the size of the step at each iteration.

Polynomial feature expansion When there is no linear relationship between the inputs and the output, we can complexify either the model or the representation of the data. Polynomial feature expansion is a technique that allows modelling of the data using a higher order polynomial. The idea is to expand the input space by adding polynomial features using a map ϕ. For example, to learn quadratic relationships with a two-dimension input of the form $\mathbf{x} = (x_0, x_1)$, we can use

the following map:

$$\phi(\mathbf{x}) = \begin{pmatrix} 1 \\ x_0 \\ x_1 \\ x_0^2 \\ x_1^2 \\ x_0 x_1 \end{pmatrix}.$$

We can then fit a linear model to the transformed data and learn non-linear relationships.

Evaluating performance To evaluate a model's performance, we can use various techniques and metrics. One can study the prediction error plot, which is a scatter plot of the predicted values $\hat{y}$ against the observed values y. An associated metric is the squared Pearson correlation coefficient, denoted r^2. It is a measure of linear correlation between two sets of data, namely the predicted values and the observed values. We have:

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{\hat{y}}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where n is the number of observations, y_i is the i^{th} observed value, $\hat{y}_i$ is the i^{th} predicted value, and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the mean of all observed values. The values of this coefficient range from 0 to 1. A value of 1 indicates a perfect linear correlation, whilst values closer to 0 indicate a weaker linear correlation.

2.3.2 Artificial neural networks

Artificial neuron Artificial neurons were conceived as mathematical models of biological neurons and are now the building blocks for artificial neural networks which are at the core of modern machine learning and deep learning. Neurons, whether artificial or biological first receive an input signal and then process it before sending an output signal. An artificial neuron receives some input $\mathbf{x}$, and performs some linear transformation using a weight vector $\mathbf{w}$. We obtain $\mathbf{w} \cdot \mathbf{x}$, called the pre-activation, where a dimension is added to $\mathbf{x}$ to account for the bias

term. It is then passed to a non-linear function f, known as the activation function, to obtain the output $\hat{\mathbf{y}}$. Formally, we have

$$\hat{\mathbf{y}} = f(\mathbf{w} \cdot \mathbf{x}) = f\left(w_0 + \sum_{i=1}^{D} w_i x_i\right),$$

as shown in Figure 2.4. The input signal is on the left, each component is multiplied by a corresponding weight w_i represented in a circle. These products are added together, and the result is passed through the activation function f, inside the unit, to obtain the output $\hat{\mathbf{y}}$.

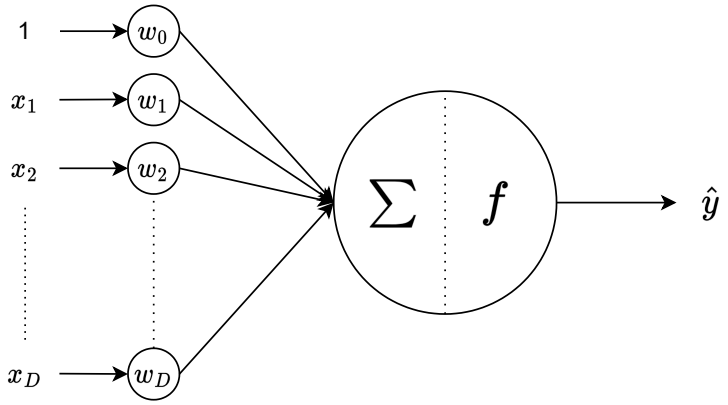


Figure 2.4: Artificial Neuron. The input signal on the left is multiplied by weights w_i, these products are added together, and the result is passed through the activation function f to obtain the output $\hat{y}$.

A famous example of an artificial neuron is the perceptron, which is a simple linear classifier invented by McCulloch and Pitts in the 1960s and implemented by Rosenblatt in the 1970s. It separates data between two classes 1 and 0 and is therefore called a binary classifier. The predictions are as follows:

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{w} \cdot \mathbf{x} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

Put more simply, it is an artificial neuron with a threshold activation function.

Feed-forward neural networks To learn non-linearly separable patterns, we can stack multiple layers of perceptrons together, creating a Multi-Layer Perceptron (MLP). More generally, we use neural networks which were inspired by the way biological nervous systems process information and are more general than MLP.

A neural network is a collection of interconnected artificial neurons, called units, organised into layers as shown in the diagram in Figure 2.5. The first layer (in **blue**) is called the input layer and processes the raw input data. The last layer (in **red**) is called the output layer and produces the prediction of the network. All the layers in between are called hidden layers (in **green**). This is where the data processing is done, and the number of hidden layers directly impacts the processing power of the network. At each layer, the units are connected using weights. Each unit, or artificial neuron, works as described earlier, computing a weighted sum of the inputs (called pre-activation) and then applying a non-linear function (called the activation function) to obtain the activation which will be passed on as input to the next layer's units.

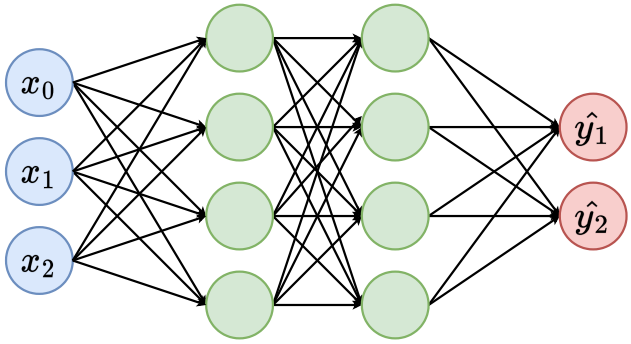


Figure 2.5: Artifical neural network with input of dimension 3 (in **blue**), 2 hidden layers (in **green**) and output of dimension 2 (in **red**). Each arrow represents a weight learnable through backpropagation and each circle represents a unit with summation and a non-linear activation function.

Backpropagation The aim is to optimise the network's weights to minimise the loss function over all training datapoints. This optimisation is often done using gradient descent, but variants are also used depending on the type of loss function, the size of the training dataset, the loss landscape and others. To "learn" the weights, we need to know how the network's predictions are affected by the input data. For this, we use the backpropagation algorithm, which is a method of learning that uses the error of the network's predictions to adjust the weights of the network [46]. The error is computed through a loss function and is used to determine the direction and the magnitude of the weight adjustment.

Activation functions Various activation functions can be used in artificial neural networks depending on the type of problem and the input. The ReLU function is very common for regression problems:

$$\mathrm{ReLU}(x) = \max(0, x).$$

The Leaky ReLU function is a variant of the ReLU function, with a small negative slope α determined by the user. This function is used to circumvent the problem of vanishing gradients, that is, gradients that are too small to be useful:

$$\mathrm{LeakyReLU}(x) = \max(0, x) + \min(0, x) \cdot \alpha.$$

The SoftMax function is often used as the last activation function of a neural network to produce a probability distribution over the output classes:

$$\mathrm{SoftMax}(\mathbf{x})_j = \frac{e^{x_j}}{\sum_{i=1}^{n} e^{x_i}}.$$

The SoftPlus function is a smooth approximation of the ReLU function, with a parameter β determined by the user:

$$\mathrm{SoftPlus}(x) = \frac{1}{\beta} \ln(1 + e^{\beta x}).$$

Ensemble methods When a single estimator is not good enough to predict the output, an ensemble method can be used to combine the predictions of multiple estimators. Two families of ensemble methods are available: bagging and boosting. Bagging stands for bootstrap aggregating and is a technique where several independent estimators are built independently using subsets of the training data. The predictions of each predictor are then combined by averaging. By contrast, boosting consists of sequentially building estimators, attempting to reduce the bias of the combined estimator.

Regularisation A widely used regularisation technique is the dropout technique which is used to prevent overfitting [23, 50]. It consists in randomly omitting, or "dropping out" a fraction of the units during the training process. Effectively, it amounts to making the training process noisy and therefore preventing co-adaptation of the units, where each layer would correct for mistakes of previous layers.

Residual connections Residual, or shortcut connections were introduced by [22] to facilitate the optimisation and improve the accuracy of very deep networks. In traditional neural networks, data flows through each layer sequentially, and the output of each layer is used as input to the next layer. Residual connections provide another path for data to reach the latter parts of the network by skipping some layers, as shown in Figure 2.6. Residual networks can still be trained end-to-end using stochastic gradient descent and backpropagation.

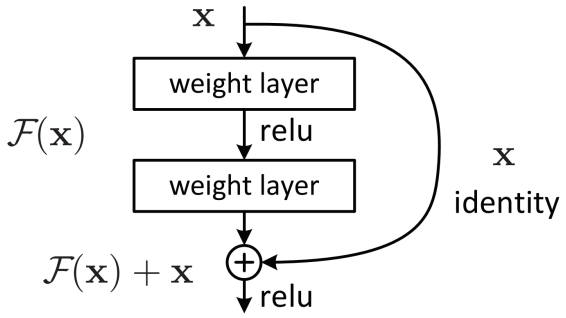


Figure 2.6: Residual network building block: residual connection. [22]

2.3.3 Attention and transformer

Attention mechanism Attention mechanisms in machine learning aim to mimic cognitive attention, the process allowing the human brain to select and focus on relevant stimuli. A typical example inspired by the real world is that of the cocktail party problem [10]. This effect describes the brain's ability to focus one's auditory attention on a single conversation in a noisy room, but the concept applies to many situations. Attention mechanisms aim to pick out salient information from noisy data. In computer science models, this is done by enhancing parts of the information whilst diminishing other parts. Attention allows us to model dependencies between sequences with respect to their relative importance rather than to their relative size or distance, thus allowing us to make use of context. It is commonly used in the field of neural machine translation (translation using deep learning) and aims to answer the question: "What parts of the input should one focus on for predicting specific parts of the output". For example, when translating sentences from one language to

another, proximity is not necessarily the best proxy for usefulness and attention allows to capture longer-range dependencies and unveil relationships between words.

Implicit attention Most neural networks have implicit attention in that they respond more strongly to certain parts of the data than others. Studying a network's Jacobian allows one to analyse the sensitivity of the network's outputs with respect to the inputs and thus study its implicit attention.

Explicit attention Explicit attention allows more flexibility, computational efficiency, scalability and interpretability. Introduced by [3], it relies on creating a context vector for every token in the input sequence. The context vector allows to capture more global information relevant to the current token and thus to the output. General attention mechanisms make use of three components: keys, queries and values. In machine translation, each word or token in an input sentence would be attributed its own query, key and value vectors, which are computed using weight matrices learnt by the network.

Hard and soft attention [56] introduces two types of attention: hard attention through glimpses and soft attention. In the context of image data, hard attention consists of fixed-size windows moving around the image. It is trained using reinforcement learning techniques since it is not differentiable. It is particularly useful for robots as they have restricted access to data and cannot "see" everywhere. Soft attention on the other hand is more flexible and can be trained end-to-end using backpropagation as it is differentiable. It relies on data-dependent dynamic weights that can change through runtime rather than fixed-size glimpses. In the context of image data, soft attention allows combining focus on multiple parts of the image at once through weighted image features.

Transformer Standard methods prior to [54] made use of attention as part of larger artificial neural network architectures such as long short-term memory [24] or gated recurrent neural networks [13]. [54] introduced the transformer which relies

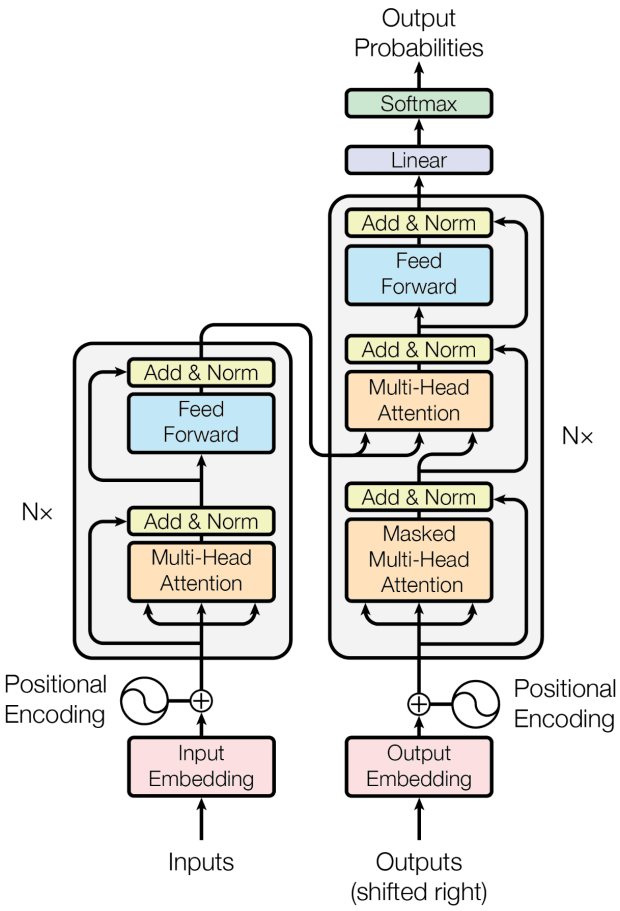


Figure 2.7: The Transformer: model architecture [54]

entirely on attention to draw global dependencies between input and output. The architecture of the transformer is represented in Figure 2.7. It is composed of an encoder (on the left of the figure) and a decoder (on the right of the figure). Encoder-decoder architectures are commonly used in machine translation, where the encoder is used to encode the input sequence into a more compact representation, and the decoder is used to decode the output sequence from the encoder's representation. Working with more compact representations improves efficiency and helps with learning longer-range dependencies. The model is said to be auto-regressive at each step, meaning that previously generated output is used as additional input during the following step. Positional encodings allow making use of the order of the sequence by injecting information about the relative and absolute position of the tokens in the sequence.

Self-attention The transformer uses a self-attention mechanism to relate different positions of a single sequence and thus compute a representation of the said sequence. Figure 2.8(a) shows the attention mechanisms used in the transformer, called the scaled dot-product attention. The input consists of queries and keys of dimension d_k, and values of dimension d_v, which are derived from tokens in the input sequence. In practice, the keys, values and queries can be packed together into matrices K, V and Q respectively, so that the attention value can be computed on multiple queries simultaneously. In mathematical terms, the scaled dot-product attention is defined as:

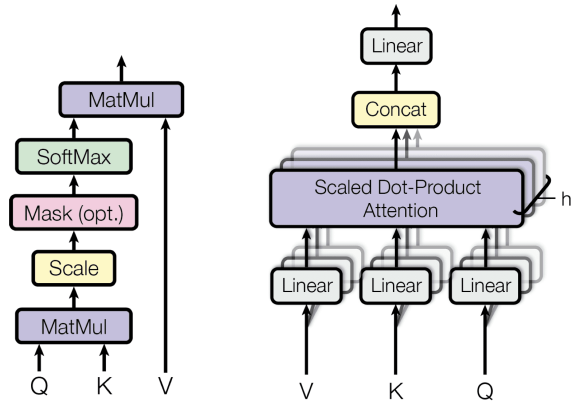
$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V.$$

Multi-head attention is a generalisation of the scaled dot-product attention, where the queries, keys and values are split into multiple heads. Having multiple heads allows attending to different parts of the sequence differently, whilst performing the computations in parallel. Figure 2.8(b) shows the mechanism, which can be written as:

$$\text{Multi-Head Attention}(Q, K, V) = \text{Concatenate}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.8}$$

$$\text{where} \quad \text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{2.9}$$

where W^O, W^Q, W^K and W^V are learnable parameter matrices.



(a) Scaled dot-product (b) Multi-head attention

Figure 2.8: Self-attention mechanisms: scaled dot-product and multi-head attention [54]

3
Problem setting

In this chapter, the problem is set formally and the research questions are stated. This chapter begins with a definition of the causal model used in this work in Section 3.1, followed by a description of the framework of the uncertainty and sensitivity analysis performed in Section 3.2, and the formulation of our research questions in Section 3.3.

3.1 Causal setting

This work is interested in unveiling the effect of aerosol (a) on cloud properties. Aerosols affect multiple cloud properties including the cloud droplet radius (r_e), the cloud optical depth (τ), the cloud water path (CWP) and the cloud fraction (CF). In the present thesis, to simplify the analysis and ease understanding, results are reported mainly for the cloud droplet radius (r_e). The simplified causal diagram is represented in Figure 3.1, with treatment in **purple** and outcomes in **red**. In this diagram and the following, arrows represent causal relationships, an arrow from A to B indicates that A causes B.

Unfortunately, aerosols are unobservable through satellites because of masking clouds. We therefore only observe the aerosol optical depth (AOD) which we use as a proxy for aerosol. AOD is however affected by local environmental processes [12],

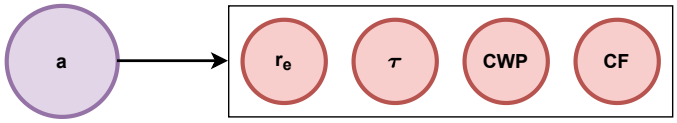


Figure 3.1: Simplified causal diagram of ACI. Aerosol concentration (a, regarded as treatment, in purple) modulates cloud properties (in **red**) including cloud optical depth (τ), cloud droplet radius (r_e), cloud water path (CWP) and cloud fraction (CF).

which also affect cloud processes. For instance, humidity affects both AOD, through aerosol swelling, and cloud micro-physics [2]. Whilst ACI is a causal problem at its core, as shown in Figure 3.1, the inability to observe aerosols directly leads to a confounding problem wherein the effects of aerosols on clouds are mixed in with other effects which results in a distortion of the true relationship. The causal diagram is represented in Figure 3.2: AOD, considered as the treatment and represented in **purple** modulates cloud properties, considered as outcomes and represented in **red**. Environmental processes, represented in **blue**, act as confounders, affecting both the treatment and the outcomes, through cloud processes. The dashed line represents the effect between aerosol and cloud properties, that is, the effect of interest.

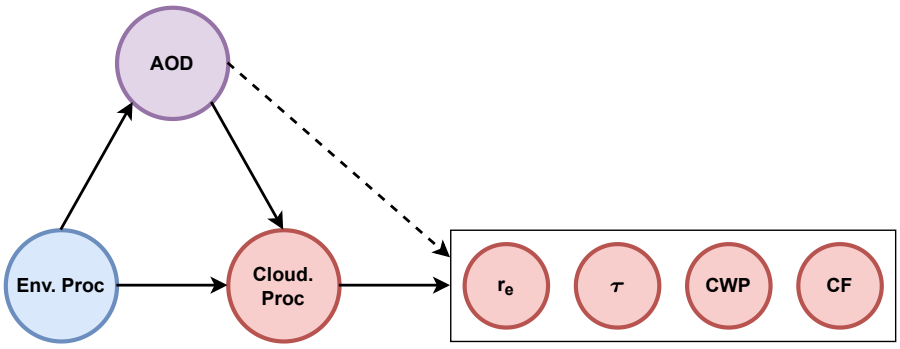


Figure 3.2: Causal diagram of ACI with confounding from environmental and cloud processes. Environmental and cloud processes confound the effect of aerosol on cloud properties. AOD (treatment, **purple**) modulates cloud properties (outcomes, **red**). Environmental processes (in **blue**) and cloud processes (in **red**) act as confounders. The dashed line is the effect of interest, measured with the APO.

Since environmental and cloud processes are impossible to directly observe from satellite data, we use meteorological proxies such as relative humidity (RH), sea surface temperature (SST), estimated inversion strength (EIS), vertical winds (ω500) and lower tropospheric stability (LTS). These meteorological proxies are

selected because they are thought to be a good approximation for temperature, pressure and saturation, which all impact cloud processes.

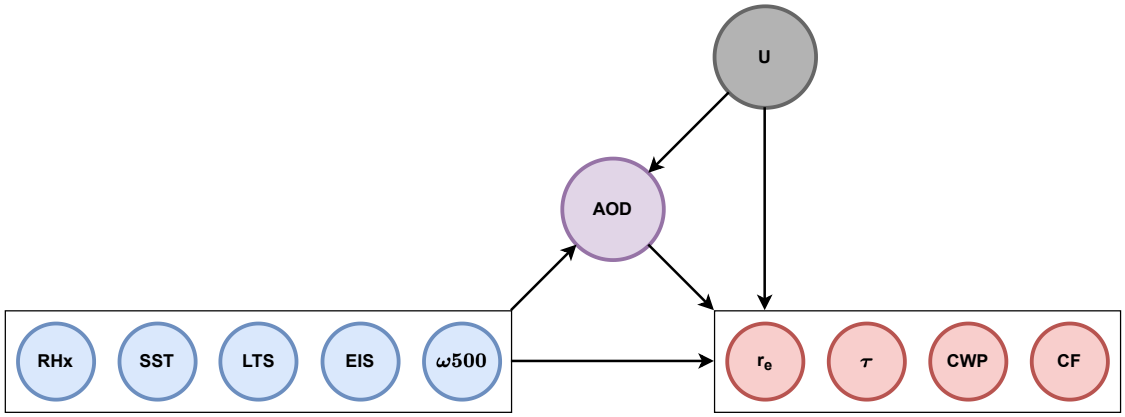


Figure 3.3: Causal diagram of ACI we report within. AOD (treatment, **purple**) modulates cloud properties (outcomes, **red**), which are confounded by meteorological proxies (covariates, **blue**) and unobserved confounding (**grey**).

This work investigates how unobserved confounding variables, such as the environment, and the use of AOD as a proxy for aerosol, change the estimates of the effects of aerosol on cloud properties whilst controlling for meteorological proxies, as shown by the causal diagram in Figure 3.3.

3.2 Uncertainty and sensitivity analysis

Using observational data requires knowledge about the characteristics of the population and generally leads to lower certainty in the estimated causal effects. This is because the assumptions needed for identifiability of the conditional average potential outcome (CAPO), such as unconfoundedness, are unrealistic in practice, and often untestable with observational data. These assumptions are especially problematic in the continuous treatment regime. Violation of these assumptions induces bias in the estimates of causal effects, that is, $\tilde{\mu}(\mathbf{X}, \mathrm{t})$ can be an arbitrarily biased estimate of the CAPO $\mu(\mathbf{X}, \mathrm{t})$.

When performing uncertainty and sensitivity analysis, we want to understand how robust our estimates are to the violation of assumptions and compute uncertainty bounds with respect to the relaxation of these assumptions. For this,

we quantify the degree of violation of the assumptions and derive intervals of possible causal effects. The width of the interval increases as the assumptions are challenged more severely.

We focus on the violations of two main assumptions: unconfoundedness and positivity. We note the trade-off between these two assumptions, where fitting as many covariates into $\mathbf{X}$ as possible ensures unconfoundedness but violates positivity due to the curse of dimensionality. We describe possible violations more precisely using the framework of the continuous treatment-effect marginal sensitivity model (CMSM) introduced by [26].

Unconfoundedness violations Confounding variables are unobserved factors that influence the treatment assignment or outcomes, and thus manifest as variance in the estimates of the outcome and propensity density for treatment. Unobserved confounding variables, that is, variables that we know impact the treatment assignment or the outcome, but that we do not observe, lead to the violation of the unconfoundedness assumption shown in Equation (2.4). The CMSM proposes a parameter Λ to explain a certain level of violation of the unconfoundedness assumption [27, 28]. This parameter relies on the observation that any divergence between the unidentifiable $\mathbb{P}\left[\mathrm{Y}_\mathrm{t} \mid \mathbf{X} = \mathbf{x}\right]$ and the identifiable $\mathbb{P}\left[\mathrm{Y} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}\right]$ is indicative of hidden confounding. Working with densities instead of probabilities and using measure theory, the authors set

$$\lambda(\mathrm{y_t}; \mathbf{x}, \mathrm{t}) = \frac{p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})}.$$

Whilst λ cannot be identified from data alone, the CMSM enables domain experts to set a degree of hidden confounding using the parameter Λ such that

$$\Lambda^{-1} \leq \lambda(\mathrm{y_t}; \mathbf{x}, \mathrm{t}) \leq \Lambda,$$

that is, hypothesising that $p(\mathrm{t} \mid \mathbf{x})$ and $p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})$ differ by at most Λ. Intuitively, Λ represents the proportion of range in unexplained outcome Y coming from unobserved confounders after observing the covariates $\mathbf{x}$ and the treatments t.

In our study, confounding comes from three main sources: aerosols, clouds, and the environment. The aerosol proxy does not allow us to access aerosol type, size, or hygroscopicity which are all confounders of ACI. We also rely on meteorological proxies to capture environmental processes which confound ACI. Clouds are moreover interconnected in real life but not in our model. All of these simplifications lead to unmeasured confounding in our model which we discuss throughout this work.

Positivity violations Since the observed data is finite and the treatment is continuous, it is almost always impossible to observe all treatment levels for every covariate value describing a set of units, thus violating the positivity (or overlap) assumption shown in Equation (2.5) [15]. The CMSM proposes that positivity violations are linked to statistical uncertainty. Indeed, statistical uncertainty is high where the overlap is weak, that is, where few treatment levels are observed for a given covariate value. They build $(1 - \alpha)$ statistical confidence intervals for the upper and lower bounds and suggest that the parameter α can be used to describe different levels of violations of the positivity assumption.

3.3 Research questions

Our work consists of the study of the artificial neural networks described in [26] (Overcast) to estimate the effects of aerosols on cloud properties. We also perform an uncertainty and sensitivity analysis to study hidden confounding using the CMSM, which allows accounting for various levels of violations of the unconfoundedness and positivity assumptions through the parameters α and Λ. Our work is articulated around the following research questions:

1. How well do the Overcast models emulate ACI in terms of predictive accuracy and treatment-effect estimates?

2. How well do the Overcast models capture geographical dependencies of ACI?

3. How does unmeasured confounding affect plausible ranges of treatment-effect estimates of ACI?

4
Experimental setup

This chapter details the experimental setup. The data is described in Section 4.1, with an explanation of the data sources, the datasets used and the pre-processing steps applied. In Section 4.2, the Overcast model architecture for both the feed-forward neural network and the transformer is described and illustrated. Details about our implementation, and the training and tuning procedures are provided for reproducibility in Section 4.3, and the general setup for our experiments is given in Section 4.4.

4.1 Data

4.1.1 Data sources

Data is in the form of tabular data that has been retrieved from re-analyses of satellite observations. The Moderate Resolution Imaging Spectroradiometer (MODIS) instruments aboard the Terra and Aqua satellites observe the Earth at approximately 1 km $\times$ 1 km resolution [4]. These observations are fed into the Modern-Era Retrospective Analysis for Research and Applications version 2 (MERRA-2) real-time model to emulate the atmosphere and its components, such as aerosols [19]. MERRA-2 calculates global vertical profiles of temperature, relative humidity, and pressure, and assimilates hyperspectral and passive microwave

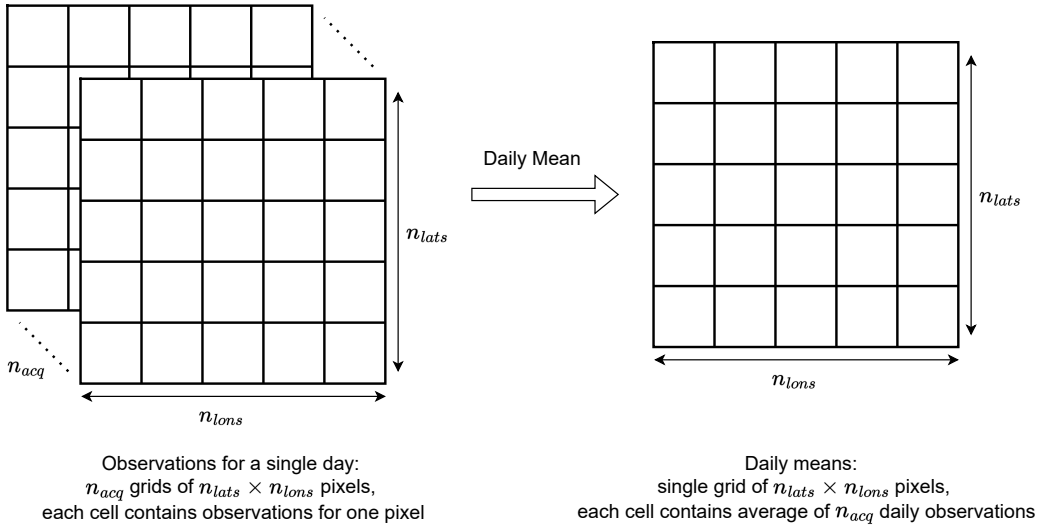


Figure 4.1: Schematic representation of the datasets. For each day, there are n_{acq} grids of satellite observations of $n_{lats} \times n_{lons}$ pixels of a given size. These grids can be averaged to obtain daily means of observations.

satellite observations to enhance its ability to model Earth's atmosphere. The data studied are MODIS observations from the Aqua and Terra satellites collocated with MERRA reanalyses of the environments.

Aerosol optical depth (AOD) at 550nm from MERRA-2 is derived from MODIS observations of aerosol from multiple satellites (Terra, Aqua, Suomi-NPP), with corrections for sunglint and near-cloud optical effects [6]. The cloud droplet radius r_e is found for pixels that are both probably or definitely cloudy according to the MODIS Cloud Mask. The NOAA CMORPH CDR precipitation product is found by integrating multiple observations of precipitation from both satellite and in-situ sources. Sea surface temperature (SST) from NOAA WHOI CDR is found using multiple observations of surface brightness temperature and incorporating precipitation estimates to better approximate the effects of the diurnal cycle on sea surface temperature. The data sources are summarised in Appendix A.

4.1.2 Datasets

Figure 4.1 is a schematic representation of the datasets. For each day, there are n_{acq} grids of observations of $n_{lats} \times n_{lons}$ pixels of a given size, where n_{acq} is the number of acquisitions in a given day, n_{lats} is the number of pixels along the latitude

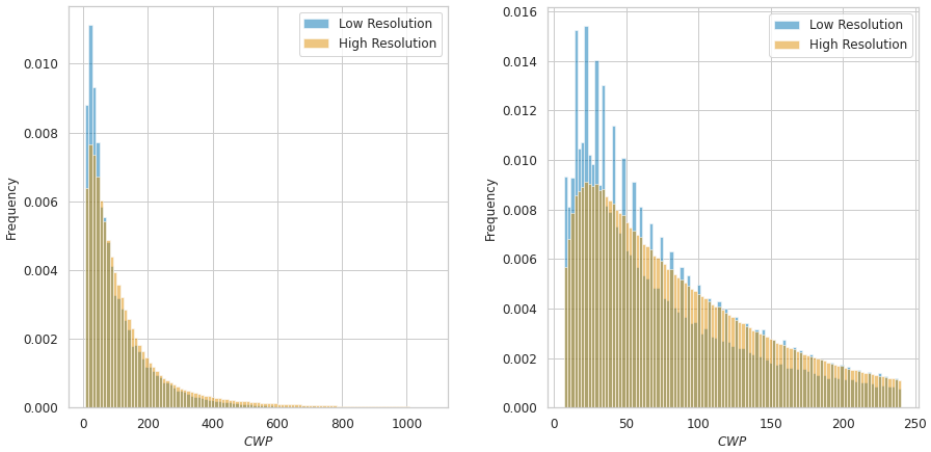
and n_{lons} is the number of pixels along the longitude. These grids can be averaged to obtain daily means of observations, as shown on the right of the figure. The size of each pixel depends on the spatial resolution of the dataset. In the native observations, the pixels have a size of 1 km $\times$ 1 km.

We use three different datasets: (1) low-resolution data from the South-East Pacific (LR Pacific), (2) low-resolution data from the South Atlantic (LR Atlantic) and (3) high-resolution data from the South-East Pacific (HR Pacific). For the low-resolution datasets, we use the $1^\circ \times 1^\circ$ daily observed means of clouds, aerosol and the environment. For reference, 1° is approximately 110 km. This allows homogenising our observations of clouds and the atmosphere. Schematically, this is the case on the right of Figure 4.1 with pixels of size $1^\circ \times 1^\circ$, and 15 years of data (from 2004 to 2019). For the high-resolution dataset, we use the gridded 25 km $\times$ 25 km resolution of the cloud products from MODIS, and 12 km $\times$ 12 km resolution of the aerosol products from MODIS, and $0.5^\circ \times 0.625^\circ$ resolution from MERRA. Schematically, this corresponds to the case on the left of Figure 4.1 with pixels of size 25 km $\times$ 25 km and 4 acquisitions per day for an entire year (2003).

4.1.3 Pre-processing

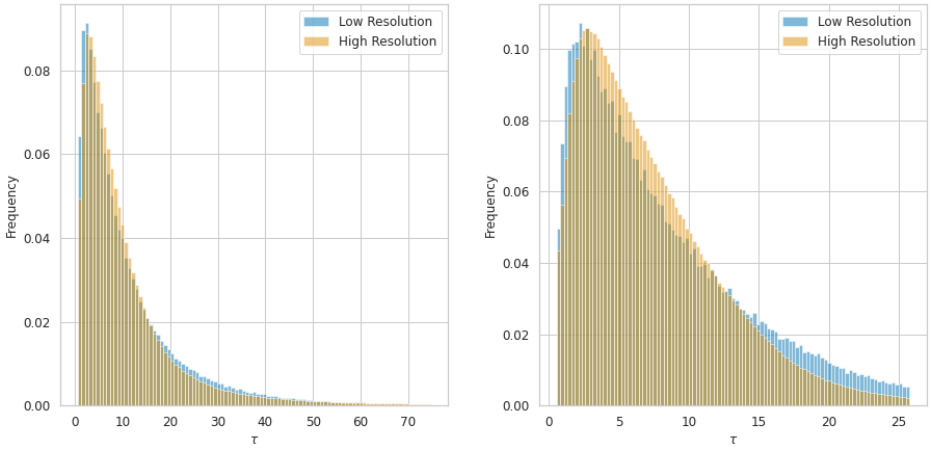
We restrict our observations to clouds in the "aerosol limited" regime by applying some filtering [30]. In "aerosol limited" regimes, we assume that cloud development is limited by the availability of cloud-condensation nuclei, and thus aerosol. Our choice of filtering is informed by domain knowledge and the comparison of the data distributions between the low- and high-resolution datasets. CWP are filtered to values below 250μm and r_e to values below 30μm. AOD values are filtered, only keeping values between 0.03 and 0.3. Below 0.03, no effect is expected, since there is not enough aerosol to lead to changes. Above 0.3, there is a chance of direct effects, whereas the focus of our study is on indirect effects. We also filter out precipitating clouds to avoid a loop in the causal graph. Finally, all features are normalised before being fed into the model.

Figure 4.2 and 4.3 show histograms for cloud water path (CWP) and cloud optical depth (τ) before and after restricting CWP values to below 250μm. We notice that filtering CWP also influences the distribution of τ because they are causally linked. It allows getting rid of the high τ and CWP values in the high-resolution dataset. We moreover study the distributions of mean cloud droplet size r_e and apply further filtering as shown in Figure 4.4. Doing so brings these distributions closer, getting rid of outliers in the high-resolution data.



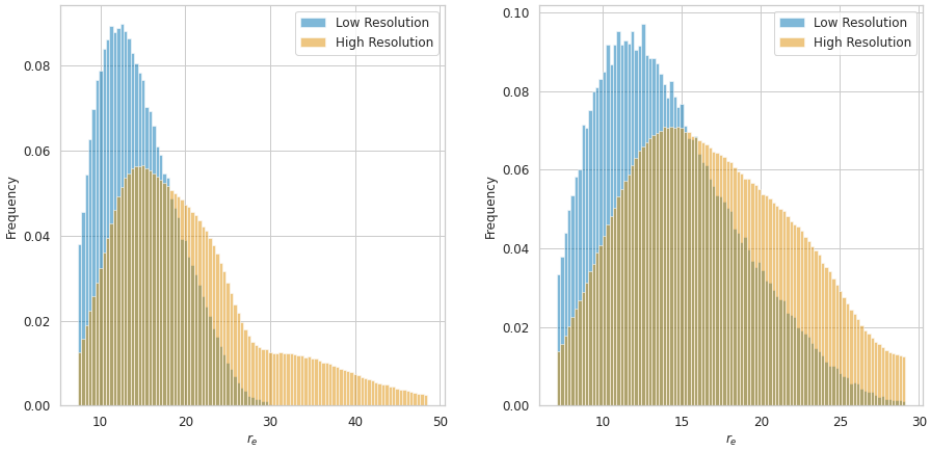
(a) Before filtering (b) After filtering

Figure 4.2: CWP histograms before and after CWP filtering



(a) Before filtering (b) After filtering

Figure 4.3: τ histograms before and after CWP filtering



(a) Before filtering (b) After filtering

Figure 4.4: r_e histograms before and after r_e and CWP filtering

These histograms moreover improve our understanding of the data. In particular, we notice in Figure 4.2 that CWP is layered for the low resolution datasets. This layering comes from the way CWP is computed and aggregated for low-resolution data, relying on assumptions from sensors which lead to CWP not being truly continuous. We do not see these effects for the high-resolution data because no aggregation is needed. These observations enlighten differences between the high- and low-resolution datasets and underlying sources of confounding in the data.

4.2 Overcast methods and models

4.2.1 Model architecture

In what follows, we denote by $\mathbf{x}$ the input data (the meteorological proxies), t the treatment (AOD), and y the output data (the cloud properties). The Overcast models' architecture is described and illustrated in Figure 4.5. The models are neural-network architectures with two basic components: a feature extractor $\phi(\mathbf{x}; \boldsymbol{\theta})$ (represented in **green** in the figure) and a density estimator $f(\phi, \mathrm{t}; \boldsymbol{\theta})$ (represented in **orange** in the figure). There are two different versions of the model: a feed-forward neural network and a transformer which differ in the ability of their feature extractor to capture context. The covariates $\mathbf{x}$ (represented in **blue**) are given as input to the feature extractor, which is concatenated with t (represented in **purple**) and

given as input to the density estimator which outputs $p(y \mid t, \mathbf{x}, \boldsymbol{\theta})$ from which we can sample to obtain samples of the outcomes (represented in **red**).

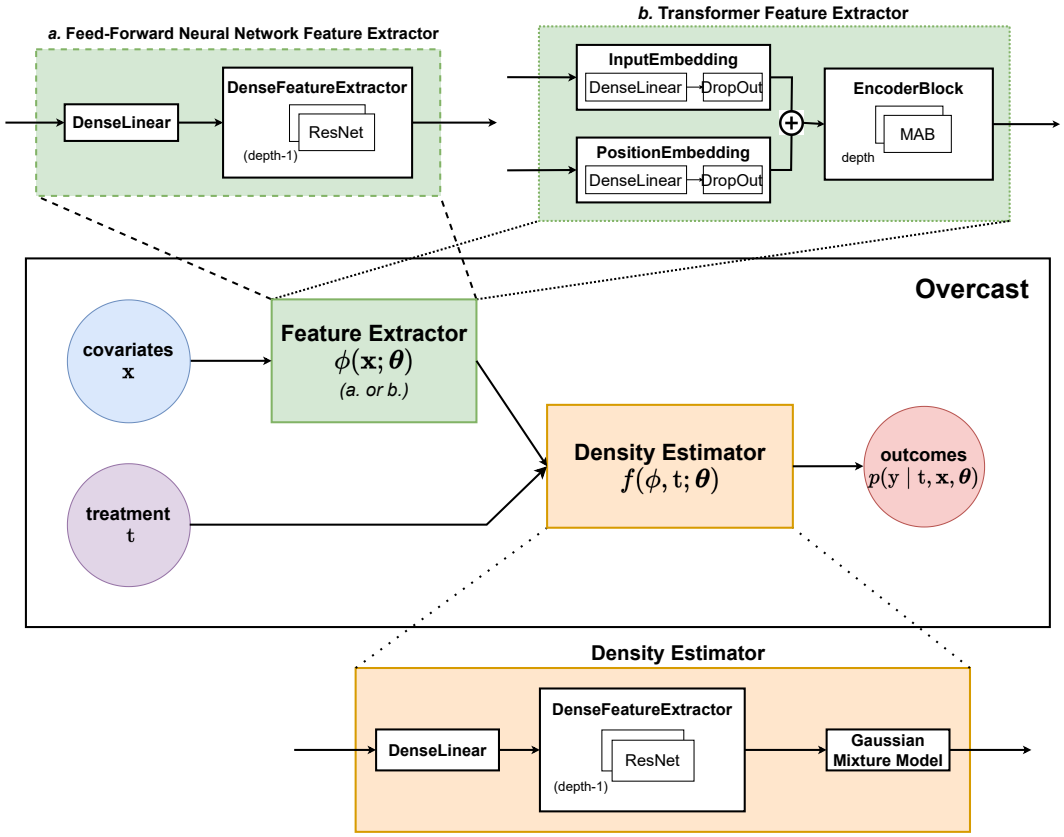


Figure 4.5: Overcast model architecture. The inputs are represented by circles, in **blue** the covariates, in **purple** the treatment. In the **red** circle is the output of the model, the outcomes distribution. The model has different feature extractors (in green) for the feed-forward neural network and the transformer. It has a single density estimator (in orange).

4.2.1.1 Feed-forward neural network feature extractor

The feature extractor for the feed-forward neural network is fairly simple. It takes as input the covariates $\mathbf{x}$, which are fed into a dense linear layer followed by a dense feature extractor. The dense feature extractor is composed of several residual network layers.

4.2.1.2 Transformer feature extractor

The main advantage of the transformer architecture in the feature extractor is that attention allows to model the spatio-temporal correlations between the covariates

on a given day. This is interesting because confounding may be latent in the relationships between neighbouring variables. Typically, environmental processes (which is one source of confounding) are dependent upon the spatial distribution of clouds, humidity and aerosol, and this feature extractor may capture these confounding effects better. It takes as input both the covariates $\mathbf{x}$, and a position vector which includes geographical positions for each pixel in the form of latitude and longitude. The input and position vectors go through separate embeddings which are composed of a dense linear layer followed by a drop-out layer. These embeddings then go through an encoder block, which is composed of several multi-head attention blocks (MAB) as described in [54] and shown in the left grey block in Figure 2.7. The input goes through multi-head attention and is then added to the original input, the result of which is normalised. It then goes through a feed-forward neural network, the results of which are added to the result of the previous step before being normalised.

4.2.1.3 Density estimator

The density estimator is composed of a dense linear layer, followed by a dense feature extractor, and a Gaussian mixture model (GMM) [5]. The dense feature extractor is composed of multiple residual network layers [22]. The Overcast GMM is represented in Figure 4.6. It outputs a Gaussian mixture density, which is an estimator of $p(y \mid t, \mathbf{x})$ with the same number of components as the number of outcomes (n_y) and is of the form:

$$p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{n_y} \tilde{\pi}_j(\phi, \mathrm{t}; \boldsymbol{\theta}) \mathcal{N}\left(\mathrm{y} \mid \tilde{\mu}_j(\phi, \mathrm{t}; \boldsymbol{\theta}), \tilde{\sigma}_j^2(\phi, \mathrm{t}; \boldsymbol{\theta})\right),$$

where $\mathcal{N}(\cdot \mid \mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. Modelling the density allows to sample y values and perform the sensitivity analysis. The mixing coefficients $\boldsymbol{\pi}$ are estimated with a linear layer and a softmax layer, to obtain $\tilde{\boldsymbol{\pi}}$, represented in **blue** in the figure. The vector of means of the Gaussian kernels $\tilde{\boldsymbol{\mu}}$ is obtained by n_y linear layers (in **green** in the diagram), whilst the vector of variances $\tilde{\boldsymbol{\sigma}}$ is obtained by n_y blocks of linear layers and SoftPlus layers (in **orange** in the diagram).

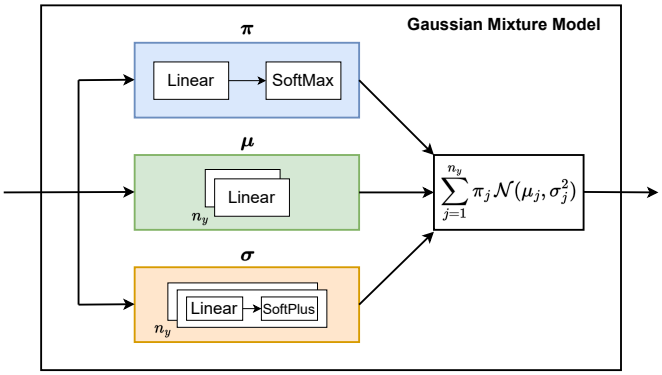


Figure 4.6: Overcast Gaussian mixture model

4.2.2 Making predictions

After model training, at inference time, we perform two main tasks. The first task is to predict the values of the studied outcomes, i.e., the cloud properties. The second task is to do causal inference and estimate the dose-response, or APO curve alongside its confidence intervals. To estimate these quantities, we make use of the following:

$$\tilde{\mu}(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}) = \sum_{j=1}^{n_y} \tilde{\pi}_j(\phi, \mathrm{t}; \boldsymbol{\theta}) \tilde{\mu}_j(\phi, \mathrm{t}; \boldsymbol{\theta}) = \mathbb{E}\left[\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \boldsymbol{\theta}\right].$$

This is the mean of the Gaussian mixture density.

We use bootstrap aggregation, that is, use $n_b = 10$ different models whose parameters are obtained by training on different subsets of the data denoted by $\{\hat{\mathcal{D}}_k\}_{k=1}^{n_b}$. We let $\boldsymbol{\theta}_k$ denote the parameters of the model of the k-th bootstrap sample of the data. The predictions of all n_b estimators are averaged to obtain the final predictions. This allows for improving the accuracy and stability of the predictions made. The predictions for the cloud properties are obtained from:

$$\tilde{\mu}(\mathbf{x}, \mathrm{t}) = \frac{1}{n_b} \sum_{k=1}^{n_b} \tilde{\mu}(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}_k),$$

where $\tilde{\mu}(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}_k)$ is computed across all datapoints characterised by $\mathbf{x}$ and t in the testing set.

To do causal inference, and obtain the APO curve, we also use $\{\tilde{\mu}(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}_k)\}_{k=1}^{n_b}$. Since the amount of data is finite, the treatment cannot be considered truly continuous and has to be quantised. The CAPO $\tilde{\mu}(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}_k)$ is then predicted for

each treatment level and each model parameterised by $\boldsymbol{\theta}_k$. It is then averaged across all n_b models and across all covariates to obtain the APO, which can be plotted as a function of t only.

To quantify the uncertainty presented by finite data and possible violations of the positivity assumption, bootstrapped uncertainty intervals are computed. When unconfoundedness is assumed, the uncertainty intervals are obtained directly from the confidence interval for the confidence level α.

To study possible violations of the unconfoundedness assumption, we work with samples from the density estimator rather than the expected value. The continuous treatment-effect marginal sensitivity model (CMSM) from [26] proposes to quantify the interval of $\mathbb{E}[\mathrm{y} \mid \mathbf{x}, \mathrm{t}]$ compatible with the data and a user-specified relaxation of the unconfoundedness assumption through the parameter Λ. The interval is then obtained by applying the CMSM to the bootstrapped predictions.

4.2.3 Evaluating performance

We use two metrics to evaluate the performance of the Overcast models, linked to the two different tasks: predicting the outcomes and predicting the dose-response. The predicted values can be compared to the observed values by plotting them in a scatter plot. The squared Pearson correlation coefficient r^2 can also be computed, as explained in Chapter 2. The average potential outcome (APO), or dose-response curves can be plotted separately for each outcome against the treatment. Using the CMSM allows us to plot uncertainty around these curves for various levels of hidden confounding. We compare the tightness of the ignorance regions and use domain expert knowledge to evaluate the shape and slope of these curves as there is no ground truth.

4.3 Implementation details

We follow the implementation from the original paper [26]. The code is written in python and is available at `https://github.com/msolal/MT-MLforACI`. The

packages used include PyTorch [36], scikit-learn [41], Ray [33], NumPy, SciPy and Matplotlib.

We use ray tune [31] with HyperBand Bayesian Optimisation [18] search algorithm to optimise our network hyper-parameters. The hyper-parameters considered during tuning are accounted for in Appendix B.1. The final hyper-parameters for each model and each dataset are given in Appendix B.2. The hyper-parameter optimisation objective is the batch-wise Pearson correlation averaged across all outcomes on the validation data for a single dataset realisation with random seed 1331.

We split the data into training, validation, and testing sets across different dates. The original paper splits data in the following way: Mondays to Fridays datapoints are in the training set, Saturdays datapoints are in the validation set, and Sundays datapoints are in the testing set. In our implementation, we keep the same ratio between datasets but we randomise the splits. We do so by using random seed 42 and having 5/7 of the data in the training set, 1/7 in the validation set, and 1/7 in the testing set. The randomisation is motivated by the fact that there is a clear weekly cycle of AOD [11]. Models are optimised by maximising the log likelihood of $p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \boldsymbol{\theta})$.

4.4 Experiments

Unless stated otherwise, we use the low-resolution Pacific data and the experimental setup is the following. The covariates considered are relative humidity at 900, 850 and 700 millibar (RH900, RH850, RH700), sea surface temperature (SST), vertical motion at 500 millibars (ω500) and inversion strengths (with both the lower tropospheric stability, LTS, and the effective inversion strength, EIS). The treatment is aerosol optical depth (AOD). The outcomes of interest are cloud droplet size (r_e), cloud optical depth (τ), cloud water path (CWP) and cloud fraction (CF). For the sake of clarity, we use the following colour code in our result plots: blue when the dataset is low-resolution South-East Pacific, and orange for any variation, for example, high-resolution Pacific data, or low-resolution Atlantic data.

5 Evaluating performance

In this chapter, our work on establishing baselines is reported. With these experiments, we address our first research question about evaluating the performance of Overcast models. We use the off-the-shelf models ridge regression, polynomial ridge regression and multi-layer perceptron to benchmark the performance of Overcast models in Section 5.1. Section 5.2 compares the performance of both Overcast models. The results from this chapter are summarised in Table 5.1 and discussed in Section 5.3.

5.1 Regression baselines

In this section, the results for off-the-shelf methods are presented. Three different models are compared: ridge regression, polynomial ridge regression and a multi-layer perceptron with a single hidden layer. We use implementations from scikit-learn [41], including cross validation which is used to set the regularisation hyper-parameter λ. The data is standardised to have zero mean and unit variance. The models are evaluated by comparing predicted outcomes to observed outcomes. This is done using prediction error plots and the squared Pearson correlation coefficient r^2. Note that these baselines are not causal and therefore do not allow us to study the treatment effect but only predictions of cloud properties. During the project, we also

studied some causal baselines, namely causal forests. These causal baselines consider the treatment as binary rather than continuous and are therefore difficult to compare to Overcast models which is why we don't include them in the present report.

5.1.1 Linear ridge regression

The results for the ridge regression model are in Figure 5.1. As expected, we notice differences in the performance for the different outcomes, that is, certain cloud properties are harder to predict than others. In particular, the cloud coverage or cloud fraction (CF) is harder to predict than the cloud cloud water path (CWP) as shown by the difference in r^2, where $r^2 = 0.154$ for CF and $r^2 = 0.289$ for CWP. This difference is likely to stem from the causal graph and physical processes. In particular, r_e, τ and CWP are physical quantities with underlying mathematical equations linking them all together [20]. CF on the contrary cannot be derived mathematically from other cloud properties. Moreover, further confounding comes from the fact that AOD is better observed for low cloud fraction [29]. Overall, ridge regression performs poorly as shown by the fact that r^2 is low across all outcomes, between 0.150 and 0.290. This could be because there are non-linear relationships between meteorological proxies and cloud properties, or because the covariates do not explain enough of the variance in the outcome. To distinguish between these cases, we perform subsequent experiments.

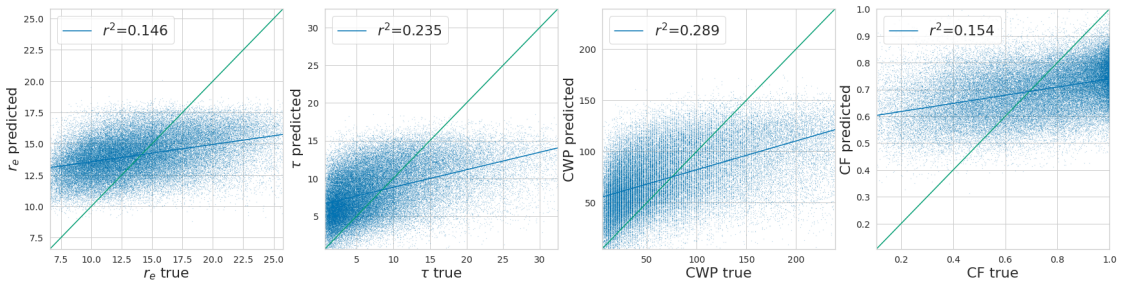


Figure 5.1: Prediction error plot for ridge regression on low-resolution Pacific data. Input RH900, RH850, RH700, LTS, EIS, ω500, SST, AOD. Output r_e, τ, CWP, CF

5.1.2 Polynomial ridge regression

To capture non-linear relationships in the data, polynomial feature expansion with degree 3 is performed before the data is fed into a linear ridge regression model. Our results for this experiment are in Figure 5.2. We choose to set the degree of our polynomial expansion to 3 as a trade-off between performance and runtime. Degree 2 yields slightly worse results with lower r^2 (by approximately 0.1), whereas degree 4 takes longer to run and may lead to overfitting. Overfitting would occur since the features are simply not good enough to explain the variance in outcomes. As expected, we notice improved performance across all outcomes compared to the linear regression baseline. For instance, for predicting the cloud droplet radius r_e, $r^2 = 0.146$ with linear ridge regression and $r^2 = 0.191$ with the polynomial ridge regression.

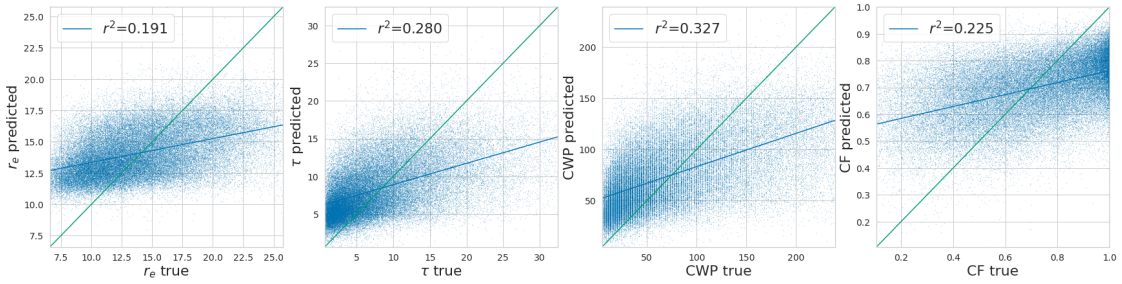


Figure 5.2: Prediction error plot for polynomial ridge regression on low-resolution Pacific data. Input RH900, RH850, RH700, LTS, EIS, ω500, SST, AOD. Output r_e, τ, CWP, CF

5.1.3 Multi-layer perceptron

We build a Multi-Layer Perceptron (MLP) with a single hidden layer with ReLU activation. The results are shown in Figure 5.3. We notice a slight improvement in performance across all outcomes compared to the polynomial ridge regression, as expected. For instance, for the cloud droplet size r_e, $r^2 = 0.213$ for the MLP and $r^2 = 0.191$ for the polynomial ridge regression. The performance could be further improved by increasing the complexity of the model, for example by increasing the number of hidden layers and units.

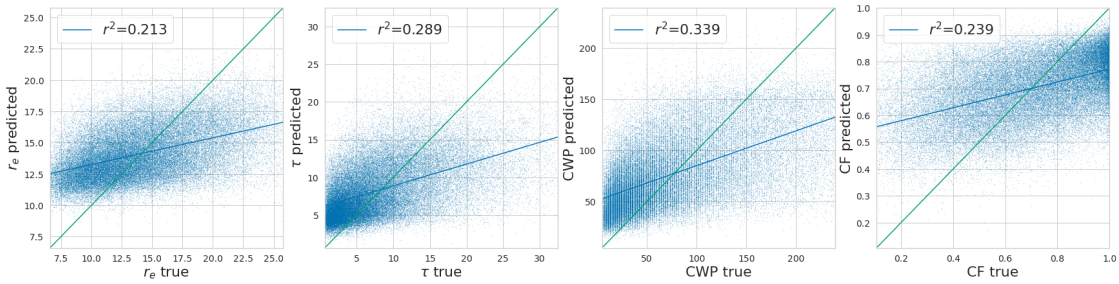


Figure 5.3: Prediction error plot for multi-layer perceptron on low-resolution Pacific data. Input: RH900, RH850, RH700, LTS, EIS, ω500, SST, AOD. Output: r_e, τ, CWP, CF

5.2 Overcast models

In this section, the performance of the Overcast models is studied. Recall that there are two Overcast models which differ in their feature extractor. The Overcast transformer can capture spatio-temporal dependencies in the data through an attention mechanism whereas the feed-forward neural network cannot, as explained in Section 4.2. The models are evaluated based on their predictive accuracy as we did for the baselines, and the shape and underlying uncertainty of the predicted dose-response curves, or average potential outcome (APO). In this section, the results for the transformer are in blue and the neural network are in orange.

5.2.1 Predictive accuracy

Figure 5.4 shows the prediction error plots for both models across all outcomes. In orange are the predictions for the feed-forward neural network, and in blue are the predictions for the transformer. We find that the feed-forward neural network performs similarly to the polynomial ridge regression and the multi-layer perceptron, whilst the transformer performs slightly better. For instance, for the cloud droplet radius, $r^2 = 0.213$ for the multi-layer perceptron, $r^2 = 0.201$ for the Overcast neural network and $r^2 = 0.281$ for the Overcast transformer.

The transformer performs better than the feed-forward neural network because its architecture allows to model spatial dependencies between meteorological proxies thanks to the use of attention mechanisms. This is done by allowing to attend neighbouring pixels to make predictions for a specific pixel as evoked in Section 4.2.

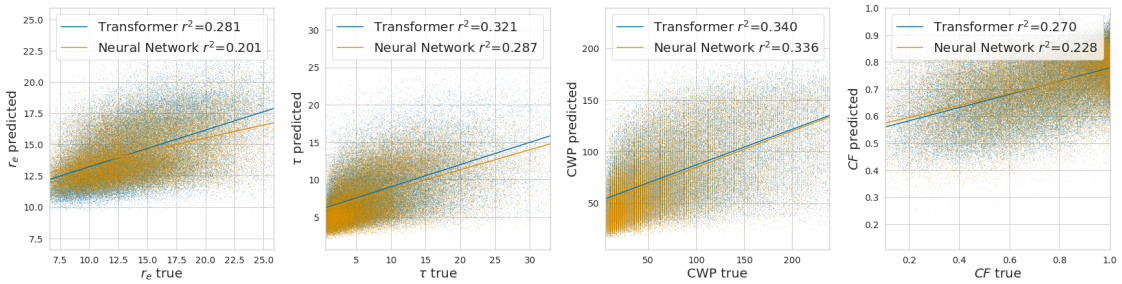


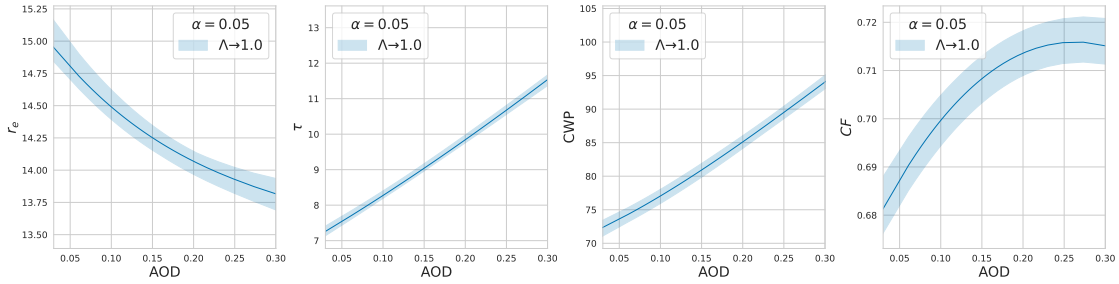
Figure 5.4: Prediction error plot: transformer and feed-forward neural network. Low-resolution Pacific dataset. Covariates: RH900, RH850, RH700, EIS, LTS, SST, W500. Treatment: AOD. Outcomes: r_e, τ, CWP, CF.

Physically, this approach is more reasonable than that of the feed-forward neural network because the pixels are not assumed to be independent.

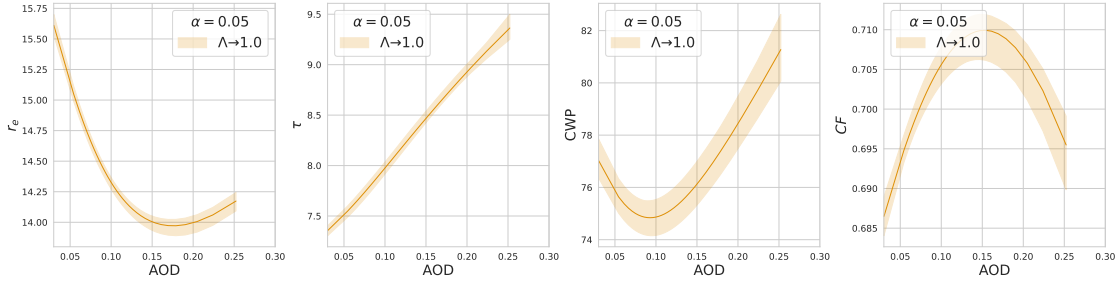
5.2.2 Dose-response curves

The dose-response curves of the two models are then compared. The dose-response curves for all outcomes are shown in Figure 5.5, with the transformer model in blue in Figure 5.5(a) and the feed-forward neural network model in orange in Figure 5.5(b). We find that the transformer's curves agree best with domain knowledge. In particular, the non-monotonicity of the neural network dose-response curves for r_e, CWP and CF are not in accordance with the underlying physical processes. Moreover, CF is highly confounded under our chosen causal graph as evoked earlier and shown by the parabolic shape of its dose-response curve. These results may be indicative of unobserved confounding that the transformer captures better than the neural network by modelling spatio-temporal dependencies. For the remainder of our work, we decide to focus on cloud droplet size r_e to simplify the analysis of the results. This choice is motivated by the fact that the effect of aerosol on r_e is described by the direct Twomey effect which is well understood.

To analyse the dose-response curves, we display them on the same plot in Figure 5.6. Recall that, to study a dose-response curve, we are interested in three aspects: (1) its shape, (2) its uncertainty bounds, and (3) its slope. Figure 5.6(a) allows to investigate the former two aspects. To compare the slopes of two curves,



(a) Overcast Transformer



(b) Overcast Feed-Forward Neural Network

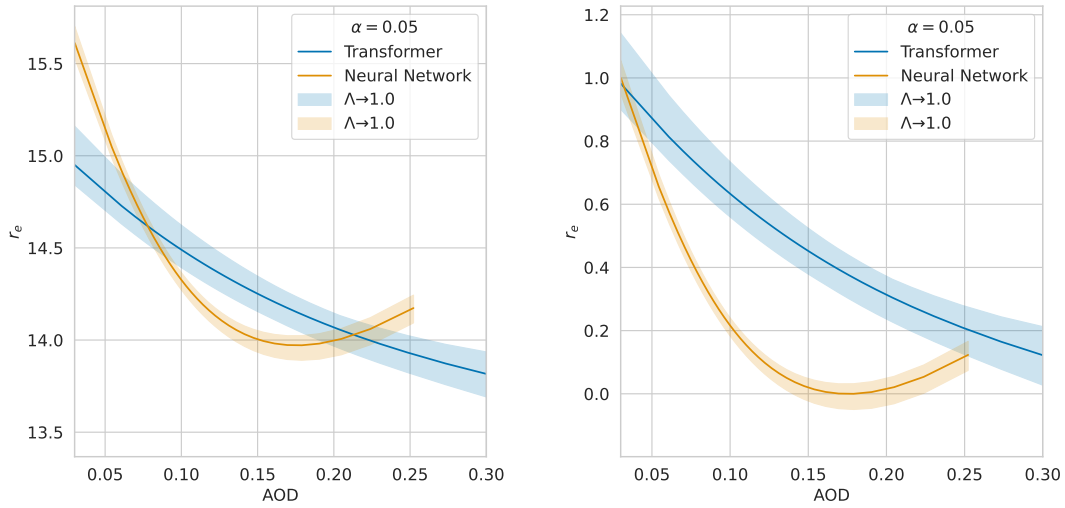
Figure 5.5: Dose-response curves on low-resolution Pacific data: transformer and feed-forward neural network. Low-resolution Pacific dataset. Covariates: RH900, RH850, RH700, EIS, LTS, SST, W500. Treatment: AOD. Outcomes: r_e, τ, CWP, CF.

we can min-max scale them:

$$x \mapsto \frac{x - \min(x)}{\max(x) - \min(x)},$$

which scales the data from domain $[\min(x), \max(x)]$ to codomain $[0, 1]$. Figure 5.6(b) shows the scaled dose-response curves and allows to compare the slope of the curves.

As expected from the results shown in Figure 5.5, we notice that the transformer's dose-response curve agrees better with domain knowledge both in terms of shape and in terms of slope. Further, the uncertainty bounds for the transformer are larger and therefore worse than that of the neural network. This is because the transformer feature extractor considers all pixels from a day as a single datapoint whereas the neural network considers every single pixel as a datapoint. It suggests that using more data, specifically data from more days, may reduce uncertainty.



(a) Raw

(b) Scaled by min-max

Figure 5.6: Dose-response curves for r_e: transformer and feed-forward neural network. Low-resolution Pacific dataset. Covariates: RH900, RH850, RH700, EIS, LTS, SST, W500. Treatment: AOD. Outcomes: r_e.

5.3 Discussion

We summarise the results for all the experiments from this chapter in Table 5.1.

Model	Squared Pearson coefficient r^2			
	r_e	τ	CWP	CF
Linear Ridge Regression	0.146	0.235	0.289	0.154
Polynomial Ridge Regression	0.191	0.280	0.327	0.225
Multi-Layer Perceptron	0.213	0.293	0.338	0.238
Overcast Feed-Forward Neural Network	0.201	0.287	0.336	0.228
Overcast Transformer	0.281	0.321	0.340	0.270

Table 5.1: Prediction accuracy r^2 for baseline and Overcast models. Dataset: low-resolution Pacific. Input: RH900, RH850, RH700, LTS, EIS, ω500, SST. Treatment: AOD. Output: r_e, τ, CWP, CF.

These experiments allow us to answer our first research question about the evaluation of Overcast models. We show that the more complex the model, the better the prediction accuracy, as expected. We find that the Overcast models perform slightly better than baseline models for all outcomes. The predictive accuracy is still quite low, with $r^2 \simeq 0.3$ so future works should focus on improving these predictions.

These experiments and observations moreover suggest that the transformer is the

better model. Whilst the uncertainty levels are higher, it agrees best with domain knowledge in terms of dose-response curves. More importantly, modelling context using attention mechanisms allows reducing sources of confounding and makes more sense physically. Let us consider this fact more carefully and attempt to provide some intuition. Recall that our low-resolution data consists of grids of daily means of observations per pixel identified through their geographical coordinates (with latitude and longitude). The transformer allows capturing dependencies between the different cells of the grid, that is, between neighbouring pixels whilst the feed-forward neural network considers each pixel or cell independently. Physically, it makes more sense to model spatial dependencies in that way since physical processes are continuous and boundaries between pixels have no physical meaning. Future work should consider improving the attention mechanism to also capture temporal dependencies.

In our experiments, we moreover notice significant differences in the predictive accuracies for the different outcomes. The relative performance for each outcome remains in the same order, with CWP being the easiest to predict, then τ, then CF, and finally r_e. This suggests that the magnitude of the confounding differs between outcomes. In other words, the meteorological proxies that we use as covariates allow explaining the variance in certain outcomes better than others.

In the experiments that follow, we focus on the Overcast transformer. Since the transformer takes very long to be trained, tuned and evaluated, we also use the degree 3 polynomial ridge regression for additional experiments. It is the best trade-off between predictive accuracy and runtime, and side experiments are helpful to improve our understanding of the topic.

6

Capturing geographical dependencies

In this chapter, our experiments on geographical dependencies are presented. We work with various datasets which differ in terms of geographical regions and spatio-temporal resolution. This work allows us to address our second research question regarding the ability of the Overcast transformer to capture geographical dependencies. This chapter is divided into four sections. Our motivations are explained in Section 6.1. Section 6.2 compares data with observations from two distinct regions: the South-East Pacific and the South Atlantic. Section 6.3 reports our work on different spatio-temporal resolutions. This chapter ends with a discussion in Section 6.4.

6.1 Motivations

Clouds are inherently local and interconnected, with a scale of interactions on the order of kilometres. Models therefore need to capture geographical dependencies to accurately predict cloud properties and emulate ACI. To study the Overcast transformer's ability to capture geographical dependencies, we compare its performance on data of different geographical regions and scales. The datasets have different levels of unobserved confounding. For instance, the Atlantic and the Pacific regions are known to have different types of aerosols. The low-resolution data may be

averaging cloud types, thus obscuring the signal compared to the high-resolution data. Our causal graph does not include these confounders directly, since they are impossible to retrieve from satellite data. Instead, they are accessed through meteorological proxies and aerosol optical depth (AOD). With these experiments, we empirically study how well the Overcast models capture these confounders by comparing their performance on the different datasets. This work contributes to our evaluation of the models as we assess their ability to generalise.

6.2 Geographical regions

In this section, our work on geographical regions is presented. We study two datasets, which respectively contain observations from the South-East Pacific and the South Atlantic.

To best understand our experiments and results, let us describe the physical differences between the two regions. The South-East Pacific and South Atlantic are regions with similar meteorology but different confounding influences. The possible confounding factors include aerosol type, aerosol hygroscopicity, aerosol size, and others, which notably impact aerosols' activation as cloud condensation nuclei. For instance, the South Atlantic has a higher concentration of sea salt and sand aerosols than the South-East Pacific which is a more polluted region. Consequently, aerosol optical depth (AOD) may be a better proxy for aerosols in certain regions. The Twomey effect of aerosol on cloud droplet radius r_e informs us that we can expect similar effects and effect sizes across both regions, and thus it makes sense to compare dose-response curves [53].

Structurally, the two datasets are the same and consist of $1^\circ \times 1^\circ$ daily means of satellite observations between 2004 and 2019. We therefore rely on the same covariates, treatment, and outcomes for both regions. In what follows, results for the South-East Pacific and the South Atlantic are respectively in blue and orange and were obtained using the Overcast transformer.

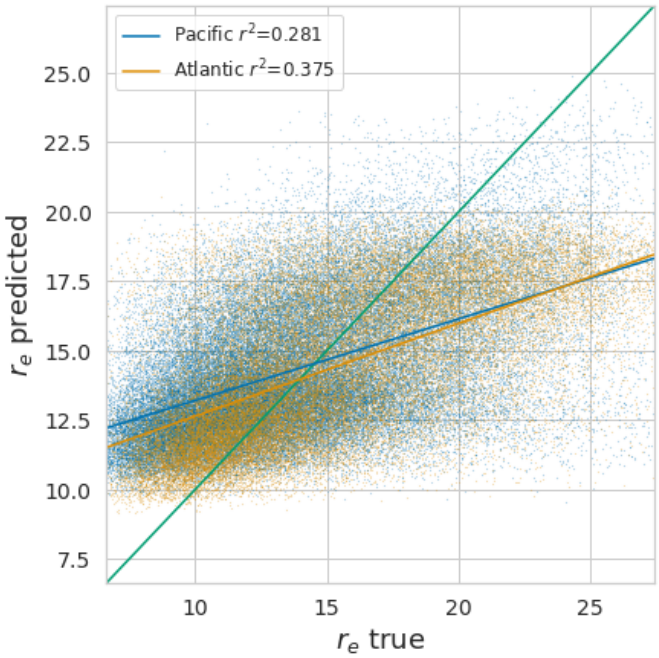
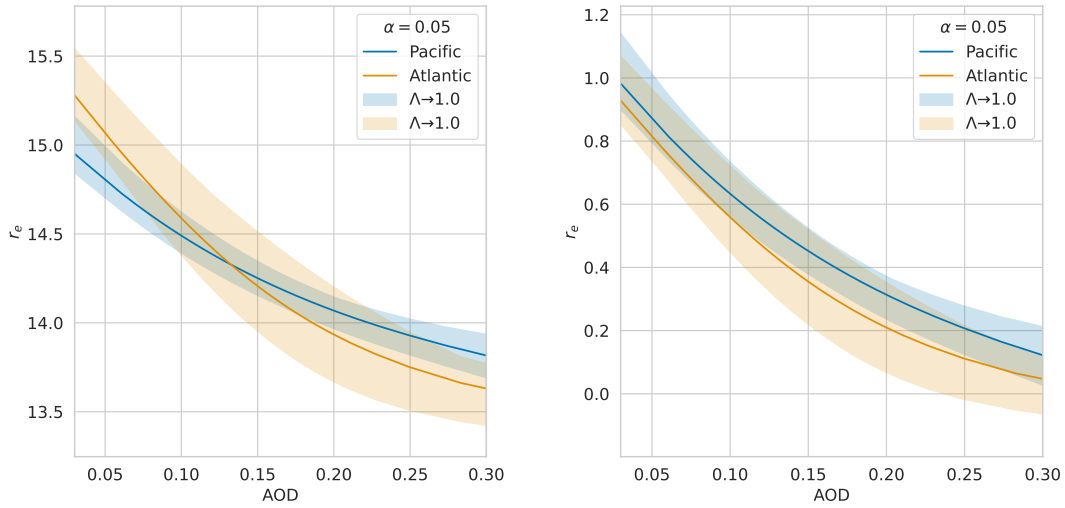


Figure 6.1: Prediction error plot for r_e: South Atlantic and South-East Pacific. Overcast transformer on low-resolution datasets. Input RH900, RH850, RH700, LTS, EIS, ω500, SST, AOD. Output r_e

6.2.1 Results

Figure 6.1 shows the prediction errors plots for both regions. We find that the model performs significantly better on data from the Atlantic, with $r^2 = 0.375$ than on data from the Pacific, with $r^2 = 0.281$. This suggests that our chosen meteorological proxies better explain the variance in cloud droplet radius in the Atlantic than in the Pacific.

Next, the dose-response curves of the two datasets are considered. These curves are shown in Figure 6.2, with the raw curves in Figure 6.2(a) and the min-max scaled curves in Figure 6.2(b). We notice that the uncertainty bounds are larger for the Atlantic compared to the Pacific and that the range of outcomes is smaller in the Atlantic, with values ranging from 15.2 to 13.7, than in the Pacific, with values between 15.5 and 13.4. We also observe that the slopes of the curves are very similar, as expected from domain knowledge.



(a) Raw

(b) Scaled by min-max

Figure 6.2: Dose-response curves for r_e: South Atlantic and South-East Pacific. Overcast transformer, low-resolution datasets. Covariates: RH900, RH850, RH700, EIS, LTS, SST, W500. Treatment: AOD. Outcomes: r_e.

6.2.2 Discussion

Overall, this experiment shows that the model generalises well to different geographical regions. We see this through the fact that r^2 is similar for both regions, and slightly better for the Atlantic than the Pacific for cloud droplet radius. The APO curves moreover have very similar slopes, as expected.

We notice better performance for the South Atlantic data compared to the South East Pacific. This means that the meteorological proxies and AOD allow us to better capture variance in cloud droplet radius in the Atlantic than in the Pacific. This could be explained by the existence of larger-scaled physical interactions in the Pacific than the Atlantic that our model is unable to capture with our chosen meteorological proxies. Namely, the Pacific is subject to tropical effects and different circulations patterns compared to the Atlantic [35]. Another possible explanation of these differences concerns the type of aerosols. In the South Atlantic region, dust and sand are prevalent aerosols, whereas in the Pacific there are more anthropogenic aerosols. Together, the meteorological conditions, the differences in aerosol types, and the differences in cloud types may impact how well aerosol optical depth

approximates aerosol concentrations and their effects.

6.3 Spatio-temporal resolution

In this section, our work on data with different spatio-temporal resolutions is presented. Both datasets contain observations from the South-East Pacific but differ in three main ways: spatial resolution, temporal resolution and timescale. The low-resolution data contains daily means of satellite observations gridded at $1° \times 1°$ resolution. The high-resolution data contains 4 acquisitions of satellite observations per day which are not aggregated, and are gridded at approximately 25km $\times$ 25km resolution. Moreover, the two datasets do not cover the same timescale as the low-resolution data spans from 2004 to 2019 whereas the high-resolution data is from 2003. We refer the reader to Section 4.1 for more detail.

Our motivation to compare these datasets stems from our will to investigate the ability of the Overcast models to capture geographical dependencies and from the original paper [26]. The experiments presented therein make use of the low-resolution Pacific dataset. In their discussion of the results, the authors suggest that the resolution of the observations could be averaging cloud types and obscuring the signal. They hint that higher resolution data could resolve some confounding influences.

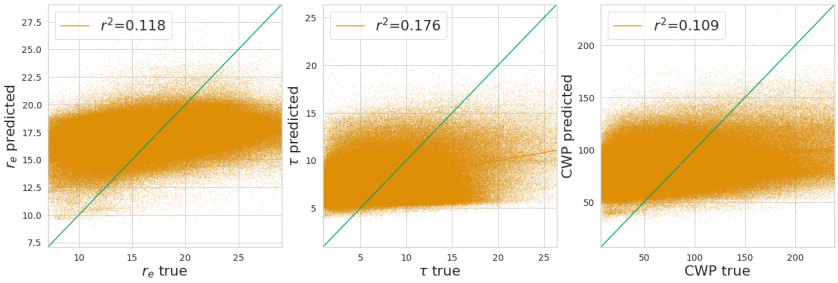
When working with high-resolution data, we consider slightly different variables. In terms of covariates, RH900 is replaced by RH950, and EIS is omitted. We expect the replacement of RH900 by RH950 to have little effect on the result, but removing EIS is expected to have a larger impact. In terms of outcomes, CF is omitted and N_d is added. We do these changes since not all meteorological proxies and cloud properties measurements are available for both datasets. To summarise, the covariates are therefore relative humidity (RH950, RH850, RH700), sea surface temperature (SST), vertical motion (ω500) and inversion strength (LTS), and the outcomes of interest are cloud droplet number (N_d), mean cloud droplet size (r_e), cloud optical depth (τ) and cloud water path (CWP).

6.3.1 Results

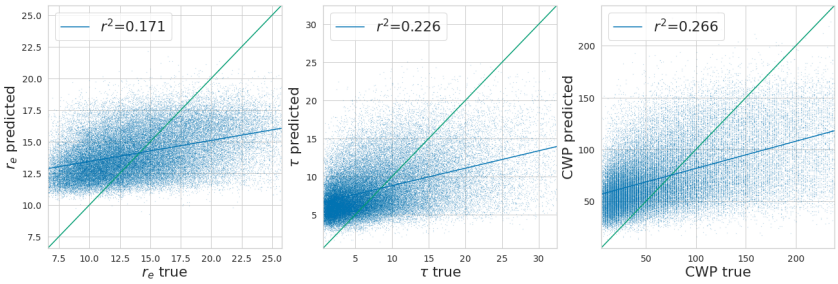
When running the Overcast transformer on the high-resolution data, we find extremely low predictive accuracy for the cloud droplet radius with $r^2 = 0.014$, which is indicative of a bug. Moreover, given the large amount of data and memory requirements, we were not able to plot the dose-response curve. Preliminary results with anterior versions of the Overcast models however confirm that these models perform worse on high-resolution data and so do baseline models. The results for the degree-3 polynomial regression are shown in Figure 6.3, with high-resolution in Figure 6.3(a) and low-resolution in Figure 6.3(b), where we select the same covariates for both datasets. We notice even more difference between the datasets if we also include RH950 and RH900 to the covariates for the high- and low-resolution datasets, as shown in Figure 6.4, with high-resolution in Figure 6.4(a) and low-resolution in Figure 6.4(b). We notice lower accuracy with high-resolution data for all outcomes, especially cloud optical depth and cloud water path. This result is surprising and there is high motivation to understand it better in the hope of improving performance. We therefore perform subsequent experiment to investigate the performance gap.

6.3.2 Investigating the performance gap

The following experiments investigate the performance gap between the high and low-resolution data. Our analysis of the aforementioned results is hindered by the threefold difference between datasets. We therefore attempt to build a more controlled environment in the following experiments, studying the different factors separately to increase our understanding of the relative importance of each factor. We use the degree-3 polynomial ridge regression baseline model throughout this study since the Overcast models require a long running time. We assume that these results are representative of the results we would obtain with the Overcast models. Since we rely on regression baselines, we can only study the prediction error and not the treatment effect.

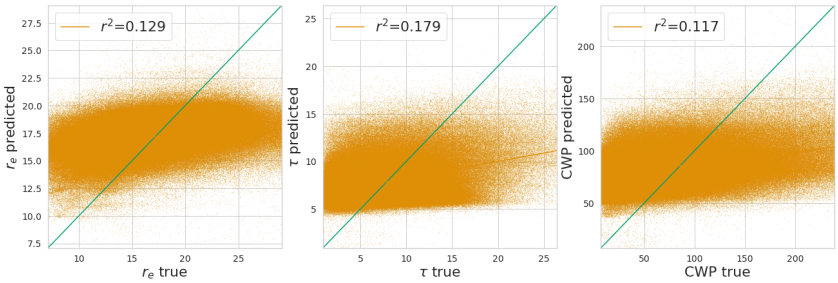


(a) High-resolution

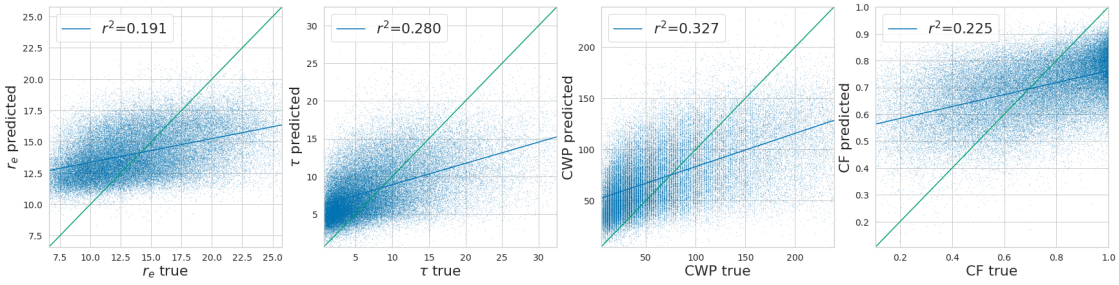


(b) Low-resolution

Figure 6.3: Prediction error plot: high-resolution and low-resolution Pacific datasets with the same covariates. Degree 3 polynomial ridge regression model. Covariates: RH850, RH700, LTS, SST, W500. Treatment: AOD. Outcomes: r_e, τ, CWP.



(a) High-resolution



(b) Low-resolution

Figure 6.4: Prediction error plot: high-resolution and low-resolution Pacific datasets with similar covariates. Degree 3 polynomial ridge regression model. Covariates: RH950 (for high-resolution) or RH900 (for low-resolution), RH850, RH700, LTS, SST, W500. Treatment: AOD. Outcomes: r_e, τ, CWP.

6.3.2.1 Treatments

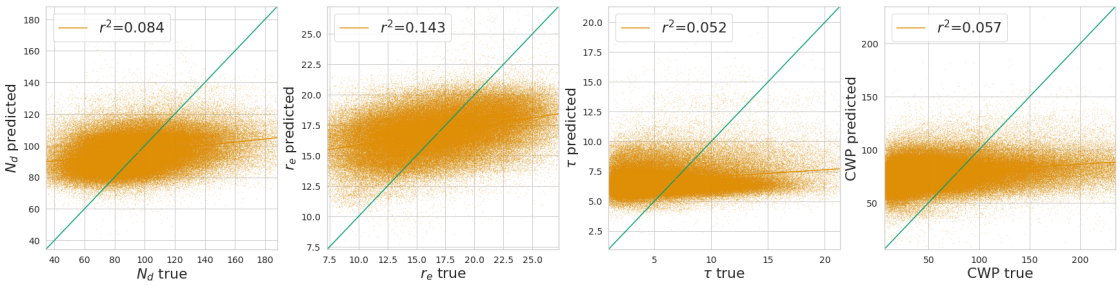
In this section, different treatments are compared, by adding them successively to the list of covariates: aerosol optical depth (AOD), cloud droplet concentration (N_d) and mean cloud droplet size (r_e). These variables intervene at different moments in the causal chain: an increase in AOD leads to an increase in CCN which leads to an increase in N_d which leads to a decrease in r_e. The next events in the causal chain are the changes in cloud optical depth (τ), cloud water path (CWP) and cloud fraction (CF). Because of the causal relationships between the variables, we can consider AOD, N_d and r_e as various proxies for aerosol. With this experiment, we study how additional features explain the variance in cloud properties better than AOD in the high-resolution Pacific dataset.

Figure 6.5 shows the results obtained for this experiment, namely the prediction error plots, with AOD as treatment in Figure 6.5(a), N_d in Figure 6.5(b), and r_e in Figure 6.5(c). There is a significant improvement in performance especially for predicting cloud optical depth (τ), with $r^2 = 0.498$ with N_d as treatment instead of 0.052 with AOD as treatment, and even larger with r_e as treatment where $r^2 = 0.917$.

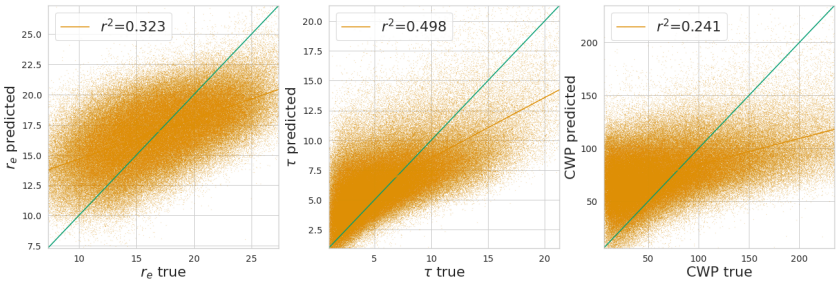
We find that the further in the causality chain the better the predictive power. By moving further in the causal chain, proxies have less influence and the observations are less confounded which helps make better predictions. We get improved predictions with N_d as treatment and best predictions with r_e as treatment. However, we note that considering N_d as treatment induces bias because N_d is computed from r_e and τ during the re-analysis in the current datasets.

6.3.2.2 Timescale

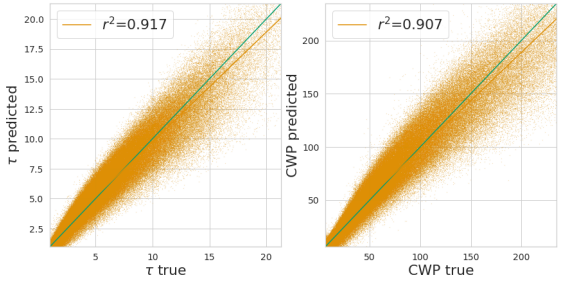
We study fifteen years of data in the low-resolution data (from 2004 to 2019), but only one year in the high-resolution data (from 2003). To best compare the impact of resolution on performance, we study the importance of timescale for the predictive accuracy of the models. We do so for both the low-resolution and the high-resolution data.



(a) AOD



(b) N_d

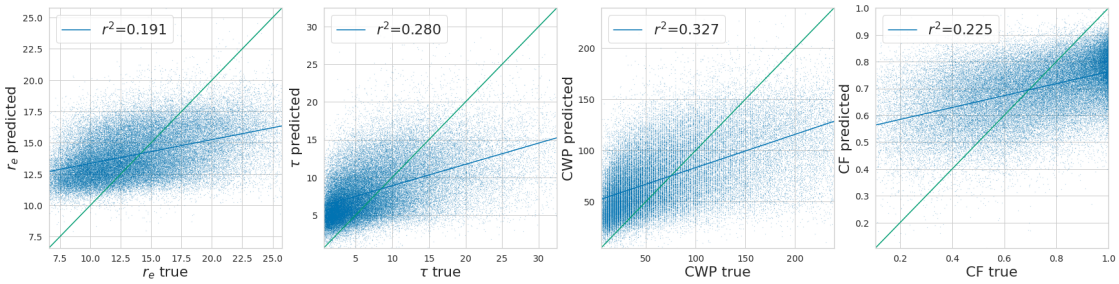


(c) r_e

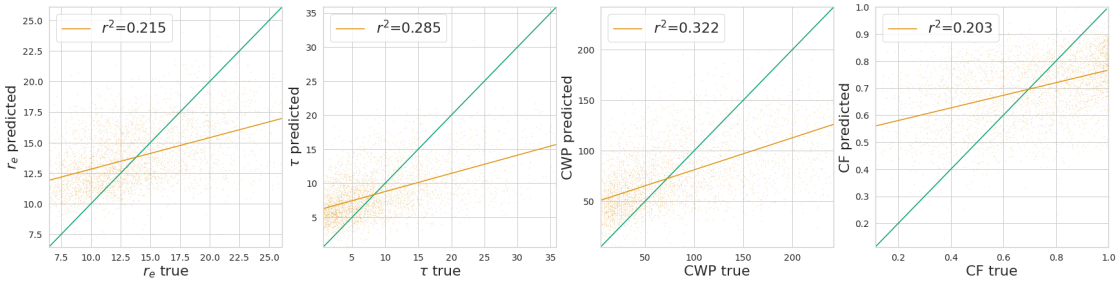
Figure 6.5: Prediction error plot for polynomial ridge regression: AOD, N_d and r_e as treatments. high-resolution Pacific data. Covariates: RH950, RH850, RH700, LTS, ω500, SST. Outcomes: N_d, r_e, τ, CWP

The results for the low-resolution data are in Figure 6.6. In Figure 6.6(a) are results for the control experiment, with the entire timescale from 2004 to 2019. Figure 6.6(b) contains results when the models are run on data from 2004 only. We notice that we obtain similar performance across all outcomes. For instance, the predictive accuracy for cloud optical depth τ is $r^2 = 0.285$ for a single year and $r^2 = 0.280$ for all 15 years.

Figure 6.7 shows the results for a similar experiment with the high-resolution data. We compare using data from January 2003 to July 2003 in Figure 6.7(a) to using data from the entire year of 2003 in Figure 6.7(b). We notice that we



(a) 2004 - 2019



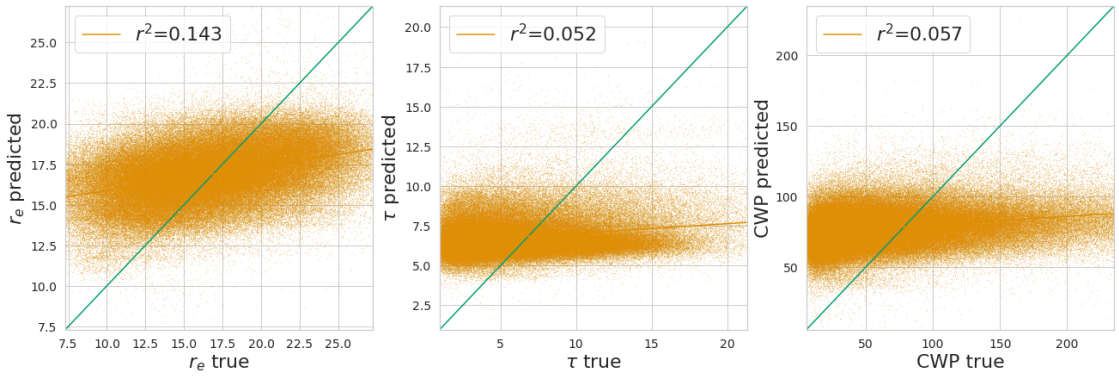
(b) 2004

Figure 6.6: Prediction error plot for polynomial ridge regression on low-resolution Pacific data: 2004 and 2004-2019 timescale. Covariates: RH900, RH850, RH700, LTS, EIS, ω500, SST. Outcomes: r_e, τ, CWP, CF

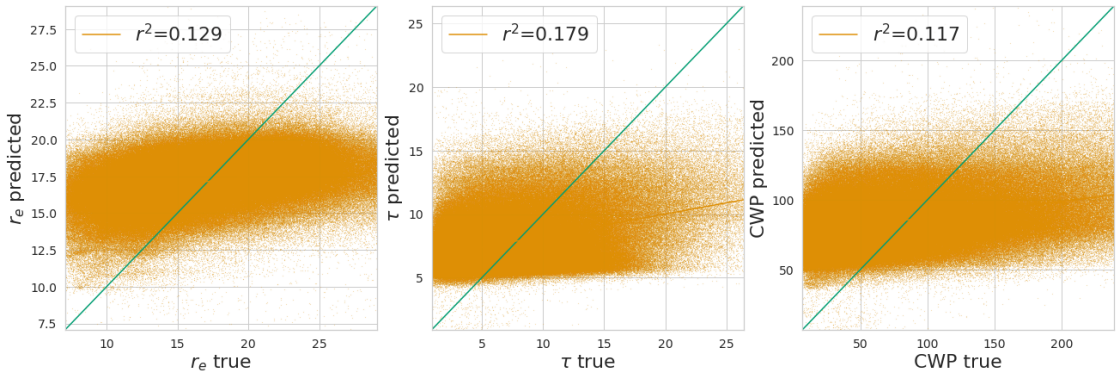
obtain similar performance in predicting the cloud droplet radius r_e and improved accuracy for τ and CWP. For instance, the predictive accuracy for τ is $r^2 = 0.052$ for 7 months, and $r^2 = 0.179$ for the entire year. This suggests that given more data, and specifically, data from a larger timescale, the performance would improve.

6.3.2.3 Discussion

The threefold difference between the high- and low-resolution datasets hinders our analysis. The timescale experiment motivates further experimentation with longer timescales for high-resolution data in the hope that the prediction accuracy would improve. It would also be interesting to study more precisely the influence of the spatial resolution of the data. In particular, the high-resolution data could be aggregated into daily means, leading to the only difference between datasets being their spatial resolution. Unfortunately, due to time constraints and the amount of time needed for the models to run, we were quite limited in the number of experiments that we could perform. Another interesting aspect to consider is the



(a) January 2003 - July 2003



(b) January 2003 - December 2003

Figure 6.7: Prediction error plot for polynomial ridge regression on high-resolution Pacific data: January 2003 - July 2003 and January 2003 - December 2003 timescale. Covariates: RH950, RH850, RH700, LTS, ω500, SST. Outcomes: r_e, τ, CWP

cloud types. For example, the high-resolution data is less likely to contain types of clouds with a smaller radius like cumulus or cirrus. Maybe adding more filtering on cloud types would help with performance. The experiments on treatments suggest that an alternative framework could be used to predict cloud properties and better capture the underlying causal chain. For instance, we could use an auto-regressive model, where r_e would be used in addition to the covariates and the treatment to predict τ, which would then be used in addition to the previously used covariates to predict CWP.

6.3.3 Alternative model architecture

To improve the performance of the model, we thought about changing another aspect of the architecture, namely the attention-based feature extractor. We would make use of both the low- and high-resolution data, with high-resolution covariates, both high- and low-resolution position vectors which include acquisition time, and low-resolution treatment and outcomes. The trick is that the high-resolution and low-resolution data do not have the same dimensions since the high-resolution data contains multiple observations for each location in space for a single day whereas the observations from the low-resolution data are daily means. We hypothesise that appending a multi-head attention block with positional encoding from the high-resolution data would help us combine the low- and high-resolution data to make predictions. This would be used instead of the transformer feature extractor shown in green and labelled (b) in Figure 4.5. Unfortunately, due to the time constraint, we did not have time to implement this idea, but a diagram is shown in Figure 6.8.

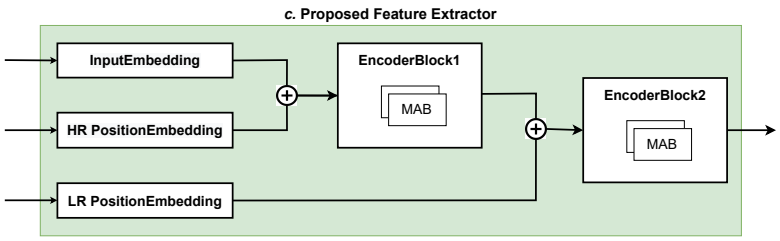


Figure 6.8: Alternative architecture for the Overcast attention-based feature extractor. Uses both low-resolution and high-resolution data, and multiple attention blocks to capture both spatial and temporal dependencies.

6.4 Discussion

The results for all the experiments from this chapter are summarised in Table 6.1.

This work allows us to answer the second research question regarding the ability of the Overcast models to capture geographical dependencies. We find that the Overcast transformer captures emulates ACI better in the Atlantic region than in the Pacific but fails to predict cloud properties using high-resolution data rather than low-resolution data.

Dataset	Covariates	Squared Pearson r^2		
		r_e	τ	CWP
High Resolution Low Resolution	RH850, RH700, LTS, SST, ω500, AOD			
High Resolution	RH950, RH850, RH700, LTS, SST, ω500, AOD	0.143	0.052	0.057
	RH950, RH850, RH700, LTS, SST, ω500, AOD, N_d	0.323	0.498	0.241
	RH950, RH850, RH700, LTS, SST, ω500, AOD, N_d, r_e	/	0.917	0.907
LR 2004	RH900, RH850, RH700, LTS, EIS, SST, ω500, AOD	0.215	0.285	0.322
LR 2004 - 2019		0.191	0.280	0.327
HR Jan - Jul 2003	RH950, RH850, RH700, LTS, SST, ω500, AOD	0.143	0.052	0.057
HR 2003		0.129	0.179	0.117

Table 6.1: Prediction accuracy r^2 for experiments on geographical dependencies. Model: degree 3 polynomial ridge regression. Experiments: treatments, timescale on low-resolution data (LR) and timescale on high-resolution data (HR).

In these experiments, we also touch upon various sources of confounding including cloud types, aerosol types, and the imperfectness of AOD as a proxy for aerosols.

Whilst the threefold difference between the high and low-resolution datasets hinders our analysis, the results suggest that using more data from more days to do predictions may improve the prediction accuracy of the high-resolution data. Furthermore, we touch upon possible changes to our model architecture to better capture geographical dependencies. The first one consists in using both low- and high-resolution data to make predictions and to allow the feature extractor to capture temporal dependencies in addition to spatial dependencies. Secondly, we suggest not assuming independence between outcomes maybe by using an auto-regressive framework where predictions for outcomes variables are made using the predictions for variables anterior in the causal chain.

7

Uncertainty-aware sensitivity analysis

In this chapter, our uncertainty-aware sensitivity analysis is described. A sensitivity analysis aims to uncover how sensitive a model is to the addition of one or more variables. We study how robust our estimates are to the violation of assumptions and derive bounds on the ignorance induced for a given degree of violation of these assumptions. The width of the interval of possible causal effects increases as the assumptions are challenged more severely. This work focuses on the relaxation of the unconfoundedness and the positivity assumptions. We pursue our investigation of sources of confounding from Chapter 6 with new experiments and attempt to provide a methodology for setting the parameter Λ proposed by [26].

Section 7.1 outlines our motivations and Section 7.2 explains our experimental setup and gives some theoretical background. Section 7.3 describes our first experiment which consists in omitting covariates thus introducing known levels of confounding in the model. The second experiment, presented in Section 7.4 is an extension to our work with the South Atlantic and South-East Pacific regions.

7.1 Motivations

Unobserved confounding variables are unobserved variables that affect both the treatment and the outcome. They are the variables which violate the unconfound-

edness assumption required to identify the CAPO and APO from observational data. As explained in Section 2.2, the most common approach to respect this assumption is to control for these confounding variables by setting them as covariates and conditioning on them. Doing so however leads to stronger violations of the positivity assumption when working with finite data.

The Overcast models make use of expert knowledge about ACI to select the covariates. Ideally, they would include pressure profiles, temperature profiles and supersaturation since these are directly involved in cloud processes and impact the quality of AOD measurements as a proxy for aerosol concentration. Unfortunately, they are impossible to retrieve from satellite data, so we rely on meteorological proxies like relative humidity, sea surface temperature, inversion strengths, and vertical motion.

The overarching aim of this work is to better understand uncertainties lying in models that study the impact of emissions on cloud properties. This fits in the grander scheme of modifying climate models to increase confidence in future projections of climate change.

7.2 Experimental setup

Recall the parameter Λ proposed by the continuous treatment-effect marginal sensitivity model (CMSM) in [26]. This parameter is set by the user to represent a belief in a certain level of violation of the unconfoundedness assumption. The idea is to relate unconfoundedness violations to the proportion ρ of the unexplained range in outcomes coming from unobserved confounders after observing the covariates $\mathbf{x}$ and the treatment t. When a user sets Λ to 1, they assume that the model has no hidden confounding, which means that the entire unexplained range of Y comes from unknown mechanisms independent of the treatment. As the user increases Λ, they attribute some of the unexplained range of outcomes to mechanisms causally connected to the treatment.

The original paper highlights the difficulty in interpreting and setting Λ. They propose a methodology where the user would sweep over values of Λ and report

bounds corresponding to a ρ value they deem tolerable. They also relate Λ to the Kullback-Leibler divergence between $\mathbb{P}[\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}]$ and $\mathbb{P}[\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}]$.

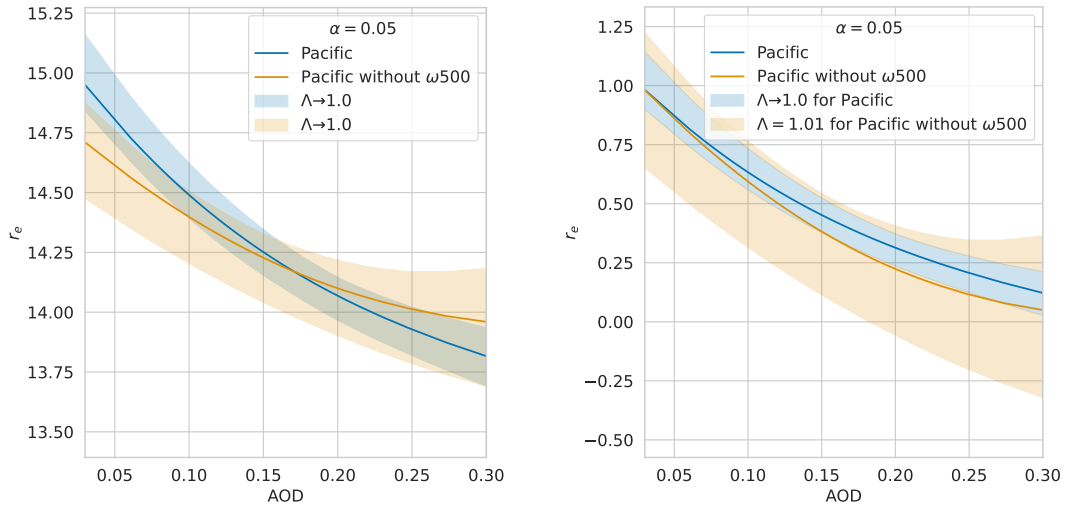
The following work attempts to set a new methodology for setting Λ. We work with two datasets, used to train two different models: (i) and (ii). The model (i) is our control model, trained on the entire low resolution Pacific data, whereas the model (ii) is our experiment. After training both models, we plot the dose-response curves for (i) and (ii) on the same plot. We can compare the shape and slope of these curves as well as their uncertainty bounds under the unconfoundedness assumption by plotting the ignorance region for $\Lambda \to 1$ for both models. Then, we are interested in setting Λ for model (ii) such that the uncertainty bounds cover the entire ignorance region of model (i) under the unconfoundedness assumption. For this, we are interested in comparing the slopes and thus min-max scale both curves.

7.3 Omitting covariates

By omitting covariates, we are removing a confounding variable from the model, therefore increasing the amount of unobserved confounding. We expect worse predictive power and changes to the APO curves and their confidence intervals. Thanks to expert knowledge, we can grasp the importance of the confounding variables we are omitting. We expect that the importance of a covariate is reflected in the distance between the predicted APO and the ground truth APO, with worse curves for important covariates. Removing covariates also results in less violation of the positivity assumption and therefore tighter uncertainty bounds around the APO since more data is available to predict $p(\mathrm{y} \mid \mathbf{x}, \mathrm{t})$. This experiment helps us gain some intuition about the influence of the parameter Λ and how it relates to the inclusion of confounding variables in the model. We perform two experiments: first omitting vertical motion at 500 millibars (ω500), a variable of moderate importance, and secondly omitting all relative humidity variables, which are highly important variables. The results of the control experiment with the low-resolution Pacific data are in blue, and the results for the experiments where we omit certain covariates are in orange.

7.3.1 Vertical motion

Vertical motion at 500 millibars (ω500) is a variable with moderate impact on the cloud properties considered in this work. Figure 7.1 reports our results for this experiment with the Pacific model in blue and the Pacific without ω500 model in orange. We notice in Figure 7.1(a) that the uncertainty bounds are larger when ω500 is omitted from the covariates. This result goes against our expectations and is indicative of a bug, which we were unfortunately not able to investigate. Figure 7.1(b) identifies the value of Λ that results in an ignorance interval around the Pacific without ω500 model predictions that covers the Pacific model predictions and its ignorance region. We find that we need to set $\Lambda = 1.01$ to account for omitting ω500 from the covariates. We also note that the slopes of the dose-response curves are slightly different, with worse predictions when omitting ω500 from the covariates, as expected



(a) Raw

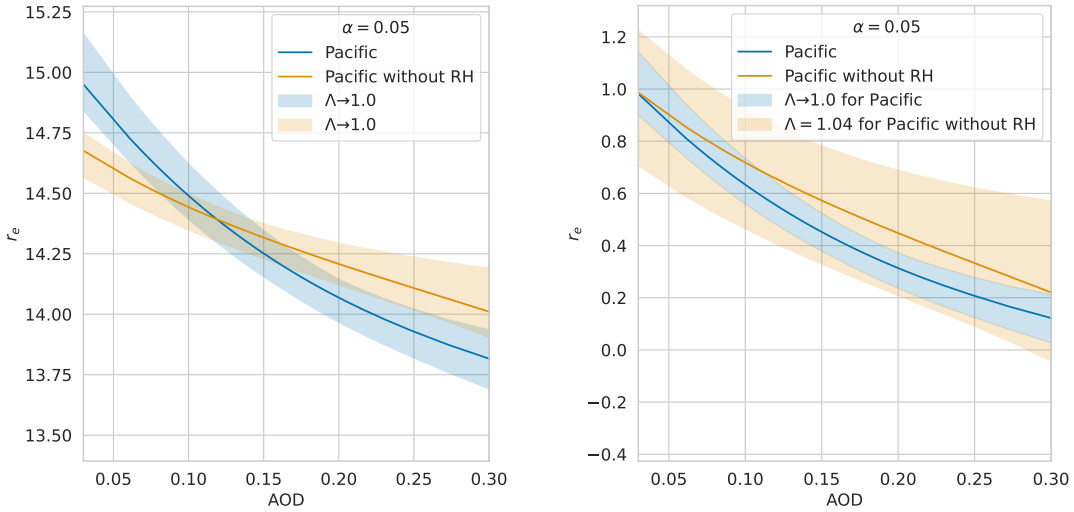
(b) Scaled by min-max with appropriate Λ

Figure 7.1: Dose-response curves for r_e: Pacific and Pacific without ω500. Overcast transformer, low-resolution dataset. Covariates: RH900, RH850, RH700, EIS, LTS, SST, (ω500). Treatment: AOD. Outcomes: r_e.

7.3.2 Relative humidity

We perform the same experiment but now omit all relative humidity variables (RH900, RH850, RH700). These variables have a higher impact on cloud properties,

and we therefore expect much worse dose-response curves. Figure 7.2 reports our results for this experiment, with the Pacific in blue and Pacific without relative humidity in orange. In Figure 7.2(a), we notice that the uncertainty bounds are similar for both the control and the test experiments as expected. In the same manner as in the former experiment, we try to identify an appropriate Λ. Figure 7.2(b) identifies the value of Λ that results in an ignorance interval around the Pacific without RH model predictions that covers the Pacific model predictions. We find that we need to set $\Lambda = 1.04$. We also see that the APO curves are quite different, especially compared to the previous experiment, and as expected from the importance of relative humidity variables.



(a) Raw

(b) Scaled by min-max with appropriate Λ

Figure 7.2: Dose-response curves for r_e: Pacific and Pacific without relative humidity. Overcast transformer, low-resolution dataset. Covariates: (RH900, RH850, RH700), EIS, LTS, SST, W500. Treatment: AOD. Outcomes: r_e.

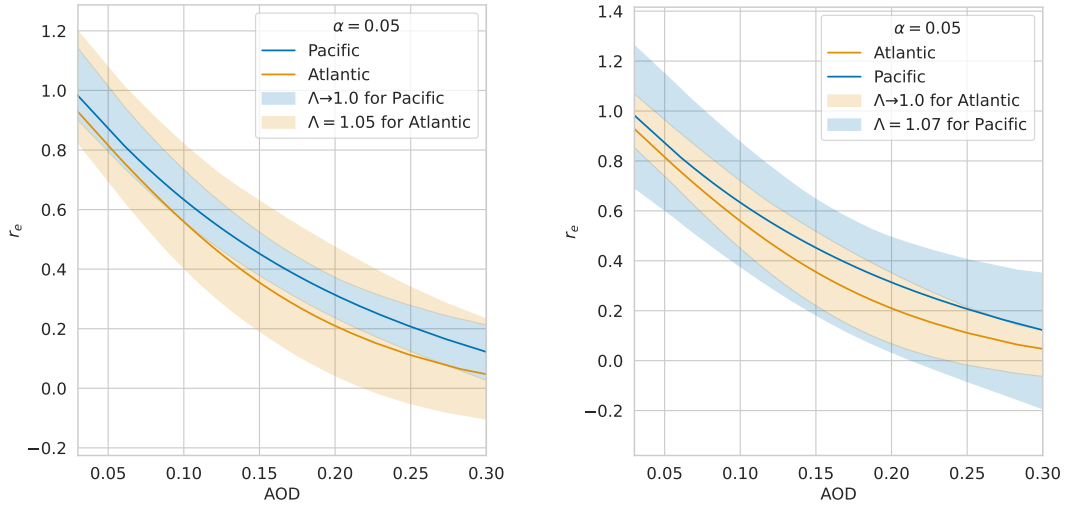
7.3.3 Discussion

We can now compare the results of both experiments. We notice that we need a larger Λ for more important covariates, as expected. We set $\Lambda = 1.04$ when omitting relative humidity variables and $\Lambda = 1.01$ when omitting ω500. Moreover, omitting covariates results in worse APO curves, especially when the covariate is important. We find that the parameter Λ can account for these shifts.

7.4 Geographical regions

Our second experiment consists of the study of data from regions with similar meteorology but different magnitudes of confounding influences. As in Chapter 6, we study the South Atlantic and the South-East Pacific regions. We are looking for the values of Λ that result in an ignorance interval explaining both the Pacific and the Atlantic data. In what follows, the results for the Pacific data are in blue and those for the Atlantic are in orange.

Figure 7.3 shows our results for this experiment. Figure 7.3(a) shows the value of Λ that results in an ignorance interval around the Atlantic model predictions that covers the Pacific model predictions under the unconfoundedness assumption. Figure 7.3(b) shows the value of Λ that results in an ignorance interval around the Pacific model predictions that covers the Atlantic model predictions under the unconfoundedness assumption.



(a) Scaled by min-max with appropriate Λ for the Atlantic to cover the Pacific

(b) Scaled by min-max with appropriate Λ for the Pacific to cover the Atlantic

Figure 7.3: Dose-response curves for r_e: South-East Pacific and South Atlantic. Overcast transformer, low-resolution datasets. Covariates: RH900, RH850, RH700, EIS, LTS, SST, W500. Treatment: AOD. Outcomes: r_e.

It is interesting to notice that the two values of Λ differ, with $\Lambda = 1.05$ for the Atlantic and $\Lambda = 1.07$ for the Pacific. This result makes sense given the fact that there are different confounding influences in the Pacific and the Atlantic.

7.5 Discussion

With these two experiments, we extend the results of the original paper by exploring possible interpretations of Λ.

This work allows us to address the third research question about the impact of unmeasured confounding on plausible ranges of treatment-effect estimates of ACI. With our first experiment on omitting covariates, we touch upon the trade-off between unconfoundedness and positivity, where increased violations of unconfoundedness lead to decreased violations of positivity and therefore worse predictions but tighter uncertainty bounds. In our second experiment, we reason about confounding in the underlying physical processes and take a more empirical approach. This work gives us more intuition about the parameter Λ relating violations of the unconfoundedness assumption to ignorance regions of dose-response curves.

It also shows the importance of controlling for confounding influences or factor violations of the unconfoundedness assumption in the parameter Λ to obtain realistic uncertainty intervals.

8 Conclusion

8.1 Summary, contributions and implications

In this work, we use machine learning approaches to estimate plausible ranges for the causal effects of aerosols on clouds and derive uncertainty bounds. Our research is based on [26], which we refer to as Overcast. The authors propose a method and models to estimate continuous treatment effects and develop a sensitivity model, the continuous treatment-effect marginal sensitivity model (CMSM) based on the potential outcomes approach to causal inference. The objective of this project is to further investigate aerosol-cloud interactions (ACI) and their uncertainties using the Overcast models. From a causal point of view, we aim to understand how unmeasured confounding can change treatment-effect estimates.

The most important aspects of this work can be summarised into three distinct objectives. First, we evaluate the method and models proposed in Overcast by implementing various baselines. Second, we study how well geographical dependencies of ACI are captured by the Overcast models. Third, we perform an uncertainty-aware causal sensitivity analysis using the CMSM to study how unmodelled confounding variables can influence the range of plausible treatment effects for a given dataset.

First, we find that whilst the predictor for cloud properties proposed by Overcast is quite weak, it agrees with off-the-shelf regression models like third-degree polynomial ridge regression. We identify that the Overcast transformer performs better than the Overcast feed-forward neural network in terms of predictive power, and agrees best with domain knowledge in terms of estimates of treatment effects. The underlying attention mechanisms allow us to model spatio-temporal dependencies between meteorological variables and capture confounding latent in the relationships between neighbouring variables. It however has larger ignorance regions which can hopefully be reduced by using larger amounts of data.

Second, our experiments with datasets from different geographical regions and resolutions show Overcast models can capture some of the geographical dependencies of ACI. We find that the model emulates ACI better in the Atlantic than in the Pacific but performs much poorly with high-resolution data compared to low-resolution data in the Pacific. This shows that the model can capture certain geographical dependencies of ACI but not others. A possible explanation for this relates to the amount of hidden confounding in the different datasets. Further investigation would be required to better understand the underlying mechanisms and improve the model.

Third, our uncertainty-aware causal sensitivity analysis allows us to study hidden confounding across different datasets. We find that omitting covariates reduces uncertainty but increases error in the treatment effect estimates that the sensitivity parameter Λ can account for. Our work also extends the original Overcast article in that it provides a further interpretation of the sensitivity parameter Λ and proposes a methodology to set it appropriately.

To a larger extent, our work contributes to understanding the climatological impacts of human emissions on cloud properties and assessing interventions that aim to reduce global warming. Among the measures considered to counteract climate change is geoengineering, deliberate large-scale interventions in the Earth's climate system. In particular, spraying seawater is a technique considered to seed larger, brighter, longer-lasting clouds, therefore enhancing clouds' ability to cool. Whilst

this method could offset some warming effects, it could also have disastrous impacts on weather patterns [16]. This project highlights the importance of uncertainty when studying climate projection models to take appropriate measures. It is of utmost importance that decision makers are made aware of proposals' underlying uncertainties and assumptions to make informed decisions.

8.2 Limitations and future works

The present work has some limitations that stem from the models used, the evaluation, the datasets studied, and time constraints.

First, let us describe the limitations of the models used. The most obvious limitation concerns the accuracy of the cloud properties predictions as shown by the low r^2. Further, the underlying causal graph contains known confounding influences that are not observed and therefore not controlled for, like aerosol types. To address these issues, future works should focus on the following areas: improving the predictor of cloud properties, accounting for the error in treatment measurement, and improving the underlying causal graph. The predictor of cloud properties should be improved significantly, probably by using another feature extractor, or by using a different density estimator. Namely, we would use both low and high-resolution data, and the feature extractor would use attention mechanisms to capture not only spatial but also temporal dependencies. The framework could be auto-regressive to capture dependencies between cloud properties. Future work could explore uncertainty arising from the fact that AOD is only a proxy for aerosol. It is currently difficult to estimate how much the use of proxies blurs out the strength of the true causal effect, especially because AOD is itself confounded by environmental processes. We suggest investigating frameworks that take into account treatment measurement errors like [57]. More generally, the underlying causal graph could account for more confounding variables if only they were observable.

Second, the evaluation performed is not sufficiently robust. This weakness comes from the fact that the ignorance levels are computed using Monte-Carlo integration which requires a large number of samples. Future work could consider

using larger sample sizes during inference time and studying the influence on the derived uncertainty bounds. Moreover, each experiment was only run once, using a specific random seed for splitting the dataset. Given the spatio-temporal variability of the data and the relatively small size of the datasets, running the experiments multiple times may lead to more reliable results. Further, the APO curves are currently plotted using only the testing set whereas using the entire dataset would likely reduce the size of the ignorance region.

Third, some limitations stem from the datasets studied. The low-resolution and high-resolution datasets differ in terms of spatial resolution, temporal resolution and timescale. This threefold difference hinders the analysis of our results. Moreover, there are confounding influences from relying on satellite data. It would be interesting to look at longer timescales, especially for high-resolution data. This should hopefully be more representative and mitigate the effects of El Niño for example. We moreover expect that increasing the amount of data will help reduce the uncertainties in the treatment-effect estimates, especially for the transformer. Further, to improve understanding of the present results, future work could investigate datasets with different spatial resolutions or temporal resolutions only.

Fourth, our work was produced under significant time constraints. Given the time needed to train, tune and evaluate the models, we were not able to run all the experiments we wanted to run to be able to fully interpret our results. Our work therefore relies on baseline models to understand the underlying mechanisms of the Overcast models which is not ideal. These limitations could be answered by allocating more time to re-implementing the model to achieve better efficiency before moving on to further experiments.

Our analysis is also impacted by the fact that our work is mostly empirical rather than theoretical. For instance, studying the underlying theory more deeply could potentially help with understanding the parameter Λ. It could also help us reason theoretically about the amount of data needed to obtain reliable results.

Appendices

A Datasets

A.1 Sources of satellite observations

Product Name	Description
Cloud Dropet Concentration N_d	MODIS (1.6, 2.1, 3.7 μm channels) [4]
Cloud Effective Radius r_e	MODIS (1.6, 2.1, 3.7 μm channels) [4]
Cloud Optical Depth τ	MODIS (1.6, 2.1, 3.7 μm channels) [4]
Cloud Water Path (CWP)	MODIS (1.6, 2.1, 3.7 μm channels) [4]
Cloud Fraction (CF)	MODIS (1.6, 2.1, 3.7 μm channels) [4]
Precipitation	NOAA CMORPH CDR [42]
Sea Surface Temperature (SST)	NOAA WHOI CDR [14]
Lower Tropospheric Stability (LTS)	MERRA-2 [19]
Vertical Motion (ω500)	MERRA-2 [6]
Estimated Inversion Strength (EIS)	MERRA-2 [19, 55]
Relative Humidity at x mb	MERRA-2 [19]
Aerosol Optical Depth (AOD)	MERRA-2 [19]

Table A.1: Sources of satellite observations

B

Additional implementation details

B.1 Hyper-parameters search space

Hyper-parameter	Transformer	Feed-Forward Neural Network
Hidden Units	tune.qlograndint(128, 512, 128)	tune.qlograndint(32, 1024, 32)
Network Depth	tune.randint(2, 5)	tune.randint(2, 6)
GMM Components	tune.randint(1, 32)	tune.randint(1, 32)
Attention Heads	tune.choice([1, 2, 4, 8])	NA
Negative Slope	tune.quniform(0.0, 0.5, 0.01)	tune.quniform(0.0, 0.5, 0.01)
Dropout Rate	tune.quniform(0.0, 0.5, 0.01)	tune.quniform(0.0, 0.5, 0.01)
Layer Norm	tune.choice([True, False])	tune.choice([True, False])
Batch Size	tune.qlograndint(32, 256, 32)	tune.qlograndint(32, 256, 32)
Learning Rate	tune.quniform(1e-4, 1e-3, 1e-4)	tune.quniform(1e-4, 2e-3, 1e-4)

Table B.1: Hyper-parameters search space for Overcast models

B.2 Final hyper-parameters

Hyper-parameter	Transformer				Neural Network
	LR Pacific	LR Atlantic	LR Pacific without ω500	LR Pacific without RH	LR Pacific
Hidden units	128	128	256	128	256
Network depth	3	4	3	3	2
GMM T components	27	7	24	27	2
GMM Y components	22	24	24	22	5
Attention heads	8	8	4	8	NA
Negative slope	0.28	0.19	0.01	0.28	0.1
Dropout rate	0.42	0.16	0.5	0.42	0.09
Layer norm	False	True	False	False	False
Batch size	128	160	32	128	224
Learning rate	0.0001	0.0001	0.0002	0.0001	0.0002
Epochs	500	500	500	500	9

Table B.2: Final hyper-parameters for each dataset and model

References

[1] Bruce A. Albrecht. "Aerosols, Cloud Microphysics, and Fractional Cloudiness". In: *Science* 245.4923 (Sept. 1989), pp. 1227–1230.

[2] O Altaratz, R Z Bar-Or, U Wollner, and I Koren. "Relative Humidity and Its Effect on Aerosol Optical Depth in the Vicinity of Convective Clouds". In: *Environmental Research Letters* 8.3 (Sept. 2013), p. 034025.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. May 2016. arXiv: `1409.0473 [cs, stat]`.

[4] Bryan A. Baum and Steven Platnick. "Introduction to MODIS Cloud Products". In: *Earth Science Satellite Remote Sensing*. Ed. by John J. Qu, Wei Gao, Menas Kafatos, Robert E. Murphy, and Vincent V. Salomonson. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 74–91.

[5] Christopher Bishop. *Mixture Density Networks*. Tech. rep. 1994.

[6] Michael Bosilovich. "MERRA-2: Initial Evaluation of the Climate". In: 43 (), p. 145.

[7] Olivier Boucher, D Randall, P Artaxo, C Bretherton, G Feingold, P Forster, V-M Kerminen, Y Kondo, H Liao, U Lohmann, P Rasch, S.K Satheesh, B Stevens, and X.Y Zhang. "Clouds and Aerosols". In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by T.F Stocker, D Qin, G.-K Plattner, M Tignor, S.K Allen, J Boschung, A Nauels, Y Xia, V Bex, and P.M Midgley. Cambridge University Press, 2013, pp. 571–658.

[8] A.I. Calvo, C. Alves, A. Castro, V. Pont, A.M. Vicente, and R. Fraile. "Research on Aerosol Sources and Chemical Composition: Past, Current and Emerging Issues". In: *Atmospheric Research* 120–121 (Feb. 2013), pp. 1–28.

[9] Y.-C. Chen, M. W. Christensen, L. Xue, A. Sorooshian, G. L. Stephens, R. M. Rasmussen, and J. H. Seinfeld. "Occurrence of Lower Cloud Albedo in Ship Tracks". In: *Atmospheric Chemistry and Physics* 12.17 (Sept. 2012), pp. 8223–8235.

[10] E. Colin Cherry. "Some Experiments on the Recognition of Speech, with One and with Two Ears". In: *The Journal of the Acoustical Society of America* 25.5 (Sept. 1953), pp. 975–979.

[11] Matthew W. Christensen, Andrew Gettelman, Jan Cermak, Guy Dagan, Michael Diamond, Alyson Douglas, Graham Feingold, Franziska Glassmeier, Tom Goren, Daniel P. Grosvenor, Edward Gryspeerdt, Ralph Kahn, Zhanqing Li, Po-Lun Ma, Florent Malavelle, Isabel L. McCoy, Daniel T. McCoy, Greg McFarquhar, Johannes Mülmenstädt, Sandip Pal, Anna Possner, Adam Povey, Johannes Quaas, Daniel Rosenfeld, Anja Schmidt, Roland Schrödner, Armin Sorooshian, Philip Stier, Velle Toll, Duncan Watson-Parris, Robert Wood, Mingxi Yang, and Tianle Yuan. "Opportunistic Experiments to Constrain Aerosol Effective Radiative Forcing". In: *Atmospheric Chemistry and Physics* 22.1 (Jan. 2022), pp. 641–674.

[12] Matthew W. Christensen, David Neubauer, Caroline Poulsen, Gareth Thomas, Greg McGarragh, Adam C. Povey, Simon Proud, and Roy G. Grainger. *Unveiling Aerosol-Cloud Interactions Part 1: Cloud Contamination Insatellite Products Enhances the Aerosol Indirect Forcing Estimate*. Preprint. Aerosols/Remote Sensing/Troposphere/Physics (physical properties and processes), May 2017.

[13] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. Dec. 2014. arXiv: `1412.3555 [cs]`.

[14] James L. Cogan and James H. Willand. "Measurement of Sea Surface Temperature by the NOAA 2 Satellite." In: *Journal of Applied Meteorology* 15 (Feb. 1976), pp. 173–180.

[15] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. "Overlap in Observational Studies with High-Dimensional Covariates". In: *Journal of Econometrics* 221.2 (Apr. 2021), pp. 644–654.

[16] Michael S. Diamond, Andrew Gettelman, Matthew D. Lebsock, Allison McComiskey, Lynn M. Russell, Robert Wood, and Graham Feingold. "To Assess Marine Cloud Brightening's Technical Feasibility, We Need to Know What to Study—and When to Stop". In: *Proceedings of the National Academy of Sciences* 119.4 (Jan. 2022), e2118379119.

[17] A. Douglas and T. L'Ecuyer. "Quantifying Variations in Shortwave Aerosol Cloud Radiation Interactions Using Local Meteorology and Cloud State Constraints". In: *Atmospheric Chemistry and Physics* 19.9 (2019), pp. 6251–6268.

[18] Stefan Falkner, Aaron Klein, and Frank Hutter. "Combining Hyperband and Bayesian Optimization". In: (), p. 6.

[19] Ronald Gelaro, Will McCarty, Max J. Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A. Randles, Anton Darmenov, Michael G. Bosilovich, Rolf Reichle, Krzysztof Wargan, Lawrence Coy, Richard Cullather, Clara Draper, Santha Akella, Virginie Buchard, Austin Conaty, Arlindo M. da Silva, Wei Gu, Gi-Kong Kim, Randal Koster, Robert Lucchesi, Dagmar Merkova, Jon Eric Nielsen, Gary Partyka, Steven Pawson, William Putman, Michele Rienecker, Siegfried D. Schubert, Meta Sienkiewicz, and Bin Zhao. "The Modern-Era Retrospective Analysis for Research and Applications,

Version 2 (MERRA-2)". In: *Journal of Climate* 30.14 (July 2017), pp. 5419–5454.

[20] Daniel P. Grosvenor, Odran Sourdeval, Paquita Zuidema, Andrew Ackerman, Mikhail D. Alexandrov, Ralf Bennartz, Reinout Boers, Brian Cairns, J. Christine Chiu, Matthew Christensen, Hartwig Deneke, Michael Diamond, Graham Feingold, Ann Fridlind, Anja Hünerbein, Christine Knist, Pavlos Kollias, Alexander Marshak, Daniel McCoy, Daniel Merk, David Painemal, John Rausch, Daniel Rosenfeld, Herman Russchenberg, Patric Seifert, Kenneth Sinclair, Philip Stier, Bastiaan van Diedenhoven, Manfred Wendisch, Frank Werner, Robert Wood, Zhibo Zhang, and Johannes Quaas. "Remote Sensing of Droplet Number Concentration in Warm Clouds: A Review of the Current State of Knowledge and Perspectives". In: *Reviews of Geophysics* 56.2 (2018), pp. 409–453.

[21] Douglas S. Hamilton, Lindsay A. Lee, Kirsty J. Pringle, Carly L. Reddington, Dominick V. Spracklen, and Kenneth S. Carslaw. "Occurrence of Pristine Aerosol Environments on a Polluted Planet". In: *Proceedings of the National Academy of Sciences* 111.52 (Dec. 2014), pp. 18466–18471.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016, pp. 770–778.

[23] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. *Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors*. July 2012. arXiv: `1207.0580 [cs]`.

[24] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780.

[25] Paul W. Holland. "Statistics and Causal Inference". In: *Journal of the American Statistical Association* 81.396 (Dec. 1986), pp. 945–960.

[26] 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*. June 2022. arXiv: `2204.10022 [cs, stat]`.

[27] Andrew Jesson, Soren Mindermann, Yarin Gal, and Uri Shalit. "Quantifying Ignorance in Individual-Level Causal-Effect Estimates under Hidden Confounding". In: *Proceedings of the 38th International Conference on Machine Learning* 139 (2021), pp. 4829–4838.

[28] Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. "Identifying Causal-Effect Inference Failure withUncertainty-Aware Models". In: *Advances in neural information processing systems* (2020).

[29] I. Koren, G. Feingold, and L. A. Remer. "The Invigoration of Deep Convective Clouds over the Atlantic: Aerosol Effect, Meteorology or Retrieval Artifact?" In: *Atmospheric Chemistry and Physics* 10.18 (Sept. 2010), pp. 8855–8872.

[30] Ilan Koren, Guy Dagan, and Orit Altaratz. "From Aerosol-Limited to Invigoration of Warm Convective Clouds". In: *Science* 344.6188 (June 2014), pp. 1143–1146.

[31] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. *Tune: A Research Platform for Distributed Model Selection and Training*. July 2018. arXiv: `1807.05118 [cs, stat]`.

[32] "IPCC, 2021: Summary for Policymakers". 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*. Ed. by Valérie Masson-Delmotte, Panmao Zhai, Anna Pirani, Sarah L. Connors, Clotilde Péan, Yang Chen, Leah Goldfarb, Melissa I. Gomis, J.B.Robin Matthews, Sophie Berger, Mengtian Huang, Ozge Yelekçi, Rong Yu, and Baiquan Zhou. Cambridge University Press, 2021.

[33] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I. Jordan, and Ion Stoica. *Ray: A Distributed Framework for Emerging AI Applications*. Sept. 2018. arXiv: `1712.05889 [cs, stat]`.

[34] Brady Neal. *Introduction to Causal Inference*.

[35] Louise Nuijens and A. Pier Siebesma. "Boundary Layer Clouds and Convection over Subtropical Oceans in Our Current and in a Warmer Climate". In: *Current Climate Change Reports* 5.2 (June 2019), pp. 80–94.

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[37] Judea Pearl. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. USA: Addison-Wesley Longman Publishing Co., Inc., 1984.

[38] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge, U.K. ; New York: Cambridge University Press, 2000.

[39] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. "Causal Inference in Statistics : A Primer". In: *Causal Inference in Statistics : A Primer*. Chichester, West Sussex: Wiley, 2016.

[40] Judea Pearl and Dana Mackenzie. *The Book of Why. The New Science of Cause and Effect*. New York: Basic Books, 2018.

[41] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. "Scikit-Learn: Machine Learning in Python". In: *MACHINE LEARNING IN PYTHON* (), p. 6.

[42] Olivier P. Prat, Brian R. Nelson, Elsa Nickl, and Ronald D. Leeper. "Global Evaluation of Gridded Satellite Precipitation Products from the NOAA Climate Data Record Program". In: *Journal of Hydrometeorology* 22 (Sept. 2021), pp. 2291–2310.

[43] Donald B. Rubin. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." In: *Journal of Educational Psychology* 66.5 (1974), pp. 688–701.

[44] Donald B. Rubin. "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment". In: *Journal of the American Statistical Association* 75.371 (Sept. 1980), p. 591.

[45] Sebastian Ruder. *An Overview of Gradient Descent Optimization Algorithms*. June 2017. arXiv: `1609.04747 [cs]`.

[46] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning Representations by Back-Propagating Errors". In: *Nature* 323 (1986), pp. 533–536.

[47] Stephen E. Schwartz. "Are Global Cloud Albedo and Climate Controlled by Marine Phytoplankton?" In: *Nature* 336.6198 (Dec. 1988), pp. 441–445.

[48] Jasjeet Sekhon. *The Neyman— Rubin Model of Causal Inference and Estimation Via Matching Methods*. Ed. by Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. Vol. 1. Oxford University Press, Sept. 2009.

[49] Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9". In: *Statistical Science* 5.4 (Nov. 1990).

[50] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: (), p. 30.

[51] Bjorn Stevens and Graham Feingold. "Untangling Aerosol Effects on Clouds and Precipitation in a Buffered System". In: *Nature* 461.7264 (Oct. 2009), pp. 607–613.

[52] S. Twomey. "Aerosols, Clouds and Radiation". In: *Atmospheric Environment. Part A. General Topics* 25.11 (Jan. 1991), pp. 2435–2442.

[53] S. A. Twomey, M. Piepgrass, and T. L. Wolfe. "An Assessment of the Impact of Pollution on Global Cloud Albedo". In: *Tellus B* 36B.5 (Nov. 1984), pp. 356–366.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need*. Dec. 2017. arXiv: `1706.03762 [cs]`.

[55] Robert Wood and Christopher S. Bretherton. “On the Relationship between Stratiform Low Cloud Cover and Lower-Tropospheric Stability”. In: *Journal of Climate* 19.24 (Dec. 2006), pp. 6425–6432.

[56] Kelvin Xu, Jimmy Lei, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. “Show, Attend and Tell: Neural Image CaptionGeneration with Visual Attention”. In: (), p. 10.

[57] Yuchen Zhu, Limor Gultchin, Arthur Gretton, Matt Kusner, and Ricardo Silva. “Causal Inference with Treatment Measurement Error: A Nonparametric Instrumental Variable Approach”. In: (), p. 11.