Using Non-Linear Causal Models to Study Aerosol-Cloud Interactions in the Southeast Pacific

Andrew Jesson*
OATML
Department of Computer Science
University of Oxford
andrew.jesson@cs.ox.ac.uk

Peter Manshausen*
Atmospheric, Oceanic, and Planetary Physics
Department of Physics
University of Oxford
peter.manshausen@physics.ox.ac.uk

Alyson Douglas*
Atmospheric, Oceanic, and Planetary Physics
Department of Physics
University of Oxford
alyson.douglas@physics.ox.ac.uk

Duncan Watson-Parris
Atmospheric, Oceanic, and Planetary Physics
Department of Physics
University of Oxford
duncan.watson-parris@physics.ox.ac.uk

Yarin Gal
OATML
Department of Computer Science
University of Oxford
yarin@cs.ox.ac.uk

Philip Stier
Atmospheric, Oceanic, and Planetary Physics
Department of Physics
University of Oxford
philip.stier@physics.ox.ac.uk

Abstract

Aerosol-cloud interactions include a myriad of effects that all begin when aerosol enters a cloud and acts as cloud condensation nuclei (CCN). An increase in CCN results in a decrease in the mean cloud droplet size (r_e). The smaller droplet size leads to brighter, more expansive, and longer lasting clouds that reflect more incoming sunlight, thus cooling the earth. Globally, aerosol-cloud interactions cool the Earth, however the strength of the effect is heterogeneous over different meteorological regimes. Understanding how aerosol-cloud interactions evolve as a function of the local environment can help us better understand sources of error in our Earth system models, which currently fail to reproduce the observed relationships. In this work we use recent non-linear, causal machine learning methods to study the heterogeneous effects of aerosols on cloud droplet radius.

*Equal contribution.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2021.

1 Clouds remain the largest source of uncertainty for future climate projections

Aerosol-cloud interactions include a myriad of effects that are initiated when aerosol, released through natural or anthropogenic activities, enters a cloud and acts as cloud condensation nuclei (CCN). Theoretically, if you hold the liquid water content of a cloud constant, an increase in CCN results in a decrease in the mean cloud droplet size (r_e). These smaller droplets increase the brightness of the cloud [38] and delay precipitation formation [3]. The resulting brighter, larger, and longer lasting cloud reflects more incoming solar radiation thus cooling the earth. The IPCC 6th report estimates an increase in the magnitude of cooling due to aerosol-cloud interactions compared to the 5th report, without any improvement to the confidence level [27]. Aerosol-cloud interactions, and in particular the adjustments due to the delay in precipitation on cloud amount and liquid water content, remain a large source of uncertainty in future climate projections of global warming.

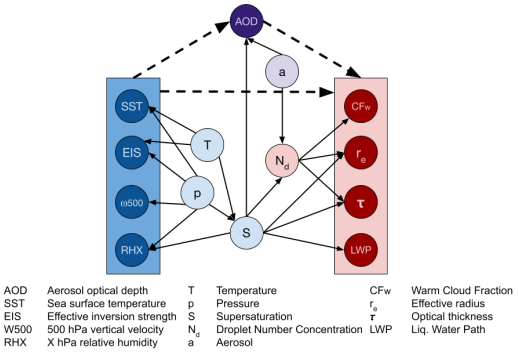


Figure 1: Causal diagram of the aerosol and aerosol proxy AOD (top, purple), affected cloud properties (red, right), and the environmental confounders (blue).

Globally, aerosol-cloud interactions work to cool the Earth, however the strength of the effect is heterogeneous. One source of heterogeneity is modulation by local meteorology [2, 35, 9, 13]. Understanding how the local environment influences aerosol-cloud interactions can help us improve our Earth system model parameterizations of these effects and outcomes. In this work we use recent non-linear causal machine learning methods [4, 20] to study the heterogeneous effects of aerosols on cloud droplet radius.

2 Background

2.1 Meteorological Influence on Aerosol-Cloud Interactions

Aerosol-cloud interactions can be labelled as a heterogeneous effect, as their sign and magnitude is modulated by a cloud's local meteorology. The local meteorology confounds the estimated effect size as it impacts both the properties of the cloud as well as the aerosol conditions in the cloud's environment. The relationships between the local meteorology, cloud properties, and aerosol are illustrated in Figure 1. Cloud properties (in red) are dependent on the amount of suitable aerosol (a, light purple) that potentially can be activated as cloud condensation nuclei (increasing Nd, light red); activation itself is dependent on the level of supersaturation within a cloud and the size, shape, and type of aerosol (S, light blue). Supersaturation itself cannot be directly observed using satellites and is dependent on other similarly unobserved meteorological variables like the temperature and pressure profiles within a cloud [23]. For these hidden variables (T, p, and S), we can rely on indirect diagnostics such as estimated inversion strength (EIS), sea surface temperature (SST), upper level convergence/divergence (w500), and relative humidities at 700 and 850 mb (RHx) (dark blue) to approximate their effects on aerosol (a).

The dependence on indirect diagnostics underscores how evaluating relationships between aerosol and observed cloud properties, without accounting for confounders, is fundamentally flawed. Furthermore, we do not have a direct way of observing the amount of aerosol in the atmosphere or the number of cloud condensation nuclei within a cloud, instead relying on aerosol optical depth (AOD), an indirect measure of the amount of aerosol in a column of atmosphere. However, aerosol optical depth is subject to hygroscopic growth, or swelling in the presence of high humidity environments such as near cloud edges [10], an additional confounding influence on the effect of aerosol on cloud properties.

Table 1: Satellite Observations and Climate Reanalysis Used

Product name	Description
Droplet Radius (r_e)	MODIS 1.6, 2.1, and 3.7 μm channels [6]
Precipitation	NOAA CMORPH Climate Data Record [32]
Sea Surface Temperature	NOAA WHOI Climate Data Record [12]
Vertical Motion at 500 mb	MERRA-2 Reanalysis [8]
Estimated Inversion Strength	MERRA-2 Reanalysis [39, 17]
Relative Humidity$_{700}$ & RH$_{850}$	MERRA-2 Reanalysis [17]
Aerosol Optical Depth (AOD)	MERRA-2 Reanalysis [17]

2.2 Causal Machine Learning for Heterogeneous Effect Estimation

The potential outcomes framework of causal inference [30, 33] provides a principled methodology for estimating the heterogeneous the effect of a binary treatment $\mathrm{T} \in \{0, 1\}$ on outcomes Y for units described by covariates $\mathbf{X}$. The treatment effect for a unit u is defined as the difference in potential outcomes $\mathrm{Y}^1(u) - \mathrm{Y}^0(u)$, where the r.v. Y^1 represents the potential outcome were the unit *treated*, and the r.v. Y^0 represents the potential outcome were the *not treated*. Realizations of the random variables $\mathbf{X}$, T, Y, Y^0, and Y^1 are denoted by $\mathbf{x}$, t, y, y^0, and y^1, respectively.

The unit level treatment effect is a fundamentally unidentifiable quantity, so instead we look at the Conditional Average Treatment Effect (CATE): $\mathrm{CATE}(\mathbf{x}) \equiv \mathbb{E}[\mathrm{Y}^1 - \mathrm{Y}^0 \mid \mathbf{X} = \mathbf{x}]$ [1] to estimate heterogeneous effects. The CATE is identifiable from an observational dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i)\}_{i=1}^n$ under the following three assumptions: **Consistency** $\mathrm{y} = \mathrm{ty}^{\mathrm{t}} + (1 - \mathrm{t})\mathrm{y}^{1-\mathrm{t}}$, i.e. an individual's observed outcome y given assigned treatment t is identical to their potential outcome y^{t}; **Unconfoundedness** $(\mathrm{Y}^0, \mathrm{Y}^1) \perp\!\!\!\perp \mathrm{T} \mid \mathbf{X}$.; **Overlap** $0 < \pi_{\mathrm{t}}(\mathbf{x}) < 1 : \forall \mathrm{t} \in \mathcal{T}$, where $\pi_{\mathrm{t}}(\mathbf{x}) \equiv \mathrm{P}(\mathrm{T} = \mathrm{t} \mid \mathbf{X} = \mathbf{x})$ is the *propensity for treatment* for individuals described by covariates $\mathbf{X} = \mathbf{x}$ [33]. When these assumptions are satisfied, $\widehat{\mathrm{CATE}}(\mathbf{x}) \equiv \mathbb{E}[Y \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x}] - \mathbb{E}[Y \mid \mathrm{T} = 0, \mathbf{X} = \mathbf{x}]$ is an identifiable, unbiased estimator of $\mathrm{CATE}(\mathbf{x})$. The population level average treatment effect (ATE) is then the expectation of the CATE over all $\mathbf{X}$.

When using satellite observations as we are within (presented in Table 1), it is common for these assumptions to fail. Nonetheless, these assumptions give us a perspective to understand why we should be uncertain about the effects we estimate from observational data [19, 20]. We use the scalable, uncertainty-aware machine learning methodology of Jesson et al. [20] to model non-linear, heterogeneous aerosol-cloud interactions.

3 Exploring Heterogeneous Aerosol Cloud Interactions using Non-linear Causal Models

First, we need to translate the problem of estimating aerosol-cloud interactions into a causal problem. We consider SST, EIS, RH$_{700}$, RH$_{850}$, and w500 as our covariates $\mathbf{X}$. The AOD is our treatment variable T. We discretize the measured AOD by applying a threshold at its median value of 0.3. Examples with raw AOD values less than 0.07 and greater than 1.0 are discarded. Finally, we look at each of r_e, CF_w, τ, and LWP as our outcome variables Y, but focus our analysis on r_e in the main text. We further focus on non-precipitating clouds by discarding examples with precipitation greater than 0.05. We look at daily averages between 2003 and 2020. We generate training, validation, and test sets by partitioning weekdays into the training set, Saturdays into the validation set, and Sundays into the test set.

We use Quince [20] to estimate the heterogeneous effects, which are validated against estimates from Bayesian linear regression [26, 37] and causal forest [4] in the appendix. Implementation details of each are given in the appendix.

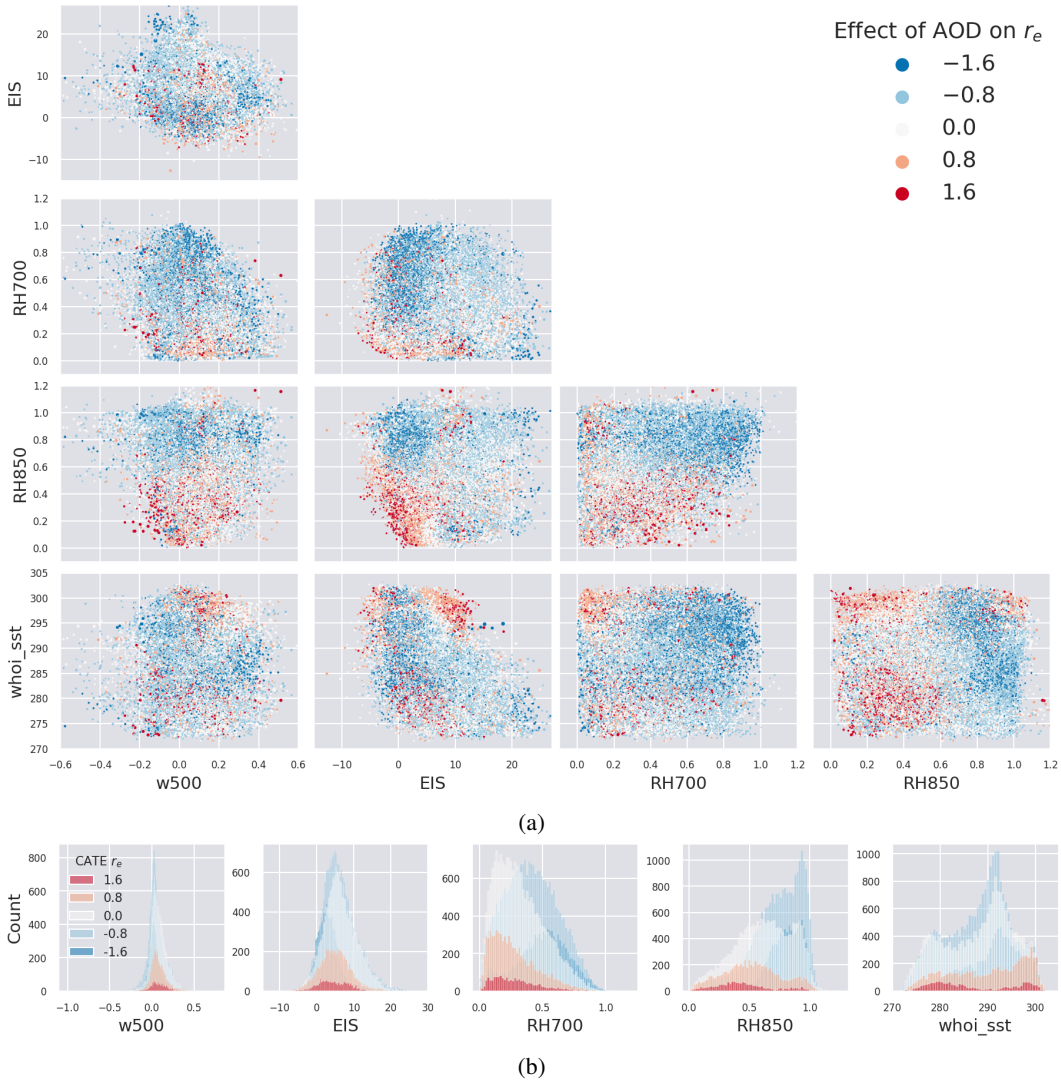


Figure 2: The effect estimates (a) of AOD on r_e using Quince for each of the ten possible environmental combinations and the histograms (b) of these effects against each of the environmental parameters.

4 Results

The average treatment effect (ATE) for AOD on r_e, measured as the difference in r_e between above average and below average aerosol cases, varies from -.359 $\pm$.533 to -.452 $\pm$.010 depending on the causal method chosen (Table 2). On average, an increase in AOD results in a reduction of r_e. Positive effects of AOD on r_e have been found to occur under certain conditions, but as expected are rare occurrences [41, 21]. The heterogeneous effects of AOD on r_e from the Quince model, or the impact of different meteorological conditions on the effect's magnitude and sign, are shown in Figure 2. The effect's magnitude and sign are directly impacted by changing environmental regime, demonstrating that aerosol effects on r_e are an example of a heterogeneous effect. By evaluating the effect as a function of different environmental parameters, the processes responsible for altering the sign or magnitude can be better understood. For example, in the RH850 vs. EIS diagram, the effect goes from negative (decreasing r_e with increasing AOD) to positive (increasing r_e with increasing AOD) as RH850 and stability both decrease. In unstable, dry conditions, convection can occur, so the effect of aerosol is dependent on the humidity transported into the cloud layer [14]. Conversely, in humid and

Table 2: Comparison of the average effect size estimates for three different causal models. The mean and standard deviation of the approximate posterior distribution (BLR) or mean and standard error of the model ensemble (Causal Forest and Quince) are reported along with the test set R^2 values for regressing the outcome given the observed AOD values.

	Estimated ATE of above vs. below average AOD on r_e			R^2
Method	Train	Valid	Test	Test
BLR	$-.287 \pm .015$			$.15\pm$
Causal Forest	$-.328 \pm .568$	$-.325 \pm .569$	$-.359 \pm .533$	$.21\pm$
Quince	$-.429 \pm .009$	$-.413 \pm .010$	$-.452 \pm .010$	$.23\pm$

stable conditions, clouds are prone to be aerosol-limited, meaning any increase in aerosol leads to a large, negative impact on the r_e [24].

In order to employ Quince to quantify aerosol-cloud interactions using satellite observations, a number of assumptions are made that may alter the overall quality of the predictions. The first being that aerosol optical depth directly impacts r_e, which we know is not the case. AOD is a proxy for aerosol and is not a direct measure of the amount of aerosol present in the atmosphere. We did not include how aerosols, clouds, and confounding effects are spatially correlated. Any spatial correlations between the environment, aerosol, and the effect size is ignored by Quince, but may be acting as a hidden confounding influence. Additionally, given we are only using proxies (EIS, SST, w500, and RHX) of the true confounders (T, p, and S), there are likely to exist other environmental features that additionally capture the true confounders influence on the effect size of aerosol on r_e. Currently, we are using daily averages at a 1° x 1° scale, which may be restricting the predictive power of our predictions, as the scale of the interactions is on the order of kilometers (cloud scale), which may not be captured by the regional, daily mean.

We do not have a true, counterfactual value to compare our predictions against, making it difficult to validate either model as the "best" model for evaluating the causal, heterogeneous relationships of aerosol-cloud interactions. In theory, Quince should better resolve the effect size as it incorporates feature extraction and inherent relationship between feature correlations with both the predictors (r_e) and treatment (AOD), however this is difficult to definitively state. Quince does show a reduced range of uncertainty compared to other methods that allow for environmental causal attribution on effect size, like a Bayesian linear regressions or causal forests (Table 2).

5 Conclusions

Herein we show how the heterogeneous effects of aerosol on r_e can be evaluated using Quince, a causal neural network. This effect, and other aerosol-cloud interactions, are primed to be untangled using causal methods; non-linear, causal models such as causal forests or Quince are the ideal tools to evaluate the complex interactions between aerosol, clouds, and the environment. These models can be exploited to flexibly model the causal relationships between treatments (aerosol loading), covariates (local meteorological influence), and outcomes (changes in cloud properties). Aerosol-cloud interactions are a high-impact science problem for new causal inference methods.

6 Acknowledgements

This project was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 821205 (FORCeS) and Marie Skłodowska-Curie grant agreement No 860100 (iMIRACLI). PS and AD were supported by 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.

References

[1] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

[2] Andrew S Ackerman, Michael P Kirkpatrick, David E Stevens, and Owen B Toon. The impact of humidity above stratiform clouds on indirect aerosol climate forcing. *Nature*, 432(7020): 1014–1017, 2004.

[3] Bruce A Albrecht. Aerosols, cloud microphysics, and fractional cloudiness. *Science*, 245(4923): 1227–1230, 1989.

[4] Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application, 2019.

[5] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML, 2019. Version 0.x.

[6] Bryan A Baum and Steven Platnick. Introduction to modis cloud products. In *Earth science satellite remote sensing*, pages 74–91. Springer, 2006.

[7] Christopher M Bishop. Mixture density networks. 1994.

[8] Michael G Bosilovich, Santha Akella, Lawrence Coy, Richard Cullather, Clara Draper, Ronald Gelaro, Robin Kovach, Qing Liu, Andrea Molod, Peter Norris, et al. Merra-2: Initial evaluation of the climate. 2015.

[9] Yi-Chun Chen, Matthew W Christensen, Graeme L Stephens, and John H Seinfeld. Satellite-based estimate of global aerosol–cloud radiative forcing by marine warm clouds. *Nature Geoscience*, 7(9):643–646, 2014.

[10] Matthew W Christensen, David Neubauer, Caroline A Poulsen, Gareth E Thomas, Gregory R McGarragh, Adam C Povey, Simon R Proud, and Roy G Grainger. Unveiling aerosol–cloud interactions–part 1: Cloud contamination in satellite products enhances the aerosol indirect forcing estimate. *Atmospheric Chemistry and Physics*, 17(21):13151–13164, 2017.

[11] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[12] James L Cogan and James H Willand. Measurement of sea surface temperature by the noaa 2 satellite. *Journal of Applied Meteorology and Climatology*, 15(2):173–180, 1976.

[13] Alyson Douglas and Tristan L'Ecuyer. Quantifying cloud adjustments and the radiative forcing due to aerosol–cloud interactions in satellite observations of warm marine clouds. *Atmospheric Chemistry and Physics*, 20(10):6225–6241, 2020.

[14] Alyson Douglas and Tristan L'Ecuyer. Global evidence of aerosol-induced invigoration in marine cumulus cloud. *Atmospheric Chemistry and Physics Discussions*, pages 1–18, 2021.

[15] Stefan Falkner, Aaron Klein, and Frank Hutter. Bohb: Robust and efficient hyperparameter optimization at scale, 2018.

[16] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

[17] Ronald Gelaro, Will McCarty, Max J Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A Randles, Anton Darmenov, Michael G Bosilovich, Rolf Reichle, et al. The modern-era retrospective analysis for research and applications, version 2 (merra-2). *Journal of climate*, 30(14):5419–5454, 2017.

[18] Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021.

[19] Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33, 2020.

[20] 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*, volume 139, pages 4829–4838. PMLR, 2021. URL `https://proceedings.mlr.press/v139/jesson21a.html`.

[21] Hongli Jiang, Huiwen Xue, Amit Teller, Graham Feingold, and Zev Levin. Aerosol effects on the lifetime of shallow cumulus. *Geophysical Research Letters*, 33(14), 2006.

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[23] Hilding Köhler. The nucleus in and the growth of hygroscopic droplets. *Transactions of the Faraday Society*, 32:1152–1161, 1936.

[24] Ilan Koren, Guy Dagan, and Orit Altaratz. From aerosol-limited to invigoration of warm convective clouds. *science*, 344(6188):1143–1146, 2014.

[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. *arXiv preprint arXiv:1807.05118*, 2018.

[26] David JC MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

[27] 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. Ipcc, 2021: Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change, 2021.

[28] 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, 2018.

[29] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814, 2010. URL `https://icml.cc/Conferences/2010/papers/432.pdf`.

[30] Jerzy S Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51, 1923.

[31] 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. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[32] 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. *Journal of Hydrometeorology*, 22(9):2291–2310, 2021.

[33] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[34] Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects, 2019.

[35] Jennifer D Small, Patrick Y Chuang, Graham Feingold, and Hongli Jiang. Can aerosol decrease cloud lifetime? *Geophysical Research Letters*, 36(16), 2009.

[36] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL `http://jmlr.org/papers/v15/srivastava14a.html`.

[37] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

[38] Sean Twomey. The influence of pollution on the shortwave albedo of clouds. *Journal of the atmospheric sciences*, 34(7):1149–1152, 1977.

[39] Robert Wood and Christopher S Bretherton. On the relationship between stratiform low cloud cover and lower-tropospheric stability. *Journal of climate*, 19(24):6425–6432, 2006.

[40] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network, 2015.

[41] Tianle Yuan, Zhanqing Li, Renyi Zhang, and Jiwen Fan. Increase of cloud droplet size with aerosol optical depth: An observation and modeling study. *Journal of Geophysical Research: Atmospheres*, 113(D4), 2008.

A Additional Results

Table 5 enumerates the estimated ATE of AOD on each of CF_w, r_e, τ, and LWP. We see that the non-linear causal models estimate AOD having a larger magnitude average effect on CF_w and r_e in the south pacific compared to the linear model estimate. Conversely we see that the non-linear causal models estimate AOD having a smaller magnitude average effect on τ, and LWP in the south pacific compared to the linear model estimate.

Table 3: Comparison of the average treatment effect (ATE) estimates of different methodologies. The mean and standard deviation of the approximate posterior distribution (BLR) or mean and standard error of the model ensemble (Causal Forest and Quince) are reported.

	Estimated ATE of AOD on CF_w			Estimated ATE of AOD on r_e		
Method	Train	Valid	Test	Train	Valid	Test
BLR	$-.002 \pm .001$			$-.29 \pm .02$		
Forest	$.003 \pm .030$	$.005 \pm .030$	$.004 \pm .029$	$-.33 \pm .57$	$-.33 \pm .57$	$-.36 \pm .53$
Quince	$.008 \pm .001$	$.008 \pm .001$	$.008 \pm .001$	$-.43 \pm .01$	$-.41 \pm .01$	$-.45 \pm .01$

	Estimated ATE of AOD on τ			Estimated ATE of AOD on LWP		
Method	Train	Valid	Test	Train	Valid	Test
BLR	2.34 ± 0.04			17.8 ± 0.40		
Forest	2.05 ± 1.37	2.03 ± 1.37	2.02 ± 1.27	13.3 ± 15.6	13.0 ± 15.6	12.8 ± 14.5
Quince	1.64 ± 0.04	1.61 ± 0.04	1.61 ± 0.04	8.49 ± 0.46	8.34 ± 0.47	8.05 ± 0.46

In Figure 3 we look at the heterogeneity of Quince estimated effect sizes of AOD on CF_w across pairs of covariates.

In Figure 4 we look at the heterogeneity of Quince estimated effect sizes of AOD on τ across pairs of covariates.

In Figure 5 we look at the heterogeneity of Quince estimated effect sizes of AOD on LWP across pairs of covariates.

B Sanity Checks

It is impossible to observe ground truth treatment effects, so validating the truth of the reported results is complicated. As a first measure of sanity, we look at the accuracy of regressing the outcome, measured by the coefficient of determination (R^2) between predicted and observed outcomes for each of CF_w, r_e, τ, and LWP. In Figure 6, we show these results for Quince across the train, validation, and test data splits.

In Figure 7, we compare the results on the test data split for each of the Bayesian Linear Regression, Causal Forest, and Quince methods. We see that the non-linear models improve regression accuracy and a further marginal gain in performance between the Quince and Causal Forest methods.

In Figure 8, we plot the estimated heterogeneous effects from both the Quince and Causal Forest methodologies side by side for each of CF_w, r_e, τ, and LWP. At a high level we see the same general trends given by both methods.

This general trend is also reflected in Figure 9 where we plot the estimated CATE values for the Quince and Causal Forest methods against one another. The Spearman rank correlation coefficient r is also reported.

However, we can see that the relationship is not perfect and that there is disagreement between the estimated ATE values of each method (Table 5). So can we go further to see if we can learn from the discrepancy between predictions?

C Dataset Details

All observations of r_e are from the daily mean, 1° x 1° resolution Cloud Product from MODIS aboard both Aqua and Terra. The effective cloud droplet radius is found for pixels that are both

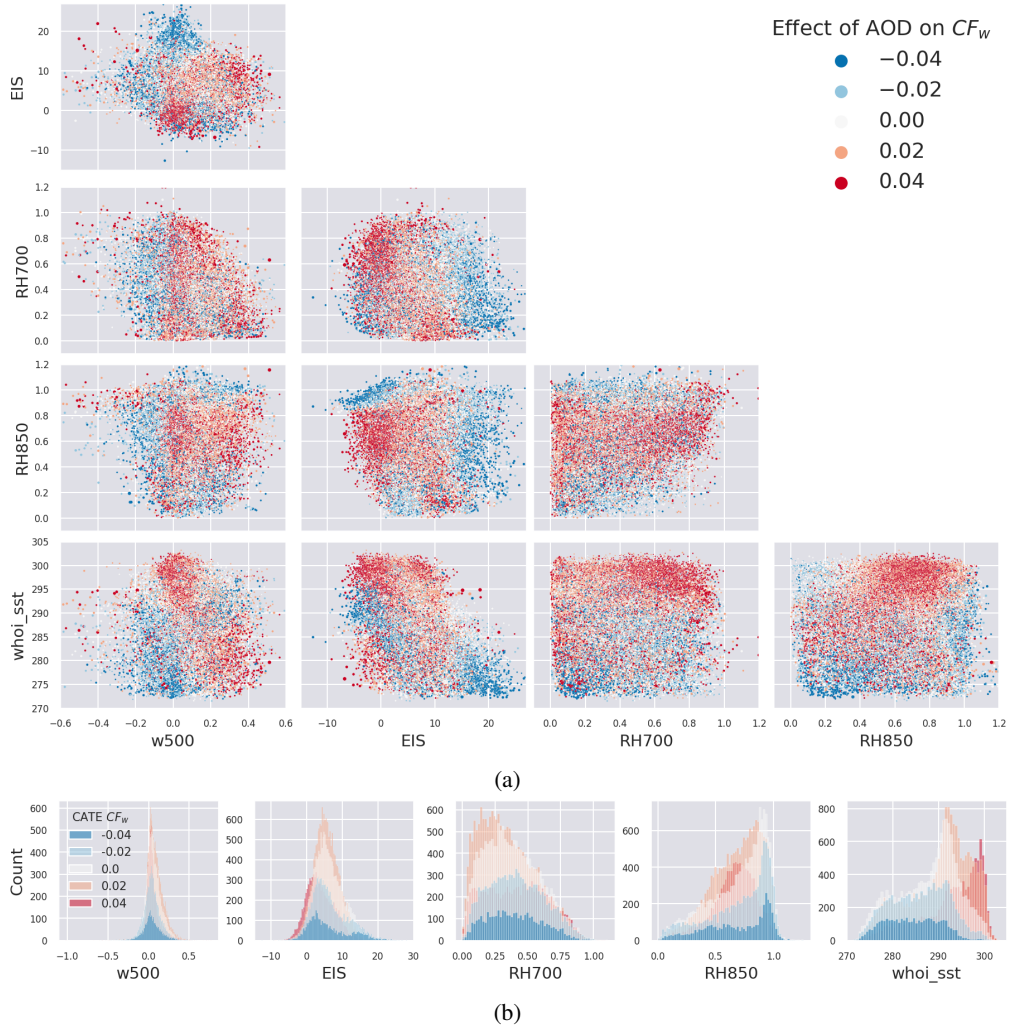


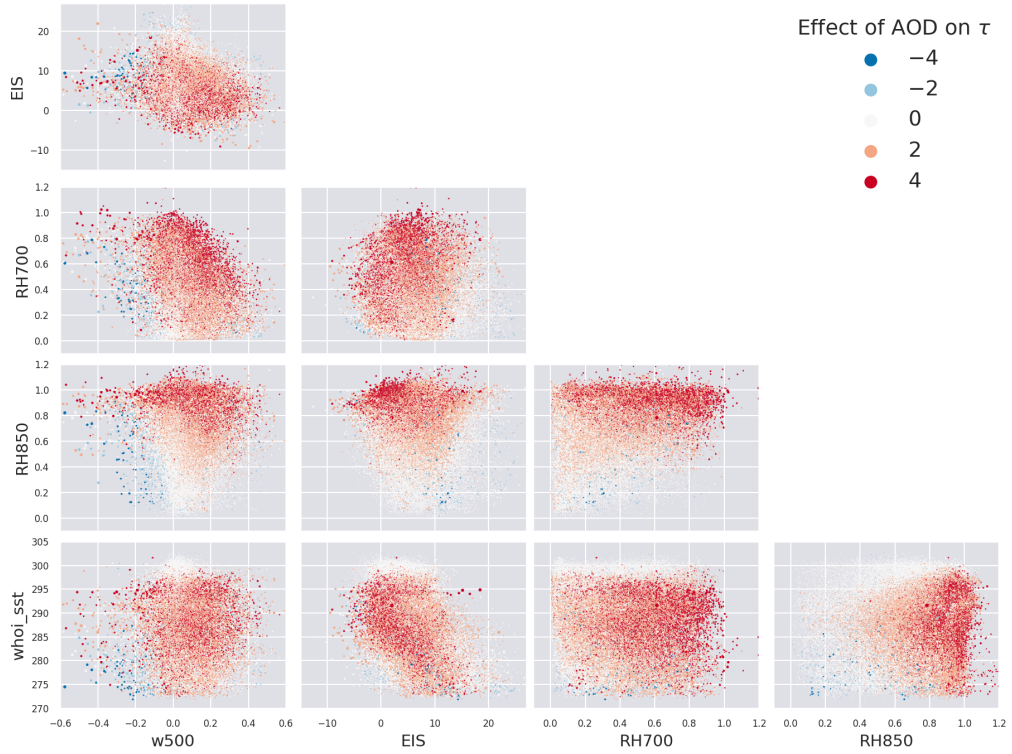
Figure 3: The treatment effect of AOD on CF_w, defined as difference between the high and low aerosol. (a) Each plot shows the effect size as the color of scatter points, while the position of the points indicates the values of the observed meteorological covariates. (b) Histograms of effect sizes plotted against each individual meteorological covariate.

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 temperatures from NOAA WHOI CDR is found using multiple observations of surface brightness temperature and incorporating precipitation estimates in order to better approximate the effects of the diurnal cycle on sea surface temperature. The MERRA-2 model, which calculates global profiles of temperature, relative humidity, and pressure, assimilates hyperspectral and passive microwave satellite observations to enhance its ability to model Earth's atmosphere.

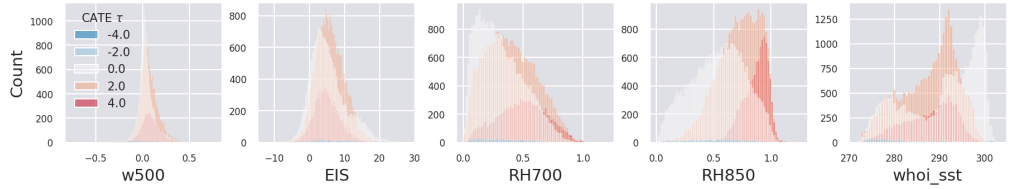
D Implementation Details

D.1 Quince

We follow Jesson et al. [19] and use an ensemble of Mixture Density Networks (MDNs) [7]. Each MDN is adapted for causal-effect inference by following the Dragonnnet architecture of Shi et al. [34]. The deep neural network producing the hidden representation has 3 hidden residual layers with 800 neurons each. Dropout [36] is applied at a rate of 0.5 after each linear hidden layer followed by



(a)



(b)

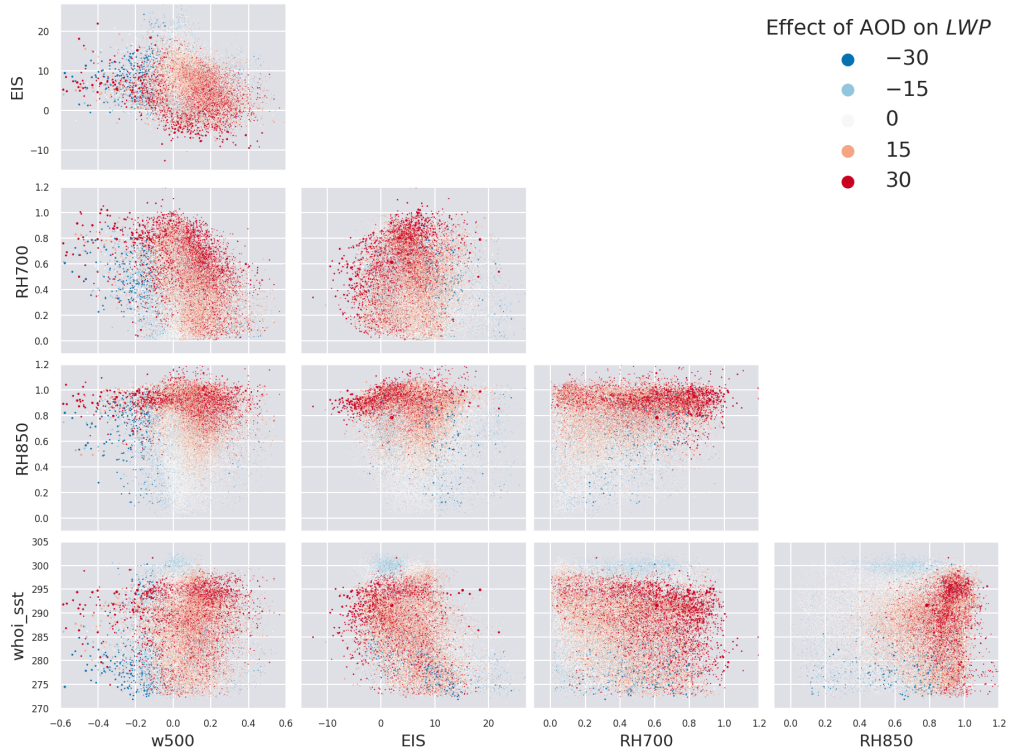
Figure 4: The treatment effect of AOD on τ (Cloud Optical Thickness), defined as difference between the high and low aerosol. (a) Each plot shows the effect size as the color of scatter points, while the position of the points indicates the values of the observed meteorological covariates. (b) Histograms of effect sizes plotted against each individual meteorological covariate.

ReLU activation functions [16, 29]. The treatment prediction head is a linear layer followed by a pytorch Bernoulli distribution layer. Each conditional outcome head is a MDN comprised of a linear layer followed by a pytorch MixtureSameFamily distribution with Normal component distributions. Each MDN has 20 mixure components. The sum of the log likelihoods for both the treatment head and each conditional outcome head multiplied by the observed treatment is minimized using Adam optimization [22] with learning rate of 0.0002 and pytorch default parameters. The model is trained on the training data split for 400 epochs with a batch size of 4096. Model selection is done by evaluating the average R2 score across the 4 outcomes on the validation set and selecting the model parameters at the epoch with the best score. We use an ensemble size of 10.

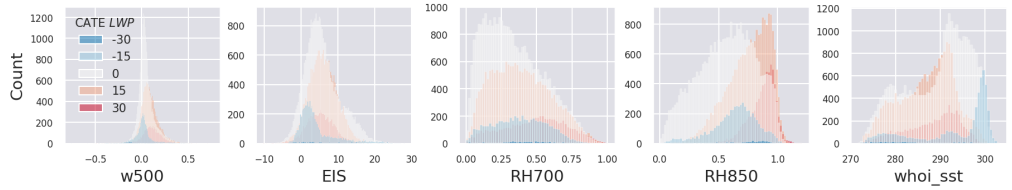
We use Ray Tune to optimize our hyper-parameters [28, 25] with the Bayesian Optimization HyperBand algorithm [15]. The space we search over is given in Table 4.

D.2 Causal Forest

We use the Causal Forest of Athey and Wager [4] as implemented in Microsoft's EconML python package [5]. We use 1000 estimators, max-samples 0.5, discrete-treatment True, and default parame-



(a)



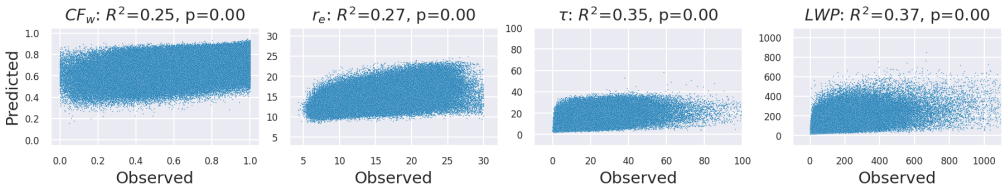
(b)

Figure 5: The treatment effect of AOD on LWP, defined as difference between the high and low aerosol. (a) Each plot shows the effect size as the colour of scatter points, while the position of the points indicates the values of the observed meteorological covariates. (b) Histograms of effect sizes plotted against each individual meteorological covariate.

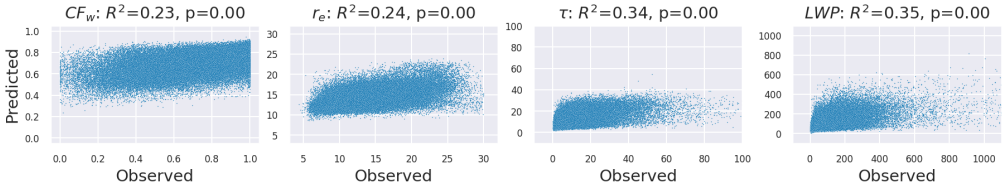
ters for the rest. We did a grid search using the validation set over the number of estimators [100, 500, 1000].

D.3 Bayesian Linear Regression

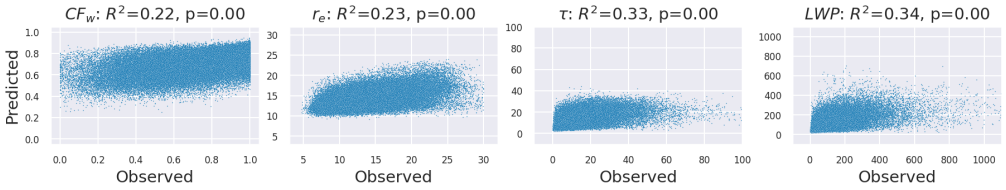
We use Bayesian Ridge regression [26, 37] from scikit learn with default parameters [31]. The model is fit on the training data. We report the coefficient for the treatment input.



(a) Train



(b) Validation

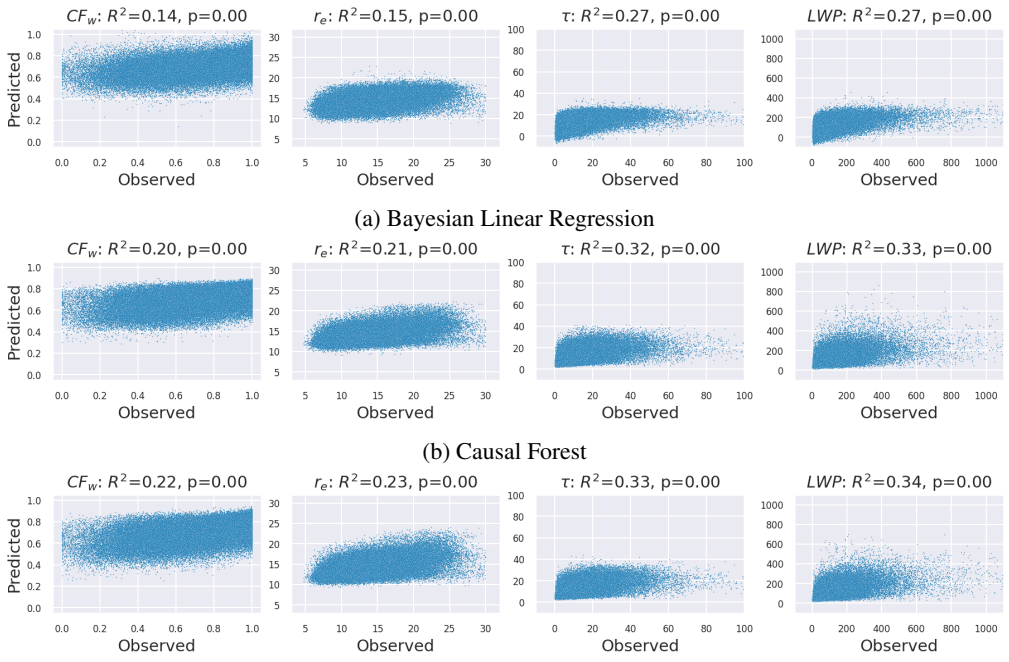


(c) Test

Figure 6: Comparing regression accuracy of Quince on train, validation and test splits. The squared Pearson R coefficient is shown with associated p-value. We can see consistent performance across all splits providing evidence that our results will generalize.

Table 4: Quince hyper-parameter search space

Hyperparameter	Space
dim hidden	[50, 100, 200, 400, 800]
depth	[2, 3, 4, 5]
num components	[1, 2, 5, 10, 20]
dropout rate	[0.0, 0.1, 0.2, 0.5]
spectral norm [18]	[0.0, 0.95, 1.5, 3.0]
negative slope [11, 40]	[0.0, 0.1, 0.2, 0.3, elu]
learning rate	[0.0002, 0.0005, 0.001]
batch size	[1024, 2048, 4096]



(a) Bayesian Linear Regression

(b) Causal Forest

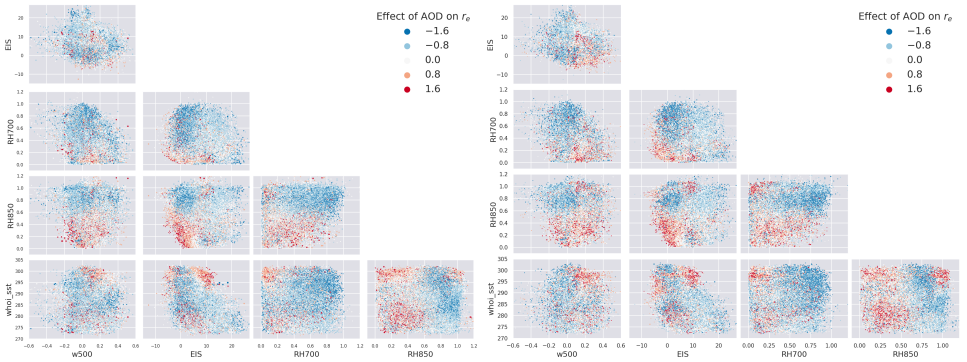
(c) Quince

Figure 7: Comparing regression test set accuracy between Quince, Causal Forest and Bayesian Linear Regression. The squared Pearson R coefficient is shown with associated p-value. We see that the non-linear models are better predictors of each outcome.

Table 5: Comparison of the average treatment effect (ATE) estimates of different methodologies. The mean and standard deviation of the approximate posterior distribution (BLR) or mean and standard error of the model ensemble (Causal Forest and Quince) are reported.

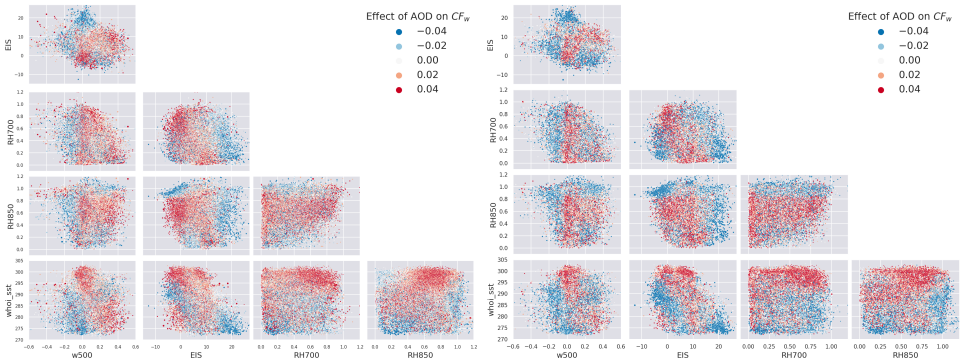
	Estimated ATE of AOD on CF_w			Estimated ATE of AOD on r_e		
Method	Train	Valid	Test	Train	Valid	Test
BLR	$-.002 \pm .001$			$-.29 \pm .02$		
Forest	$.003 \pm .030$	$.005 \pm .030$	$.004 \pm .029$	$-.33 \pm .57$	$-.33 \pm .57$	$-.36 \pm .53$
Quince	$.008 \pm .001$	$.008 \pm .001$	$.008 \pm .001$	$-.43 \pm .01$	$-.41 \pm .01$	$-.45 \pm .01$
Context	$.001 \pm .003$	$.001 \pm .003$	$.001 \pm .003$	$-.06 \pm .06$	$-.04 \pm .06$	$-.10 \pm .06$

	Estimated ATE of AOD on τ			Estimated ATE of AOD on LWP		
Method	Train	Valid	Test	Train	Valid	Test
BLR	2.34 ± 0.04			17.8 ± 0.40		
Forest	2.05 ± 1.37	2.03 ± 1.37	2.02 ± 1.27	13.3 ± 15.6	13.0 ± 15.6	12.8 ± 14.5
Quince	1.64 ± 0.04	1.61 ± 0.04	1.61 ± 0.04	8.49 ± 0.46	8.34 ± 0.47	8.05 ± 0.46
Context	1.58 ± 0.21	1.56 ± 0.21	1.52 ± 0.22	12.8 ± 2.25	12.4 ± 2.26	11.7 ± 2.30



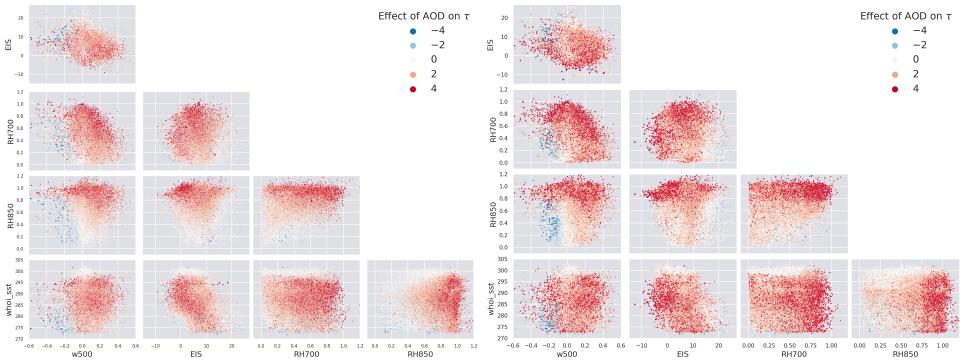
(a) Quince: Effect of AOD on r_e

(b) Causal Forest: Effect of AOD on r_e



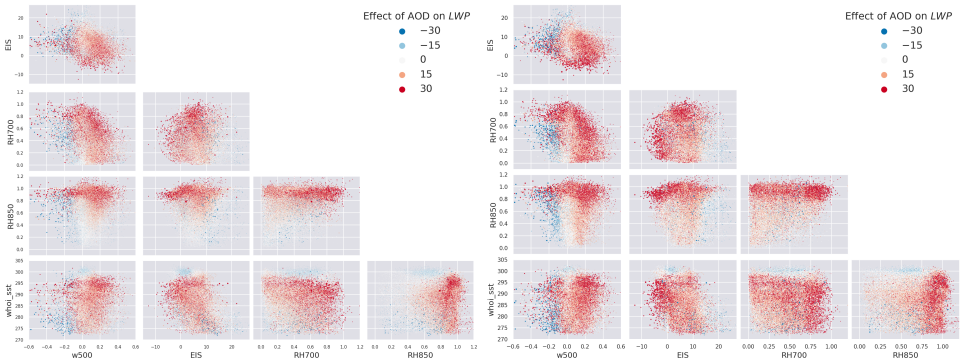
(c) Quince: Effect of AOD on CF_w

(d) Causal Forest: Effect of AOD on CF_w



(e) Quince: Effect of AOD on τ

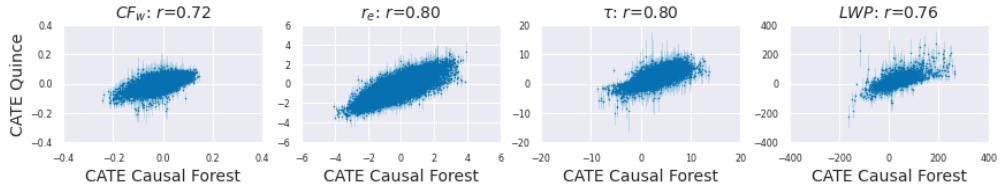
(f) Causal Forest: Effect of AOD on τ



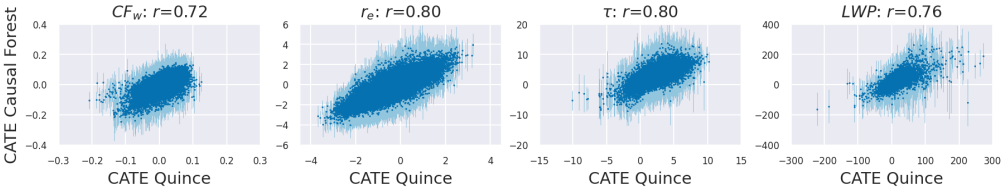
(g) Quince: Effect of AOD on LWP

(h) Causal Forest: Effect of AOD on LWP

Figure 8: Qualitative comparison of heterogeneous effect estimates of AOD on outcome variables between Quince methodology and Causal Forest. We see the same general trends using two different methods which lends evidence to support the treatment effect estimates reported.



(a) CATE Quince vs. Causal Forest with Quince Uncertainty



(b) CATE Causal Forest vs. Quince with Causal Forest Uncertainty

Figure 9: TODO