Using uncertainty-aware machine learning models to study aerosol-cloud interactions

Maëlys Solal
Department of Computer Science
University of Oxford, UK
maelys.solal@ens.psl.eu

Andrew Jesson
OATML
Department of Computer Science
University of Oxford, UK

Yarin Gal
OATML
Department of Computer Science
University of Oxford, UK

Alyson Douglas
AOPP
Department of Physics
University of Oxford, UK

Abstract

Aerosol-cloud interactions (ACI) include various effects that result from aerosols entering a cloud, 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 cool the Earth. The strength of the effect is however heterogeneous, meaning it depends on the surrounding environment, making ACI one of the most uncertain effects in our current climate models. In our work, we use causal machine learning to estimate ACI from satellite observations by reframing the problem as a treatment (aerosol) and outcome (change in droplet radius). We predict the causal effect of aerosol on clouds with uncertainty bounds depending on the unknown factors that may be influencing the impact of aerosol. Of the three climate models evaluated, we find that only one plausibly recreates the trend, lending more credence to its estimate cooling due to ACI.

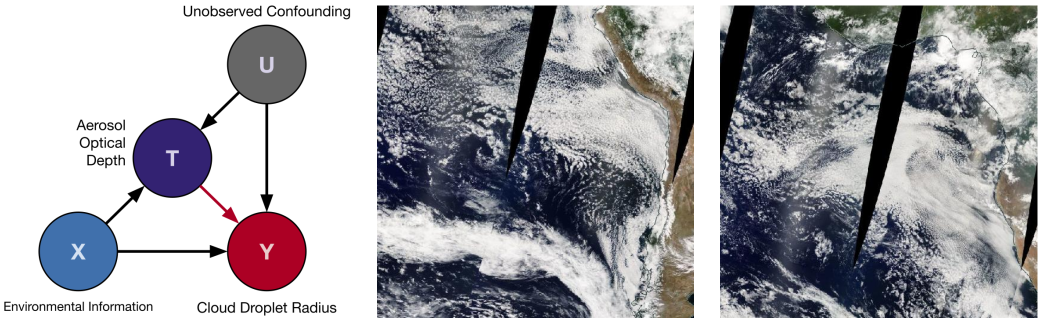
1 Introduction

Aerosol, in the form of pollution from human emissions, enters the atmosphere and eventually interacts with a cloud leading to aerosol-cloud interactions (ACI). As aerosol enters the cloud, a causal chain of events catalyzes. It begins with aerosol particles activating as cloud droplet nuclei, which increases the number of droplets within the cloud, reducing the mean radius of cloud droplets to redistribute the water vapor, and eventually increasing the cloud's brightness (Figure 1(a)) [1]. Overall, an increase in atmospheric aerosol leads to larger, brighter, longer-lasting clouds that reflect more incoming sunlight. ACI are thus a net cooling process and offset some fraction of warming due to rising levels of CO_2. The strength of the effect is however dependent on the local environment surrounding the cloud. ACI remain one of the most uncertain effects in our current climate models, as current models are limited in their ability to simulate ACI with such environmental heterogeneity [2, 3]. Climate models can only approximate ACI given their low spatial resolution and limited parameterizations, often dependent on only a few environmental parameters, such as the relative humidity within a grid cell. These factors lead to increased uncertainty in future projections. Currently, state-of-the-art climate models estimate that the range of cooling due to ACI may offset 0%-50% of the warming due to greenhouse gas emissions.

This work uses causal machine learning to estimate ACI from satellite observations, by reframing the problem as a treatment (aerosol) and outcome (change in droplet radius). We predict the causal effect of aerosol on clouds and provide uncertainty bounds that we compare to the parameterizations of

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2022.

climate model ACI. We consider uncertainty arising from violations of two assumptions: positivity (or overlap) and unconfoundedness (or no hidden confounding). Positivity violations are due to insufficient representation within the data for all treatment levels, such as "treating" cloud with aerosol. Unmeasured confounding are unobserved factors which influence both the treatment and outcomes, such as humidity causing aerosol swelling and altering cloud properties. To better understand these individual sources of uncertainty, we use Overcast [4], a prime example of the needs of a community such as ACI leading to methodological contributions in machine learning. Compared to prior work such as [5], we consider aerosol optical depth (AOD), our proxy for aerosol concentration, as a continuous treatment rather than discrete and perform an uncertainty-aware sensitivity analysis to study the consequences of possible violations of positivity and unconfoundedness.



(a) Simple Causal Graph of ACI (b) The Southeast Pacific (c) The South Atlantic

Figure 1: The causal graph underlining our knowledge of ACI and satellite imagery of the two regions analyzed, chosen due to their unique aerosol-cloud interactions and breadth of past studies to pull knowledge from.

2 Methods

Following [4], we use the potential outcomes framework to estimate the effect of a continuous treatment $\mathrm{T} \in \mathcal{T}$ (aerosol), on outcomes of interest $\mathrm{Y} \in \mathcal{Y}$ (cloud property), for a unit described by covariates $\mathbf{X} \in \mathcal{X}$ (environmental information) as shown in Figure 1(a) [6, 7, 8, 9]. We call a potential outcome and denote by Y_{t} what the outcome would have been if the treatment were t. The covariates considered are relative humidity at 900, 850 and 700 millibar, sea surface temperature, vertical motion at 500 millibars, lower tropospheric stability, and effective inversion strength. The treatment is aerosol optical depth (AOD), a proxy for aerosol concentration. The outcome considered is the cloud droplet size (r_e). To estimate the treatment-effect, we study the conditional average potential outcome (CAPO) and the average potential outcome (APO)

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

which can be identified from the observational distribution $P(\mathbf{X}, \mathrm{T}, \mathrm{Y})$ using

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

and further assumptions (unconfoundedness, positivity, no-interference and consistency). Here, we study the robustness of treatment-effect estimates to positivity and unconfoundedness violations (see Appendix A for more detail). We compute uncertainty bounds corresponding to user-specified relaxations of these assumptions. The parameter Λ, for example, is set by the user to explain an assumed level of unmeasured confounding [4, 10, 11]. Some confounding influences are impossible to measure directly with satellites, such as humidity causing aerosol swelling and altering cloud properties, and the parameter Λ can be used to encode an expert's belief in the influence of such confounders.

We use daily mean, 1° x 1° of satellite observations in order to homogenize the data from the southeast Pacific and south Atlantic (Figures 1(b) and 1(c)). Mean droplet radius (r_e) from the MODIS instrument is used as our outcome for all experiments shown within. We employ aerosol optical depth from MERRA-2 to approximate the concentration of aerosol. Our environmental confounders are the relative humidity at 900, 850 and 700 millibars, the stability of the atmosphere, the sea surface temperature, and the vertical motion at 500 mb, all also from MERRA-2. For more detail about data and implementation, please refer to Appendix B and Appendix D.

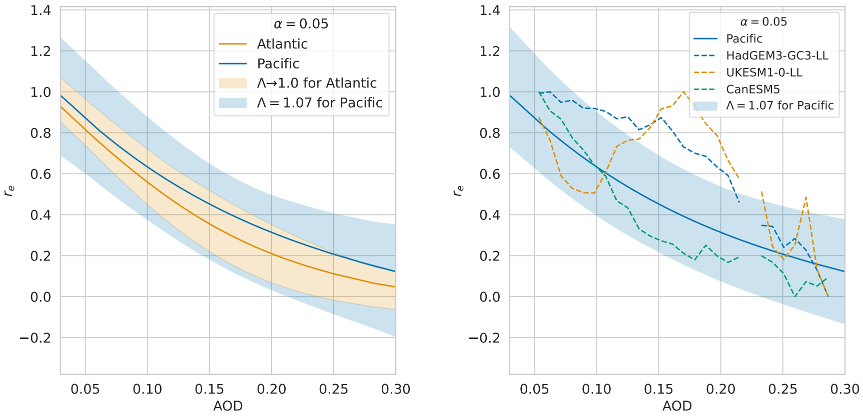
3 Results

3.1 Deducing reasonable treatment-effect bounds using domain knowledge

Unlike past studies which only crudely estimate an uncertainty range due to quantifiable effects, we are able to derive confidence intervals dependent on the influence of confounding by varying Λ. Since it is impossible to know the strength of the confounding effect from observed data alone, we propose a method to select a reasonable Λ by contrasting two geographical regions. We contrast the South-East Pacific and the South Atlantic because these regions have different environmental confounders of ACI, for example aerosol type, aerosol hygroscopicity, aerosol size. These are important confounders, but are unfortunately not included in the available data. So we select the parameter Λ for the Pacific region such that the treatment effect bounds for the Pacific region cover the effect bounds for the Atlantic region under the assumption of no hidden confounding ($\Lambda \to 1$) for the Atlantic region, as shown in Figure 2(a). Setting Λ to 1.07 gives bounds for the pacific region that reasonably account for the potential bias induced by the unmodeled confounders. While a larger Λ could still be sensible due to other drivers of confounding, domain knowledge informs us that these are the main missing physical mechanisms.

3.2 Evaluating climate models using machine learning

As we now have a possible range of ACI derived using the real, observed outcomes, we can judge how well climate models recreate this observed trend by seeing if their responses lie within our derived interval (Figure 2(b)). We find that the Canadian model CanESM5 simulates ACI better than the UK models HadGEM3-GC3-LL and UKESM1-0-LL. Our trained machine learning model not only uses the real, observed relationships to derive the magnitude of the effect, but can consider the environmental context and confounding influences to derive real, quantifiable bounds of uncertainty. Therefore, by using the curves found by Overcast as the true response, we know those models which lie outside of our bounds found by contrasting different regions are likely unphysical and highly unlikely to occur in the real-world. CanESM5 currently estimates the total cooling effect due to ACI to offset approximately half of the warming due to greenhouse gases; based on our results, we would say it is likely that this estimate is closest to the true value observed on Earth [12].



(a) Choosing Λ using two geographical regions (b) Comparison with ESMs in the Pacific

Figure 2: Plausible range of effects of aerosol (AOD) on mean droplet radius (r_e).

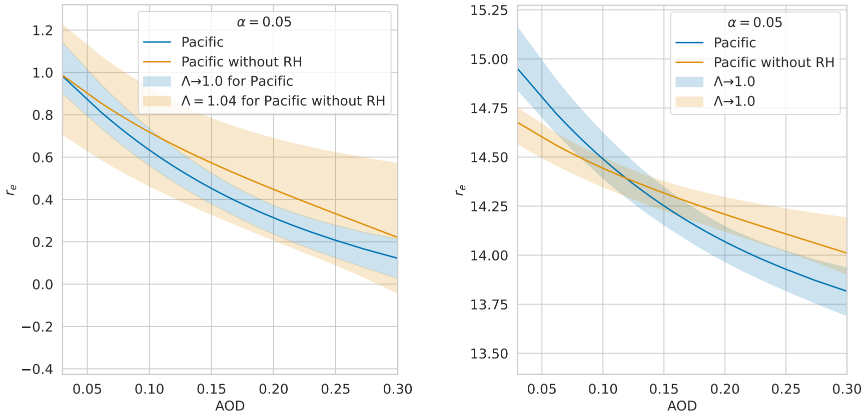
4 Discussion and conclusions

4.1 Machine learning's place in climate model verification

In this work, we show that machine learning methods offer viable ways to objectively judge how well global climate models reproduce climate processes such as the effect of aerosol on mean droplet radius. A drawback of historical studies which utilize satellite observations is their inability to quantify how the surrounding environment may affect the magnitude of the aerosol-cloud interactions. Overcast accounts for such contextual confounding and communicates bounds on the treatment effect due to an expert-informed influence of hidden-confounding. Utilizing this method gives us insight into whether climate models reproduce the observed relationships between AOD and r_e. Climate models currently only reasonably recreate large scale processes that can be explicitly calculated, leaving processes like aerosol-cloud interactions, which occur on scales smaller than the grid scale, poorly parameterized and approximated. In order to improve our climate models, we must understand in more relatable terms how well they are doing, such as by comparing their outcomes to those from observations. Machine learning provides not only a way to judge these outcomes, but the relationships learned by Overcast and similar models could in the future be fine-tuned to replace our current parameterizations [13].

4.2 Collaboration across domains vs. purely data driven

While different sources of confounding due to regional differences alter the outcomes, the choice of which environmental factors are the main sources of confounding can also be investigated using Overcast. We perform two experiments, with and without relative humidity at 900, 850 and 700 millibars, to derive varying outcome shapes and fit Λ to both dose-response curves. When Λ is set to 1.04, both curves are captured by the bounds of uncertainty, allowing us to view how the response may vary within those bounds due to meteorological uncertainty rather than regional uncertainty, where $\Lambda = 1.07$ was required (Figure 3(a)). In the absence of ground truth, purely data-driven techniques cannot decide between the model with and without relative humidity, but as domain knowledge is brought in, it is known that the curve with humidity included is the true response curve (Figure 3(b)). Purely data-driven approaches may not be the most appropriate for studying climatological processes such as aerosol-cloud interactions as domain knowledge is essential to select the correct inputs and verify the outcomes. The most robust model arises from combining data and theory, bringing together experts in machine learning and climate processes.



(a) Scaled, with appropriate Λ

(b) Unscaled

Figure 3: Plausible range of effects when omitting relative humidity from the covariates.

4.3 Limitations and future works

This work uses aerosol optical depth as a proxy for aerosol concentration, which could bias treatment-effect estimates. For example, it is known that bias can arise from measurement-error in the treatment [14]. Further, we rely on low resolution data that does not perfectly capture the microphysical processes. Future work could consider different assumptions on the underlying causal model, and attempt to include other aerosol properties like size, hygroscopicity and type.

5 Acknowledgements

This project was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 821205 (FORCeS), the European Research Council (ERC) project constRaining the EffeCts of Aerosols on Precipitation (RECAP) under the European Union's Horizon 2020 research and innovation programme with grant agreement no. 724602, and the Turing Institute post-doctoral enrichment award.

References

[1] Sean A Twomey, M Piepgrass, and TL Wolfe. "An assessment of the impact of pollution on global cloud albedo". In: *Tellus B* 36.5 (1984), pp. 356–366.

[2] 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, Bex V., and P.M. Midgley. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2013, pp. 571–658.

[3] IPCC. "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 V. Masson-Delmotte, P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2021, pp. 3–32.

[4] Andrew Jesson, Alyson Douglas, Peter Manshausen, Maëlys Solal, Nicolai Meinshausen, Philip Stier, Yarin Gal, and Uri Shalit. *Scalable Sensitivity and Uncertainty Analysis for Causal-Effect Estimates of Continuous-Valued Interventions*. 2022.

[5] Andrew Jesson, Peter Manshausen, Alyson Douglas, Duncan Watson-Parris, Yarin Gal, and Philip Stier. "Using Non-Linear Causal Models to Study Aerosol-Cloud Interactions in the Southeast Pacific". In: *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*. 2021.

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

[7] 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.

[8] 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.

[9] 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).

[10] Andrew Jesson, Sören 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.

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

[12] Christopher J Smith, Ryan J Kramer, Gunnar Myhre, Kari Alterskjær, William Collins, Adriana Sima, Olivier Boucher, Jean-Louis Dufresne, Pierre Nabat, Martine Michou, et al. "Effective radiative forcing and adjustments in CMIP6 models". In: *Atmospheric Chemistry and Physics* 20.16 (2020), pp. 9591–9618.

[13] Andrew Gettelman, David John Gagne, C-C Chen, MW Christensen, ZJ Lebo, Hugh Morrison, and Gabrielle Gantos. "Machine learning the warm rain process". In: *Journal of Advances in Modeling Earth Systems* 13.2 (2021), e2020MS002268.

[14] Yuchen Zhu, Limor Gultchin, Arthur Gretton, Matt Kusner, and Ricardo Silva. *Causal Inference with Treatment Measurement Error: A Nonparametric Instrumental Variable Approach*. 2022.

[15] 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.

[16] Ronald Gelaro et al. "The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2)". In: *Journal of Climate* 30.14 (July 2017), pp. 5419–5454.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019, pp. 8024–8035.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[24] 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.

[25] 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.

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

[27] Matthew W. Christensen et al. "Opportunistic Experiments to Constrain Aerosol Effective Radiative Forcing". In: *Atmospheric Chemistry and Physics* 22.1 (Jan. 2022), pp. 641–674.

A Theoretical background: unconfoundedness and positivity assumptions

Confounding variables are factors that influence both the treatment T and the outcomes Y. The unconfoundedness assumption states that all confounding variables are observed and controlled for using $\mathbf{X}$, so that the treatment groups are comparable, that is, $\mathrm{Y}_{\mathrm{T}} \perp\!\!\!\perp \mathrm{T} \mid \mathbf{X}$.

The positivity assumption states that all subgroups of the data with different covariates have a non-zero probability of receiving any dose of treatment, that is, $p(\mathrm{t} \mid \mathbf{x}) > 0$ for any $\mathrm{t} \in \mathcal{T}$ and for any $\mathbf{x} \in \mathcal{X}$ such that $p(\mathbf{x}) > 0$.

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

B Data and pre-processing

We work with data which is 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 × 1 km resolution [15]. 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 aerosol [16]. MERRA-2 calculates global vertical profiles of temperature, relative humidity, and pressure, and assimilates hyperspectral and passive microwave 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. We work with two different datasets which are $1^\circ \times 1^\circ$ daily means of observations over the South Atlantic and the South East Pacific from 2004 to 2019. The sources are given in Table 1.

Table 1: Sources of satellite observations

Product Name	Description
Mean Droplet Radius (r_e)	MODIS (1.6, 2.1, 3.7 μm channels) [15]
Precipitation	NOAA CMORPH CDR [17]
Sea Surface Temperature (SST)	NOAA WHOI CDR [18]
Lower Tropospheric Stability (LTS)	MERRA-2 [16]
Vertical Motion at 500 mb ($\omega 500$)	MERRA-2 [19]
Estimated Inversion Strength (EIS)	MERRA-2 [16, 20]
Relative Humidity at x mb (RHx)	MERRA-2 [16]
Aerosol Optical Depth (AOD)	MERRA-2 [16]

We restrict our observations to clouds in the "aerosol limited" regime by applying some filtering [21]. 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. 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. We also filter out precipitating clouds to avoid a loop in the causal graph. Finally, all features are normalized before being fed into the model.

C Model architecture

The models are neural-network architectures with two components: a feature extractor $\phi(\mathrm{x}; \theta)$ and a density estimator $f(\phi, \mathrm{t}; \theta)$, represented in Appendix C. The covariates $\mathbf{x}$ are given as input to the feature extractor, whose output is concatenated with the treatment t and given as input to the density estimator which outputs a Gaussian mixture density, $p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \theta)$, from which we can sample. The feature extractor uses attention mechanisms to model the spatio-temporal correlations between the covariates on a given day using the geographical coordinates of the observations. The model architecture is represented in Figure 4.

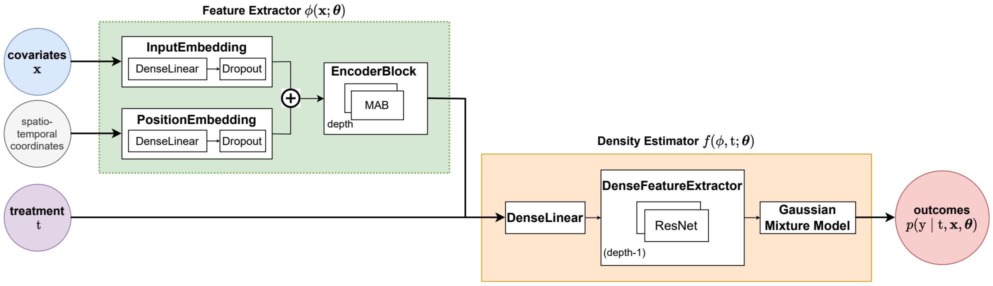

Feature Extractor $\phi(\mathbf{x};\boldsymbol{\theta})$
covariates
$\mathbf{x}$
spatio-temporal coordinates
treatment
t
InputEmbedding
DenseLinear
Dropout
PositionEmbedding
DenseLinear
Dropout
EncoderBlock
MAB
depth
Density Estimator $f(\phi, \mathrm{t}; \boldsymbol{\theta})$
DenseLinear
DenseFeatureExtractor
ResNet
(depth-1)
Gaussian Mixture Model
outcomes
$p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \boldsymbol{\theta})$

Figure 4: **Overcast model architecture.** The inputs are represented by circles, in **blue** the covariates, in grey the spatio-temporal coordinates, in **purple** the treatment. In the **red** circle is the output of the model, the outcomes distribution. The model has a feature extractor in green and a density estimator (in orange).

D Implementation details

We follow the implementation from [4]. The code is written in python. The packages used include PyTorch [22], scikit-learn [23], Ray [24], NumPy, SciPy and Matplotlib.

We use ray tune [25] with HyperBand Bayesian Optimization [26] search algorithm to optimise our network hyper-parameters. The hyper-parameters considered during tuning are given in [4]. The final hyper-parameters for each dataset are given in Table 2. The hyper-parameter optimization objective is the batch-wise Pearson correlation averaged across all outcomes on the validation data for a single dataset realization with random seed 1331.

We split the data into training, validation, and testing sets across different days. [4] splits data in the following way: datapoints from Mondays to Fridays are in the training set, from Saturdays in the validation set, and from Sundays in the testing set. In our implementation, we keep the same ratio between datasets but we randomize the splits, 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 randomization is motivated by the fact that there is a clear weekly cycle of aerosol optical depth [27]. Models are optimized by maximizing the log likelihood of $p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \theta)$.

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

Hyper-parameter	South-East Pacific	South Atlantic
Hidden Units	128	128
Network Depth	3	4
GMM T Components	27	7
GMM Y Components	22	24
Attention Heads	8	8
Negative Slope	0.28	0.19
Dropout Rate	0.42	0.16
Layer Norm	False	True
Batch Size	128	160
Learning Rate	0.0001	0.0001
Epochs	500	500