Technology, Information, and the Governance of Environmental Risk

Lucas Kruitwagen

Mansield College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2020

Abstract

This DPhil is motivated by the impact that public data can have on the practise of sustainable finance. Disclosure guidance for environmental information is increasingly being used to compel companies to better manage their environmental risk exposure and environmental impacts. Data-driven technologies can 'unravel' company information related to environmental risk and are changing the information asymmetries between company managers, their investors, and society.

Three novel and rigorous data science papers have been prepared which provide new access to company environmental risk information. The first paper develops a global inventory of photovoltaic solar energy facilties using machine learning and earth observation and then uses this dataset to assess land cover impacts due to renewable energy siting decisions. The second paper develops asset-level data arrangements of the coal, oil, and gas supply chains, and identifies concentrations of environmental transition risk in the complex network topology. The third paper advances earth observation and computer vision technical methods, assessing the power of these techniques for producing company asset-level data even with limited training data.

In developing these papers, this DPhil has identified the opportunities and limits that technology might have in the provision of information for public and private governance of environmental risk and impact. These limits add clarity to the role that policy must play in meeting the diverse information needs of both company shareholders *and* stakeholders. A discussion chapter provides commentary on these limits; on how the needs of both shareholders and stakeholders can be met with company risk disclosure at the asset-level; and how innovation might be fostered in markets for environmental information. Policy recommendations are summarised and an agenda for further research is provided.

Technology, Information, and the Governance of Environmental Risk



Lucas Kruitwagen

Mansield College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2020

This thesis is dedicated to those who see a sustainable future, equitably enjoyed, not as a matter for speculation, but one of choice.

Acknowledgements

This work represents a personal milestone after six years of postgraduate research. It is beyond time that I make good on the enormous debt of gratitude I have accumulated!

First, I'd like to thank my supervisors Ben Caldecott and Cameron Hepburn. Ben first welcomed me to Oxford and showed me that research could be exciting and impactful. He encouraged tremendous ambition and gave me opportunities to pursue it. Cameron taught me to temper novelty with rigour and helped me navigate the arcane customs of the academy. He helped me find my peers who made me feel at home in Oxford. His messages of encouragement were uncannily timed — arriving precisely when I needed them the most.

I'd also like to thank my mentors Mark Workman and Alex Money. They provided kind ears and advice ranging from the mundane ("what's the dress code?") through to the professional and deeply personal.

In preparing this DPhil I had the immense privilege to meet and collaborate with a wide range of individuals and organisations. This DPhil would not have been possible without the support of friends and colleagues at the Smith School of Enterprise and the Environment, the Oxford Sustainable Finance Programme, the Institute for New Economic Thinking, the Oxford Martin School, the Stanford Global Projects Center, the Stanford Global Energy Atlas Working Group, the World Resources Institute, the Global Energy Monitor, the Natural Resources Governance Institute, DescartesLabs, QuantumBlack, the Frontier Development Lab, the Google GeoforGood community, and the Microsoft AI for Earth community. I'd like to especially thank my coauthors on my DPhil papers: Kyle Story, Sam Skillman, Johannes Friedrich, Logan Byers, Jannes Klaas, Arash Baghaei Lakeh, and Jessica Fan. I thank them for the risk they took in working with me, for teaching me and for collaborating with me, and for the excess of patience they have had in anticipating the publication of our work.

To my friends in Toronto, Montreal, Oxford, London, California, and elsewhere, I miss you and hope to see you soon.

To my siblings, Mariel, Natasha, and Lindsay, and my parents Godwin and Lisa, this journey was possible only with your constant support and warm welcomes home.

To my partner, Rebecca, I love you and would not have been able to do this without you.

My DPhil research was funded by the Social Sciences and Humanities Research Council of Canada. I am immensely grateful for their support. I am also grateful for smaller grants received from Mansfield College, the School of Geography and the Environment, and the Grand Union Doctoral Training Programme. In-kind support was given by both DescartesLabs and QuantumBlack. These companies gave me valuable time with their professionals which significantly enriched my DPhil.

The three papers contained herein have excessive computational components. DescartesLabs granted me extensive access to their cloud computation infrastructure. Further computing resources were provided by Microsoft's AI for Earth programme, DescartesLabs Impact Science Programme, a Google Cloud Platform Research Credits grant, and an Amazon Web Services credits grant.

Finally, I'd like to thank Sophie Lopez-Welsch and the Oxford University Counselling Service, Ruth Collins and the Oxford Mindfullness Centre, and Mental Health England for their support during my DPhil.

Abstract

This DPhil is motivated by the impact that public data can have on the practise of sustainable finance. Disclosure guidance for environmental information is increasingly being used to compel companies to better manage their environmental risk exposure and environmental impacts. Data-driven technologies can 'unravel' company information related to environmental risk and are changing the information asymmetries between company managers, their investors, and society.

Three novel and rigorous data science papers have been prepared which provide new access to company environmental risk information. The first paper develops a global inventory of photovoltaic solar energy facilties using machine learning and earth observation and then uses this dataset to assess land cover impacts due to renewable energy siting decisions. The second paper develops asset-level data arrangements of the coal, oil, and gas supply chains, and identifies concentrations of environmental transition risk in the complex network topology. The third paper advances earth observation and computer vision technical methods, assessing the power of these techniques for producing company asset-level data even with limited training data.

In developing these papers, this DPhil has identified the opportunities and limits that technology might have in the provision of information for public and private governance of environmental risk and impact. These limits add clarity to the role that policy must play in meeting the diverse information needs of both company shareholders *and* stakeholders. A discussion chapter provides commentary on these limits; on how the needs of both shareholders and stakeholders can be met with company risk disclosure at the asset-level; and how innovation might be fostered in markets for environmental information. Policy recommendations are summarised and an agenda for further research is provided.

A Motivating Anecdote

In 2015, my colleagues in the Oxford Sustainable Finance Programme and I were preparing an analysis of environmental risk in the global coal supply chain for a very large sovereign wealth fund. We developed 'local risk hypotheses' to assess company exposure to, among other things, climate change-induced heat and water stress, and certain types of technology transitions. Assessment of company exposure to these hypotheses required us to know the precise latitude and longitude of coal-fired power stations, coal processing facilities, and coal mines. We approached the only data vendor with a substantially complete dataset of the world's coal mines. The data vendor was unwilling to sell us the data we required. They told us that the price for their data was well beyond what we could afford for our public-goods research project. More troublingly, they insinuated that even if we could afford the data, they would not sell it to us on the grounds that our purpose for it would be to the detriment of the companies who form their primary market.

This suggested to me that the existence and price of information was not fit for the purpose of mitigating environmental risks and impacts. Moreover, the very availability of information for public-sector decision makers was political. To me, these problems were worth investigating with a DPhil. As my colleagues and I sought alternative methods for acquiring this asset-level data, I would also use my DPhil to develop these methods and contribute to emerging open data alternatives.

I hypothesised that publicly-accessible satellite-derived remote sensing imagery might contain the information needed to localise these large fixed-capital assets. My first paper was a crash introduction to machine learning, remote sensing, and computer vision. My coauthors and I achieved what was, at the time, the single largest deployment of a computer vision model on remote sensing imagery. With my second paper, I sought to demonstrate the use of asset-level data in downstream analyses of environmental risk in supply chains. With my third paper, I returned to remote sensing, preparing a general-purpose tool that would streamline the localisation of large fixed-capital assets and their features. I then used this tool to produce, at long last, an initial global dataset of the world's surface coal mines. All the code and resulting datasets of this DPhil have been made publicly available with the exception of the code and data for Paper A which are embargoed pending publication.

DPhil Papers

A **A Global Inventory of Commercial-, Industrial- and Utility-Scale PV Solar Energy Facilities**
Kruitwagen, L., Story, K., Friedrich, J., Byers, L. Skillman, S., & Hepburn, C.
Presented at: NeurIPS Climate Chain AI Workshop 2019 (Talk, Award: Runner-up, Best Paper); Global Research Alliance for Sustainable Finance 2019 (Award: Most Innovative Paper)
Submitted to *Nature*, currently in 4th-round peer review
Draft (embargoed): `https://drive.google.com/drive/u/0/folders/1s_kx1qP9kyHS_IIq1C7zBJlUgdu2LUDx`
Repository (embargoed): `https://github.com/Lkruitwagen/solar-pv-global-inventory`
Repository (alternative): `https://drive.google.com/file/d/17Tj9dN3n3qnqTARO1rlVXWjtPTDymSkZ/view?usp=sharing`
Data (embargoed): `https://drive.google.com/file/d/1RUJvGmog1OmNo_xEUjEk8UzoXApcK5Il/view?usp=sharing`
Data Viewer: `https://lucaskruitwagen.users.earthengine.app/view/solarpvdemo`

B **Transition Risk in the Coal-, Oil- and Gas Supply Chains**
Kruitwagen, L., Jannes, K., Lakeh, A., & Vitarana, A.
Submitted to SSRN preprint.
Draft: `https://papers.ssrn.com/abstract=3783412`
Repository: `https://github.com/Lkruitwagen/global-fossil-fuel-supply-chain`
Data: `https://drive.google.com/file/d/1LWXT3WyNpMS8xmdFzStbUyQlzdPLGhv_/view?usp=sharing`

C **Towards DeepSentinel: Sentinel-1, Sentinel-2 Sensor Fusion for General-Purpose Semantic Embedding**
Kruitwagen, L.
Presented at: NeurIPS Climate Change AI Workshop 2020 (Poster); European Space Week Phi-Week (Award: Copernicus Master's Finalist)
Submitted to ArXiv preprint.
Draft: `https://<link-to-ArXiv> (available 2021-02-15)`
Repository (main): `https://github.com/Lkruitwagen/deepsentinel`
Repository (osm server): `https://github.com/Lkruitwagen/deepsentinel-osm`
Data: `https://console.cloud.google.com/storage/browser/deepsentinel`

Policy Recommendations

I offer a policy recommendation based on my finding that technology alone is unlikely to naturally unravel the exhaustive and unbiased information company shareholders and stakeholders require to mitigate environmental risks and impacts. I recommend a subversive model of mandatory disclosure of the underlying physical assets of a company. This asset-level data should entail:

- The geospatial arrangement of physical plant and equipment assets at their natural spatial agglomeration;
- Nameplate information of the asset, such as facility name, address, production capacity, etc.;
- A measure of materiality of the asset, such as capital valuation, operating revenue or profit, annual production, worker headcount, etc.

Mandatory asset-level data disclosure, as a generalised model, will achieve the following, with benefits for social welfare:

- Provide general-purpose data with which company shareholder and stakeholders can assess their exposure to environmental risks and impacts using sophisticated tools beyond the purview of company managers.
- Avert a competition between various company stakeholders for bespoke company disclosures from time-constrained company managers. This will also better delineate the role of risk analysis and mitigation at the shareholder and stakeholder level versus at the company management level.
- Reduce information asymmetry between company managers and non-material stakeholders, empowering the latter, potentially to the extent that their welfare becomes material to the company.
- De-monopolise markets for financial data where innovation has stagnated due to data monopoly rent-seeking.

Contents

List of Figures **xii**

List of Tables **xiv**

1 Introduction **1**
1.1 Motivation 1
1.2 Context 3
1.3 Contribution 7
1.4 DPhil Structure 10

2 Literature Review **11**
2.1 Introduction 11
2.2 Environmental Risk 12
2.2.1 Review of Environmental Risk 12
2.2.2 Corporate Social Responsibility 13
2.2.3 Climate Risk and New Data Needs 16
2.3 Information and Corporate Governance 18
2.3.1 Information Asymmetry between Companies and Shareholders 19
2.3.2 Information Asymmetry between Companies and Stakeholders 23
2.3.3 Disclosure of Environmental Information 25
2.4 The Production of Environmental Information 28
2.4.1 Information as a Public Good 28
2.4.2 Information Production in Markets 32
2.5 Technology 34
2.5.1 Introduction to Machine Learning 35
2.5.2 Earth Observation and Computer Vision 38
2.5.3 Web Scraping and Natural Language Processing 39
2.5.4 Crowd Sourcing Data and Computing 40

3 Paper Synthesis **42**
3.1 DPhil Contribution and Paper Synthesis 42

4 Technology, Information, and the Governance of Environmental Risk **49**
4.1 The Limits of Technology in Unravelling Company Information . . 50
4.1.1 Limits and Biases of Earth Observation 51
4.1.2 Limits and Biases of Web Scraping 54
4.1.3 Limits and Biases of Crowd-Sourcing Methods 55
4.2 Disclosure for Shareholders vs Stakeholders 56
4.3 Fostering Innovation in Markets for Environmental Information . . 59
4.4 Policy Recommendations 61
4.5 Opportunities for Further Research 63

5 Conclusion **65**

Appendices

A A Global Solar Photovoltaic Inventory. **68**
A.1 Motivation 70
A.2 Machine Learning Pipeline 72
A.3 A Global Census 75
A.4 Land Cover Analysis 79
A.5 Conclusion 83
A.6 Supplemental Methods 84
A.6.1 Machine Learning Pipeline Description 84
A.6.2 Dataset Completion 93
A.6.3 Machine Learning Pipeline Performance 94
A.6.4 Machine Learning Model Interrogation 96
A.6.5 Feature Enrichment and Analysis 98

B Asset-Level Transition Risk in the Global Fossil-Fuel Supply Chain. 111
B.1 Motivation 113
B.2 Related Work 115
B.3 Data Sources and Preparation 116
B.4 Network Topography 121
B.5 Simulated Energy Flow 123
B.6 Transition Risk Analysis 125
B.7 Future Research Directions 137
B.8 Supplemental Methods 140
B.8.1 Network Assembly 140
B.8.2 Flow Calculation 143

C Towards *DeepSentinel* **145**
C.1 Earth observation for climate change mitigation 147
C.2 Sensor fusion 148
C.3 Data preparation 152
C.4 Experimentation 153
C.5 Results 158
C.6 Application 161
C.7 Discussion 161
C.8 Conclusions 163

Bibliography **164**

List of Figures

A.1 Out-of-sample PV solar energy detections in remote sensing imagery 76
A.2 Aggregated arrangement of our PV solar energy dataset 80
A.3 Pre-existing land cover for new solar PV installations 81
A.4 Machine learning pipeline diagram. 86
A.5 Distribution of polygon areas in training, validation, and inferred corpuses. 88
A.6 PV solar energy deployment area 89
A.7 RNN-1 schematic diagram 92
A.8 RNN-2 schematic diagram 92
A.9 Overall machine learning pipeline validation performance 95
A.10 Geographic covariate shift in pipeline deployment 97
A.11 Remote sensing band oblation studies 99
A.12 Inferred installation date distribution 100
A.13 Pre-existing land cover for PV solar energy installations by country 106
A.14 Pre-exising land cover for PV solar enegy installations by size . . . 107
A.15 Aggregate dataset comparison 108

B.1 The global fossil fuel supply chain. 117
B.2 Network arrangements for coal, oil, and gas supply chains. 123
B.3 Global Coal Flow. 126
B.4 Global Gas Flow. 127
B.5 Global Oil Flow. 128
B.6 Coal flow cross-validation with production and trade data. 129
B.7 Gas flow cross-validation with production and trade data. 130
B.8 Oil flow cross-validation with production and trade data. 131
B.9 Coal energy flow under a demand shock to the IEA Sustainable Development Scenario. 133
B.10 Gas energy flow under a demand shock to the IEA Sustainable Development Scenario. 134
B.11 Oil energy flow under a demand shock to the IEA Sustainable Development Scenario. 135
B.12 Directed acyclic computational graph for the network assembly. . . 142

C.1 *DeepSentinel* summary 149
C.2 Samples from *DeepSentinel* training corpus 151
C.3 Out-of-sample land cover classification inference samples 159
C.4 Out-of-sample land cover classification Jaccard index scores 160

List of Tables

2.1 Typologies of environmental risk 13
2.2 Public Goods 28
2.3 Example Information Public Goods 29
2.4 Design Principles for Effective Institutions 29
2.5 Forms of the Efficient Market Hypothesis 34
2.6 Machine Learning Typology 36

4.1 Shareholder and Stakeholder Risks 57

A.1 PV solar energy dataset comparison 79
A.2 RNN-1 Architecture 93
A.3 RNN-2 Architecture 93
A.4 Pipeline Validation Performance 95
A.5 Detection Confidence Levels 99
A.6 PV solar energy installation capacity uncertainty 102
A.7 Dataset Metadata 105
A.8 Landcover Class Mappings 109
A.9 Aggregate Statistics by Country 110

B.1 Data Description 118
B.2 Freight and Modal Transfer Costs 122
B.3 Energy Technology Properties 122
B.4 Network Summaries 122
B.5 Demand Shock Resulting from a Transition to a Sustainable Development Scenario 132
B.6 Supply Shocks resulting from Interdictions due to Local Policies . . 138

C.1 Sensor Fusion Datasets 152
C.2 Convolutional Block Architecture (DownBlock) 156
C.3 Convolutional Block Architecture (UpBlock) 156
C.4 Attention Block Architecture 156
C.5 ResNet18 157
C.6 ResNet34 157

C.7 ResNet18 with Attention 157
C.8 Deconvolutional Header 158
C.9 Land Cover Classification weighted mean Jaccard Index 160
C.10 Copernicus CORINE Land Cover Classes 161
C.11 Out-of-sample AUC-ROC for Coal Mine Classifier 161

Publicity is justly commended as a remedy for social and industrial diseases. Sunlight is said to be the best of disinfectants

— Louis Brandeis' *Other People's Money and How Bankers Use It, 1914*

1 Introduction

1.1 Motivation

This DPhil thesis is motivated by the impact that data can have when it is held in public trust for the benefit of all. Capital markets are increasingly called upon[1] to provide the disciplining constraint on company behaviour that has averse social or environmental impacts. The question of what and how environmental risks should be disclosed by companies has been thrust into the foreground of urgent climate change mitigation.

Disclosure as a regulatory instrument has been used to mitigate both systemic financially-material risks for company shareholders and non-financially-material risks for company stakeholders. Environmental advocacy groups have played a major role in establishing the urgent treatment that climate change mitigation is now receiving by both policy makers and institutional investors. For whom, then, will emerging disclosure regimes be designed for — a company's stakeholders sensitive to the large diversity of environmental impacts that a company might have? Or company shareholders who have a narrower interest in financially-material environmental risks? And what of company environmental and social impacts which might not be material risks to the company?

Such a drama might play out along expected lines: advocacy groups, marshalling the political will of their constituents, might appeal to policy makers for more penetrating disclosures while company managers, reluctant to disclose proprietary information, resist such measures through political channels of their own. Emerging technologies, however, are poised to disrupt this political stalemate.

Data-driven technologies are empowering diverse stakeholders to produce and broadly distribute information pertaining to social and environmental impacts of companies. Examples of such technologies include earth observation combined with computer vision for localising and monitoring physical assets; web crawling with natural language processing to automate the compilation and analysis of primary and secondary reporting; and digital connectivity platforms empowering distributed stakeholders and citizen scientists to make remote environmental and social impacts salient in boardroom and social media discussions. The scalability and open-source nature of these technologies makes these interventions accessible to small teams and even individuals with only modest philanthropic or policy support, or reputational or altruistic incentives. Meanwhile, market intelligence companies, who intermediate the information asymmetry between companies and their investors, are being called upon to offer more sophisticated insight into environment and social risks. They find themselves in competition (for talent and customers) with NGOs partnered with major technology companies, individual entrepreneurs, and emerging specialist technology companies. The environmental scientists and technologists they must now employ come with different expectations of cultures and norms of their employers.

In this DPhil, I study and participate in the unravelling of company information relating to environmental risk using data-driven technology. I prepare three data science papers which make fundamental data and methods for assessing company environmental risk more accessible. I articulate the role that technology plays in moderating information asymmetry, between both companies and investors, and companies and society. This helps close the ontological gap between innovation as a paradigmatic socially-embedded process; public, private, and social governance of companies; the role of information intermediaries in markets; and disclosure

as a policy mechanism. I conclude with recommendations targeted at emerging disclosure regimes for climate change risk, to help them be suitable for the needs of all of society's diverse stakeholders.

1.2 Context

It is only in very recent history that climate change risk has been able to capture the attention of the finance industry (UNEP Financial Inquiry 2015). Following the Fourth Assessment Report of the IPCC (Pachauri & Reisinger 2007), the Copenhagen Accord of the UNFCCC confirmed a broad coalition of support for a 2°C limit to anthropogenic climate change (UNFCCC 2010, Gao et al. 2017). With the baseline science and a policy target established, scientists and advocacy groups could begin interrogating what particular warming constraints might mean for different industrial sectors. A notional 'carbon budget' was developed based on the more-or-less linear response of warming to atmospheric greenhouse gas concentrations (Meinshausen et al. 2009).[2] Of particular interest was the exposure of fossil fuel companies and their investors to the risk of 'unburnable carbon': fossil fuel reserves which were already accounted for on their balance sheets and valued in their share prices but which in aggregate exceeded carbon budgets consistent with 2°C of warming (CTI 2011, 2013, McKibben 2012, McGlade & Ekins 2015). These reserves were and are at risk of becoming 'stranded assets': assets which "suffer from unanticipated or premature write-offs, downward revaluations or are converted to liabilities"(Caldecott et al. 2013).

The role of advocacy groups and NGOs in popularlising the notion 'unburnable carbon' should not be understated. Environmental justice groups organised around a divestment campaign who advocated for divestment from fossil fuel companies on financial and moral grounds (McKibben 2012, Rusbridger 2015). The campaign sought to stigmatise fossil fuel companies and weaken their social license to operate, analogous to how a popular divestment campaign was able to stigmatise the apartheid regime in South Africa (Ansar 2013, Hunt et al. 2017). It found leverage

by appealing to the beneficiaries of pension funds, university endowments, religious groups, NGOs, and governments (gofossilfree.org 2020).

Skeptics and detractors were forced to respond. University governors (e.g. Faust 2013), fossil fuel lobbyists (e.g. Davidson 2014), government advisors (e.g. Skancke et al. 2014*a*), and companies (e.g. Cohen 2014*a*) took issue with various aspects of the moral and financial case for fossil fuel divestment. This discourse stimulated the need for a more sophisticated typology of environmental risk as fund managers were forced to explain how they managed these risks in the absence of divestment and screening practises. Where engagement with companies was the preferred strategy, investors formed coalitions to socialise among themselves the task of engaging with companies, particularly fossil fuel companies (Wildsmith 2012). These efforts came to fruition in 2015 when the management teams of prominent European oil and gas companies endorsed shareholder proposals brought by these coalitions to better disclose and manage climate risks (Clark 2015).

Policy makers also sought to interrogate systemic exposure to climate change risks. At the request of the UK's Department for Rural Affairs,(Carney 2014, Adams 2014) the UK's Bank of England Prudential Regulatory Authority published a landmark typology for climate change risk in the UK insurance sector (The Bank of England 2015, Carney 2015). The typology established the framework of *physical risk*, a company's direct exposure to the changing environment, and *transition risk*, a company's exposure to society's response to a changing environment. Companies, investors, intermediaries, and other stakeholders could use this framework to bring rigour to the measurement and mitigation of diverse spectra of environmental risks.

In 2015, agreed at the 21st Conference of Parties of the UNFCCC (COP21), the Paris Agreement increased international ambitions, commiting parties to hold "the increase in global temperatures to well below 2°C and pursuing efforts to limit the temperature increate to 1.5°C above pre-industrial levels." (UNFCCC 2015). The SDGs (2015) were ratified at the UN general assembly less than two months earlier. This combination of multilateral agreements, despite being multi-objective in nature, brought coherence to climate change and development priorities,

the dual challenges of the early 21st-century (Stern 2015). COP21 was widely attended by private industry representatives which represented a step-change in the acceptance of climate change as an urgent priority and the role of private industry in its mitigation (Benabou et al. 2017).

As the management of climate risk was beginning to be operationalised, it became clear that the data required to effectively measure and mitigate climate change risk was not readily available. Disclosure initiatives had previously been established to help stakeholders assess company exposure to carbon pricing and were not suitable for a broader typology of climate and other environmental risks (e.g. Hummel et al. 2016, Woods 2003). In 2015, the Chair of the Financial Stability Board, an institution comprised of the Finance Ministers of the G20, commissioned the creation of the Task Force on Climate-Related Financial Disclosures (TCFD) (G20 Finance Ministers and Central Bank Governors 2015). The TCFD was formed comprising of major private-industry stakeholders including investment and retail banks, pension fund trustees, insurers and reinsurers, corporates, financial consultants, and credit ratings agencies. It issued its final recommendations for the disclosure of climate-related risks in 2016 (TCFD 2016). These recommendations now play a central role in the design and implementation of climate disclosure programmes worldwide, including, e.g., mandatory adoption by UK firms (HM Treasury 2020), and mandatory disclosure aligned with Art. 173 for French asset managers and institutional investors (Forum Pour L'Investissement Responable 2016). The TCFD continues it's convening and norm-setting work on climate-related risk disclosure, issuing periodic updates on adoption and assorted detailed guidance, including on scenario analysis, forward-looking climate metrics, and risk management integration.(Financial Stability Board 2020).

Data played a key role in the development and popularisation of 'unburnable carbon' as a financial risk. While, e.g., Meinshausen et al. (2009) and McGlade & Ekins (2015) were able to make characterisations of the extent of the 'carbon bubble' using country-level fossil fuel reserve data, the Carbon Tracker Initiative (CTI) used firm- and asset-level data to ascribe this risk first to markets (CTI 2011) and

then to companies (CTI 2013, 2014, 2015). Data of global coal-fired power stations maintained by the NGO CoalSwarm (e.g. Shearer et al. 2017*a*) was used to study climate risk in the global power sector, for example in Pfeiffer et al. (2016) and in Caldecott et al. (2016). These analyses were able to be made without the cooperation of the management of the companies affected. The data which enabled them share a specific characteristic: the unit of reporting is the individual physical asset.

Asset-level data describes the underlying physical assets of a company (Caldecott & Kruitwagen 2016, Caldecott 2016, Weber et al. 2017, Lord et al. 2019). Definitions vary, but in general asset-level data are geospatially localised and tie an underlying asset to an owner company. They provide the crucial link between the real and financial economies. Asset-level data have been used to produce multi-sector assessments of climate change risk and cumulative committed greenhouse gas emissions (e.g. Thomä et al. 2015, Caldecott, Dericks, Bouveret, Schumacher, Pfeiffer, Tulloch, Kruitwagen & McCarten 2018*b*, Tong et al. 2019).

In their survey of asset-level data providers, Weber et al. (2017) find that most asset-level data is obtained via government data, press reports, and surveys and interviews directly with companies, operators, suppliers, and equipment manufacturers. Traditional asset-level data they describe is resource-intensive to collect and collate; is produced by a social network of surveryors/reporters and respondees, introducing substantial bias; and is necessarily collected with the consent of industry.

One of the most promising technologies for obtaining better asset-level data is the combination of remote sensing imagery with machine-learning based computer vision (e.g. Caldecott, Kruitwagen, McCarten, Zhou, Lunsford, Marchand, Hadjikyriakou, Bickel, Sachs & Bohn 2018). This technology-driven method is agnostic to company- and country-reporting, exchanging bias in surveyor/surveyee social network topology for bias in the availability of training data. Significantly, such methods do not rely on the cooperation of company management, alluded to by CTI in the title of their report 'nowhere to hide' on measuring greenhouse gas emissions from space (CTI 2018). These methods coincide with the ongoing emergence of the 'new space' paradigm, a shift in both the commercialisation and unbundling of the space value

chain, and a technological transition to substantially smaller spacecraft ("cubesats") (Liddle et al. 2020). Under this new paradigm, the number of spacecraft in orbit is doubling approximately every 3 years, driven by launches of new earth observation satellites (Union of Concerned Scientists 2020).

Of course, not all features of interest can be detected in satellite imagery, but nor is satellite imagery the only technology that is disrupting the traditional provision of asset-level data. Web-scraping combined with machine-learning enhanced natural langauge processing can automate the collection and organisation of unstructured web data, such as that employed by the DOE in their preparation of a global asset-level oil and gas database (Rose, Bauer, Baker, Barkhurst, Bean, DiGiulio, Jones, Jones, Justman, Miller et al. 2018*a*). The NGO CoalSwarm has now turned its focus to building a Global Energy Monitor (e.g. Nace et al. 2019), digital connectivity allowing their devolved network of NGO collaborators to collectively produce a global database of existing, planned, and under construction energy infrastructure.

These efforts highlight the roles that NGOs and academics are playing in the provision of public goods for the purposes of sustainable finance. It is unclear that data of this type will be otherwise forthcoming from company disclosures, even with (mandatory) adoption of TCFD recommendations. A gap, then, exists between the data that being disclosed by companies and the data that is required to manage climate change risks for the benefit of all company stakeholders. This DPhil will examine the existence of this gap, and propose and develop the methods for closing it.

1.3 Contribution

This DPhil provides four major contributions to environmental risk and corporate governance scholarship. The first three contributions are technical in nature and correspond to the three DPhil papers. These novel and rigorous data science papers make substantial contributions to the data and methods available for environmental risk analysis of companies. The fourth contribution is in the synthesis and discussion that follows, drawing on the experiential process of developing the prior three papers. I discuss the role of technology in moderating markets for information, and the

implications thereof on corporate governance. I develop recommendations for disclosure policy based on these observations.

The first DPhil paper, *A Global Inventory of Commercial-, Industrial-, and Utility-scale PV Solar Energy Generating Units*, provides a new globally-exhaustive inventory of solar photovoltaic (PV) generation stations in excess of 10kW. This inventory is developed using a longitudinal corpus of remote sensing imagery, machine learning, and a very large cloud computation infrastructure. This work represents the first time a computer vision model has been used to search the entire planet for a specific type of infrastructure. Data on the diffusion of solar PV generating installations is extremely relevant to the ongoing energy transition, and thus providing a globally-complete asset-level inventory is a most urgent project with substantial social benefit. My coauthors and I develop policy-relevant analysis of this new dataset, interrogating the land cover impacts of PV solar energy deployment. We find that PV solar energy facilities above 10kW in generating capacity are most commonly sited on croplands and deserts. The former has implications for the stability and resilience of food supply, while the latter are uniquely fragile ecosystems.

The second DPhil paper, *Asset-level transition risk in the global coal, oil, and gas supply chain*, implements a geospatial network arrangement of the global fossil fuel supply chain using only publicly available data. Fossil fuel production, demand, and transmission are estimated for over five hundred thousand real-economy infrastructure assets in the complex supply chain. This minimum cost flow is estimated using a non-linear cost-scaling algorithm and validated using international trade data. My coauthors and I prepare a counterfactual demand-shock scenario, wherein global energy demand becomes aligned with the Paris Agreement-compliant IEA Sustainable Development Scenario, and we assess the impact this transition has on asset-level energy flows and costs. We also prepare supply-shock scenarios, where regional energy supplies are made more costly, for example by a local biodiversity conservation or conventional pollution policy. With these interdictions we identify geospatial sensitive intervention points where a relatively small intervention might have an outsized impact on the transition to a post-carbon economy.

The third DPhil paper, *Towards DeepSentinel*, develops a novel computer vision architecture to unlock substantially more use-cases from medium resolution remote sensing imagery. DeepSentinel combines Sentinel-1 synthetic aperture radar (SAR) imagery with Sentinel-2 multispectral imagery to capture the broadest range of input signals for general-purpose semantic embedding. The training curriculum combines large-scale self-supervised pretraining followed by supervised training on generic land-cover data. I assess the generaliseability of this training curriculum by fine-tuning on a very small corpus of surface mine commodity data. The result is a pretrained computer vision model suitable for a broad range of earth observation tasks.

In the interest of reproduceability and to maximise research impact, the code repository for each paper is being publicly published. The repositories for the second and third papers are already publicly available, the repository for the first paper is embargoed until paper publication, at which point it will be made publicly available. Likewise, all supporting and derived datasets are being publicly published. For all papers, input and supporting datasets are available on cloud storage. Results datasets are also made publicly available on cloud storage, which include, respectively for the three papers, the novel global dataset of PV solar energy facilities, the resulting interdicted flow results datasets, and the sensor fusion computer vision models. The PV solar energy dataset is also being publicly rehosted by the World Resources Institute, Descartes Labs Inc., and is also available via an Earth Engine App.

The work contained in this DPhil has been achieved with extensive support from coauthors, collaborators, and institutional sponsors. While full acknowledgements are made in the DPhil pretext, it is worth again acknowledging the extensive support this DPhil has received from these institutional partners. The first paper was supported by in-kind donation of computation time from Descartes Labs Inc as well as data from the World Resources Institute. Additional computation resource for model training was provided by an Amazon Web services computing grant. The second paper was supported by a pro-bono engagement with QuantumBlack, a McKinsey & Company data science subsidiary, and by an intern sponsored by

the Ogden Trust. Computation resource was provided by Microsoft Corporation under their AI for Earth program. The third paper is also supported by Descartes Labs Inc as part of their Impact Science Program. Additional training data were provided by the Natural Resources Governance Institute. Computational support for the third paper is provided by both Microsoft's AI for Earth Program and a Google Research Credits grant.

1.4 DPhil Structure

This DPhil proceeds with the following structure. Following this introduction is a detailed literature review. The literature review establishes key concepts for understanding environmental risk, used throughout, and provides an overview of information economics and corporate governance scholarship. A synthesis of the three DPhil papers is included in the main text, that describes the unified programme of inquiry that is articulated by the three works. A discussion chapter is provided wherein the findings of the DPhil are contextualised in ongoing debates on the form of environmental disclosure policy. A brief conclusion is provided. The three papers which form the bulk of the contributions of this DPhil are enclosed in appendices.

Notes

[1]To suggest something is 'called upon' naturally raises the question of 'by whom?'. Whether the increasing role capital markets play in disciplining company behaviour represents an innovation of or capitulation by policy makers is discussed in the literature review.

[2]Earlier estimates of a carbon budget are available, including Krause et al. (1989) and even Swedish Nobel Laureate Svante Arrhenius (1896)

2
Literature Review

Contents

2.1 Introduction 11
2.2 Environmental Risk 12
2.2.1 Review of Environmental Risk 12
2.2.2 Corporate Social Responsibility 13
2.2.3 Climate Risk and New Data Needs 16
2.3 Information and Corporate Governance 18
2.3.1 Information Asymmetry between Companies and Shareholders 19
2.3.2 Information Asymmetry between Companies and Stakeholders 23
2.3.3 Disclosure of Environmental Information 25
2.4 The Production of Environmental Information 28
2.4.1 Information as a Public Good 28
2.4.2 Information Production in Markets 32
2.5 Technology 34
2.5.1 Introduction to Machine Learning 35
2.5.2 Earth Observation and Computer Vision 38
2.5.3 Web Scraping and Natural Language Processing 39
2.5.4 Crowd Sourcing Data and Computing 40

2.1 Introduction

The purpose of this literature review is to give the reader some familiarity with the scholarship within which this thesis situates itself. It begins with a review of

the environmental risk literature and the historical scholarship on the relationship between companies and their investors, and nature and society. It proceeds with a review of how information features in corporate governance: asymmetries thereof between investors and company management; information asymmetries between companies and society; the role of disclosure in addressing these asymmetries; and the role of other market intermediaries. In order to provide a theoretical grounding for discussion to follow, the literature review then explores the production of information: it's nature as a public good, and it's production within markets. The literature review concludes with a section on technology, general introductions to the three technologies identified in the DPhil.

2.2 Environmental Risk

2.2.1 Review of Environmental Risk

Environmental risk is a category of risk related to the changing natural world (usually with anthropogenic causes) and society's responses to those changes. 'Risk' is used here and throughout in the broadest sense: as a shorthand for Knightian risk and uncertainty, including stochastic phenomena with known probability distributions (risks) and unknown phenomena or those with unknown probability distributions (uncertainties) (Knight 1921, Keynes 1921); as an imprecise shorthand capturing both aleatoric and epistemic uncertainty (e.g. Der Kiureghian & Ditlevsen 2009); and also contemporary understandings of risk as the product of hazard magnitude, exposure level, and vulnerability to impact (e.g. Koks et al. 2015). Following a 'traditional' style credit analysis (e.g. Standard & Poor's Ratings Services 2014), environmental risk poses threats (and opportunities) to the income of a company (business risk, economic risk) and its financial position (financial risk, balance sheet risk). Income risks include changes in revenue and expenditures (particularly energy, water, and resources costs), reputational risks, loss of licence to operate, regulatory compliance costs, and loss of business risks (TCFD 2016, Buhr 2017). Financial risks include risks to operating costs and margins, stranded asset risks, liability provisions, short-term revenue risks, and capital risks (TCFD 2016, Buhr 2017).

Table 2.1: Typologies of environmental risk

Category	Description
Physical Risks	First-order risks arising from weather events (e.g. floods and storms), damages resulting directly from weather events and indirect damages such as the disruption of supply chains
Transition Risks	Financial risks arising from a transition to a lower-carbon economy, in particular the rapid re-pricing of carbon intensive financial assets.
Liability Risks	Risks arising from the recovery of losses and damages due to climate change from the parties who may have been responsible, e.g. via third-party, officers' liability, and professional indemnity insurance.

(a) The Bank of England (2015)

Category	Set	Subset
Operational		Regulatory Compliance
		Other Regulatory
		Litigation
Climate	Mitigation	Carbon Pricing
		Other Carbon
		Other Mitigation
	Adaptation	Adaptation Cost
		Event Risks
		Societal Risks
Natural Capital	Depletion	Water
		Fossil Fuels
		Renewable Resources
		Non-Renewables Resources
	Geopolitical Event	Resource-Based
		Stability-Based
		Global Warming Impacts
		Subsidy Risks

(b) Buhr (2017)

A wide range of environmental factors impose these business and financial risks, see Table 2.1 (a)-(d) for example typologies.

2.2.2 Corporate Social Responsibility

The modern formulations of environmental risk shown in Table 2.1 have been developed with the benefit of a history of research into the relationship of the corporation with the natural environment. The role of companies in society has

Category	Set
Transition	Policy & Legal Technology Market Reputation
Physical	Acute Chronic

(c) TCFD (2016)

Category	Set
Environmental Change	Climate change; natural capital depletion; biodiversity loss; air, land, and water contamination; water and heat stress.
Resource Landscapes	Price and availability of resources such as oil, gas, coal and minerals (e.g. shale gas revolution, phosphate availability).
Government Regulations	Carbon pricing; subsidy regimes; pollution regulation; disclosure requirements; licensing conditions.
Technology Change	Falling clean technology costs (e.g. solar PV, onshore wind, electric vehicles)
Social Norms and Consumer Behaviour	Fossil fuel divestment campaign; product labelling and certification schemes; changing consumer preferences
Litigation and Statutory Interpretations	Carbon liability and litigation; changes in statutory interpretation

(d) Caldecott et al. (2016)

been the topic of interest and debate for decades, born out of the social sensibility of business people and need for optimality in the management of the resources of their companies (for an early history see Carroll 1979).

A theory of Corporate Social Responsibility (CSR) developed as a response to the contentious extreme of Chicago School of Economics thinkers, exemplified by Milton Friedman who wrote in 1970 that "the social responsibility of business is to increase its profits". CSR by contrast describes the 'enlightened self interest' (e.g. Baumol 1970, Keim 1978) that a company may take in advancing social interests which align with the interests of the company. Pollution abatement and environmental conservation were included even in early formulations of CSR (e.g. Committee for Economic Development 1971).

In response to the shareholder primacy model of Friedman (1962, 1970) and others, Stakeholder Theory (Freeman & Reed 1983, Freeman 1984, Mitroff 1983) argues that the welfare of many of society's stakeholders is material to the performance of

a company (Branco & Rodrigues 2007, Brown & Foster 2012). Companies acting in their own interest should also ensure the welfare of, as examples, local communities on whose labour they depend, and the natural environment from which they extract primary resources. Writing in 2003, Smith says the two theories are at a 'standoff'. But in 2020, the private sector seems to be reaching a consensus on stakeholder forms of company management, with endorsement from the likes of management consultants McKinsey & Company (2020), US$7tn asset manager BlackRock (2020), the World Economic Forum (2019), and the Business Roundtable of CEOs (2019).

CSR and Stakeholder Theory researchers have spent decades examining the materiality of stakeholder welfare on company performance. Various indices and indicators of corporate social performance (CSP) were developed, examples including the extent of CSR disclosure (e.g. Ingram 1978, Cowen et al. 1987); the extent of disclosed pollution control expenditures (e.g. Freedman & Jaggi 1982, Belkaoui 1976); and reputational scores (e.g. Conine & Madden 1987, Simerly 1994). These indicators were compared with the more conventional indicators of corporate financial performance (CFP) (e.g. net income, return on assets, return on equity, price-to-equity ratio, market-to-book value). Orlitzky et al. (2003) developed a metastudy of 52 studies relating CSP to CFP and empirically discredit arguments about any efficiency trade-offs between CSP and CFP. Controlling for measurement and sampling errors, CSP (and to a lesser extent corporate environmental responsibility) is universally correlated to CFP. Eccles et al. (2014) delineate between CFP and corporate operational performance (COP) and reinforce that formal governance of sustainability issues leads to outperformance in both. Clark et al. (2015) provide an updated metastudy of over 200 studies and find CSP does not impair COP, CFP, or the firm's cost of capital.

Among investors, environmental risk is studied as a dimension of environmental, social, and governance (ESG) risk. ESG studies have examined the performance of portfolios which are constructed with consideration of ESG criteria relative to benchmark portfolios. The synthesis of evidence shows, as a minimum, no penalty for the integration of sustainability criteria into investment practises and

in many cases outperformance relative to a relevant benchmark (Khan et al. 2015, Kotsantonis et al. 2016). Ethical investment strategies also rely on ESG data to inform security selection. Strategies like Socially Responsible Investment (SRI) and Impact Investing describe a continuum of fund management techniques ranging from passively screening the universe of investment options to meet ethical criteria (e.g. excluding 'sin stocks' like equities of tobacco and weapons companies) through to the explicit targeting of investment opportunities to maximise social or environmental outcomes (Kruitwagen et al. 2016). Generally, ethical investment strategies demonstrate no significant difference in risk-adjust returns relative to comparable portfolios (Bauer et al. 2005, Revelli & Viviani 2015). The popularity of the fossil fuel divestment campaign drove commensurate increases in the the studies of ethical portfolios, particularly ones which exclude all or some fossil-fuel extractive or consumptive universe of companies (e.g. MSCI ESG Research 2013, IMPAX Asset Management 2013, Balkissoon & Heaps 2014, Geddes et al. 2016).

These mixed findings show that not all stakeholders are material to all companies (Khan et al. 2016, SASB 2018, Cheng et al. 2014). Non-material stakeholders may still, however, be impacted by company decisions. Edmans (2020) argues that company managers should naturally favour stakeholders that they have a comparative advantage in serving, and the ones who are most material to the company, because the benefits will be most shared with the company. The role the disclosure should play for *all* company stakeholders, both material and non-material, is developed in the discussion.

2.2.3 Climate Risk and New Data Needs

Global climate change policy coalescence occured first in 2008-2010 around a 2°C warming limit with successive G8 summits (Graham 2009), the Copenhagen Accord (2009), and the Cancun Agreements (2010), and then continued with commitments to "pursue efforts" towards a 1.5°C warming limit with the Paris Agreement (2015). With a policy benchmark, scientists were able to articulate a range of warming and policy scenarios and what these might mean for the future of energy systems and the

companies operating within them (e.g. Meinshausen et al. 2009, IPCC 2014, McGlade & Ekins 2015). Global fossil fuel reserves were found to far exceed a notional 'carbon budget' that would constrain warming to less than 2°C. The market valuation of fossil fuel companies is largely based on the assumption that their reserves base will be extracted and sold (Howard & Harp 2009, Bryan 2012, Kaiser & Yu 2012). These valuations then, are in conflict with a future constrained to a 2°C or 1.5°C warming limit. This mis-valuation of extractive companies was called the 'carbon bubble'.

A coalition of civil society groups, led notably by 350.org, called for divestment from fossil fuel companies on ethical and financial grounds (McKibben 2012, Rusbridger 2015, gofossilfree.org 2020). Civil society and private sector organisations published extensive analyses of stranded fossil fuel asset risk exposure (e.g. CTI 2011, 2013, 2014, Spedding et al. 2013, Caldecott & Elders 2013, Lewis & Voisin 2014). Companies, investors, and other fiduciaries were forced to articulate their response to the divestment campaign and stranded asset risk (Faust 2013, Skancke et al. 2014*b*, Cohen 2014*b*, ExxonMobil Inc 2014). The impetus of the popular divestment campaign demanded a new sophistication in the management of climate change risks.

Climate change risk was substantially legitimised by the Bank of England and Financial Stability Board of the G20 (FSB) under the leadership of Bank of England Governor Mark Carney. In April 2014 the Prudential Regulatory Authority (PRA) of the Bank of England accepted an invitation from the Department of Environment, Food & Rural Affairs (DEFRA) to report on climate change adaptation with particular focus on the UK insurance sector (Adams 2014). The PRA released its report in September 2015 (The Bank of England 2015) coincidentally with a speech by Governor Carney at Lloyd's of London entitled 'Breaking the tragedy of the horizon - climate change and financial stability'(Carney 2015). Shortly thereafter the FSB announced the creation of the Task Force on Climate-Related Financial Disclosures (FSB 2015). In September 2016, Governor Carney spoke again on the threat of climate change to financial stability and the role climate change mitigation in resolving those risks and driving prosperity (Carney 2016). The TCFD delivered its final recommendations to the FSB in December 2016 (TCFD 2016).

The findings of the TCFD call for a new sophistication in response to climate change risk. They reinforce a growing momentum of interest in ESG risk more broadly (Kotsantonis et al. 2016), brought on by the urgency of environmental change and the data requirements of society's response thereto. New metrics are in development which will allow the closer monitoring and management of environmental risk (e.g WFE Sustainability Working Group 2015, 2 Degrees Investing Initiative 2015, Menou & Nishikawa 2015). These new metrics join an array of new high-precision analyses of environmental risk including of portfolio alignment with decarbonisation scenarios (Mercer LLC 2015, Thomä et al. 2015, Faria et al. 2015), stranded assets (Caldecott et al. 2016), and committed emissions (Pfeiffer et al. 2016).

Many of the historic and emerging metrics and analyses of environmental risk rely on voluntary or self-reported data (as in the CSR literature examined above). CDP, the leading voluntary environmental information disclosure programme (formerly the Carbon Disclosure Project), indicates that it receives a response rate of only about 60% to its information solicitations, with the worst response rates (<40%) in the energy and utility sectors (Hummel et al. 2016). The ability to study environmental risk, therefore, is substantially limited by the information that is disclosed by companies and cross-sectional comprehensive studies are all but impossible. New technology, however, shows promise in reducing the cost of information to the extent that it may be accessible to investors, policy makers, and other public and private-sector stakeholders without the cooperation of intransigent company managers.

2.3 Information and Corporate Governance

The balance of power between shareholders and company managers is a key determinant of company performance (Gompers et al. 2003). Corporate Governance describes how private sector organisations and individuals transfer risk and decision making authority among each other (see, e.g., Shleifer & Vishny 1997), creating direct outcomes on how environmental risk is managed. Private sector organisations in turn have risk and transactional relationships with the government, public

sector organisations, and the public (see, e.g., Polanyi 1944, Ramanathan 1976). Accountability for the management of environmental risk relies substantially on the flow of information between these different stakeholders. This section will focus on the information asymmetries in corporate governance relationships, treating private-sector separately to public sector, and disclosure as a policy tool for reducing information asymmetry.

2.3.1 Information Asymmetry between Companies and Shareholders

The theory of change of 'regulation by disclosure' is that reduced information asymmetry between companies and their shareholders will lead companies to better manage environment risks on behalf of their investors, with beneficial outcomes for company stakeholders (Cohen & Santhakumar 2007, Gupta & Mason 2014). The incentives that govern the revelation of and demand for information between private parties are described in this section.

Agency Theory

Agency Theory describes the misalignment of incentives and risk bearing when one party (the 'principal') delegates work to another (the 'agent') (see Ross 1973, Eisenhardt 1989). While the theory is ubiquitous, the most common application examines the separation of ownership and decision making between shareholders and company managers (e.g. Berle & Means 1932, Jensen & Meckling 1976, Fama 1980). The theory formalises the 'adverse selection' and 'moral hazard' problems of asymmetric information which had long been a topic of research in the insurance industry (for a history see Laffont & Martimort 2001).

'Moral Hazard' describes action by an agent which is misaligned with the interests of the principal. A canonical example, the purchaser of an auto insurance policy may adopt more risky driving behaviour (e.g Cummins & Tennyson 1996). The principal is only able to contract for ideal behaviour to the extent that the principal can observe the action of the agent. In capital markets, company managers may adopt more risky behaviour than their creditors had priced into financing terms

(Park 2000). In corporate governance, shareholders have imperfect information concerning the effort with which managers execute their duties - this information is privately held by the managers (e.g. Stoughton 1993).

'Adverse Selection' describes the selection of an agent which is misaligned with the interest of the principal. A canonical example, a person with a high risk of medical-related mortality is more likely to buy life insurance (e.g. Subramanian et al. 1999). In capital markets, adverse selection refers to the incomplete information of investors relative to that of company managers in the affairs of the company. Company managers may use this information favourably when issuing debt and equity (Meyers & Majluf 1984). In corporate governance, adverse selection is present in the labour market for company executives (Hart 1983). The shareholders have incomplete information regarding the skill of their prospective managers - information which is privately held by the managers.

Agency Theory provides foundations which are resolved by Signalling and Contract Theories. Contract Theory explicitly delineates adverse selection and moral hazard based on whether the information asymmetry occurs prior to or following the establishment of a contract (Laffont & Martimort 2001).

Signalling Theory

Signalling theory develops from the seminal works of Ackerlof (1970) and Spence (1973). Ackerlof (1970) considers a market in which the sellers have private information about the quality of their goods. The buyers, being unable to distinguish the quality of goods ex-ante, set a price reflective of the uncertainty in the quality of the goods. Sellers of low-quality goods will accept this price, while sellers of high-quality goods will leave the market. Spence (1973) considers how sellers in a market may signal to the market their 'type' (i.e. either high-quality or low-quality) so the market may price them accordingly, even if the sellers must suffer a cost to do so. Connelly et al. (2011) suggest that "the profundity of the theory ... lies in *ascribing costs* to information aquisition processes".

Ross (1977) considers the signal embedded in the capital structure of a firm. Firm dividends convey signals to markets (Watts 1973), albeit through a lens of corporate governance (La Porta et al. 2000). Actions by corporate insiders contain strong signals to company shareholders (Filatotchev & Bishop 2002). Signallers may of course signal falsely if they perceive it to be to their advantage (e.g. Cohen & Dean 2005). Reputation, developed over repeatedly verified signals, becomes a valuable resource to the signaller (Deephouse 2000).

Contract Theory

Contract Theory builds upon the developments of Signalling and Agency Theories and the foundations of risk, uncertainty, and game theory to describe a complete theory for the analysis of incentives, information, and economic institutions (Bolton & Dewatripont 2005). Incentives can be designed which resolve the asymmetric information problems of both adverse selection (Mirrlees 1971) and moral hazard (Holström 1979). Contract Theory extends to considerations of repeated contracts (Laffont & Tirole 1988), and incomplete contracts (Grossman & Hart 1986, Hart & Moore 1990). The theory of incomplete contracts was critical in the development of a network-of-contracts theory of the firm (Jensen & Meckling 1976).

Adverse selection is generally resolved by Contract Theory by positing the ability of a principal to offer an agent a menu of contracts from which to choose which are optimised for the range of the agent's potential type (Spence 1973). In this manner, employers can screen for the skill, training, or productivity of a target hire (Bolton & Dewatripont 2005). Holström (1979) shows that the ideal contract for avoiding moral hazard does not directly incentivise outcome - rather creates an additional information system that incentivises the agent based on observations of the agent's action or the state of nature. Bonner & Sprinkle (2002) delineate between an agent's effort and a wide range of personal, task, and environmental variables in incentivising task performance.

Information Economics

Information economics primarily examines the outcomes of imperfect information between actors in markets and builds on the foundations of agency, singalling, and contract theory. Information theory was required to reconcile the failure of micro-foundations to generalise to observable macro trends, such as consumption trends, inventory cyclicality, and the existence of unemployment (Stiglitz 2002). While there was a role for the state in information economics, it was to resolve information failures which prevented pareto efficiency and equity outcomes (Stiglitz 2002). Externalities like pollution received low interest — research focussed on the endogenous (voluntary) disclosure of firms and individuals rather than the efficacy of mandatory disclosures as a policy intervention. Negative externalities were thought to result from imperfect contracting due to transaction costs (Coase 1960), principal among which was information 'search costs' (Stigler 1961). Farrell (1987) showed that when there was important private information, Coasian private negotiation fails to achieve 'first-best' efficient outcomes.

The extensions of game theory and mechanism design drew information economics closer to the needs of environmental policy makers. A contracting paradigm of environmental regulation was developed whereby regulators design contracts to induce the desired behaviour from individual facilities, firms, or sectors (Karkkainen 2006). The more general Mechanism Design describes how institutions (including makrets) can be designed to efficiently produce goods (including public goods) under various paradigms of private and public information (Prize Committee of the Royal Swedish Academy of Sciences 2007). Mechanism design has been used, for example, to develop multilateral frameworks for addressing climate change (Martimort & Sand-Zantman 2013, Nordhaus 2015).

Market-based regulatory mechanisms (e.g. via taxes, prices, or quotas on emissions) are themselves a response to information challenges. As regulatory problems become large, a finite regulator is unable to ingest and process information as efficiently as a distributed market (Karkkainen 2006). The private information held by market agents prevents markets from operating efficiently (Prize Committee

of the Royal Swedish Academy of Sciences 2007). Further, choosing the appropriate tax or quota in the design of any regulatory response to an environmental problem requires ex-ante information about the welfare impact of the problem and the aggregate cost of its mitigation — information which may be privately held by companies (Karkkainen 2000). A baseline information environment is required to identify and diagnose where market externalities are resulting in public welfare issues requiring policy intervention.

2.3.2 Information Asymmetry between Companies and Stakeholders

Information asymmetry between companies and stakeholders is not as well studied as between companies and shareholders. In this section, it is considered under two broad categories: first, the specific role of information in granting companies their social license to operate, and second, the positive externalities of information more generally.

Information as a Social Good

Because information is non-excludeable (i.e. it's use does not deplete its availability for future use), its production and use often has positive externalities. In many cases it is both a public and a social good. Robust information environments are required to empower private entities to bilaterally resolve environmental damages, either with bilateral contracting (Coase 1960) or legal tort (Esty 2004). Scientific research has been historically recognised as a social good (Callon 1994), as has scientific literacy (Stiglitz 1999). The former is required for ex-ante information about the impact of policy on society ("what is the impact of the pollution we produce?"), the latter to empower policy makers via public deliberation ("how much pollution do we want?") (Karkkainen 2000). Conventional wisdom calls for a policy process which appropriately applies scientific evidence in policy decisions (Rothman & Poole 1985). Pragmatism recognises the reflexive relationship between science and politics — that policy decisions dictate the available evidence ("policy-based evidence" rather than "evidence-based policy") (Funtowicz & Ravetz 1993, Marmot 2004). The ex-ante

information gap in policy design is then political in nature, subject to the competing interests of the public, civil society, companies, investors, and other stakeholders.

Legitimacy Theory

Legitimacy theory is a tangential theory of corporate governance which argues that corporates must justify to society that their actions are "desireable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions" (Burlea & Popa 2013). Companies must maintain a "social license to operate" (Joyce & Thomson 2000, Gunningham et al. 2004) which is granted by the public and can be costly to maintain (e.g. Neale 1997, Moore 2001). This social licence to operate is the primary target of the Fossil Free Divestment Campaign (Mattauch 2015). The maintenance of a social license to operate is proposed as an explanation for why corporates may disclose environmental information beyond what is required for regulatory compliance even when such disclosure is costly (McClelland & Horowitz 1999).

Legitimacy Theory evolves from liberal political economy, wherein society benefits optimally from the maximisation of economic growth and market competition (Smith 1776, Ricardo 1817). Formulations of the role of business in society thereunder supposed that the law described the boundaries of morality. Private firms were free to operate within those bounds, any such activity being definitionally 'moral', but that they should not interfere with the political process that defined those bounds themselves (Friedman 1962). Contemporary thinking in deliberative democracy recognises the agency of the firm in the political and regulatory process (Schere & Palazzo 2011), even to the extent that the firm is able to secure policy which may be contrary to the public interest (Grossman & Helpman 1994, Chomsky 1999). The public and civil society groups are also endowed political agency under such a model, resulting in a system of power relationships between companies, the public, civil society organisations, and policy makers (Schere & Palazzo 2011). The relative power of these stakeholders appears to determine the extent of a corporate's environmental disclosures (Neu et al. 1998).

2.3.3 Disclosure of Environmental Information

The study of financial reporting and disclosure predates its recent coupling with CSR. Researchers have empirically (see Healy & Palepu 2001) and analytically (see Verrecchia 2001) studied the extent to which companies disclose their information and the impact of these choices on capital market behaviour. Theories of information supply are used to describe the incentives for firm managers to disclose information relevant to their activities, and to develop optimal policy interventions for correcting market failures due to insufficient or inappropriate information supply.

A general typology of the voluntary disclosure literature might adopt Dye's (1985) 'disclosure principle' which categorizes management-held information as either proprietary or non-proprietary. Proprietary information includes all information that could alter a firm's future earnings prior to management compensation, including all regulatory and legal action, reduced consumer demand, negotiation with labour unions, suppliers, clients, etc., and other strategic information (Dye 1985). As a baseline, the disclosure principle holds that all non-proprietary information held by managers will naturally be 'unraveled' by market actors as market actors rationally discount the value of the asset's quality until such point that the manager is incentivised to release the relevant information (see Grossman & Hart 1980, Grossman 1981, Milgrom 1981). No straightforward principle is available for the disclosure of proprietary information, however Verrecchia (2001) argues that the cost of disclosing proprietary information (a tautological formulation in itself) is the most compelling reason for managers to not do so. Environmental information, including direct costs for environmental services and current or potential costs as a result of regulatory action, fall in this category.

Legitimacy theory (described above) posits that company managers might make disclosure of proprietary information in order to secure their social licence to operate (Burlea & Popa 2013). The Political Cost Hypothesis describes almost the exact opposite of this incentive.

The Political Cost Hypothesis emerges out of Positive Accounting Theory, a body of literature that seeks to predict and explain choices in the practise and

application of accounting (see Watts & Zimmerman 1978, 1990). The theory posits that managers choose to reveal information in a strictly rational way that maximises their private benefit. The political cost hypothesis suggests that managers will refrain from revealing information that may attract government intervention in the form of an explicit or implicit tax or regulation (such as environmental impact information) (Healy & Palepu 2001). This policy perspective was more of a corollary to the theory than integral to it - the original theory focussed only on policy interventions to correct monopoly market failures (Watts & Zimmerman 1978). Efforts to graft the theory into the mainstream of CSR disclosure literature have been met with substantial critique (see Healy & Palepu 2001, Milner 2002).

The Proprietary Cost Hypothesis posits that company managers have a natural reluctance to reveal information that would compromise the future earnings of the firm (Verrecchia 1983, Jovanovic 1982, Dye 1985). Even for non-proprietary information, managers may withhold information in spite of any 'unravelling' (per Grossman & Hart 1980, and others). A risk adverse manager may refrain from revealing even non-proprietary information due to his/her uncertainty with how the revelation of such information will cause a revaluation of his/her human capital (e.g. Nagar 1999). A manager may also reveal or withhold information pursuant to his/her interest in protecting his/her credibility (e.g. Teoh & Hwang 1991).

For investors faced with systemic environmental risk, and policy makers attempting to understand and correct environmental externalities, it is even more important that environmental information be readily forthcoming from corporates. Neu et al. (1998) find that "firms utilize a combination of acquiescence, compromise, and defiance strategies within their environmental disclosures to respond to the concerns of relevant publics". The publics in question are explicitly company shareholders, government regulators, and "First Nations and other environmentalists"(Neu et al. 1998). Clark (2017) writes that company managers are "reluctant to disclose additional information, or may treat the incremental information desired as a chesspiece in a net-zero sum game between company directors and shareholders".

By the 1990s, the study of CSR had become intrinsically linked with the study of social and environmental disclosure (Gray et al. 1995). Disclosed information had been understood to be a public good among financial stakeholders (e.g. Healy & Palepu 2001), but as CSR efforts turned to voluntary and mandatory disclosure regimes, the notional public good nature of disclosed information was extended to become social good for all society's stakeholders. CSR disclosure was posited to have a have a 'self-regulating' effect whereby corporates performing well in CSR would disclose their performance and be rewarded by the market and with affirmed social legitimacy, creating a virtuous cycle of CSR performance competition (Karkkainen 2000, Matten & Moon 2008, Jackson & Apostolakpou 2010). Companies not performing well would be disciplined by capital markets. Policy makers, seeking to reinforce this virtuous cycle and correct market failures in information markets, looked to mandated disclosure as a policy instrument. (Cohen 2001, Mol 2006, Cohen & Santhakumar 2007, Gupta & Mason 2014).

Proponents suggest that dislosure-as-regulation, employing the devolved decision making of markets, is more efficient than the command-and-control regulatory intervention (Hayek 1945). Empirical studies of the success of disclosure intitatives show that they are at least comparably successful relative to more conventional regulatory interventions (Konar & Cohen 1997, Foulon et al. 2002, Lee 2010). It might be understood that disclosure regulation creates the baseline information environment which regulators require in order develop more conventional market-, contract- or command-and-control based regulation. The disciplining effects of capital markets in response to information disclosure might simply be in anticipation of the ensuing traditional regulation. It is little surprise then that disclosure initiatives meet stiff resistance from companies (per, e.g., Clark et al. 2017): they are politicaleconomy flashpoints which indicate the impending intervention of a regulator on the activities of a company.

A final category of disclosure is mandated 'right-to-know' disclosure. This is policy-mandated disclosure that seeks to empower those impacted by companies (Sand 2002, Bennear & Olmstead 2008), whether they are material or non-material

stakeholders. Civil society organisations use both 'right-to-know' and 'self-regulating' environmental disclosures to provide 'civil regulation' or 'civil accountability' engagements with corporates (Murphy & Bendell 1999, Knox-Hayes & Levy 2011) — legitimacy theory in action.

2.4 The Production of Environmental Information

Environmental information, as well as being disclosed by companies, is also produced by intermediaries in markets. These intermediaries take various forms: consulting companies, credit ratings agencies, and pure data vendors. Their incentives for producing information are relevant to the existence of environmental information.

This section first established the public goods nature of information. 'Information goods' may be either non-rivalrous or non-excludeable, so the theory of public goods must be developed first. Then the modes of production of common-pool and public goods can be examined: first the 'old commons' — the non-excludeable fish, timber, and meadows of Ostrom (1990); and then the 'new commons' — non-rivalrous open-source software (Benkler 2006, von Hippel 2012).

2.4.1 Information as a Public Good

Public goods theory (Samuelson 1954) is used to categorise goods as to whether they are *rivalrous* and/or *excludeable* see Table 2.2. Excludability means it is difficult or impossible to exclude individuals from their use. Rivalry means that consumption by one party reduces its availability for consumption by others.

Table 2.2: Public Goods

	Excludeable	Non-Excludeable
Rivalrous	**Private Goods** Food, Cars, Parking Spaces	**Common Pool Resources** Fish, Timber, Meadows
Non-Rivalrous	**Club Goods** Private Gardens, Movie Theatres	**Pure Public Goods** Weather Forecasts, National Defense

Information displays unique characteristics that allow it to occupy multiple types of public goods (Hall 1981, Hofmokl 2010), see Table 2.3

Table 2.3: Example Information Public Goods, adapted from Hofmokl (2010)

	Excludeable	Non-Excludeable
Rivalrous	**Private Goods** Domain Names	**Common Pool Resources** Radio Spectra
Non-Rivalrous	**Club Goods** Closed Software, Drug Formulae	**Pure Public Goods** Radio Broadcasts, Open Software

Information may also be *appropriable*, if, by using the information, you allow others to acquire it, or capture some of the economic returns (Leland & Pyle 1977). The environmental information goods of interest in this research are non-rivalrous, meaning they can be produced as either club or public goods.

The Old Commons: Governing Non-Exludable Consumption

Ostrom (1990) writes that the conventional approach to governing non-excludeable commodities is to invoke one of two Hobbesian (1651) *Leviathans*: quotas enforced by the state, or by division into private property (with property rights ultimately enforced by the state (per, e.g. Polanyi 1944)). Ostrom's work shows that institutions can appropriately govern common-pool resources without the resorting to imposed government quotas or the division of the resource into private holdings. She describes self-governing institutions as sets of rules: sequential, contingent, and frequency-dependent decisions where simultaneous, noncontingent, frequency-independent actions had prevailed, and observes the emergence of institutions to govern common pool resources across countries, resource types, and centuries. Ostrom's framework for effective institutions is summarised in Table 2.4.

Table 2.4: Effective Institutions, adapted from Ostrom (1990)

Design Principles	Threats
• Clearly defined boundaries and roles • Provision rules must be congruent with appropriation • Collective-choice governance • Monitors are accountable to appropriators or are appropriators themselves • Graduated sanctions • Low-cost conflict resolution mechanisms • Rights to organise • Multi-layer nesting of institutions as appropriate	Endogenous: • Low capacity to communicate • Insufficient trust • No sense of a shared common future • Powerful incumbents Exogenous: • Malign external powers

Most from Ostrom's work concern governance of the consumption of non-excludeable goods. She also, however, applies her principles for effictive institutions to the governance of the production of non-excludeable goods, which is analagous to the production of public information.

The New Commons: Governing Non-Rivalrous Production

The internet is recognised as a 'new commons' — the non-rival nature of many information goods drawing comparison to the non-excludeable bio-physical commons of conventional study. The conventional approach to governing the production of non-rival commodities is to again invoke the leviathans of the state or privatisation (Benkler 2002). In the case of privatisation, strong extended property rights (copyright, intellectual property) need to be granted to the producers of information to allow them to recover their investment costs. As the marginal cost of digital-resource sharing has become negligible, the quality of 'openness' challenges the merits of business models built on strong extended property rights (Carroll 2013).

With near-zero communication costs, conventional assumptions about the mode of production of economic agents become less valid. Market and heirarchal production structures give way to peer production (Benkler 2002); non-substitutable capital goods give way to general purpose technologies (Hofmokl 2010); the networked information economy enables a gift economy alongside a traditional exchange economy (Hofmokl 2010); and free-rider losses give way to passive or even positive

freeriding (Weber 2000) (e.g. whereby a user of a free software generates value by reporting bugs).

New licensing regimes have been developed to support content and intellectual property sharing in the new commons (e.g. Apache, MIT, GNU, Creative Commons). Permissions are granted not just to specific entities but to the unrestricted public (Carroll 2013). The diversity of licenses allows creators to specify how their work is used, shared, altered, and commercialised, some with the explicit purpose of promoting social solidarity, sharing, and collaboration (e.g. Stallman 2016). Software companies increasingly use and contribute to open source software as part of their core business operations (Sverdlik 2014).

Open source collaboration is a key institution in the new commons. Commercial software had traditionally been favoured for its stability, security, and on-demand customer support. Open-source software, on the other hand, began as a niche for advanced users and tinkerers contributing out of interest, and for non-profits and academics with no ability to pay (e.g. van Rooij 2007). These relative merits are diminishing (Chubirka 2014). The growth of open-source collaborators has created a large community of peer-support, and almost all open source software now also has affiliated fee-for-service companies offering product support and integration (Hillis 2016). Commercial software-as-a-service business models meanwhile have traded version stability for flexibility and the ability to iteratively meet the needs of advanced users (Hillis 2016). Security vulnerabilities appear across both institutional environments (Chubirka 2014).

Open-source software is now ubiquitously embedded in commercial software products, as is the development time of companies allocated to open-source projects (Linux Foundation 2014). Socialised production of software may ultimately be more efficient than in-house production: software business leaders see open-source development as reducing engineering costs and shortening product development cycles (Linux Foundation 2014). Von Hippel (2012) argues that "given modularity, heterogeneous users innovating independently and freely revealing [their work] can produce more and better designed work ... than can individual producers who protect

their private innovation". Countering arguments concern the quality control (e.g. Kemp 2010) and governance (e.g. Adewumi et al. 2015) of collective projects. Facing competition from the open-source Wikipedia, encyclopaedia Britannica's website is telling: "explore the *fact-checked* online encyclopedia from Encyclopaedia Britannica with hundreds of thousands of *objective* articles..." (Britannia 2017, emphasis added). Jim Zemlin, Executive Director of The Linux Foundation, anticipates a wholesale shift to open-source software, whereby all commercial products will be comprised of at least 80% open-source software with only the remaining 20% value-added proprietary code (Sverdlik 2014).

2.4.2 Information Production in Markets

Information Markets and Intermediaries

Principals have a third-party alternative for procuring information relevant to their contracting with an agent. Information is produced as a commodity by intermediaries to assist principals in the management of moral hazard in contracts with agents (Leland & Pyle 1977, Campbell & Kracaw 1980, Allen 1990). The existence of intermediaries in capital markets has also been posited as a response to the asymmetry of information in those markets. Intermediaries might exist because principals are otherwise unable to produce information themselves due to the 'appropriability' condition of information (Leland & Pyle 1977).

Information displays characteristics of both public and private goods (Rose 1999). Even in cases where information is exclusively obtained, the use of such a piece of information, typically to make an observable transaction in a market, allows the observers to infer the content of the information (called the 'appropriability' of information) (Leland & Pyle 1977). Regulators endow different private property rights to information owners for different items and contexts (e.g. restrictions on insider trading)(Armour et al. 2016).

A commodity that cannot be made into private property cannot be efficiently provided through market competition (per Arrow 1962) leading to monopolistic conditions in the provision of commodity. The production of information has a

high fixed cost and barriers to entry but almost negligble marginal costs (McCain 1988). Competition in information markets quickly reduces prices, dissuading new entrants (Fishburn et al. 2000). Information markets might then be said to be monopolistic. The concentrated condition of information markets results in inflated prices and sub-optimal levels of production, quality, and innovation (Arrow 1962, Hirshleifer 1971, Hall 1981, Armour et al. 2016). These market failures have frequently attracted regulatory correction. One of the purposes of mandatory regulated disclosure, for example, is to compensate for the underproduction of information (Hall 1981, Armour et al. 2016).

Beyond monopoly pareto-suboptimality, information failures are endemic to the existence of socially externalised costs, such as those caused by pollution. Information asymmetries impose large 'search costs' - the costs of seeking out the information necessary for a decision and parsing it into known information (Stiglitz 2000). Search costs are transaction costs which frustrate both effective environmental policy making (Esty 2004) and multilateral contracting which might otherwise capture environmental externalities (Coase 1960).

Market Efficiency

Market efficiency describes the rate at which information moves through markets (i.e. how quickly market actors appropriate information)(Fama 1970). Market efficiency is relevant in this context because it determines how 'appropriable' information is — how quickly the arbitrage opportunity of a new piece of information deteriorates. This affects the incentives available to information producers. Fama (1970) describes three forms of the Efficient Market Hypothesis which are summarised in Table 2.5. The proliferation of environmental information vendors suggests that markets are not efficient per Grossman & Stiglitz (1980).

Table 2.5: Forms of efficient market hypothesis (Fama 1970)

Form	Description
Weak Form	Market prices reflect all past market prices; Technical analysis may not be used to predict price activity.
Semi-Strong Form	Market prices reflect all publicly available information; New public information is quickly reflected in market prices.
Strong Form	Market prices reflect all private and publicly available information; Any new information is quickly incorporated into market prices. Fundamental analysis may not be used to predict price activity.

2.5 Technology

Much has been made of the 'disruptive' potential of technology, whether its role in the creative destruction within markets (Schumpeter 1949), the supplantation of industry giants by smaller, nimbler companies (Bower & Christensen 1995), or even the wholesale destruction of labour markets by automation (Frey & Osborne 2013). Technology is surely the largest factor of production (Solow 1957, Magee & Devezas 2016) but otherwise resists categorisation and typological discovery.

Technology changes are often framed in Kuhnian (1962) 'paradigms': paradigmatic shifts describing periods of rapid change interceding periods of relative stable orthodoxy (see also Perez 1985). The prime instigator of these changes can range from market forces (Schumpeter 1949) to institutions, social needs, and policy decisions (Dosi 1982, Mazzucato 2018). The structure of society is embedded in and reflexive with technology (Mishra 1979); imagine the Reformation without the printing press, or open source software without the internet (Dewar 1998, Goldin & Kutarna 2016).

Writers at the turn of the millennium identified the emergence of an 'information age' (Castells 1996, Dewar 1998). Moore's Law (1965), the casual observation that the density of transistors in an integrated circuit doubles every year or two, showed no sign of slowing. In 1977, the Cray-1 Supercomputer was sold for $8.86mn with a processor capable of performing 133mn floating point operations per second (133 MFLOPS) (The History of Computing Project 2013). In 2016, NVidia's newest graphics processing unit (GPU), the Tesla P100, sold for $5,899

and delivered 4.7 TFLOPS (Microway Inc. 2017). Adjusted for inflation, the price of computing power has reduced by an order of 10^8 in 40 years. This improvement removed many constraints on the complexity of models and the computational intensity of algorithms.

Shortly into the new millennium, it became clear that automation would be a defining hallmark of the emerging information age (Frey & Osborne 2013, Bostrom 2014, Harari 2016, Tegmark 2017). The coincident emergence of smartphones, cloud data infrastructures, commoditised web services, and computation boosted by GPUs led to an explosion of interest in machine learning (Hardy 2016). The 'AI winter' since the 1990s came to an end, and data was hailed 'the new oil' (Toonders 2014, Economist 2017). Many incumbent disciplines continue to be taught 'the bitter lesson' that Moore's law driven improvements in *search* and *learning* eventually outperform hard-won heuristical human knowledge (Sutton 2019).

This DPhil has identified three data-driven technologies which will meaningfully change the availability of environmental information — two of which are AI technologies, one of which is driven by digital connectivity. For the purposes of this literature review, an introduction to machine learning is also provided.

2.5.1 Introduction to Machine Learning

A typology of machine learning might split the field into several genera: unsupervised learning, supervised learning, and reinforcement learning. This typology is sufficient for this introduction to the field, but in practise algorithms of several categories can be combined in a single learning machine. This typology is extended in Table 2.6.

Unsupervised learning algorithms detect patterns in data without the assistance of target labels. Use cases include pattern discovery, for the purposes of, e.g. data mining; clustering of samples; or dimensional reduction prior to the data's input to another algorithm or learning machine. Popular clustering algorithms include gaussian mixture models (Pearson 1894), k-means (MacQueen 1967) and its variants, DBSCAN (Ester et al. 1996), and CURE (Guha et al. 2001). Common

Table 2.6: Machine Learning Typology

Learning Type	Description
Supervised learning	Machine learning mapping an input sample to an output label using a corpus of input-output pairs.
Unsupervised learning	Pattern learning in a data corpus without data labels.
Weakly supervised learning	Supervised learning where data labels are noisy, limited, or imprecise.
Semi-supervised learning	Mixed supervised/unsupervised learning where only some data have labels.
Self-supervised learning	Supervised learning where data are automatically labelled by (some subset or function of) the data itself.
Online learning	Supervised learning where data are only used a single time (as in a live stream of data).
Active learning	Supervised learning with live input from a (human) user.
Transfer learning	Supervised learning initialised with model weights trained on a similar (usually much larger) data corpus.
Reinforcement learning	Agents learn from a reward function dictated by its actions in an environment.

dimension reducion algorithms include principle component analysis (F.R.S. 1901) and more-generalised manifold learning variants (e.g. Roweis & Saul 2000).

Supervised learning algorithms are trained with data samples and target labels. They are generally used for prediction problems such as classification and regression. Popular algorithms include support vector machines (Vapnik & Lerner 1963), random forest variants (Tin Kam Ho 1995, Breiman 2001, Chen & Guestrin 2016, Dorogush et al. 2017), Bayesian network models (Pearl 1985), gaussian processes (Williams & Rasmussen 1996), and neural networks (Rosenblatt 1958, Kelly 1960). Neural networks combine small *neuron* computational units, comprised of a simple linear function and a non-linearity, into an information processing system. Multiple layers of connected neurons, *deep neural networks*, can, in theory, be parameterised to approximate any non-linear function (Zador 2000). Machine learning with deep neural networks, *deep learning*, has become synonymous with the development of artificial intelligence.

Deep learning arranges a graph of neurons into an architecture effective for minimising a loss function (usually a function of the error between predictions and labels in a regression or classification problem). Graphs are normally constructed of layers: tensors of parameters and their gradients encoded to use linear algebra and

harware parallelisation (e.g. GPU computation) for efficient optimisation. Deep learning research focuses on developing novel architectures, optimisers, and loss functions for problems of interest, and on developing intuition of how information propagates during network training.

Deep learning has been particularly effective in developing architectures specialised for dimensionalised data, both one-dimensional sequence data (e.g. text and audio) and higher-dimensional images and video. These methods are discussed further below so are omitted here.

Reinforcement learning is a form of supervised learning which can be used when no training labels are available, but an environment (or model) can be sampled to generate sample-label pairs. It is used in dynamic control problems wherein an autonomous agent seeks to optimise a "reward" issued by the environment. Reinforcement learning has an extensive history across disciplines (see Sutton & Barto 199), but became an area of intensive deep learning research when deep neural networks were used to generate policy selections ("deep Q-networks", DQN) (Mnih et al. 2015). Using DQN, Google's Deepmind has mastered classic Atari games (Mnih et al. 2015); Chess, Shogi, and Go (Silver et al. 2018); and Blizzard/Activision's Starcraft II (Vinyals et al. 2019).

Several more novel deep learning techniques and architectures are worth highlighting in this literature review. Transfer learning (Pratt 1993) allows models that have been pretrained on very large datasets to be fine-tuned for specific use cases. Generative adversarial networks (GANs) (Goodfellow et al. 2014) pit two or more neural networks in competition with each other to complete a given task, and has led to major improvements in many computer vision applications. *Transformers* (Vaswani et al. 2017*a*) have broken records in NLP tasks using an *attention* mechanism similar to that found in LSTMs. This mechanism is now being adapted to computer vision tasks (Wang, Jiang, Qian, Yang, Li, Zhang, Wang & Tang 2017, Parmar et al. 2018, Bello et al. 2019).

2.5.2 Earth Observation and Computer Vision

Computer vision describes a broad category of problems with imagery or video data as input, such as object or facial recognition and tracking, semantic segmentation, or image superresolution. These techniques take advantage of the spatial arrangement of pixels (and temporal, in the case of video) by using convolutional neural nets (CNNs) to encode data. CNNs are composed of kernels of neurons which encode a receptive field into a representation of the original data, an idea originating from Hubel & Wiesel (1962). Modern CNNs began with LeCun et al. (1989, 1998) who applied back-propagation to train CNN kernels to identify handwritten digits.

ImageNet (Deng et al. 2009) is a benchmark dataset of 14mn red-green-blue images labeled with 20,000 unique categories. The ImageNet Large Scale Visual Recognition Challenge invited participants to compete on the error rate of their best algorithms. In 2012, Krizhevsky et al. (2012) submitted AlexNet, a machine using CNNs trained, for the first time, on a GPU. GPU training has since become the norm for machine learning research and practise. "Human-level" performance on the ImageNet Challenge (around 5% error) was surpassed in 2015. Top-performing computer vision architectures are published open-source with model weights to enable their application in diverse use cases via transfer learning (i.e. fine-tuning a pre-trained model on a task-specific dataset). Popular pre-trained models include ResNet50 (He et al. 2016), VGG-16 and -19 (Simonyan & Zisserman 2015*a*), and GoogleNet (Szegedy et al. 2015). U-Net (Ronneberger et al. 2015*a*) showed success on a small data corpus with limited class label categories.

These deep-learning methods were quickly adopted to earth observation. Early applications focused on object detection, for example human-made objects such as roads (Zhang et al. 2018), buildings (Ishii et al. 2016), automobiles (Audebert et al. 2018), aircraft (Zuo et al. 2018), and ships (Bentes et al. 2018). CNNs have also been used with remote sensing imagery to determine coal-fired power station utilisation rates (Gray et al. 2018), crop yield prediction (Wang et al. 2018), and wind power forecasting (Wang, Li, Wang, Peng, Jiang & Liu 2017).

Emerging applications seek to better mitigate the differences between earth observation imagery and conventional imagery. Earth observation imagery in often obfuscated by atmospheric interference and clouds, the spatial resolution is usually quite coarse relative to the objects being detected, there is no horizon and the focus of the object is always a fixed distance away. Earth observation imagery also comes in many spectra other than standard red-green-blue. Sensor fusion research seeks to combine sensors of different resolutions or radiographic spectra (e.g. He & Yokoya 2018, Bermudez et al. 2019, Fuentes Reyes et al. 2019). Other studies seek to mitigate the impact of atmospheric interference (e.g. Eckardt et al. 2013, Grohnfeldt et al. 2018, Bermudez et al. 2018, Meraner et al. 2020). Still other seek novel CNN architectures and training curricula specially developed for earth observation imagery (Birch et al. 2020).

2.5.3 Web Scraping and Natural Language Processing

Natural langauge processing (NLP) is a branch of machine learning concerned with problems relating to natural language: translation, text comprehension and transcription, speech-to-text, and many more. Sequence models, where the model input is a series of words, or tokens, are used for these problems. To propagate information information forward and backward through the data sequence, recurrent neural network (Rumelhart et al. 1986) architectures are used. Similar to GPU-trained CNNs, the RNN long short-term memory (LSTM) architecture (Hochreiter & Schmidhuber 1997) broke records in handwriting recognition (Graves & Schmidhuber 2009), speech recognition (Li & Wu 2014), and machine translation (Sutskever et al. 2014). Now NLP models have evolved to use Transformer architectures (Vaswani et al. 2017*a*) and an 'arms race' is ongoing with rival companies building general-purpose language models with ever larger numbers of parameters.

Most large NLP models are built using web scraping. BERT (Google, Devlin et al. 2018, 340mn parameters), GPT-2 (OpenAI, Radford et al. 2019, 1.5bn parameters), Turing-NLG (Microsoft, Rosset 2019, 17bn parameters), and GPT-3 (OpenAI, Brown et al. 2020, 175bn parameters) were all trained at least partially

on web-scraped data. Web scraping and web parsing is a prime application for these language models. Applications include sentiment analysis (Cambria et al. 2017), clinical and health data processing (Dreisbach et al. 2019), and financial data parsing (Evans et al. 2019).

2.5.4 Crowd Sourcing Data and Computing

Technology is empowering the public to participate in the generation of science in an unprecedented way. Citizen science has a long history, for example in amateur astronomy or ornithology (Silvertown 2009), however the recent diffusion of the internet and the smartphone has enabled an expansive increase in citizen science and associated activism and monitoring. The tradition of amateur naturalism has modernised into community-based environmental monitoring, e.g. of water quality (Kim et al. 2011), biodiversity (Couvet et al. 2008), forest management (Kelly et al. 2012). Several of these initiatives directly target the environmental impacts of the activities of companies. Examples include community monitoring of Rio Tinto's nickel mining in Michigan (Kozacek 2014) and Murray Energy Corporation's mountaintop removal coal mining in Virginia (SAMS Admin 2017). The success of these community monitoring activities give credibility to theories of corporate legitimacy and social license to operate.

The enabling role of technology in citizen science is well documented. Advances in technology have provided both the low-cost, ubiquitous, and networked devices for data capture, and the 'back-end' platforms for storing and processing community-submitted data (Gouveia & Fonseca 2008, Silvertown 2009, Newman et al. 2012). Amateur communities are also contributing to the processing and analysis of data. Community-driven computing was famously popularised by the SETI@home and Rosetta@home projects, which allow users to donate their computing resource and spare time to signal processing and protein folding problems respectively (Hand 2010). The connectivity of the internet has allowed the focus of the human brain to be commodified, enabling humans to donate their computational power to causes they care about.

Relative to the institutional production of science, open- and citizen-sourced data is vulnerable to critique on the grounds of quality and veracity (Sadiq & Indulska 2017). The design of new data initiatives must successfully capture generated metadata and quality information in order to credibly generate science (e.g. Wiersma et al. 2016). Governments are also increasingly participating in open information and open data environments. They increasingly recognise the private and social value unlockable from government data, leading to initiatives like the United States' Data.gov (2016) and the United Kingdom's Data.gov.uk (2012).

3
Paper Synthesis

Contents

3.1 DPhil Contribution and Paper Synthesis 42

3.1 DPhil Contribution and Paper Synthesis

The novel, significant, and rigorous contribution of this DPhil is contained in the three data science papers which are attached as appendices. While each paper is novel and significant in its own right, together they provide a compelling evidence base for not only how asset-level data might be obtained using earth observation, but that asset-level data itself is necessary for understanding environmental risk.

The first paper, *A global inventory of commercial-, industrial-, and utility-scale photovoltaic (PV) solar generating units*, provides, for a the first time, a global asset-level inventory of PV solar energy facilities. Data on the diffusion of PV solar generating installations is extremely relevant to the ongoing energy transition and achieving a net-zero carbon energy system. The precise location of PV energy installations is needed in order to accurately forecast and now-cast PV energy generation, helping grid operators dispatch flexible electrical loads, send stronger pricing signals to incentivise curve-flattening, and maintain narrower margins of

(fossil-fuel) based operating reserve due to higher confidence short-term renewables projections. Increased generation of renewables exposes incumbent generators to transition risks, as their generators are displaced by must-run solar energy and, even where quantities are not effected, renewables cannibalise power market prices by running at zero-marginal-cost. Because PV generation is dependent on the precision geolocation of the generating unit, these applications are not possible without precisely-localised data.

The inventory is developed using computer vision, earth observation, and a large cloud computation infrastructure. Training data are obtained from OpenStreetMaps. As a technical acheivement, this deployment was, to our knowledge and at the time it was conducted, the largest deployment of a machine learning model on remote sensing imagery for the purposes of object detection. The extra challenges involved in deploying at such a scale are also noteable, including managing covariate shift of training and test data across geographies, as well as the excessive optimisation and parallelisation of algorithms to reduce computing costs. Although the computer vision models are taken from state-of-the-art literature, their application to remote sensing is new and novel, and can help inform other researchers who are working with this unique imagery.

We develop additional analysis using our novel dataset of PV energy facility sites. We sample the sites from a global land cover product to understand how the growth of PV energy is impacting land cover worldwide. We find that the most common land cover type for PV solar installations is croplands, followed by aridlands and grasslands. This analysis shows how policy makers must consider trade-offs between renewable energy generation, food production, and conservation. We also develop a novel method for estimating nameplate generating capacities using the geometries resulting from our machine learning pipeline. These contributions are all crucial for developing a full understanding of energy transition risk and would not be possible without geolocalised assets.

In the context of this DPhil, this is the contibution of the first paper: a demonstration of the necessity of asset-level data for assessing climate change

transition risk and also impacts on environment and community stakeholders. This paper also serves as a case study for the opportunities and challenges associated with obtaining global asset-level data using these earth observation and machine learning methods. It helps to establish what might possible and what is likely not possible with these methods prior to discussion of their role in producing public goods for the purposes of sustainable finance.

The second paper, *Transition Risk in the Global Coal, Oil, and Gas Supply Chains*, presents analysis of transition risk in fossil fuel supply chains using networks of asset-level data. A wide range of sources are drawn upon to build a corpus of asset-level data describing fossil fuel production, transport, and consumption. These assets are arranged into a directed graph, where geospatial proximity between supply chain components is used to form the graph edges. The complex networks, one for each of the three commodities, contain 6.09mn nodes and 15.70mn edges after simplification. We use these complex networks to solve for minimum-cost flow between energy supply and energy demand, and compare these results to ground-truth energy trade data. We solve for energy flow through the supply chain by implementing a novel non-linear cost scaling algorithm (Goldberg & Tarjan 1990). We analyse supply chain risk by imposing a number of shock scenarios. First we consider a demand-shock in the form of a transition to energy demand levels consistent with the IEA Sustainable Development Scenario — a scenario aligned with the Paris Agreement and the UN SDGs (International Energy Agency 2019*c*). Next, we analyse a number of supply-shock scenarios caused by a doubling of regional energy costs. In each scenario, we assess the system-wise and asset-level impacts on energy flow and costs. In so doing we identify geospatial *sensitive intervention points* (SIPs) (Farmer et al. 2019*a*) — elements in the supply chain where a small intervention can have an outsized impact on decarbonisation.

This paper contains some novel technical achievements. The geospatial operations used to assemble this network were parallelised in a novel way, achieving state-of-the-art performance in the spatial join operations. The supply chain networks prepared in this paper are very large, requiring novel approaches to

memory management and algorithm design. The entire paper is implemented using a data science framework for maximum reproduceability. The computational graph for network assembly, prior to any flow calculation, features 47 computational nodes and 72 dataset edges.

In the context of the DPhil, this paper demonstrates that asset-level data are required for complete analysis of transition risk. While previous studies of transition risk have ascribed these risks to sectors and markets, we ascribe transition risks to assets and their location in the topology of a complex network. The elements of this complex network are geospatially localised. Only with globally exhaustive asset-level data can this supply chain be assembled and transition risk be interrogated.

The third paper, *Towards* DeepSentinel: *An extensible corpus of labelled Sentinel-1 and -2 imagery and a general purpose sensor-fusion semantic embedding model* is an earth observation computer vision paper. With this paper I make progress against two perennial problems in earth observation for object detection: first, the presence of atmospheric noise and clouds, and second, a lack of training data labels. Atmospheric noise and clouds are very common challenges in earth observation applications. New approaches use Sentinel-1 Synthetic Aperature Radar imagery which has an active sensor which can pierce clouds and atmosphere. This data can add a lot of information alongside more conventional Sentinel-2 multispectral data in sensor fusion applications. Sentinel-1 contains many novelties, however, that make it difficult to use. In this paper I experiment with model architectures that optimally fuse Sentinel-1 and Sentinel-2 imagery for general-purpose object detection.

The lack of training labels is a common problem in machine learning with earth observation. Although there are plentiful earth observation samples, the number of known objects to use as training data often tends to be small. In our work with PV solar facilities, as an example, we were fortunate that such a large sample of facilities was available from OSM (~10,000s), allowing us to use supervised learning directly. For other asset types, the number of training samples available might be considerably fewer (~100s). In the computer vision literature, this challenge is overcome by first pre-training a model, either by self-supervision, or by using a much

larger more general imagery corpus. In this paper I adapt both these approaches to earth observation computer vision. I develop a training curriculum which involves first self-supervised pre-training on pairs of Sentinel-1 and Sentinel-2 imagery, then coarse-tuning on a large landcover dataset. These coarse-tuned models are then ready for final fine-tuning specific to a given problem.

As a test of this method, I develop commodity labels for a large global dataset of the planet's surface mine sites. Using only a small sample of 608 hand-labeled surface mines, I fine-tune a multi-class classifier to estimate the commodities produced in the 22,000 surface mines localised by Maus et al. (2020*a*). Unfortunately the accuracy of the coarse-tuned models on landcover classification and fine-tuned models on the mines data was not sufficient as to be 'production ready'. The experiment developed in this paper, comparing combinations of pretraining methods and encoding architectures, has created valuable intuition for further iterations towards a general-purpose embedding model.

In the context of the DPhil, this paper demonstrates that earth observation with computer vision is a highly extensible method and that some of the major technical concerns around the availability of training data and atmospheric interference might be able to be mitigated. This study uses 10m Sentinel-2 and Sentinel-1 data, which, relative to leading private satellites with 30cm ground resolution, might be considered 'medium resolution'. Even with medium resolution imagery it is possible localise assets (as in Paper 1) and determine their features (as in Paper 3). This medium-resolution imagery is publicly available, so these papers demonstrate that these methods might be accessible to public sector organisations who are unable to pay for premium imagery.

It is worth describing, as experiential evidence to be drawn upon in the DPhil, who the coauthors and supporting external intuitions are on these projects. The first paper was developed in collaboration with Descartes Labs Inc. and the World Resources Institute. Descartes Labs is an earth observation technology company spun out of NASA's Los Alamos lab in the United States. Descartes Labs was working on a similar project and we agreed to combine our efforts. This project was

among the first supported by what would become Descartes Lab's Impact Science Program. Under this program, Descartes Labs supplied the project with in-kind computation support, the cost of which was approximately US$20,000. The World Resources Institute (WRI) is an NGO based in Washington DC that maintains the Global Power Plant Database (GPPDB) that was used for cross-validation data. WRI has agreed to rehost our final, peer-reviewed dataset on their platform ResourceWatch and also add a selection of it to the GPPDB. The model training and hosting was supported by a research cloud credits grant from Amazon Web Services.

The second paper was accelerated substantially by a pro-bono engagement with QuantumBlack. QuantumBlack joined the project after approaching the Smith School of Enterprise and the Environment looking for a sustainability-related project they could get involved with. Interest in the project by QuantumBlack employees was oversubscribed. The project took another step forward when it was supported by an undergraduate intern on an Ogden Fellowship. Supporting the computational requirement of the project was Microsoft's AI for Earth program. Microsoft's AI for Earth supports research with substantial in-kind grants for use with their Azure cloud computation platform.

The third paper is supported by both the Descartes Labs Impact Science Program and a Google Research Credits grant. The work was presented at the Google GeoforGood Virtual Conference in 2020 and was also selected as a Copernicus Masters finalist with the invitation of presenting at the European Space Agency's Phi-Week, a machine learning and earth observation conference. The NGO the Natural Resources Governance Institute provided the labelled training data which were used to finetune the final commodity classfication model.

These institutional arrangements demonstrate a few of the themes which are drawn upon in this DPhil. First, paper 1 and paper 3 demonstrate the interest of public sector organisations in developing better asset-level data for the purposes of assessing environmental risk and impact. This data is mission-critical for these NGOs who represent the diverse interests of company stakeholders. Second, all three papers were supported by technology company partners, either as direct

collaborators and coauthors, or by providing cloud computation credits. This speaks to the interest of technology companies in supporting and participating in sustainability and public sector research. Lastly, the necessity of obtaining third-party research credits to perform this research provides further evidence to show that large scale computational research is becoming less possible from within academia. The scale of computation required for applied machine learning methods has consequences for types of research questions which are accessible to public sector research institutions, with consequences for policy and public welfare.

4

Technology, Information, and the Governance of Environmental Risk

Contents

4.1 The Limits of Technology in Unravelling Company Information 50
4.1.1 Limits and Biases of Earth Observation 51
4.1.2 Limits and Biases of Web Scraping 54
4.1.3 Limits and Biases of Crowd-Sourcing Methods 55
4.2 Disclosure for Shareholders vs Stakeholders 56
4.3 Fostering Innovation in Markets for Environmental Information 59
4.4 Policy Recommendations 61
4.5 Opportunities for Further Research 63

In the literature review, I established that disclosure is a policy mechanism which is being implemented to meet the needs of both company shareholders and stakeholders. I interrogated the economic nature of information, its role in moderating the relationships between companies and investors and companies and society, and its modes of production. In the paper synthesis, I showed the increasing role that technology might be expected to play in unravelling company information, and the crucial role of asset-level data in assessing transition risk. In this section, I develop policy recommendations based on the literature review, the

evidence exhibited in the three DPhil papers, and the experiential evidence of developing these papers.

This discussion will seek to address three questions which are of interest to a social planner:

- What are the limits and biases of the role of technology in unravelling company information?
- How can disclosure be made fit for the purposes of all society's stakeholders?
- How might the innovation of data providers and intermediaries be shaped to provide better information relevant to environmental risk?

4.1 The Limits of Technology in Unravelling Company Information

The first question I examine is what might be the limits of technology in unravelling company information, and what role does this imply for technologists and policy makers?

I have shown, in my first and third DPhil papers, how earth observation might be used to localise company physical assets and ascertain some of their features. In the literature review, I have also drawn on examples where web scraping and crowd-sourcing data has been used to produce structured data relevant to environmental risk. In this section, I consider these three technology pathways, the limits these technologies might have in producing novel datasets relevant to environmental risk, and the potential new biases that they will introduce. I consider bias using the typology of Crawford (2017) which distinguishes between *allocative* and *representative* harms resulting from biased data and models. I ultimately find that critical biases and limits remain which will hinder the universal unravelling of key company information.

4.1.1 Limits and Biases of Earth Observation

Various industry observers have noted a step-change in the structure and pace of earth observation activity (Patel 2020, McKie 2020). It is clear enough that, for the foreseeable future, increasingly more earth observation spacecraft will be launched (Union of Concerned Scientists 2020), leading to more frequent revisit times, and sensor spatial resolutions across a variety of spectra will improve inline with Moore's law. What limits, then, might there be to company information obtainable using earth observation and computer vision?

As a baseline, it can be expected that inventories of asset-level data will be increasingly available. Currently-available inventories of asset-level data derived from earth observation include, as examples, our work on PV solar energy (Paper A), as well as that of Yu et al. (2018), Maus et al.'s (2020*a*) global mining data, and the US oil and gas infrastructure inventory of Sheng et al. (2020). While the types of assets in question might initially be limited to large fixed-capital assets, increased spatial and spectral resolution will mean that increasingly more and diverse assets types will be reliably detectable.

Features (such as production quantities, equipment type, pollution control technology, etc) of assets that might be detected will only be those which are unobscured to a satellite's line-of-sight. Earth observation will not be able to determine, for example, the contents of a warehouse or the type of equipment manufactured at a manufacturing plant. Other features unrelated to physical objects, such as the owner or operator of the asset, will not be able to be detected with this method. Certain features of assets might be able to be determined not from direct object detection, but may be proxied by other objects or spectral signatures, such as plume detection to determine power station utilisation rates (Couture et al. 2020), or the use of inSAR to measure water table extraction (Vasco et al. 2019).

The computer vision algorithms that are and will be used for these tasks will be biased. The training data for developing these machine learning systems will be drawn from a distribution which does not represent the true distribution of the objects being detected. Because exhaustive datasets for these assets are unavailable,

their true distribution is unknown, so conventional bias-control techniques employed in machine learning cannot be used. Covariate shift the input space (i.e. remotely-sensed pixels) would be expected, and shift might also occur in the latent embedding space. Latent features, corresponding to patterns of detections representing, for example, pieces of equipment, might be present in some distributions of assets but not in others, leading to biased outcomes. (For example, if a computer vision system was only trained on power stations with large round cooling towers, it will misclassify powerstations without these features.)

The needle-in-a-haystack nature of localising assets and their features in earth observation implies an extreme class imbalance for model training. While there are methods available to manage class imbalances, the imbalance means that in most cases 'bootstrapping' (i.e. human labelling of test-data to add to the training dataset) will be so inefficient so as to be impractical. Additionally, while a human may be able to 'ground-truth' optical imagery with visible-spectrum bands, a human will not be able to ground-truth non-visible spectra imagery, such as multispectral, thermal, and synthetic aperture radar imagery.

The end result is that earth observation and computer vision methods will biased according to their training data. For many assets, the available training data is significantly biased by geography — asset-level data being most available in the Global North. This will lead to substantially deteriorated model performance in geographies in the Global South, as we found in the deployment of our PV solar energy detection model. The geographical distribution of information is the inverse of the geographical distribution of financial risk, return, and growth opportunities, most of which are in the Global South (Melas 2019).

Imagery quality is also biased. Certain geographies have substantially more atmospheric interference and clouds, adding noise or decreasing the availability of samples in those regions. In my third paper, I attempt to control these geographic biases by pre-training on a large global corpus of imagery. Even if geographic covariate shift can be marginalised, the model will still be trained on positive samples drawn from a biased distirbution of training data. As such, in a model

deployment, the performance will still be worse in geographies where training data was not available. As an example, in detecting PV solar energy facilities, we found that facilities had substantially different size distributions in different geographies. Our model was mostly trained on moderate-sized installations in Europe and Japan, and fared worse at detecting very large installations such as those in Chile and California.

This bias in representation will result in model impairments where reporting and public data is the most scarce, most commonly in the Global South. These impairments could lead to allocative harms depending on how these models are used. Company managers, for example, might be grateful that biased training data and model impairments have meant that their assets are underrepresented in a public data product, if the purpose of this product is to better draw attention to environmental impacts. The same company's stakeholders and shareholders, meanwhile, will not have access to the data they need to drive meaningful engagements with company managers. It bears noting that, given a counterfactual where no information is available, low-quality data with known bias is still preferable to no data at all.

While agnosticism to company- and country-level reporting is a desireable feature of earth observation methods because it's able to produce more exhaustive datasets, technologists must be responsible in their development and deployment. There is a risk that, if these methods are used exclusively in the service of financial interests concentrated in the Global North, the result will be to further erode the sovereignty of Global South nations and to entrench a neo-colonial and extractive global finance industry whose benefits are not equitably distributed. Only if this data is held in trust as a public good can its benefits be equitably distributed.

The use of earth observation for this and other humanitarian purposes would suffer a catastrophic setback if military weapons such as those currently under development were used to destroy rival nations' satellites. In an extreme scenario, humanity might be 'locked in' by a space debris field of our own making (Kessler & Cour-Palais 1978). The hegemonic power of the United States military has so far

kept this scenario from materialising. As rival nations grow their space capability, countries will need to cooperate to continue to mitigate this risk.

4.1.2 Limits and Biases of Web Scraping

The production and analysis of company data is a common application for NLP. Its use for the creation of detailed asset-level data, however, is still nascent. Applications have so far focussed on the timeliness, rather than feature-richness or collective exhaustion of data. Natural language models are becoming increasingly sophisticated and continue to break records in terms of training corpus and model size. While skeptics argue that bigger has not led to better (McCoy et al. 2019, Bender & Koller 2020), the arms race for better human language models will continue. Web scraping powered by NLP allows technology developers to add structure to unstructured website and web query data. Typical data sources include company filings (e.g. Rogers et al. 2017), business media and press (Hawley 2017), and company websites (Berardi et al. 2015).

Parsing entity relationships and knowledge graphs from natural language text is an active niche of NLP research (e.g. Wang et al. 2014, Luan et al. 2018). Putting these methods to work in the service of obtaining data relating to companies is a common task (e.g. Liu et al. 2019). Few efforts, however, have been made to use these methods to collect asset-level data and their specific features. While this offers a great opportunity for future research, it means that the discussion here is more speculative.

As with earth observation, NLP methods are biased by the availability and distribution of training data. In this case, base corpuses are developed from natural language to develop pre-trained models and then further corpuses are called upon or developed to suit the specific use case. While English-language corpuses are substantial, corpuses in other languages are less available. Combined translation-training is possible, but this still introduces the bias of pretraining on a more narrow corpus. English language corpuses are now drawn extensively from across the internet, but non-English corpuses may be biased in the text sources they draw from. The bias in models trained on English language corpuses is

becoming better studied, with consensus emerging on a bais typology and research agenda (Blodgett et al. 2020).

The success of this technology in populating detailed asset-level data will depend ultimately on the availability of this data in publicly accessible fora on the internet. As such, the bias derived datasets will depend on broader cultural factors, such as (technological) literacy rate, access to the internet, freedom of and instutitional structure of the press, conventions in company reporting, and so on. These factors may vary greatly from country to country. Combined with the underdeveloped work on non-English NLP technologies, it is unclear the extent to which this technology will be able to reliably populate asset-level data. Drawing on the strongest use of this technology to date, Rose, Bauer, Baker, Barkhurst, Bean, DiGiulio, Jones, Jones, Justman, Miller et al. (2018*b*) document the geographical shift in their pipeline performance, showing much lower performance in the Global South.

As with earth observation, allocative harm from these methods arises from the purposes they are put to. If data or features are present in one geography but not another, this may come as a benefit to the managers of companies who are less represented in the data, at the expense of their shareholders and stakeholders.

An institutional concern that would prevent these methods from having a positive impact would be if net neutrality isn't upheld by governments around the world. Net neutrality prevents the privitisation of the internet — internet service providers are prevented from contracting with third parties for preferential traffic or even censorship of certain IP addresses over others. A less-free internet means that information will be increasingly concentrated and less diverse in the insight it provides.

4.1.3 Limits and Biases of Crowd-Sourcing Methods

Both of the previous two methods — earth observation and web-scraping — can also include crowd-sourced data in their design. As an example, many NLP applications commonly draw on Wikipedia as a training corpus, particularly as it is available in many non-English languages. In our work detecting PV solar energy facilities, we

drew on crowd-sourced OSM data as training data. We found that OSM data reliably labeled PV solar facilities, but conformed to different definitions of what composed the footprint of an installation. Various studies have interrogated OSM data quality and have recommended assessment criteria with ground-truth both available or not available (e.g., respectively, Zhang & Malczewski 2019, Sehra et al. 2017).

Crowd-sourcing need not necessarily imply co-production among all contributors, the 'crowd' can be curated as in the NGO CoalSwarm's network of collaborators and data contributors, or the 'expert ameteurs' invited to participate in citizen science projects. There is a growing body of literature that catalogs crowd-sourced and citizen science research and data quality. Crowd-sourced data collection can be designed such that precision and accuracy, sample sizing, and spatial and temporal sampling techniques can match the quality of 'professional' science. Lukyanenko et al. (2016) write that inviting citizens to merely act as human sensor or annotators misses the exploratory opportunities of more participatory science — citizen science might, in some cases, *enhance* data quality rather than necessarily impair it.

Crowd-sourcing often relies on participants' digital connectivity. In this sense it is biased by access to a free and open internet and the affordability of devices and telecommunications network access. As with the NLP applications discussed above, crowd-sourced data will share the biases of the contibuting public, which may lead to underperformance or allocative or representational harms.

4.2 Disclosure for Shareholders vs Stakeholders

Disclosure is used by policy makers for the purposes of both ensuring the efficient mitigation of financial risks and for achieving policy goals relating to social welfare. For shareholders, disclosure reduces information assymetry with company managers to promote better governance of companies and more efficient capital markets. For company stakeholders, disclosure descreases their information asymmetry with company managers and allows them to engage constructively on risk mitigation affecting their welfare. Not all stakeholders in a company are material to the company however (SASB 2018, Edmans 2020). Non-material stakeholders are

Risk Bearer	Hazard	Exposure	Vulnerability
Shareholder	Environmental change and societal responses as a result of company actions	Asset Locations	Vulnerability of company performance
Material Stakeholder	As above PLUS direct company impacts	Asset Locations	Vulnerability of company performance and stakeholder welfare
Non-material Stakeholder	Direct company impacts	Asset locations	Vulnerability of stakeholder welfare

Table 4.1: Shareholder and Stakeholder Risks

impacted by companies but do not share in a company's upside risk, nor is their welfare a liability to the company, so they have no position of leverage with company management. Using the definition of risk as the product of exposure, hazard, and vulnerability, Table 4.1 shows the key differences between shareholders and material and non-material stakeholders.

To mitigate environmental risk for company shareholders, disclosure must meet the needs of all of a company's material stakeholders. Company stakeholders often embody a more diverse group of interests, and meeting their needs might require tradeoffs between different stakeholder groups. As an example from Paper 1, stakeholders in PV solar energy siting decisions include the local residents who might find the installation unappealing, local land owners whose property value might be affected, local biodiversity impacted by habitat loss, local community and indigenous groups whose public or traditional use of land is affected, water resource and food production systems which might be stressed, and stakeholders in the global climate (which is to say, all members of all species) who may benefit from displaced greenhouse gas emissions from renewable energy generation. Only some of these stakeholders, however, might be material to a company.

To mitigate environmental impact, disclosure must go further and also meet the needs of non-material company stakeholders. Disclosure guidance to-date has advised companies to report their managers' opinion of their risk using commonly-accepted scenarios within narrow categories of risks. Company managers cannot be exhaustive in risk reporting, and will not be able to meet the needs of all

stakeholders. Neither might it be desireable to shareholders or material stakeholders that company managers to spend an increasing portion of their time reporting exhaustively on their risk. Rather, it should be the role of investors and stakeholders, who have wider risk management remits, to allocate capital efficiently. In this liberal model of politicaleconomy, the public sector might then be responsible for making a broader envelope of risks material to companies, whether by demonstrating impact of stakeholders on company performance, or challenging company legitimacy when they act against the social interest. A natural equilibrium might then develop whereby companies are held to account for their impacts in such a way that also improves their performance.

Universally-disclosed asset-level data could provide a crucial ingredient for acheiving this well-functioning model. As shown in Table 4.1, asset locations are a common need for assessing risk for shareholders and material and non-material stakeholders. Asset materiality is also required in order to assess the scale of the exposure, whether for assessing risk to company performance on behalf of shareholders, or impact on a material or non-material stakeholder. This general-purpose disclosure will also help company managers avoid excessive reporting on indiviudal risks — avoiding the stalemate of demands by shareholders and stakeholders for increasingly sophisticated disclosures against the limited time of company managers, which is ultimately corrosive to company value.

Company managers will find calls for mandatory disclosure of their asset locations and materiality subversive. They will likely resist such calls on the grounds that this information, like other supply chain information, is part of their competitive advantage. They will argue that their competitive advantage should be protected for the benefit of shareholders and stakeholders alike. This argument is narrow and shortsighted.

First, there is emerging evidence to show that supply chain collaboration can lead to increased value for companies (Liao et al. 2017). Information sharing is a key component of establishing collaborative relationships (Panahifar et al. 2018). A strong public information environment can facilitate these relationships and catalyse

their formation. There are many examples where supply chain collaboration is already being used to manage environmental and social impacts, for example in precompetitive tendoring and responsible procurement coalitions (Chen et al. 2017).

Second, in the face of systemic risks like climate change, and other risks that threaten planetary boundaries, the collective exhaustion of available information is required to mitigate these risks. When the size of an industry threatens to overwhelm social or ecological limits, the priority for shareholders should no longer be the efficiency of the industry but the sustainability of the society and biosphere the industry is enclosed within. Exhaustive public data is required to assess the risk to these limits and should supersede the inclination of company managers to withhold private information.

Finally, the subversive disclosure of information closely held by company managers may be crucial in helping to address structural destabilising injustices associated with 'managerial', or 'agency' capitalism (wherein company managers and financial intermediaries act in their own interest rather than the principal's) (Bogle 2005). A better public information environment may be required to help the other actors in the liberal politicaleconomy drama — policy makers, NGOs, and the general public — better hold company managers and financial intermediaries accountable. A more equitable distribution of information and power may help all actors become willing participants in the transition to a sustainable economy.

4.3 Fostering Innovation in Markets for Environmental Information

Existing asset-level data provision by financial intermediaries shows a monopolised configuration, with only one or two firms providing asset-level data for each industrial sector, and charging excessive rents for data access. Environmental risk analysis, on the whole, has been underserved by financial services — it has only been with a norms shift to an interest in sustainable finance that these offerings are now becoming available. How can more innovation and competition be fostered in private markets for environmental innovation?

Without intervention, the economic nature of information as a non-rivalrous public good will lead to its monopolisation and underproduction (Arrow 1962). Environmental information firms follow a conventional model of innovation in information products: first secure access to a valuable piece of information, and then establish a deep 'defensive trench' around the business to prevent the entry of competitors (McCain 1988, Fishburn et al. 2000). The result is a concentrated information provision market with inflated prices and sub-optimal levels of production, quality, and innovation (Arrow 1962, Hirshleifer 1971, Hall 1981, Armour et al. 2016).

This conventional model is perhaps partly responsible for why the provision of financial services related to environmental risk has significantly lagged the maturity of environmental science and technology. As an example, long-term climate modelling projections have been available since the mid-2000s with CMIP3 and AR4 2007, but it is only recently that financial services products considering the impact of long-term climate change on companies have become available. It has taken a significant change in cultural norms around climate change risk to overcome the barrier of inaccessible asset-level data.

Other information-based sectors of the economy have been more effective in self-organising the production of public goods. Software development, an analogous industry with information as its primary product, self-organises around the production of open-source products. Individuals and companies have a number of incentives for developing open-source software, some of which are based in norms, culture, and marketing, but also include advantages in cost efficiency (Benkler 2006). Companies adopt the 'open innovation' model of innovation whereby they are able to efficiently socialise development costs among multiple companies and benefit from positive free-rider effects (von Hippel 2012). The growth of open-source software has demonstrated that this collaborative mode of production is ultimately more efficient for many companies.

Collaborative production in finance is difficult to identify. One reason for this might be due to the appropriately of information. In software development, this might amount to the revelation of which underlying technology or method a software

product is using — a revelation which might invite competition but rarely does. In finance, appropriated information may immediately reveal valuable insider or unique information, the competitive advantage of which will disappear as quickly as markets can price correct. This qualitative difference might explain the reluctance of financial actors to participate in collaborative production, their respective competitive positioning being much more sensitive to information appropriation.

A robust public option for asset-level data, provided by mandatory company disclosures, will help invalidate monopolistic data provision business models, and provide a common baseline from which co-production might develop. Because asset-level data are the missing link between the science and public research community and the finance industry, the public provision thereof may allow the collaborative production norms of the science community to enter the finance industry, as it has in software development.

4.4 Policy Recommendations

As policy recommendations, for the purposes of enhancing general welfare, improving the efficiency of capital allocation, and reducing monopolisation of information markets, I call for a subversive model of mandatory company disclosure. I recommend the development and implementation of regulations for the mandatory disclosure of the underlying physical assets of a company. This asset-level data should entail:

- The geospatial arrangement of physical plant and equipment assets at their natural spatial agglomeration;
- Nameplate information of the asset, such as facility name, address, production capacity, etc.;
- A measure of materiality of the asset, such as capital valuation, operating revenue or profit, annual production, worker headcount, etc.

In many industries, there is a natural and obvious agglomeration of physical assets that underly the value of the company, typically a production facility at a

single address. Within a single facility there may be many assets, with intermediary, shared (e.g. joint-venture), or third-party owners (e.g. leased equipment). These assets will be similarly exposed to both physical and transition risk. The location and materiality of assets can be spatially agglomerated in a sensible fashion. As examples, an agribusiness may spatially aggregate its operations with areal polygons for different land plots; a manufacturing facility may use a single latitude and longitude point as a sensible aggregation of a single-address facility; a shipping route might provide a linestring describing the typical path of the vessel.

Mandatory asset-level data disclosure, as a generalised model, will achieve the following, with benefits for social welfare:

- Provide general-purpose data with which company shareholder and stakeholders can assess their exposure to environmental risks using sophisticated tools beyond the purview of company managers.
- Avert a competition between various company stakeholders for bespoke company disclosures from time-constrained company managers. This will also better delineate the role of risk analysis and mitigation at the shareholder and stakeholder level versus at the company management level.
- Reduce information asymmetry between company managers and non-material stakeholders, empowering the latter, potentially to the extent that their welfare becomes material to the company.
- De-monopolise markets for financial data where innovation has stagnated due to the rent-seeking of data monopolies.

In developing these data science projects, I have found that technology does play a role in helping non-material stakeholders achieve access to some data that would otherwise be beyond their means. The NGOs that build these platforms, however, struggle to maintain open data programs when government and philanthropic support favours new research. I further recommend that policy initiatives support the maintenance and upkeep of public datasets and platforms. Often grant or

philanthropic funding is attached to a specific study or outcome and funds cannot be spared for what amounts to deferred maintenance on datasets and platforms.

4.5 Opportunities for Further Research

This DPhil has laid the groundwork for a broad portfolio of impactful research. Some future directions are laid out here.

Surveying Asset-Level Data Providers and Disclosers

Social science research to survey the providers of asset-level data across various industries to discover market concentration, and the current level of disclosure by companies. How much do companies already disclose at the asset level?

Asset Localisation using EO and Computer Vision

Continuing the deployment of computer vision models with EO imagery for asset localisation in more industrial sectors.

Populating Asset Attributes with Natural Language Processing

Answer the critical question of what information is already publicly available on the internet, and what bias does it carry?

Experimentation with crowd sourcing information

Answer the critical question of what data can be obtained using crowd sourcing alone, and the quality of that information.

Experimentation with new information markets

What might the design considerations be for a microtransaction market for company information?

Agglomeration of Asset-Level Data

Bring available asset-level data together in one place where it can be accessible to multiple stakeholders

Surveying Norms, Culture, and Collaborative Production in Finance vs. Software

Isolate the differences in collaborative production better finance and software. How much is due to skills and norms vs. qualitative differences in the information products produced.

Financial Data Clubs

What might the design considerations be for a 'financial data club'? How might a financial data club overcome appropriability concerns in order to build better collaborative information products?

5
Conclusion

This DPhil developed novel methods for democratising access to company information crucial for mitigating environmental risks and impacts. Three papers were presented. The first paper documents the development of a new global inventory of PV solar energy facilities, crucial asset-level data for assessing energy company transition risk and environmental impacts, particularly those related to land use. The second paper uses asset-level data to demonstrate how transition risk emerges from the topology of a supply chain network, assembling an implementation of the global fossil fuel economy using only open-source data. The third paper develops general-purpose earth observation computer vision methods to assess how transferable these methods are to new domains and new problems in company- and asset-level environmental risk assessment.

Based on these contributions and the experiential evidence of developing these data science projects, a discussion is offered concerning the opportunities and risks of data-driven technologies in the provision of environmental risk information. While these technologies demonstrate opportunities for empowering public and private governance of company activities, they are not without limits, biases, and potential allocative harms. A policy recommendation is offered, calling for a subversive model of mandatory disclosure of general-purpose asset-level information. Such a policy would have the effect reducing the complexity of environmental

disclosures being asked of company managers; better delineate the risk management responsibilities of company managers and investors; erode monopolistic rent-seeking on environment-related financial data by market intermediaries; and secure more equitable information asymmetries between companies, their shareholders, and their material- and non-material stakeholders.

A future research programme is proposed which continues the material work of democratising company information with data-driven technologies. This programme would study, and participate in, emerging forms of institutions in the 'new commons'. It asks whether self-organising, collaborative production of financial information, based on a frictionless abundance of public goods, might prove more efficient in addressing the critical social and environmental challenges of the 21st century than today's keep-and-rampart stewards of capital. And further, what new expressions of solidarity, generosity, and choice might come as a result.

Appendices

A

A Global Solar Photovoltaic Inventory.

Kruitwagen, L.,[1,*] Story, K.,[2] Friedrich, J.,[3] Byers, L.,[3] Skillman, S.,[2] & Hepburn, C.[1]

[1] University of Oxford, Smith School of Enterprise and the Environment
[2] Descartes Labs Inc.
[3] World Resources Institute
* Corresponding Author

Contents

A.1 Motivation 70
A.2 Machine Learning Pipeline 72
A.3 A Global Census 75
A.4 Land Cover Analysis 79
A.5 Conclusion 83
A.6 Supplemental Methods 84
A.6.1 Machine Learning Pipeline Description 84
A.6.2 Dataset Completion 93
A.6.3 Machine Learning Pipeline Performance 94
A.6.4 Machine Learning Model Interrogation 96
A.6.5 Feature Enrichment and Analysis 98

Abstract

To mitigate climate change and alleviate global poverty, energy system projections show a nearly ten-fold increase in photovoltaic (PV) solar energy generating capacity by 2040 (International Energy Agency 2018, International Renewable Energy Agency 2019). Geospatial data describing the energy system is required to manage generation intermittency, mitigate climate change risks, and identify adverse trade-offs with biodiversity, conservation, and land protection priorities caused by land use and land cover change. Here, we provide a global inventory of commercial-, industrial-, and utility-scale PV solar energy generation stations (i.e. PV generating stations in excess of 10kW nameplate capacity) using a longitudinal corpus of remote sensing imagery, machine learning, and a large cloud computation infrastructure. We search 550TB of imagery to locate and verify 68,661 facilities with a cumulative estimated power output of approximately 379 GW, an increase of 253% on the previously best-available asset-level data. We enrich our dataset with estimates of facility installation date, historic land cover classification, and proximity to protected areas and indigenous and community lands. We find that the majority of the PV solar energy facilities we identify are sited on cropland, followed by aridlands and grassland. Further analysis shows that PV solar energy development is biased towards towards agricultural areas and developers disfavour forest and shrubland for development. As international climate policy efforts turn towards assessing progress and debating monitoring mechanisms, this analysis demonstrates the potential for remote sensing to provide climate-relevant data that is unaffected by biases in reporting from companies and countries. Fundamental datasets such as this inventory are crucial for the operation of increasingly renewables-based electricity systems, and for the design, implementation, and validation of effective policy regimes and markets to deliver the diffusion of PV solar energy generation aligned with the UN Sustainable Development Goals. Our enriched data helps decision makers navigate trade-offs between a 1.5°C constraint to anthropogenic climate change and other biodiversity, conservation, and land protection goals. We make our dataset publicly available to facilitate more research of this type, and to jump-start the availability of a globally exhaustive, feature-rich, and publicly available PV solar energy asset-level data.

A.1 Motivation

Since 2009, installed photovoltaic (PV) solar energy generating capacity has grown at an average rate of 41% per year while costs have fallen by 12% per year over the same period (BP plc 2018, BloombergNEF 2018). The rapid deployment of solar energy must continue if climate and Sustainable Development Goals are to be met, particularly in the global south where power systems must simultaneously grow and transition to clean energy. In the International Energy Agency's (IEA) Sustainable Development Scenario, 4240GW of PV solar generating capacity is projected to be deployed by 2040 (International Energy Agency 2018), a ten-thousand-fold increase from 385 MW in 2000 (BP plc 2018). In some countries, power generating companies and their investors already face risks and opportunities as PV solar energy reduces residual electricity demand with embedded generation and cannibalises market prices with zero-marginal-costs. Grid operators must now balance uncertain electricity demand with intermittent supply. Policy makers need to properly incentivise the adoption of clean energy in line with economic development and climate goals, while being sensitive to the impacts renewable energy facilities might have on biodiversity, ecosystem services and vulnerable lands.

Currently-available inventories of solar generating capacity are insufficient to address these needs because they are either aggregated summary statistics (e.g. of the IEA (2018), IRENA, (2019) or BP (2018)); limited in geographical scope (e.g. the World Resources Institute's Global Power Plant Database (GPPD) (2018), OpenPV US database (2019), Yu et al.'s DeepSolar (2018), Huo et al.'s. SolarNet (2020)); are not geospatially localised (e.g. S&P Global World Electric Power Plant Database (WEPP) (2018)); and/or are not publicly available for the research and policy community (e.g. IHS' Electric Plants (2020)). The generating behaviour of PV solar energy (and wind energy) reflects the uncertainty and complexity of the natural world; detailed asset-level data, including the spatial arrangement of installations, are required particularly to address the challenges of generation now-casting and forecasting faced by electricity system operators and electricity markets operators and participants.

In this work, we produce such a dataset by mapping commercial-, industrial-, and utility-scale PV solar energy facility footprints in remote sensing imagery using deep learning. We adapt our definitions of PV installation sizes from the IEA (2019*b*), where 'commercial and industrial' installations are defined as those between 10kW and 1MW in nameplate capacity, 'utility-scale' as installations in excess of 1MW. For the purposes of our study, we also include 'off-grid' installations (as defined by the IEA) in excess of 10kW nameplate capacity. Bolinger et al. (2019) categorise 'utility-scale' installation as those in excess of 5MW capacity. We adapt these two categorisations and use bins of 10kW to 1MW; 1MW to 5MW; and >5MW nameplate generating capacity as categories in our analysis and refer to our scope collectively as 'non-residential PV'. Our study is therefore comparable to 89% of the capacity reported by the IEA (2019*b*) and includes both the conventional generating stations and 'behind-the-meter' installations that electricity system operators might otherwise have poor visibility of. We prepare a machine learning pipeline comprising of three convolutional neural networks, two recurrent neural networks, and a number of heuristical filters and deploy it on a cloud computation infrastructure. The resulting dataset expands the best publicly-available facility-level data for PV solar energy by 420%, including 18,449 new installations in China, 9,906 in Japan, 4,525 in the United States, 2,021 in India, and 17,918 in the European Economic Area. Our method is agnostic to political borders, and so provides new insight into the diffusion of PV solar energy in reporting-poor areas.

We enrich our data with estimates of generating capacity, installation date, historic land cover, and proximity to protected areas and indigenous and community lands. Our observed installation dates can be used to generate insight into the gap between facility-level final investment decisions, construction start, construction completion, and facility operation. Our land cover analysis provides insight into global trends for PV siting decisions. Sampling from a global landcover map, we observe that non-residential PV is most commonly installed on croplands, followed by deserts and grasslands. We compare PV solar energy land cover to local and national land cover distributions to observe the bias in regional and local PV siting

decisions. These findings highlight the need for further coherence on policy efforts seeking to navigate trade-offs between climate change mitigation, biodiversity and ecosystem health, conservation and land protection, and food production.

A.2 Machine Learning Pipeline

Machine learning combined with earth observation is the only method capable of producing a dataset like ours on a global scale. We use earth observation to overcome the barriers imposed by the limited publicly reported data and use machine learning to scale our inference beyond what teams of human analysts might be capable of. The maturation of computer vision using convolutional neural networks (CNNs) fukushima:80,lecun:89,krizhevsky:2012 has unlocked new approaches for localising real economy objects in remote sensing imagery. Object detection methods using CNNs have delivered convincing results for several human-made object classes, including roads zhang:18, buildings ishii:16, automobiles audebert:18, aircraft zuo:18, and ships bentes:18. CNNs have also been used with remote sensing imagery to determine coal-fired power station utilisation rates cti:18, crop yield prediction wang:18, and wind power forecasting wang:17.

Similar methods have shown promise in localising PV solar energy installations in multispectral remote sensing imagery. Imamoglu et al. (2017) demonstrate the ability to localise PV solar energy in Landsat 8 imagery by adding a class activation map to the shallow CNN architecture of Ishii et al. (2016). Malof et al. (2016) and Camilo et al. (2018) develop models that use hand-crafted features and later the deep CNN SegNet (2015) for semantic segmentation of aerial imagery containing PV solar energy installations. Yu et al. (2018) fine-tune the deep CNN VGG-16 (2015*b*) and add a class activation map to segment solar PV in aerial and very-high resolution satellite imagery, mapping the continental United States. Hou et al. (2020) use a UNet model with attention, similar to our primary inference models, to map China. These studies are limited in geographic extent and have not produced any publicly-available census for further study.

We prepare a machine learning pipeline that uses CNNs and recurrent neural networks (RNNs) to detect solar PV installations in two globally-available satellite imagery sources: high resolution (1.5m pixel resolution) 4-band SPOT-6/7 imagery and medium resolution (10m pixel resolution) 12-band Sentinel-2 imagery. The wide range of spectral bands captured by Sentinel-2 (and to a lesser degree by SPOT) give our algorithms sensitivity to the spectral signature of PV panels, which are designed to absorb light across a wide range of optical wavelengths including those beyond the human-visible spectrum. The combination of Sentinel-2 and SPOT6/7 makes use of their respective advantages - the increased spatial resolution of latter complimenting the spectral resolution of the former. We develop a semantic segmentation pipeline to estimate facility collecting areas, an important variable for estimating facility generating capacity.

We first train two custom semantic segmentation CNN models with U-Net architectures (Ronneberger et al. 2015*b*, Zhang et al. 2018), one each for imagery source, to predict which image pixels contain solar PV, see Figure C.2. The models are trained with sample data obtained from crowd-sourced OpenStreetMaps (OSM). OSM was queried for geometries with metadata tags for power generating stations sourced by solar energy and was filtered to remove small area polygons corresponding to rooftop installations. The resulting dataset of 36,882 polygons was global, with substantial concentrations in Germany (66%), the United States (22%), China (4%), the United Kingdom (3%), and Japan (2%). We interrogate the quality of the OSM data and find that the annotated geometries are reliable but that they capture different concepts of PV solar energy land cover (i.e differing according to whether just the PV modules, the bounding area, or the entire site is annotated). Our training data is therefore 'noisy' in that it includes a portion of pixels which are incorrectly labeled due to the differences in annotating convention. See Supplemental Methods for further details.

False positive detections from Sentinel-2 are eliminated using RNNs to classify timeseries derived from longitudinal inference of the backcatalogue of imagery. Each generating unit's installation date is also estimated from the backcatalogue inferences.

False positive detections from SPOT6/7 are eliminated using a fine-tuned Imagenet ResNet50 (2015) network with a dense final layer to output binary classification. The remaining Sentinel-2 and SPOT6/7 candidates are merged and hand-verified. Figure C.2 shows inference examples from the primary computer vision models.

We assess the performance of our machine learning pipeline using the performance metrics of precision, recall, and Jaccard index (i.e. and henceforth, intersection-over-union, IoU). We develop a validation set of hand-labelled polygons seeded from the locations of PV solar energy installations in the WRI GPPD for the purposes of pipeline tuning. The validation set was used to tune the hyperparameters of the CNNs and RNNs, as well as the vectorisation of the SPOT6/7 detections. We are unable to develop a test set because the true distribution of PV solar energy installations is unknown. As a proxy for test recall, we instead measure aggregates of our dataset against known aggregates for countries around the world - see Table A.1 and Figure A.15. Figure A.15 shows that for many countries, our estimated aggregate generating capacity approximates the aggregate generating capacity reporting by IRENA (2019). Adjusting our aggregate generating capacity estimates by our validation recall, many of our country-level aggregates even exceed the totals reported by IRENA. Further work is required to reduce uncertainty and bias in bottom-up estimates like ours, and clarify whether the statistics reported by IRENA might be under-representing the true deployment of PV solar energy. It is our hope that our dataset can be used as a minimally-biased test set for similar future studies.

Our machine learning pipeline was designed to first maximise the recall of the semantic segmentation convolutional neural networks for the SPOT and S2 branches respectively, and then introduce additional information and filters to reduce false positives and increase precision. Separating the Sentinel-2 and SPOT6/7 branches, while less 'end-to-end', was necessary to respect the use and licensing conditions of the SPOT6/7 imagery and was driven by the differences between the two imagery corpuses, with a distribution of acquisition dates and qualities in SPOT6/7 imagery, and the full back-catalogue of Sentinel-2 imagery available. This also had the upside of facilitating model interrogation, simplifying the injection of heuristical

filters and features into the pipeline, and simplifying back-catalog inference. For installations over 10,000m^2 (approximately 600kW), we achieve precision of 98.5% relative to our validation set, with a modest trade-off in recall which drops to 76%, see Figure A.9. The intersection-over-union of our final dataset is 66% – sufficient for the wide range of use cases that our global dataset enables. The precision of our pipeline shows covariate shift across geographies, illustrated in Figure A.10. In many geographies the precision of the pipeline at deployment underperforms validation performance suggesting the pipeline as a whole might be overfit to the validation set. Because the true distribution of PV solar energy across the planet is unknown, an unbiased test set for verifying final pipeline performance is unavailable. We instead produce a comparison of country-level aggregates of our data to IRENAInternational Renewable Energy Agency (2019) country-level data to develop insight into the collective exhaustion of our data, see Figure A.15. See Figure A.9 and Table A.1 for more information regarding pipeline performance.

To develop further intuition about the performance of our primary semantic segmentation models, we perform band perturbation, band dropout, and feature activation experiments on our trained models. We find that our primary inference models learn a sophisticated topology of spectral and spatial features. We show that non-visible bands contain important information for the determination of the presence of a PV installation, see Figure A.11. This suggests that a trained machine might achieve super-human performance at this task. These results and further details on the machine learning pipeline are available in the Supplementary Methods and should help the future design of computer vision models for build-environment object detection.

A.3 A Global Census

We deployed our machine learning pipeline on the global corpus of Sentinel-2 and SPOT6/7 imagery using Descartes Labs cloud computation infrastructure. Masking with a dilated human population density map, we searched 72.1mn km^2 or 48.4% of the earth's land surface area. We assumed that PV solar energy installations must be

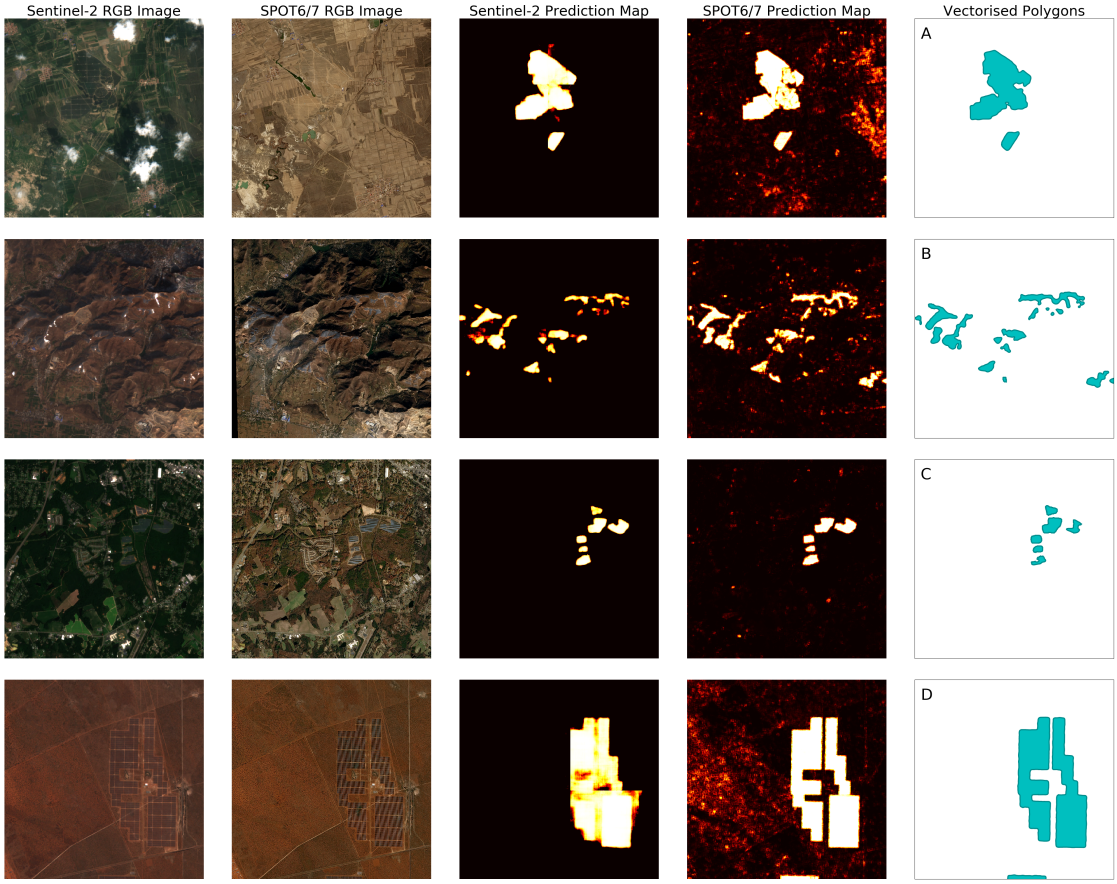


Figure A.1: We detect solar PV facilities with machine learning and remote sensing. Here, we show out-of-sample examples showing SPOT6/7 and Sentinel-2 optical imagery, corresponding U-Net prediction maps, and vectorised output. A) a "panda" array in China (lat/lon: 39.98,113.48); B) covered hills in China (34.34,113.38); C) a new array detected in Sentinel-2 but missing from SPOT6/7 in the US (35.46,-79.18); and D) a 120MW array in South Africa (-27.58,22.93).

located with reasonable proximity to human populations, and used a validation set of spatially-localised installations available from WRI's GPPD (2018) to determine that 99.9% of installations are located within 7km of areas with human population density greater than zero. The date range for our study was 2016-06-01 through 2018-09-30 and we assume that no solar PV installations have been decommissioned or relocated. Deployment processed 550 TB of imagery, used in excess of 1mn CPU-hours and 20,000 GPU-hours, and took approximately 2 months in real time.

We add properties to the polygon detections to produce a feature-rich dataset. Nominal alternating-current (AC) generating capacity of installations are estimated using simple assumptions about installation tilt angle and ground coverage ratio.

We obtain facility installation dates using the backcatalog of Sentinel-2 detections. Historical land cover classification data is added for each installation by querying a global land cover product. Political administration codes (country-level and state/province-level) are added for faster aggregation and filtering. We also add proximity to protected areas and indigenous and community lands within 10km. Where feasible, we match installations to unique identifiers from other publicly available datasets. The resulting dataset contains 68,661 detections located in 131 countries, 30% of which are complete with installation dates, 43% are matched to existing known installations in WRI's GPPD and the EIA's public data. Our new dataset expands the coverage of publicly available solar PV facility-level data by 38,852 installations, 91 countries, and an estimated 271 GW.

We observe that over the study period, estimated installed capacity of non-residential PV increases by more than 81% to 379 GW, led by increases in China (120%), India (184%), the EU-27+UK (20%), the United States (58%), Japan (119%), South Africa (19%), Thailand (15%), Chile (60%), South Korea (58%), and Turkey (143%). With over half of new installations located in China, the portion of global installed capacity in China grew from 36% to 44%. We detect spatial-temporal hotspots in the deployment of solar PV, for example in Turkey and the Netherlands through 2017, Mexico, Hungary, and Australia through 2018, and ongoing regional deployments in the United States, India, and China. See Figure A.2 for an aggregated arrangement of the dataset, Table A.9 for country-level aggregated statistics, and the Supplemental Methods for a link to an interactive map of our dataset.

We compare aggregate statistics of our data to countries where data on the size distribution of deployed solar PV is well known. Table A.1 shows installed capacity of solar PV at the end of 2018 net of residential-scale installations (<10kW) and compares our dataset against other well-known bottom-up asset-level inventories and top-down aggregate statistics. We show that with simple assumptions for installation type, tilt angle, and literature-obtained distributions for panel efficiency, ground coverage ratio, and inverter loading ratio, we obtain gross installed capacities that are comparable to top-down aggregate statistics, and far exceed the currently

best-available asset-level inventories. Using area-binned recall values from our validation set (see Table A.4), we obtain a best-estimate of 494 GW for global installed non-residential PV generating capacity at the end of 2018. We are able to quantify our uncertainty in ground coverage ratio, inverter loading ratio, and module efficiency, which gives an error of ± 6GW with 95% confidence. We do not quantify a range of other uncertainties so this figure is provided only for illustrative purposes. Nonetheless, this figure approximates the non-residential PV figures published by IRENA (2019) and the IEA (2019*b*). This analysis highlights the relative advantages of both top-down estimates and bottom-up inventories: while the former can serve as a placeholder datum for collectively exhaustive capacity, asset-level data is ultimately required to bring transparency, veracity, and feature-richness to these very important statistics.

Aggregate capacity estimates from our bottom-up survey are impacted by variation in detected geometries, ground coverage ratios, inverter sizing, module efficiency, panel orientation, and the presence of single or dual axis tracking equipment. To reduce the uncertainty in generating capacity estimates, and improve inventory yield relative to top-down statistics, further work is now required to improve coverage in targeted geographies, to add residential (i.e. <10kW) installations, obtain better installation geometries, and develop installation-level estimates of installation type (i.e. tracking technology), module efficiency, ground coverage ratio, panel orientation and inverter sizing. These tasks are aided by remote sensing imagery and machine learning, for which our dataset provides an initial training sample. See Supplemental Methods for discussion of generating capacity uncertainty and further work required to develop global feature-rich asset-level PV solar energy installation data. We make our dataset publicly available and invite the research community to join us in answering these questions.

We match our detections to localised data from the WRI GPPD and EIA using location and estimated generating capacity. We compare our estimates of installation dates to matched data from the EIA. We find, on average, our estimated installation dates (based on the beginning of construction) predates the reported operating

Table A.1: PV solar energy dataset comparison

(GW, $N_{inst.}$)	Aggregate		Asset-Level Inventories		
Country	IRENA, IEA[A]	IHS[B]	WEPP[C]	GPPD[D]	Ours (% improvement)[E]
Global	422.4 (n/a)	90.8 (4,004)	107.4 (12,915)	54.3 (5,289)	379.4 (68,661, +253%)
China	173.5 (n/a)	28.9 (436)	13.9 (411)	0.0 (0)	167.7 (18,449, +333%)
Japan	46.5 (n/a)	5.7 (411)	6.0 (905)	2.0 (131)	16.7 (10,504, +178%)
United States	40.2 (n/a)	14.0 (468)	27.1 (3,522)	21.4 (1,790)	53.7 (7,639, +98%)
Germany	38.7 (n/a)	3.7 (380)	6.0 (1,388)	3.9 (408)	15.7 (4,702, +161%)
Italy	15.9 (n/a)	1.4 (179)	2.8 (1,079)	0.4 (74)	11.7 (5,796, +318%)

A: Solar PV installations >10kW at end 2018.(International Renewable Energy Agency 2019, International Energy Agency 2019*b*)
B: (IHSMarkit 2020) C: Latest release in March 2018, includes operating plants and planned or in-construction plants for 2018. (Platts 2018)
D: (Byers et al. 2018) E: Assumes mean ground coverage ratio, inverter loading ratio, and panel efficiency, respectively 51%, 116%, and 14.6%. Percent improvement relative to generating capacity of most complete dataset.

date of the EIA by approximately 2 months, matching our intuition and providing insight into the time delay between final investment decisions, construction, and facility operation. We conclude that our dataset provides an initial global census of commercial-, industrial-, and utility-scale solar PV installations, and can be used as a starting point for a more exhaustive, feature-rich inventory of global solar PV. See Supplementary Methods for further details.

A.4 Land Cover Analysis

With our dataset of installation geometries we are able to generate novel insight into global land cover patterns of PV solar energy sites. Land use for renewable energy is an urgent area of study, as the land chosen for the deployment of renewable energy must navigate impacts on and tradeoffs between the costs of renewable energy transitions (Wu et al. 2019, Konadu et al. 2015); greenhouse gas emissions due to land cover and land use change;(Turconi et al. 2013) ecosystem health and biodiversity (Hernandez, Easter, Murphy-Mariscal, Maestre, Tavassoli, Allen, Barrows, Belnap, Ochoa-Hueso, Ravi & Allen 2014); water resources and food

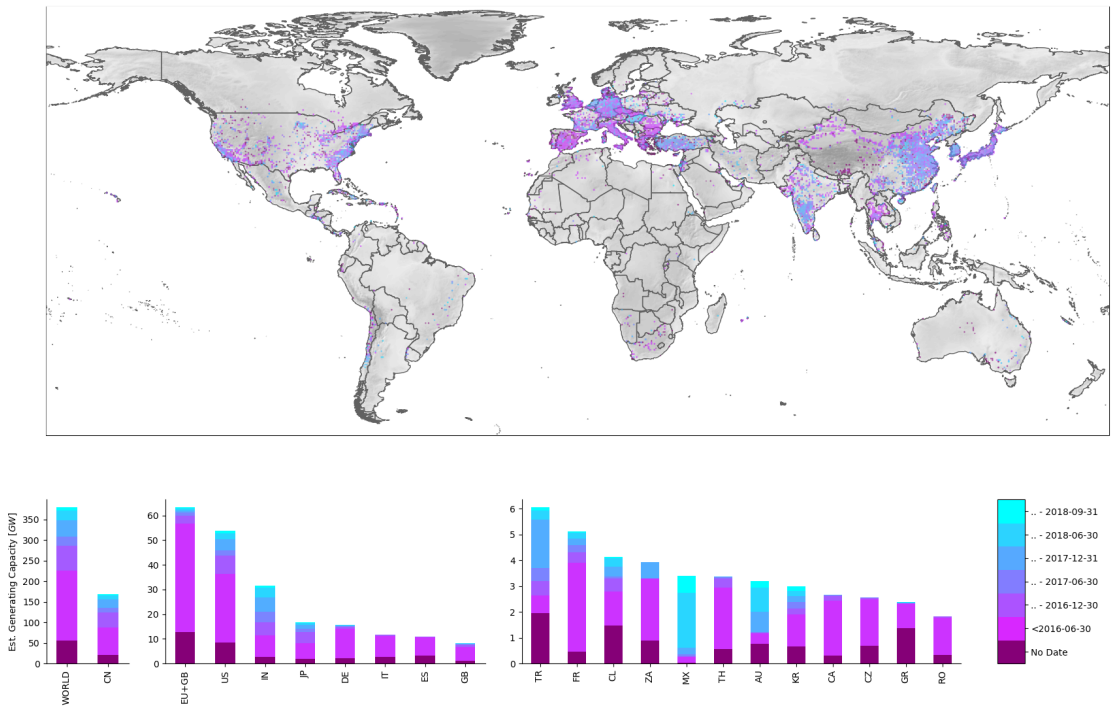


Figure A.2: Aggregated arrangement of the global dataset, with generating capacity by installation date for major countries. Note the last period is only three months.

production;(Bukhary et al. 2018, Dias et al. 2019) indigenous and community land use (Grodsky & Hernandez 2020); land and property values (Carlisle et al. 2015); and political acceptability (Mulvaney 2017). These trade-offs are captured by the multi-criteria challenge of the SDGs themselves: while renewable energy development might be aligned with SDGs 7 (clean energy), 8 (economic growth), 9 (infrastructure), and 13 (climate action), it might have detrimental affects on SDGs 2 (zero hunger) by displacing croplands, 3 (good health) by impairing ecosystem health benefits, 10 (reduced inequality) by displacing community land use, and 15 (life on land) by impacting biodiversity. We develop insight into impact that the diffusion of PV solar energy has on land use by sampling land cover prior to installation. Our dataset brings transparency to PV solar energy land cover trends at the global scale, and can help policy makers navigate trade-offs in policy objectives at the multilateral, national, and subnational level.

We obtain pre-existing land cover for all installations in our dataset back to 2006, sampling from the European Space Agency Climate Change Initiative 300m land

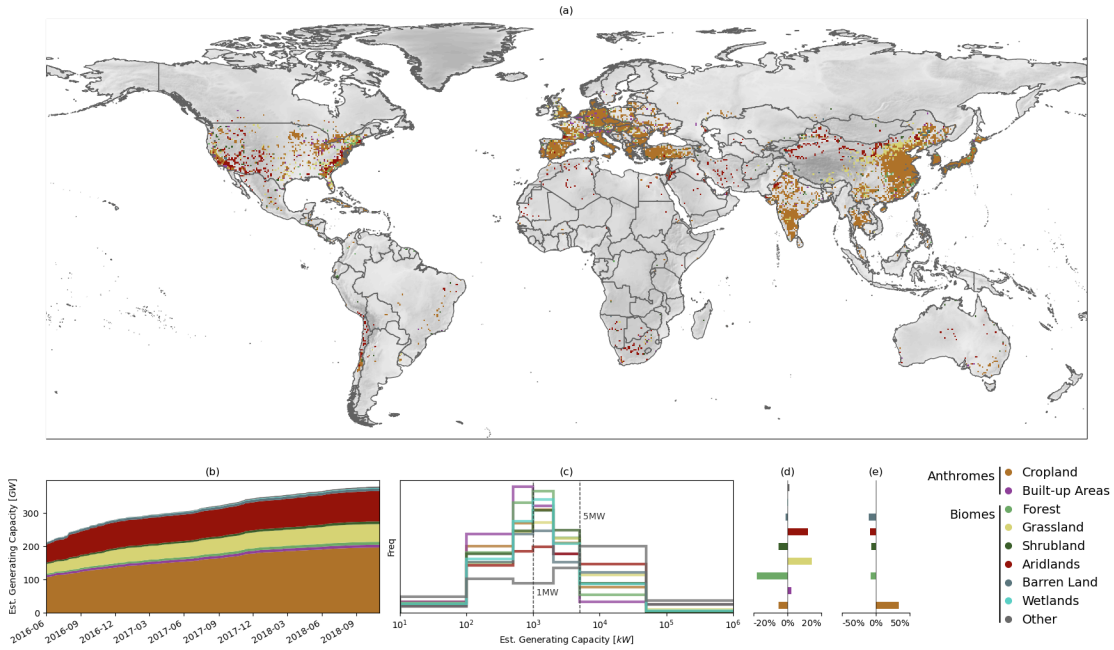


Figure A.3: Pre-existing land cover for new solar PV installations, showing the time series of installations (b); the distribution of installation sizes by land cover (c); local bias (d) between PV land cover and local land covers, where the local local land covers are those lying within the same 0.5°x0.5° grid pixel; and global bias (e) between the global land cover distribution and the distribution of all PV-containing pixels. A positive bias indicates PV is preferentially installed on this land cover type; a negative bias indicates this land cover type is avoided. Country-level analysis is available in Figure A.13.

cover data (Lamarche et al. 2017). We reduce the land cover classfication system to 8 classes, 2 anthrome classes: cropland and built-up areas; and 6 biome classes: forests, grassland, shrubland, barren land, wetlands, and other, see Table A.8. To capture the impact of PV solar energy sitings on vulnerable desert ecosystems, we also add a biome class for aridlands. Following the definition of the World Desertification Atlas, (Cherlet et al. 2018) aridlands are defined as a natural biome with an average aridity index less than 20%, which we calculate using ERA5 reanalysis data. (Hersbach et al. 2020) Figure A.3 shows how land cover used for solar PV deployment has changed over the study period, Figure A.13 shows these trends for the top 20 countries in our dataset, and Figure A.14 shows these trends for our three nominal generating capacity bins.

We observe that, globally, PV solar energy installations are most often sited on land covers indicating significant anthropogenic land use (i.e. 'anthromes'), namely

croplands. The second two most frequent land covers for PV solar energy sitings indicate biological and ecosystem land use (i.e. 'biomes'): aridlands and grasslands. We compare the global land cover distribution with the distribution of land cover for PV solar energy sites, finding a significant bias towards siting on cropland. By contrast, the distribution of PV sites compared to the local distribution of land cover, where 'local' is defined as a 0.5°x0.5° grid cell, shows a bias towards grassland and desert and away from cropland, forest, and shrubland. This difference suggests that while rural anthromes are the most frequent targets for PV solar energy sitings, this might be driven by proximity to human populations more than a preference for previously-developed land. We posit that the local bias towards deserts and grasslands may be due to the reduced cost of preparing the land for PV solar energy installation. We note that siting decisions in anthromes may still cause land use and land cover changes in biomes. As an example, cropland displaced by PV solar energy may be re-developed on an biome site elsewhere. These trends vary significantly by country, see Figure A.13.

Considering European Economic Area countries, for example, we find the distribution of PV solar energy to be similarly biased towards agricultural areas in Germany, Italy, Spain, the United Kingdom, the Czech Republic, and Romania, however *within* PV-containing localities, Germany and France show a heavy bias towards built-up areas for installation, while the others further enforce the bias towards installation on croplands. Japan and South Korea show almost no bias in the country-level distribution of PV solar energy, but at a local level show considerable bias towards cropland sites. China, the United States, India, Spain, France, South Africa, Mexico, and Chile show significant deployment of PV solar energy on aridlands, uniquely vulnerable ecosystems. Further country-level analysis can be found in the Supplementary Methods and Figure A.13.

These trends highlight the different policy approaches to PV development, developer incentives, and the constraints of geography and existing land cover. As the reduction in the cost of solar PV continues to drive diffusion and adoption, policy makers must carefully consider trade-offs between food supply, ecosystem

and biodiversity impacts, land protection for indigenous and community uses, and climate change mitigation. Our dataset and analysis shows the fundamental changes ongoing in the geography of energy resources and is available to help policy makers navigate these trade-offs.

A.5 Conclusion

Our global survey of non-residential PV solar energy installations, using machine learning and remote sensing, has generated a novel public global database of 68,661 spatially-localised facility footprints with installation dates, land cover assessments, estimated generating capacity, and other metadata. This is the first time, to our knowledge, that machine learning and remote sensing has been used to search the entire planet for a specific type of infrastructure asset. We enhance the utility of our dataset with an analysis of pre-existing land cover for PV installations, and show that PV installations most commonly are sited on croplands, followed by deserts and grasslands. Future work to enhance the dataset should include temporal updates, improved installation geometries, better estimates of installation type, and its combination with residential solar PV datasets. Opportunities for subsequent analysis might include solar PV forecasting and now-casting, further land use and land cover impact studies, future infrastructure planning pathways, socioeconomic spatial-temporal diffusion studies, and policy-targeted counterfactual scenarios. We are encouraged by the success of our pipeline and our use of noisy crowd-sourced training data, and suggest that our method may be applied elsewhere to stimulate the creation of robust public asset-level data. We acknowledge the fundamental public-goods nature of such asset-level data and make dataset publicly available to facilitate future research.

Data Availability

The dataset will be visualised and available for download at the World Resources Institute ResourceWatch: `https://bit.ly/2JSTxVt`, and via the Descartes Labs platform.

Code Availability

A code repository and supporting data to reproduce our work is available here for the purposes of peer review: `https://bit.ly/2VymKLj`. The repository will be published publicly concurrent with publication.

Energy and Emissions Footprint

Our machine learning pipeline was deployed on computation infrastructure powered by net-zero carbon electricity. The majority of the training was powered by net-zero carbon electricity, the Sentinel-2 model was trained on average generation mix electricity in the Eastern United States. We calculate training and deployment consumed approximately 71 MWh but emitted only 14kg of CO_2eq in greenhouse gas emissions. The energy footprint of our study is approximately the same as the energy required to drive an electric vehicle the distance between the earth and moon. The carbon footprint of our study is approximately the same as driving a petrol vehicle from New York City to Philadelphia.

Author Contributions

L.K. designed and implemented the machine learning pipeline, designed and implemented the dataset analysis, and wrote the paper draft. K.S. designed the machine learning pipeline and dataset analysis, implemented the SPOT6/7 branch of the machine learning pipeline, and wrote the paper draft. J.F. and L.B. contributed to the acquisition of training data and analysis of the dataset. S.S. contributed to the deployment of the machine learning pipeline. C.H. contributed to the analysis of the dataset and wrote the paper draft.

A.6 Supplemental Methods

A.6.1 Machine Learning Pipeline Description

Our pipeline for detecting solar PV installations in SPOT and S2 remote sensing imagery and estimating their installation dates uses a combination of deep-

learning models, model-stacking, and human-in-the-loop techniques which are shown schematically in Figure A.4. The pipeline features a shared corpus of training data, a spatial mask to limit the deployment area, and two separate branches, one for each primary sensing platform employed. The development of two separate branches allows us to employ the relative advantages and mitigate the relative downsides of both sensing platforms.

Data Description

Solar PV installations are detected in both Sentinel-2 and SPOT-6/7 multipectral remote sensing imagery. Sentinel-2 is a multispectral earth observation mission of the European Space Agency and European commission. The platform is designed for interoperability with other multispectral remote sensing missions, and provides imagery consisting of 12 bands ranging from ultraviolet, through the visible spectra, to short wave infrared. The maximum ground resolution of Sentinel-2 is 10m for red, green, blue, and near-infrared bands, increasing to 60m for narrow-bandwidth ultra-violet and shortwave infrared bands. Sentinel-2 data were corrected for effects of the atmosphere using the Descartes Labs Surface Reflectance (DLSR) algorithm. DLSR is a spectral processing model that corrects pixel intensity for each spectral band to remove the effects of the Earth's atmosphere for each scene.

The SPOT-6/7 satellites, operated by Airbus, produce a global composite of images that are selected to minimize cloud cover and maintain seasonal consistency. The four-band images (red, green, blue, and near-infrared) are pan-sharpened from 6m to 1.5m, then enhanced to adjust luminosity, contrast, and color balance(Airbus Intelligence" 2013). Image quality and acquisition date is not consistent across the globe. The most recent available image up to December 31, 2018 was used for each location.

Training sets for both the S2 and SPOT solar PV prediction models were developed from OpenStreetMap (OSM) data. The polygon labels are reliable but noisy due to the crowd-sourced nature of OSM data. Drawing a small sample, we identify that 9% annotate the 'total area' as defined by Ong, S. et al. (2013) and

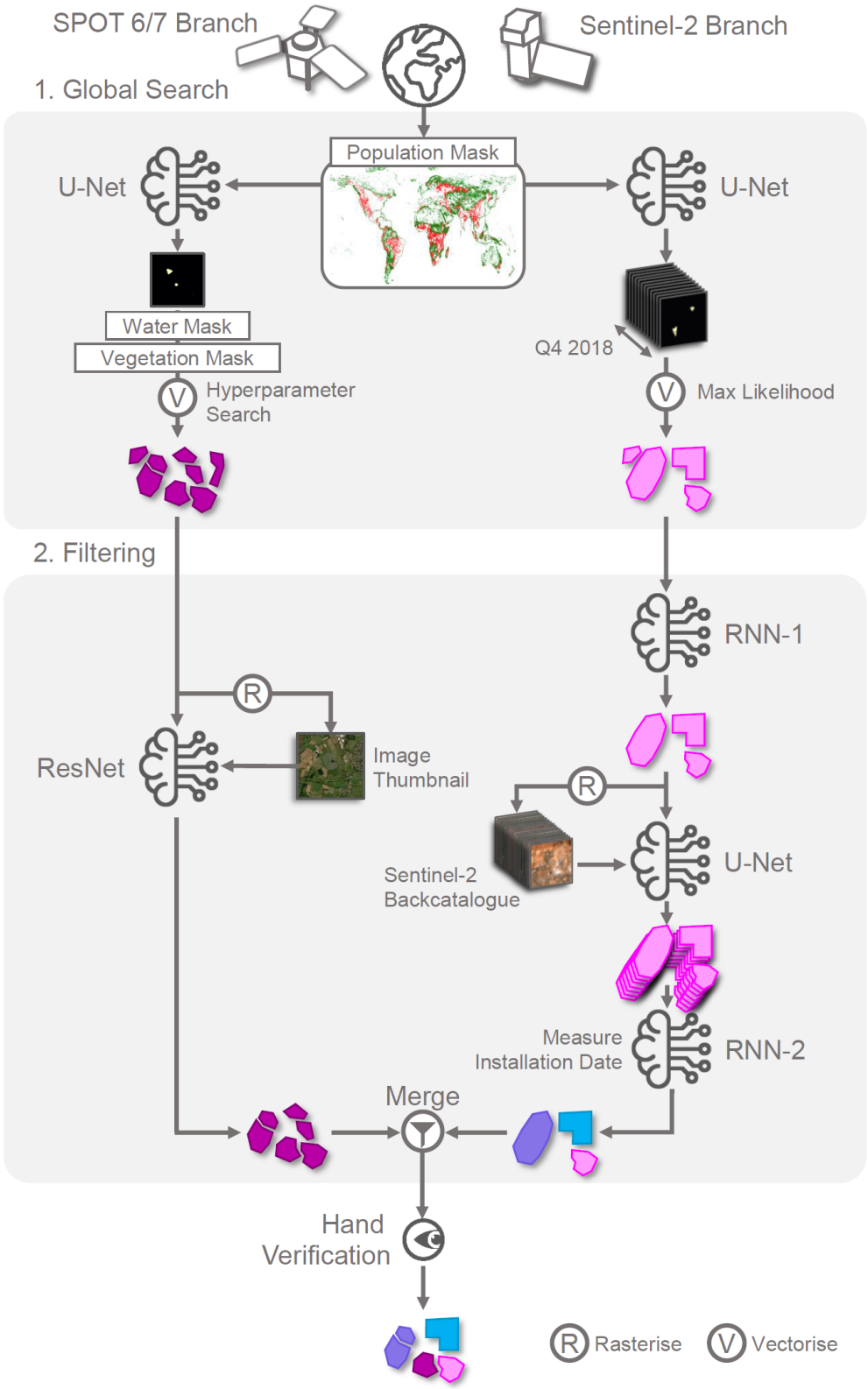


Figure A.4: Machine learning pipeline diagram.

Hernandez, R. R. et al. (2014) (i.e. the full area enclosed by the site boundary), 18% annotate the 'direct area' (i.e. the area covered by the solar arrays, land in between them, and supporting equipment), and 73% annotate the 'array area' (i.e. the area directly under the PV modules). Sets of imagery samples $4km^2(200^2$ pixels) and $0.6km^2$ (512^2 pixels) for S2 and SPOT respectively were generated capturing the complete OSM polygon set. The size distribution of training polygons is shown in Figure A.5. Pixel labels, i.e. single-band class-map images, were rasterised from OSM polygons intersecting tile geometries, converted to the pixel coordinate system. For the S2 training, an additional 11,862 negative samples were randomly generated from points on the earth's surface area not intersecting any existing tiles. A final set of 4219 'bootstrapped' tiles was added to provide hard-negatives and positives based on early model deployments over the UK, China, India, California, and North Carolina.

For the SPOT6/7 training data, a number of hand-labeled polygons were added in China, giving a resulting dataset of 38,541 polygons of positive training examples, with 67% in Europe, 28% in the United States, and 5% in Asia. Training imagery was developed from these polygons and hard-negative samples 'bootstrapped' from early iterations of the model.

A validation dataset was developed from hand-labelled polygons seeded by latitude-longitude locations of solar PV generating stations in the World Resources Institute Global Power Plant Database(2018). The validation set locations were chosen randomly to provide a more geographically-diverse sample. The most represented countries in the validation set include Japan (%28), India (%24), the United States (%23), and the United Kingdom (%10). The size distribution of the 6,281 validation polygons is shown in Figure A.5. Tiles were generated for the hand-labelled polygons. Tile imagery (both up-to-date S2 imagery and the very-high resolutution Google Maps basemap) was then inspected for any additional solar PV installations visible in that tile, which were subsequently labled. We proceed with the assumption that this validation set captures the ground-truth distribution of solar PV generating stations at the beginning of Q4 2018.

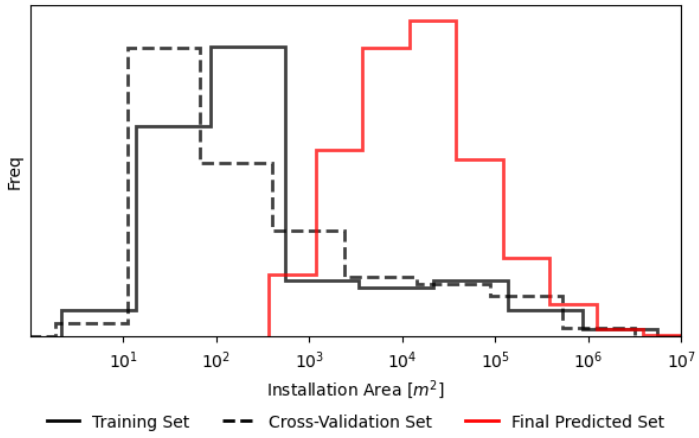


Figure A.5: Distribution of polygon areas in training, validation, and inferred corpuses.

Primary Semantic Segmentation Models

Two semantic segmentation models were trained as primary computer vision models to accept remote sensing imagery as inputs and to output a per-pixel class prediciton map. The U-Net architecture was chosen for its proven performance on similar problems with relatively small training corpuses, few output classes, and fixed-perspective imagery.(Ronneberger et al. 2015*b*) One model each was trained for the Sentinel-2 and SPOT6/7 imagery respectively. The Sentinel-2 model also incorporated the residual innovations of He et al, (2015) adapted to U-Net by Zhang (2017).

The Sentinel-2 Model was trained on a corpus of 22,087 200x200 images at 10m resolution. Images were randomly augmented with flips and rotations. The SPOT6/7 model was trained on a corpus of 9,846 512x512px images at 1.5m resolution. Images were randomly augmented with flips and rotations, 10% additive noise, and 10% multiplicative noise. Both models were trained on NVIDIA Tesla P100 GPUs, on commercial cloud computing platforms. The SPOT6/7 model was trained for 82 epochs, taking approximately 9 hours; the Sentinel-2 model was trained for 150 epochs, taking approximately 28 hours. All machine learning models were trained with binary cross-entropy loss and Adam optimizersKingma & Ba (2014).

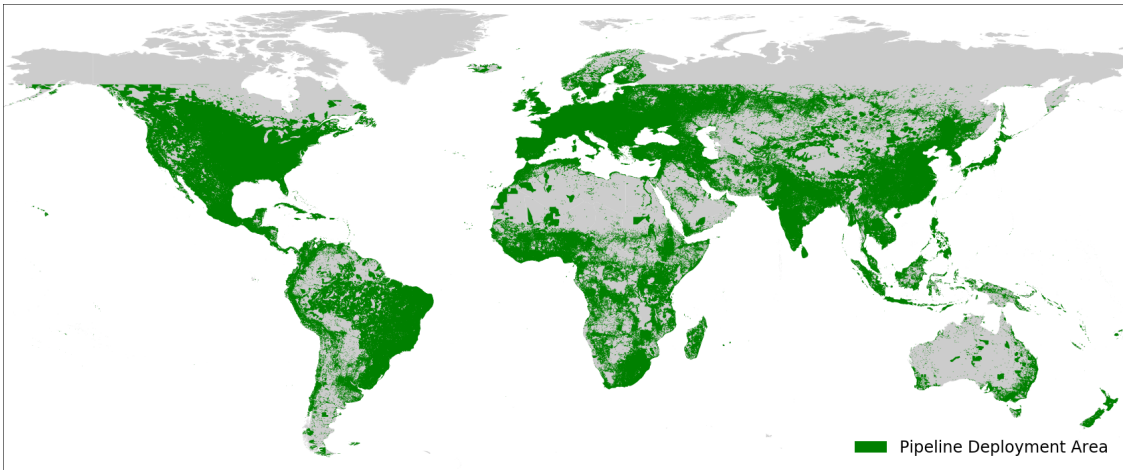


Figure A.6: Deployment mask used to limit search area, with on-scope areas indicated in green. Land area with a population density of greater than 1 person per km^2 was searched, plus a dialation of 7km.

Deployment Area

The computer vision models were deployed to a global corpus of remote sensing imagery. To sensibly narrow the search area, only land area in close proximity to human population was searched. This was accomplished by first making a binary raster mask of where population density exceeds 1 person per km^2 derived from a global high-resolution population density map taken from the EU Joint Research Council GHS Population Grid (GHSL) (2015). This mask was dilated until it included 99.9% of pixels from the validation polygons, a 7km dilation. We remove certain very large but low-density population areas from Australia and Canada, artefacts of the coarse census data for those areas used to construct the original population raster. We also clip the search area to above 60° North for the United States, Canada, and Russia, very large land areas with coarse census data where no PV is expected. Using the Descartes Labs tile server to tile the earth, we search the set of image tiles that completely cover this dilated mask, an area of more than 72.1mn km^2, 48.4% of the land surface on Earth. We do not exclude inland water bodies and the tile generator also extents slightly beyond the coast geometry, so we also expect to include any near-land water-born 'floatovoltaic' installations. Figure A.6 shows this search area.

Sentinel-2 Branch

The Sentinel-2 branch was designed to identify PV solar energy in sentinel-2 imagery with high confidence, and to identify the installation date of each facility using the extensive back-catalogue of imagery. The pipeline includes three key steps using three deep learning models which we now describe; see Figure A.4.

In step 1, the primary segmentation model (U-Net) was run over all S2 scenes in the fourth quarter of 2018 (at most 15 observations for any given location), see "U-Net" in the *Global Search* box of Figure A.4. These multiple looks improve detection performance by normalizing for cloud coverage and atmospheric conditions, among other things. The 3-dimensional prediction stack is reduced to two dimensions by taking the maximum prediction value along the stack for each pixel. This maximum-prediction map is then segmented into separate installation masks using a watershed algorithm. For each separate identified installation, a scalar timeseries is produced by spatially reducing each scene of the 3-dimensional prediction stack using the maximum-prediction installation mask. Two spatial reduction functions are used: the mean of the unmasked pixel values and the count of the unmasked pixels exceeding a threshold value normalised by the total unmasked pixel count of the scene-wise stack maximum. The result is a two-feature timeseries of at most 15 observations associated with each installation detection. The original maximum-prediction installation masks are vectorised into polygons for further processing.

The second step uses an RNN filter model to separate true detected installations from false positives, see "RNN-1" in Figure A.4. An RNN model is required because they are capable of handling variable size inputs, like the timeseries above. RNN-1 first encodes the scalar timeseries using a Long Short-Term Memory (LSTM) neural network. We use an LSTM architecture to take advantage of network's feedback connections, recognising that signals in the timeseries are confounded by the intermittent and cyclic nature of remote sensing data availability and atmospheric conditions. The LSTM embedding is then concatenated with several hand-crafted geometric features, see Figure A.7 and Table A.2 for architecture details.

The geometric features chosen are the detection's minimum rotated bounding-box aspect ratio (i.e. length-to-width ratio), the detection's perimeter-to-area ratio, and the base-10 logarithm of the detection's area in m^2. These features capture, respectively, the overall shape of the installation, the 'roughness' of the installation boundary, and the size of the installation. This heuristical knowledge is introduced to increase pipeline precision. The combined RNN-1 model predicts if candidates are true installations.

The third step uses the entire Sentinel-2 catalogue, from 2016-01-01 through to 2018-12-31, to further eliminate false positives. For each installation found in step 2, we run the primary U-Net model over all available scenes, resulting in a 3-dimensional prediction stack with up to 160 observations; see "U-Net" in the *Filtering* box of Figure A.4. The resulting timeseries of scene-wise relative coverages and means for each installation is fed into a final RNN filter (see "RNN-2" in Figure A.4 Figure A.8, and Table A.3 for architecture details) to determine whether the detection is a true or false positive. This second RNN increases pipeline precision by further mitigating remote sensing intermittency and atmospheric noise and cyclicality. The longer scope of this RNN also allows it to mitigate for seasonality. An RNN is again used for this step to manage the variable size of the input, and an LSTM is chosen for its feedback connections. The purpose of RNN-2 is again to predict if candidates are true installations. RNN-1 and -2 are separated to reduce the excessive computational requirements of running UNet inference on the entire Sentinel-2 backcatalog.

For true detections, we heuristically obtain the installation date from the final timeseries which is the input to RNN-2. We estimate the installation date from the first scene in the timeseries where the relative coverage exceeds 20%.

SPOT6/7 Branch

The SPOT6/7 branch of the pipeline was designed to produce accurate installation geometries, taking advantage of the imagery's higher spatial resolution. The probability raster output from the U-Net model is fed into two watershed algorithms

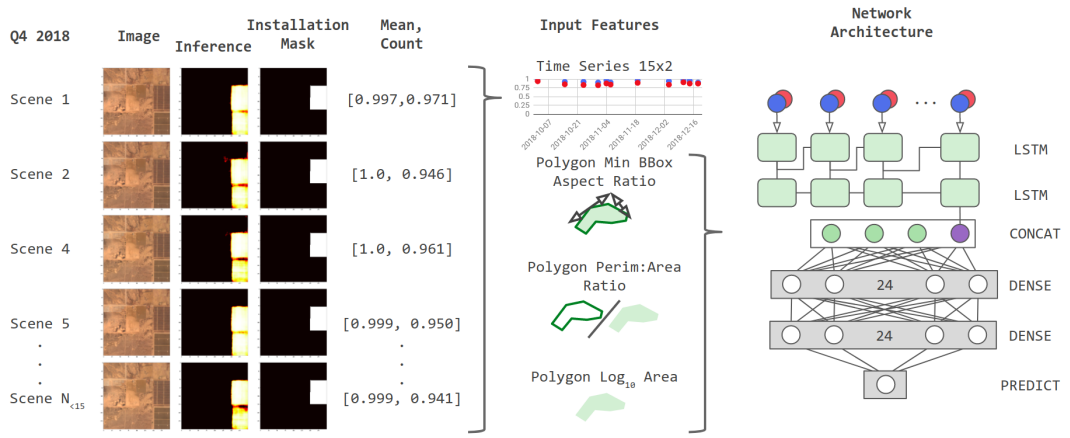


Figure A.7: RNN-1 schematic diagram. RNN-1 filters candidates from the primary U-Net model to remove false-positives.

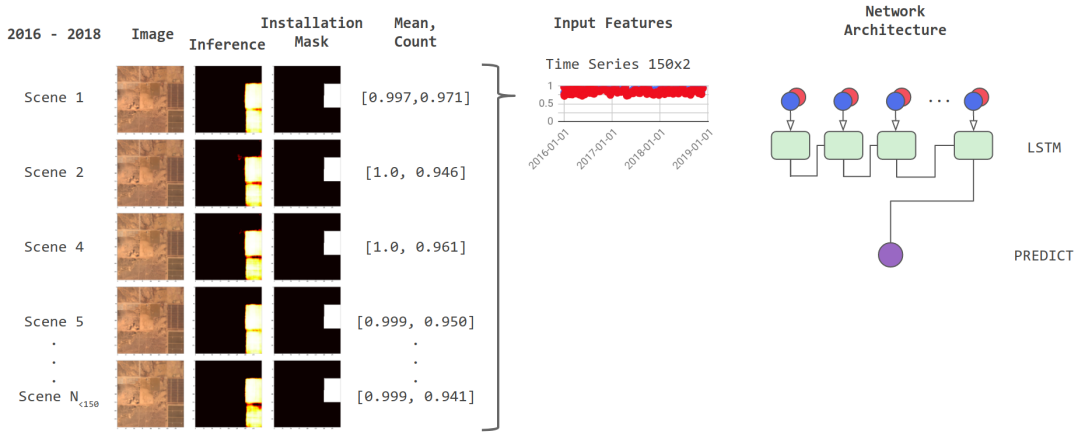


Figure A.8: RNN-2 schematic diagram. RNN-2 is used to further eliminate false positives from candidates that pass RNN-1, now drawing on the full Sentinel-2 corpus.

that eliminate false positives and separate individual installations. The first watershed includes lower- and upper-bound filters, as well as a gaussian filter, whose hyperparamters were chosen using Bayesian search. The Bayesian search identified hyperparameters which optimised the sum of precision and recall when applied against the validation set of polygons. The second watershed algorithm identified spatial extents of individual installations using local maxima seeded from the first algorithm. Hyperparameters for the second watershed are chosen again using Bayesian search optimising for Jaccard Index (IoU) against the validation set. An edge detection algorithm was used to vectorise the watershed output, which was then transformed from the pixel coordinate space to latitude and longitude.

Table A.2: RNN-1 Architecture

Layer	Name	Architecture	Output Shape	Parameters
0	Timeseries_Mean_Precision	INPUT	[: , 15 , 2]	0
1	LSTM_0	LSTM	[: , 15 , 8]	352
2	LSTM_1	LSTM	[: , 8]	544
3	Min_BBox_Aspect_Ratio	INPUT	[: , 1]	0
4	Area_Perimeter_Ratio	INPUT	[: , 1]	0
5	Log10_Area	INPUT	[: , 1]	0
6	Dense_0	DENSE	[: , 24]	288
7	Activation_0	RELU	[: , 24]	0
8	Dropout_0	DROPOUT	[: , 24]	0
9	Dense	DENSE	[: , 24]	600
10	Activation_1	RELU	[: , 24]	0
11	Dropout_1	DROPOUT	[: , 24]	0
12	Dense_2	DENSE	[: , 24]	25
13	Activation_2	SIGMOID	[: , 24]	0
			Total Parameters:	1,809

Table A.3: RNN-2 Architecture

Layer	Name	Architecture	Output Shape	Parameters
0	Timeseries_Mean_Precision	INPUT	[: , 160 , 2]	0
1	LSTM_0	LSTM	[: , 12]	720
2	Dense_0	DENSE	[: , 1]	13
3	Activation_0	SIGMOID	[: , 1]	0
			Total Parameters:	**733**

A single-class classifier with a ResNet-50 (He et al. 2015) architecture was used to further eliminate false positives from the SPOT6/7 detections. The ResNet-50 was fine-tuned from ImageNet weights using sample images drawn from the initial training set.

A.6.2 Dataset Completion

Sentinel-2 and SPOT pipeline branches are combined into a final vector dataset using a rules-based filter. Where Sentinel-2 and SPOT polygons intersect with a

Jaccard index (Intersection-over-union) in excess of 30%, the geometry of the SPOT polygon is retained, inheriting the installation date from the Sentinel-2 detection (confidence level "A"). Detections from only the SPOT and S2 branches are retained with confidence levels "B" and "C" respectively. Where the IoU does not exceed 30%, the union of both geometries are retained, inheriting the S2 installation date (confidence "D"). The distribution of each confidence level is shown in Table A.5.

As a final check of the dataset, the merged detections were hand-verified by the authors. We provide this final step in order to present a clean, high-quality dataset for downstream applications. The authors inspected all the samples in both very-high resolution satellite imagery and up-to-date medium resolution imagery to verify the presence of PV solar energy at the location of the detection. Detection geometries were not changed in final hand verification. Any false positives remaining in the dataset are a product of human error. Final precision statistics in Figure A.9 and Table A.4 we obtained with a secondary round of hand-verification, with 200 samples in each area range. We find human error in hand labelling reduced final precision to approximately 98.5%.

The final dataset includes 68,661 detections in 131 countries with a mean detection area of approximately 70,000m^2. Country-level aggregates are shown in Table A.9.

A.6.3 Machine Learning Pipeline Performance

We design our machine learning pipeline to produce a dataset of installations suitable for a variety of applications spanning finance, policy, planning, and engineering. As such, we optimise the pipeline for precision, as false positives must be expensively hand-filtered. The first steps of both the S2 and SPOT6/7 branches optimise for recall. Then the subsequent filters eliminate false positives to improve precision. Performance benchmarks through the pipeline are shown in Figure A.9. The performance of our pipeline is comparable to other non-residential PV solar energy remote sensing computer vision studies. Yu et al. (2018) achieve a precision of 93.7% and a recall of 90.5% for 'non-residential solar'. Imamoglu et al. (2017) achieve a

Table A.4: Pipeline Validation Performance

Area Range $[m^2]$	Precision [%]						Recall [%]					IoU [%]			
	S2-V1	S2-V4	SPOT-V1	SPOT-V2	PRE-HANDV.*	FINAL**	S2-V1	S2-V4	SPOT-V1	SPOT-V2	FINAL	S2-V1	S2-V4	SPOT-V1	SPOT-V2	FINAL
10^2-10^3	17	80	38	56	100	**98.0**	25	13	18	9	**18**	0	0	0	0	**0**
10^3-10^4	31	100	49	77	90	**98.6**	34	15	37	26	**25**	4	1	7	8	**5**
10^4-10^5	74	91	75	87	94	**99.4**	91	70	65	30	**71**	27	20	21	14	**31**
10^5-10^6	89	96	88	91	93	**97.5**	98	89	69	29	**82**	38	35	28	18	**50**
$>10^6$	94	96	83	75	96	**97.5**	100	88	61	16	**88**	66	62	37	16	**66**

* Precision for validation set prior to hand-verification.
** Precision shown for random spot-checks after hand verification for the entire dataset, not validation.

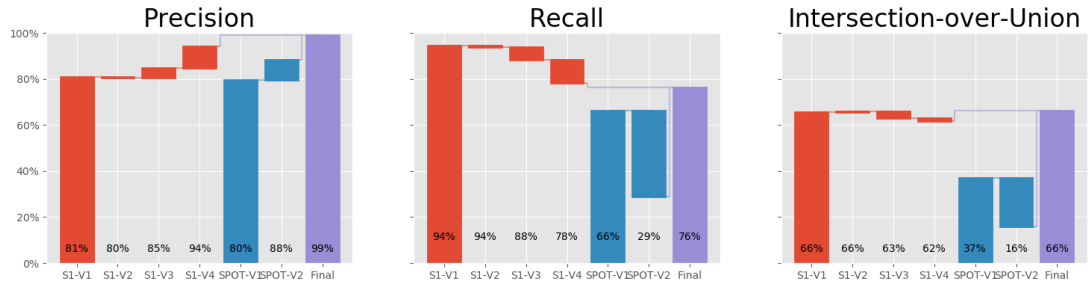


Figure A.9: Machine learning pipeline validation performance for S2, SPOT, and combined pipeline branches for installations over 10,000 m^2, capturing 97% of installation generating capacity. Final precision is shown for the entire dataset, not validation.

recall of 93.3% and an IoU of 56% for PV solar energy installations with generating capacities in excess of 5MW. Hou et al. (2020) report a mean IoU of 94.21% but do not provide other performance criteria nor the size of PV solar energy systems that they detect. The performance of our pipeline is comparable to these studies, and the slight impairment in key metrics might be ascribed to the considerably larger and more diverse geography we trained and deployed our model on.

In general, smaller installations were more difficult to detect, with both false negatives and false positives skewed towards smaller installations in validation. Increasing the precision of detections of smaller installations in the SPOT6/7 branch came with considerable tradeoff in SPOT6/7 recall. The combination of

the SPOT6/7 and S2 branches took advantage of their comparative advantages: SPOT6/7 boosted the pipeline's precision for small installations, while the S2 branch ensured near-complete coverage of large installations.

Covariate shift between our training and validation sets, and our imagery corpus for deployment introduces uncertainty into our pipeline performance. Our training and validation data is biased by geography, the size of detections (see Figure A.5), and the installation date of the samples. The imagery corpuses themselves are biased - both in natural phenomena, e.g. amospheric conditions and land cover, and also measurement. While Sentinel-2 imagery is consistent, the quality and timeliness of SPOT6/7 imagery varies across the planet, with imagery in the US being more recent and of higher quality. We observe the geographic shift in our machine learning pipeline precision by comparing the distribution of our dataset before and after hand-verification (respectively, S2-V4, SPOT-V2 and FINAL in Table A.4), see Figure A.10. Across many geographies, the precision of pipeline at deployment underperforms validation performance, suggesting the pipeline as a whole has been overfit to the validation set.

A.6.4 Machine Learning Model Interrogation

We interrogate our semantic segementation models to study the relative importance of each spectral band and susceptibility to noise, see Figure A.11. We test for impairment of the model‘s IoU performance against the validation set under the conditions of one-vs-all band dropout, and additive and multiplicative noise of 10%, 20% and 30%.

For the S2 ResUNet model, we find that additive noise has a similar affect on model performance when added to any band other than cirrus and the alpha channel. We hypothesise that the model‘s variance sensitivity to the cirrus band has been nullified by the heuristic atmospheric correction algorithm used to preprocess the S2 imagery. Additive noise has a larger affect on model performance than multiplicative noise, both of which are consistent across bands. Additive noise is

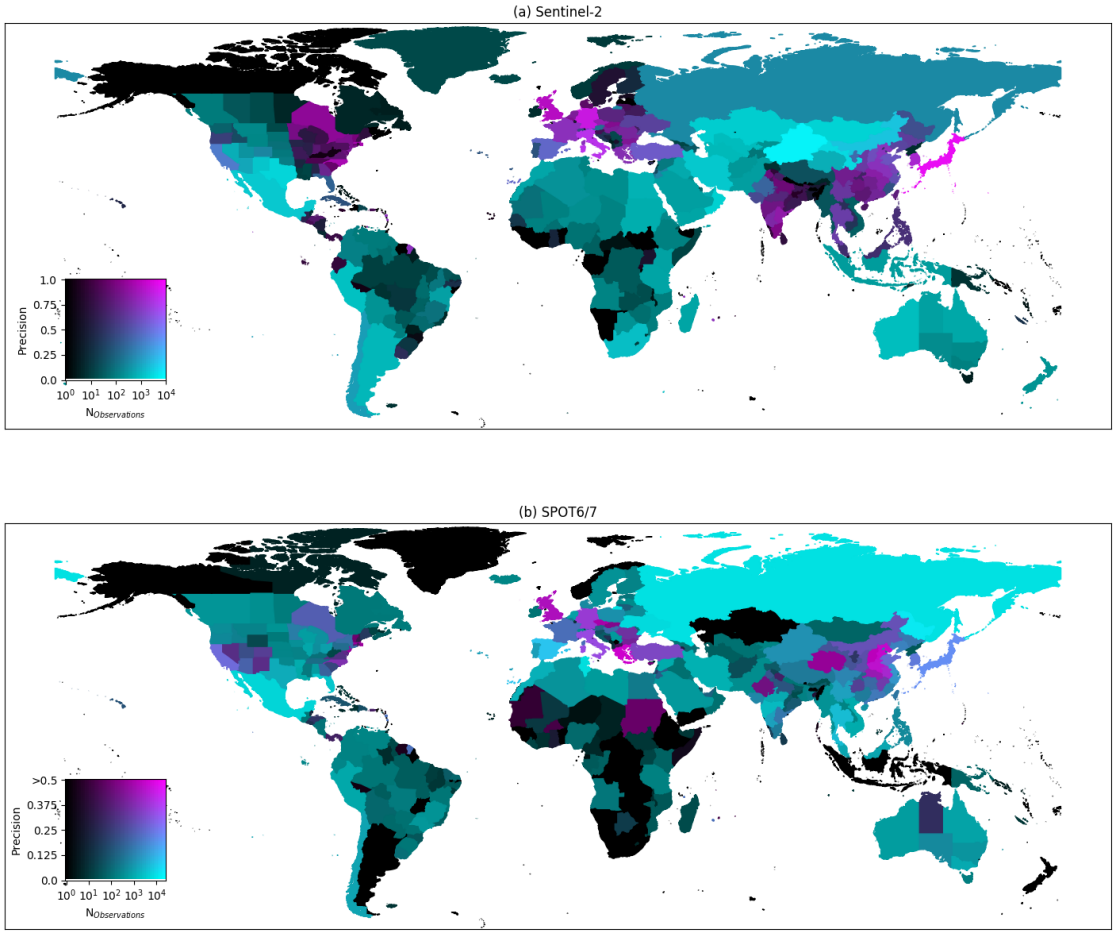


Figure A.10: Pipeline precision for S2 (a) and SPOT (b) branches prior to hand verification. Magenta areas indicate high model precision and many detections; cyan areas indicate low model precision and many detections; dark areas indicate few detections. S2-branch precision was high in geographies with a large amount of PV solar energy not represented in the training set (e.g. CN, IN, JP, IT), but produced excessive false positives in desert and mountainous areas (e.g. MX, ZA, IR, KZ, Texas, Western Australia) and reduced precision in areas where PV solar energy was otherwise abundant (e.g. ES, CL, TR, Xinjiang, California). SPOT precision was driven by confusion with objects and shadows. As examples, confusion with tree farms, crop patterns, chicken coups, and mountain shadows drove underperformance in Tunisia, Mexico, and Russia.

likely larger in absolute magnitude than multiplicative noise due to the skew of the data to the lower end of the sensor‘s operating range.

With band dropout, we find IoU performance of the S2 ResUNet is not discernibly biased to either the spectral or spatial features of the input bands. The coastal-aerosol ultraviolet band at 60m resolution shows a similar importance to model performance as the 10m near-infrared band. Blue and green 10m bands are highly important for model performance, as are narrow bandwidth red-edge bands and broad bandwidth short-wave infrared bands.

For the SPOT6/7 model, we find it is more sensitive to lower wavelength green and blue bands than red and near-infrared bands. This matches our intuition based on the blue-green appearance of PV panels in the visual spectrum. The SPOT6/7 model was not as affected by additive noise, likely due to the noise augmentations used in training.

We qualitatively examine false positive detections in both S2 and SPOT branches. In the SPOT branch, false positives were often generated by objects: greenhouses, row crops, and regularly-spaced trees, as well as mountain shadows and water, suggesting the dominance of spatial and geometric features, and color. In the S2 branch, false positives corresponded to both spatial features (chicken coups, greenhouses) and spectral features (deserts and high aerosol areas, and green sports fields and tennis courts). We conclude that the both computer vision models makes use of the whole spectrum of available bands, and robustly learn feature representations across a sophisticated spectral and spatial topography.

A.6.5 Feature Enrichment and Analysis

We enhance our dataset with a number of metadata properties for each detection, see descriptions in Table A.7.

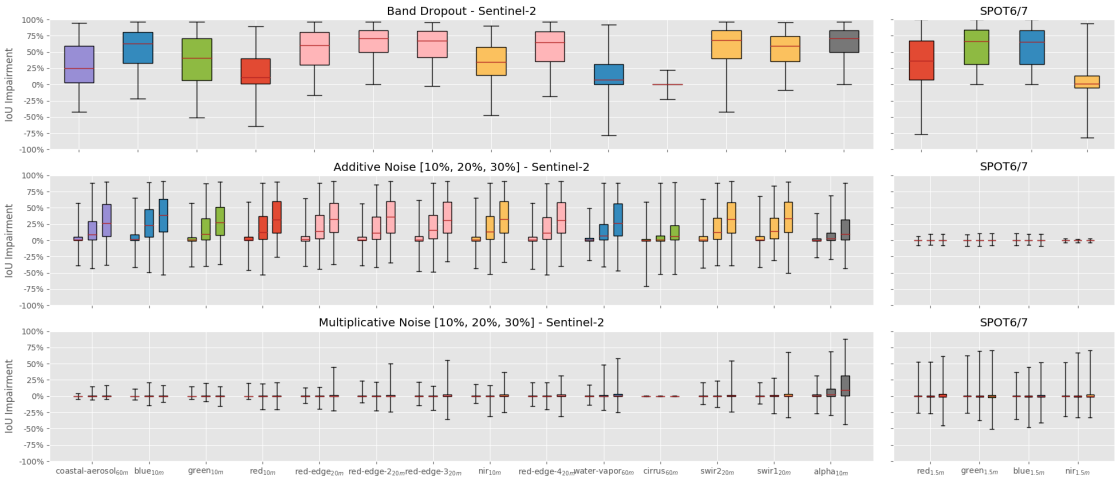


Figure A.11: Impairment of IoU for S2 and SPOT semantic segmentation models under band perturbation, additive noise, and multiplicative noise. S2 model shows sensitivity to all bands except the cirrus band, indicating the importance of both spectral and spatial features. The SPOT model shows increased sensitivity to the lower-wavelength green and blue bands, indicating an increased reliance on color.

Table A.5: Detection Confidence Levels

Level	Geometry	Install Date	Description	Fraction of Polygons	Fraction of Capacity
A	SPOT	S2	Intersecting SPOT and S2 detections where IoU>30%	50%	36%
B	SPOT	None	SPOT-only detections	15%	42%
C	S2	S2	S2-only detections	17%	14%
D	Both	S2	Intersecting SPOT and S2 detections where IoU<30%	18%	8%

Installation Date

We obtain installation dates as described in the machine learning pipeline. For positive detections determined by RNN-2, we obtain the date of the first scene in the S2 back-catalog where the pixel-wise primary inference mean is greater than 20%. To evaluate the accuracy of this heuristical rule, we compare our data to the installation dates of matched PV solar energy installations in the United StatesUS Energy Information Administration (2020). Accurate installation dates for other geographies were not available. The distribution of our estimated installation dates

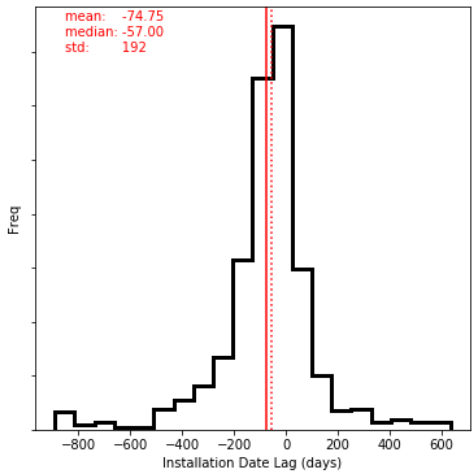


Figure A.12: Distribution of difference in days between our estimated installation dates and US EIA (2020) installation dates for matched plants in the United States. Our observed installation date, corresponding to 20% construction completion, leads facility 'operational date' by approximately two months.

against the EIA sample is shown in Figure A.12. Our observed installation dates tend to lead the US EIA 'operational dates', by approximately two months. This matches our expectations of the gap between the beginning of construction and facility operation. Our uncertainty in these dates mean that our study may have captured facilities which, at the end of 2018, might not have yet been entered into official statistics. Further, while insight might be drawn at the aggregate level, this uncertainty is too high for reliable conclusions at the level of individual PV solar energy facilities. We show, however, that earth observation might provide a valid method for more detailed studies of PV installation dates and construction progress, and might be able to provide more specific insight into the time gap between investment decisions and facility operation.

Generating Capacity

Our study includes all installations over $10kW$ in size, a cross-sectional definition encompassing the IEA definitions of 'utility-scale', 'distributed commercial/industrial', and 'off-grid' and over 89% of cumulative installed capacity in 2018 (International Energy Agency 2019*b*). This definition captures both conventional generating

stations and 'behind-the-meter' installations that electricity system operators might otherwise have poor visibility of. We focus on these non-residential PV solar energy installations because they comprise the majority of installed PV generating capacity, suiting our target policy and engineering use cases. Our dataset does not include residential solar which is too small to detect reliably in our imagery sources and makes up 11% of total installed capacity(International Energy Agency 2019*b*).

We develop a simple estimate of generating capacity for illustrative and comparative purposes, drawing on the novelty of our spatially-localised dataset. We address the non-trivial challenge of converting geospatially-localised polygons to AC generating capacities by obtaining global maps of incident irradiance at optimal tilt angle, GTI, and PV solar energy production intensity, $PVOUT$, from the *Global Solar Atlas*(2016). We obtain very rudimentary distributions of the installation ground coverage ratio (also called the packing-factor), GCR, and the module efficiency, η, from Ong, S. et al. (2013)Ong et al. (2013), who obtain data for 193 sample projects in the United States. To obtain a distribution for inverter loading ratio, ILR, we cross-reference their survey with data obtained from the US EIA.(2020) We use gamma distributions for GCR, ILR, and η to capture the skew of the data. Values for the shape parameters k, the location parameters loc, and the scale parameter θ are shown in Table A.6. We estimate nominal peak AC generating capacity, $gencap$, using the array area, $area$:

$$gencap\,[kW_{p,AC}] = area\left[m^2\right] \times GTI\left[\frac{kWh_{irr}}{m^2}\right] \times GCR\left[\frac{m^2}{m^2}\right] \times \eta\left[\frac{kWh_{prod}}{kWh_{irr}}\right] \div PVOUT\left[\frac{kWh_{prod}}{kW_{DC,p}}\right] \div ILR\left[\frac{kW_{p,DC}}{kW_{p,AC}}\right] \quad \text{(A.1)}$$

With the area-binned recall values shown in A.4 we are able to obtain a best estimate of 494GW gross installed generating capacity, with an uncertainty due to the distribution of ILR, GCR, and η of $\pm 6GW 95\% CI$. We do not quantify our uncertainty in GTI, $PVOUT$, $area$, or our recall values. To use GTI and $PVOUT$ in this simple calculation of nominal peak AC generating capacity, we must assume

Table A.6: Installed capacity calculation uncertainty (Ong et al. 2013, US Energy Information Administration 2020)

Metric	Symbol	$N_{samples}$	k	loc	θ
Ground Coverage Ratio	GCR	39	19.5	0.030	0.025
Inverter loader ratio	ILR	61	1.8	0.087	0.032
Module efficiency	η	109	26.3	0.68	0.018

all the generating units in our dataset are fixed-tilt and fixed-orientation. Our inference, vectorisation, and measurement of *area* best corresponds to the 'direct area' metric of Ong et al. (2013) and Hernandez, R.R. et al (2014), but we do not quantify our uncertainty in these measurements. Our purpose in providing this estimate is to show indicative progress towards globally-complete asset level data, which must be expressed in terms of generating capacity in order to be compared to other data sources. We therefore stress that the method we present here is very rudimentary and we advise against using these figures for detailed design engineering or local policy purposes.

There is considerable scope to improve this methodology in future work. The type of PV solar energy installation (e.g. fixed, single-axis, and dual-axis tilt systems) can be determined in remote sensing imagery. Estimates of the ground coverage ratio, panel count, installation orientation, and even tilt angle can be obtained from high resolution imagery. The type of cells (single- and poly-crystalline, thin film) might also be ascertained from high resolution imagery. Installation date, cell type, and other meta-data might be combined to obtain an estimate of panel efficiency. Our dataset provides a starting point for the development of a more detailed and accurate dataset.

Land Cover

Land use change for renewable energy production has climate and conventional environmental impacts and trade-offs with food production systems, community and indigenous land use, and other aspects of the human economy. We identify

the pre-existing land cover type for PV solar energy installations using publically available historical land cover datasets. We sample PV solar energy sites again European Space Agency Climate Change Initiative 300m land cover map for the years 1992 through 2018.(Lamarche et al. 2017) We choose this land cover product for its global availability, minimising geographical bias in land cover classification. (We recognise that for certain geographies there are higher resolution land cover products available, however we choose not to introduce the bias of sampling from multiple land cover products.) For sites where the installation date has been determined, the pre-isntallation land cover is obtained from the 2012 global land cover map. Where the installation date is unknown, or is known to be prior to 2016, we choose 2007 as a representative pre-installation date. This date is sufficiently long ago to ensure the sample pre-dates the installation of the PV facility, while mitigating land cover changes which may have happened between 1992 – the first year for which data is available – and the eventual (unknown) installation date of the facility. This introduces a small uncertainty for the 8.5GW of installations (per IEA (2019*b*)) built before this time period.

Figure A.13 shows the diversity of land cover, installed capacity growth, and installation size distributions in the top 20 countries in our dataset. Over the study time period, installed capacity in the top four countries (China, the United States, India, and Japan) grows considerably, while EEA countries (Italy, the United Kingdom, Germany, Spain) do not grow as fast, with the exception of France. The size histograms of installations show that larger installations are installed preferentially on barren, grassland, and desert landcovers, while smaller installations are situated on forests or built-up areas.

The land cover class distribution of PV solar energy sites varies significantly by country, indicating land cover is not the single driving factor for PV siting decisions. PV installations in most countries are located on cropland, but China, the United States, and France showing a larger diversity of pre-existing land covers at PV sites. Local land cover bias varies by country, suggesting that land cover considerations are secondary to other factors. We hypothesise that the bias to agricultural sites is

driven by unobserved confounding whereby both PV solar energy and agriculture favour sites in close proximity to human populations and access to markets, labour, infrastructure, etc. Most countries show a large bias against forest-covered sites and a moderate bias against shurbland sites. We hypothesise that these land covers impose expensive land clearing operations prior to PV solar energy installation. Only the Czech Republic, France, and Germany show a concerted bias for the development of PV solar energy on built-up land. PV mega-projects in excess of 250MW have been built in China, the United States, Mexico, India, and South Africa. Many of these mega-projects are also built on aridlands, shown in the strong positive bias for aridlands in large facilities in excess of 5MW, see Figure A.14.

We detect a number of 'other' installations which includes installations on water. Due to the coarse 300m resolution of the land cover data, we hand-verify these to ensure that they are not mis-classified coastal installations, for example. In doing so we identify 84 floatovoltaic sites in our data.

Figure A.14 shows the growth in generating capacity and land cover bias binned by generating capacity. Despite being fewer in number, large utility-scale projects make up the largest portion of gross generating capacity. Unlike installations smaller than 5MW, these large installations have a bias away from croplands and towards aridlands. Installations smaller than 5MW have a positive bias towards croplands and a stronger positive bias towards built-up areas. These two trends suggest the fundamental drivers for PV solar energy siting decisions are dependent on the size of the installation and there is a meaningful difference between large and small PV energy installations. All size classifications show a bias away from forest and shrubland.

Protected Areas and Indigenous and Community Lands

Recognising the impact that non-residential PV solar energy development can have on protected areas and community and indigenous lands (see, e.g. Hernandez, R.R. et al. 2015), we also add proximity to protected areas and community and indigenous lands as additional features of our data. For each detection, protected

Table A.7: Dataset Metadata

Property	Description
unique_id	A unique identifier for each detection.
geometry	The detection geometry in WGS84 longitude/latitude coordinates.
confidence	Installation confidence classification, see Table A.5.
iso-3166-1	Country-level political administrative ISO-A2 code.
iso-3166-2	Province/state-level political administrative code.
area	Installation geodesic area in m^2.
capacity_mw	Estimated nominal peak AC generating capacity in MW.
install_date	Comma-separated estimates for installations and major expansions in 'YYYY-mm-dd' format.
gti	Incident solar irradiance for optimally tilted angle, in $\frac{kWh_{irr}}{m^2}$.
pvout	Photovoltaic power potential, in $\frac{kWh_{prod}}{kW_{p,DC}}$.
wdpa_10km	Protected area records from the World Database of Protected Areas (2016) within 10km, arranged as {distance, WDPA_ID, name, type}
ind_comm_10km	Indigenous and community land records from Landmark.org(2017) within 10km, arranged as {name_1, name_2, etc.}
lc_20XX	Land coverage assessment for 1992 through 2018 for ESA Climate Change Initiative 300m land cover, see Table A.8.*

areas and indigenous and community lands within 10km of the site are identified. We obtain protected areas from the World Database of Protected Areas(2016), and retain the distance, identification number, name, and type of protected area for each of our detections. For indigneous and community lands, we obtain proximate lands from Landmark.org(2017) and note the record name. Neither of these datasets are perfect, so absence of these records should not indicate the absense of a protected, community, or indigenous area, but providing convenient access to this data alongside our detections might help researchers better identify and study impacts on these vulnerable lands.

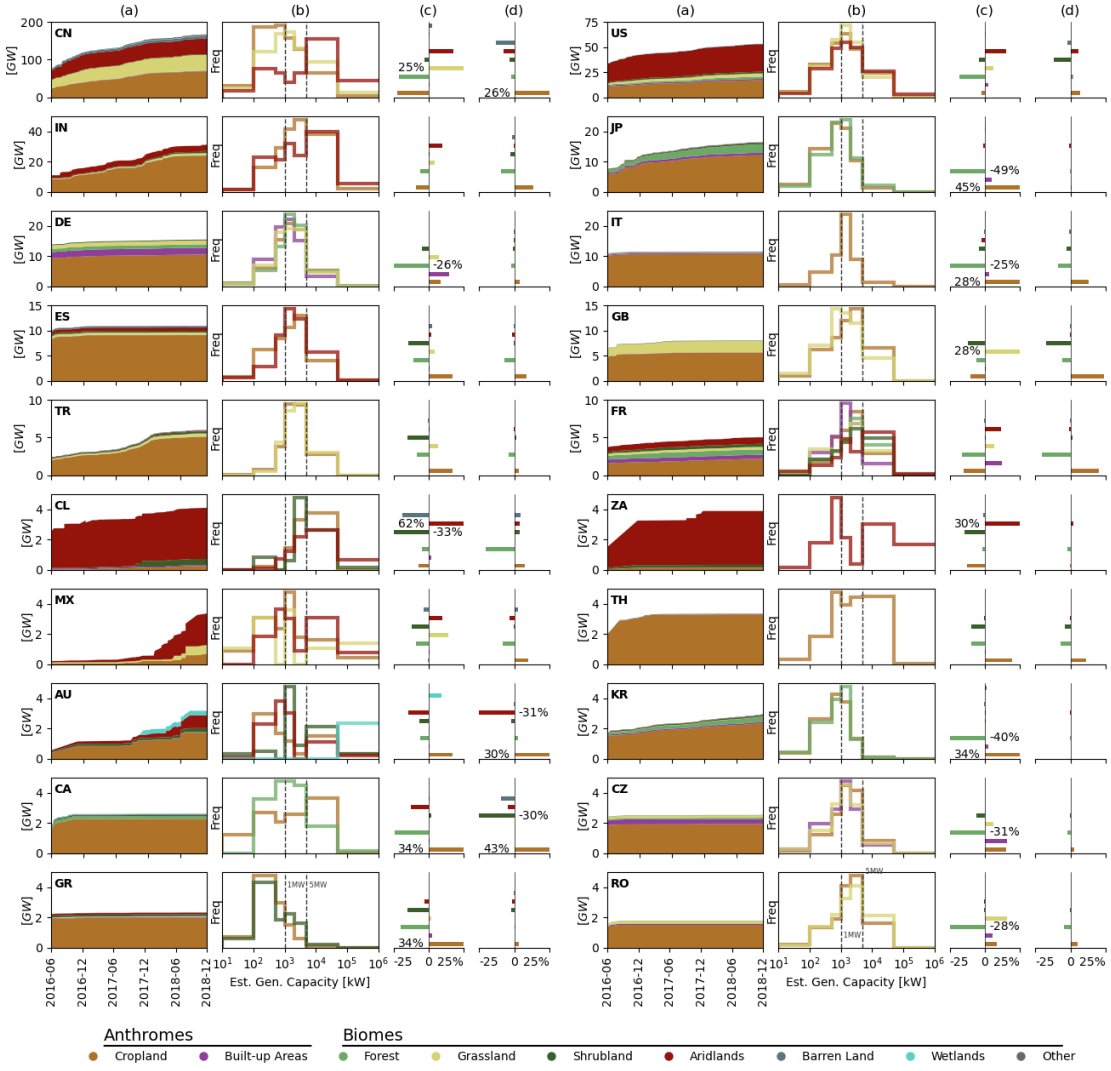


Figure A.13: Pre-existing land cover for PV solar energy installations, showing the time series of installations (a); the distribution of installation sizes by land cover (b); local bias (c) between PV land cover and the distribution of local land covers within a 0.5°x0.5° grid pixel; and country-level bias (d) between country land cover distribution and distribution of local land covers containing PV detections.

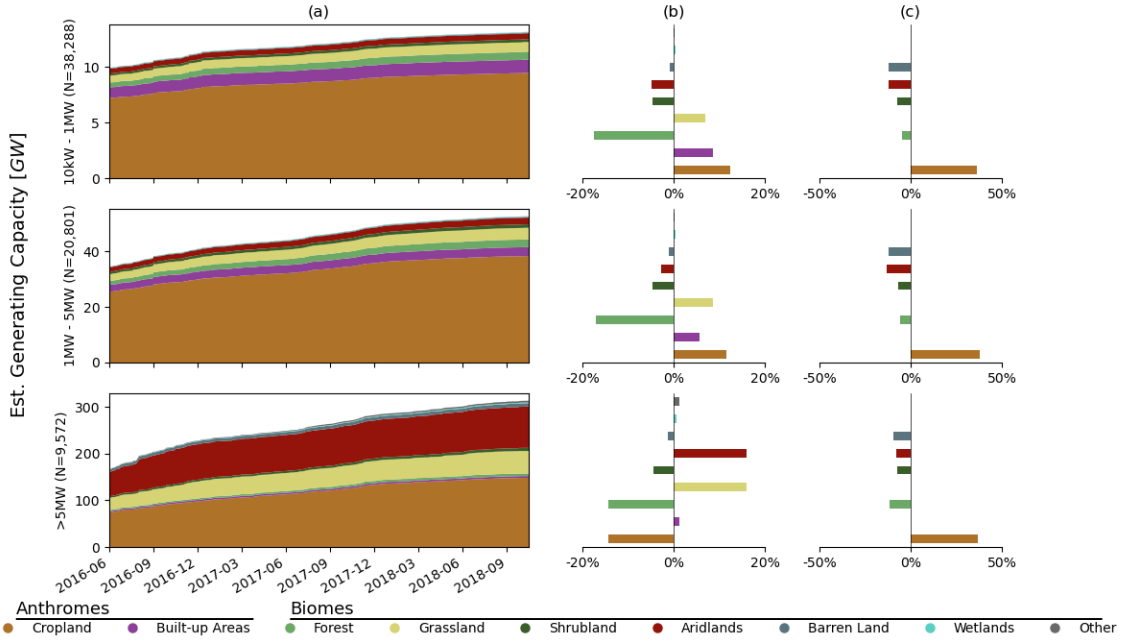


Figure A.14: Pre-existing land cover for PV solar energy installations for three bins of nominal generating capacity: 10kW to 1MW; 1MW to 5MW; and >5MW. Panels show the time series (a) of installations; local skew (b) between PV land cover and local land covers, where localities are described by the pixels of a 0.5°x0.5° grid containing PV detections; and global bias (c) between the global land cover distribution and locality land cover distributions for localities containing that size of PV solar energy installations.

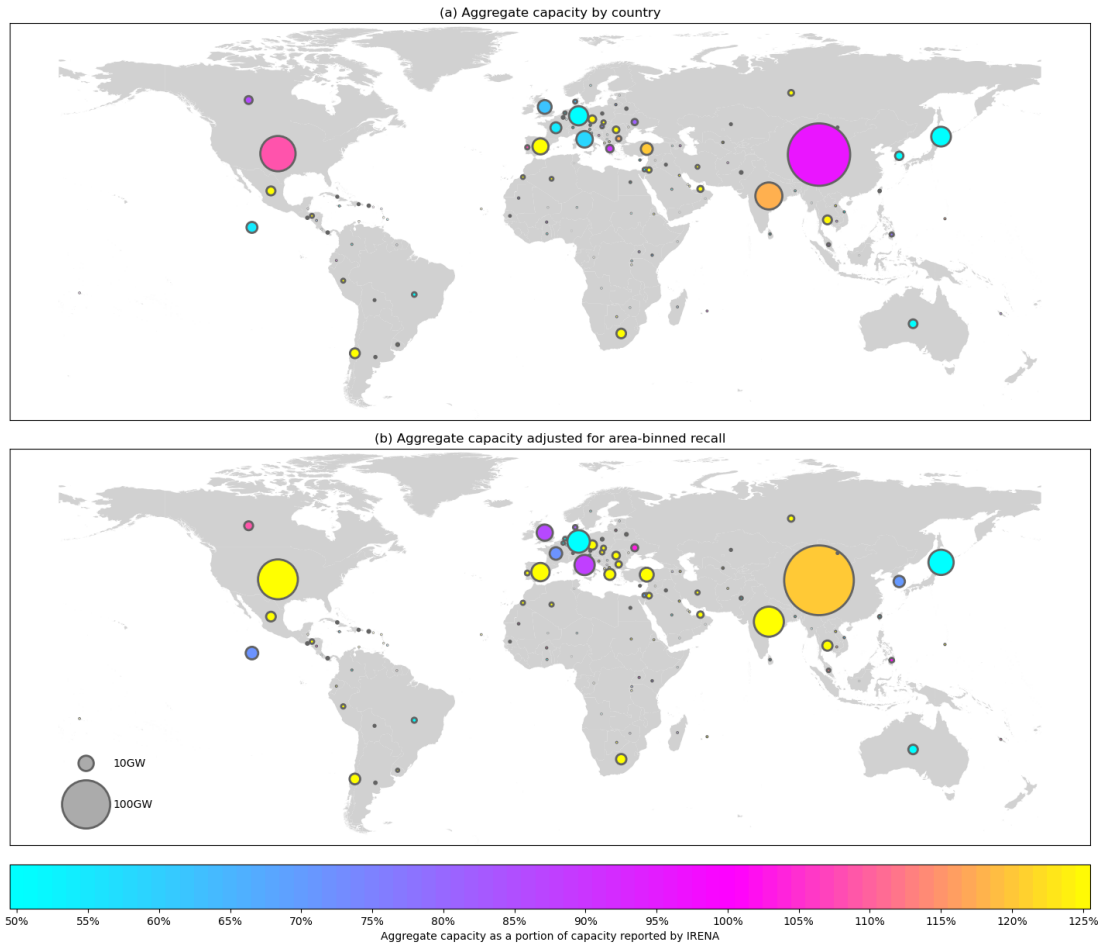


Figure A.15: Comparison of our data aggregated by country to country-level data from IRENAirena:2019. Panel (a) shows direct comparison between our data and IRENA totals. Panel (b) shows our data adjusted using area-binned validation recall figures (see Table A.4). Note that IRENA data is inclusive of residential PV solar energy.

Table A.8: Landcover Class Mappings - ESA Climate Change Initiative Land Cover (Lamarche et al. 2017)

Class	Mapping Classes
Cropland	10: Cropland, rainfed; 11: Cropland, rainfed, herbaceous cover; 12: Cropland, rainfed, tree or shrub cover; 20: Cropland, irrigated or post-flooding; 30: Mosaic cropland 50% / natural vegetation 50%; 40: Mosaic natural vegatation 50% / cropland 50%
Built-up areas	180: Urban areas;
Forest	50: Tree cover, broadleaved, evergreed; 60: Tree cover, broadleaved, deciduous; 61: Tree cover, broadleaved, deciduous, closed; 62: Tree cover, broadleaved, deciduous, open; 70: Tree cover, needleleaved, evergreen; 71: Tree cover, needleleaved, evergreen, closed; 72: Tree cover, needleleaved, evergreen, open; 80: Tree cover, needleleaved, deciduous; 81: Tree cover, needleleaved, deciduous, closed; 82: Tree cover, needleleaved, deciduous, open; 90: Tree cover, mixed leaf type;
Grasslands	130: Grassland;
Shrub, herbaceous, and sparse vegetation	100: Mosaic tree and shrub 50% / herbaceous cover 50%; 110: Mosaic herbaceous cover 50% / tree and shrub 50%; 120: Shrubland; 121: Evergreen shrubland; 122: Deciduous shrubland; 140: Lichens and mosses; 150: Sparse vegetation; 151: Sparse trees; 152: Sparse shrub; 153: sparse herbaceous cover;
Barren areas	200: Bare areas; 201: Consolidated bare areas; 202: Unconsolidated bare areas;
Aridlands	Forest, Grasslands, Shrub/herbaceous/sparse, or barren areas where the aridity index is less than 20%
Wetlands	160: Tree cover, flooded, fresh or brackish water; 170: Tree cover, flooded, salien water; 180: Shrub or herbaceous cover, flooded;
Other	0: No data; 210: Water bodies; 220: Permanent snow and ice;

Table A.9: Aggregate Statistics by Country

ISO-3166-1	Number of Detections	Est. Total Gen. Capacity [MW]	Mean Detection Size [MW]	ISO-3166-1	Number of Detections	Est. Total Gen. Capacity [MW]	Mean Detection Size [MW]
CN	167,698	18,449	9.09	BY	223	38	5.88
US	53,727	7,638	7.03	SV	213	24	8.89
IN	31,503	2,277	13.84	NA	207	18	11.52
JP	16,674	10,504	1.59	AR	195	8	24.34
DE	15,651	4,702	3.33	PL	193	127	1.52
IT	11,701	5,796	2.02	CU	158	40	3.94
ES	10,994	1,970	5.58	DO	152	13	11.73
GB	8,094	2,222	3.64	IQ	145	6	24.22
TR	6,057	1,543	3.93	KZ	142	6	23.73
FR	5,119	936	5.47	BO	137	4	34.26
CL	4,137	150	27.58	MN	126	6	20.97
ZA	3,926	1107	35.69	EG	110	24	4.60
MX	3,396	110	30.87	SN	109	9	12.16
TH	3,382	529	6.39	AT	102	70	1.45
AU	3,207	164	19.56	UZ	85	6	14.24
KR	2,995	2,680	1.12	LK	70	10	7.03
CA	2,661	353	7.54	MR	68	11	6.19
CZ	2,573	951	2.71	CY	62	78	0.79
GR	2,378	3,645	0.65	CH	61	48	1.27
RO	1,833	398	4.61	KE	61	1	60.53
UA	1,624	166	9.79	LT	59	31	1.91
AE	1,535	15	102.30	PS	55	18	3.08
RU	1,403	51	27.52	BF	53	4	13.22
BG	1,210	340	3.56	GT	51	7	7.30
JO	1,150	84	13.69	VN	51	5	10.19
BR	889	63	14.11	LA	42	4	10.40
PT	720	145	4.96	KW	41	5	8.18
SK	718	279	2.57	GU	40	2	20.04
PH	691	82	8.43	UG	38	3	12.60
PE	646	14	46.13	SI	33	22	1.49
HN	637	64	9.95	AZ	32	7	4.61
DK	619	119	5.20	BD	31	14	2.21
IR	593	47	12.62	MU	28	11	2.58
MA	588	44	13.36	CO	28	6	4.71
DZ	575	30	19.16	SA	28	7	4.00
IL	566	66	8.57	JM	28	2	13.78
HU	509	301	1.69	MK	27	15	1.82
NL	396	130	3.04	NC	27	4	6.81
MY	374	59	6.33	EC	25	8	3.11
UY	319	20	15.93	KH	24	5	4.82
BE	303	156	1.94	GH	23	3	7.74
PK	276	72	3.83	BW	22	4	5.61
TW	238	161	1.48	MG	21	1	21.35
PA	231	20	11.57	NI	21	5	4.25
PR	224	36	6.23	AF	20	7	2.92

B

Asset-Level Transition Risk in the Global Fossil-Fuel Supply Chain.

Kruitwagen, L.,[1,*] Klaas, J.,[2] Baghaei Lakeh, A.,[2] & Fan, J.[2]

[1] University of Oxford, Smith School of Enterprise and the Environment
[2] QuantumBlack, a McKinsey Company
[*] Corresponding Author

Contents

B.1 Motivation 113
B.2 Related Work 115
B.3 Data Sources and Preparation 116
B.4 Network Topography 121
B.5 Simulated Energy Flow 123
B.6 Transition Risk Analysis 125
B.7 Future Research Directions 137
B.8 Supplemental Methods 140
B.8.1 Network Assembly 140
B.8.2 Flow Calculation 143

Abstract

Climate change risks manifest in the real economy, with grave consequences for welfare of human populations and the biosphere; the economic returns of industrial sectors; and geopolitical stability. Understanding the diffusion of risks in real infrastructure networks is an urgent priority for delivering climate change mitigation, adaptation, and resiliency. The oil, gas, and coal supply chains are the most salient progenitors and inheritors of these environmental risks. We prepare a geospatial arrangement of the global oil, gas, and coal supply chains using open-source asset-level data. The resulting complex network has 6.09mn nodes and 15.70mn edges and is implemented in a graph database. With estimates of annual coal, gas, and oil demand in 13,229 global population centres and 8,165 global power stations, we use a minimum-cost flow method to estimate global asset-level energy flows. We develop a method for cross-validating and tuning our network flow simulation using aggregate country-level import and export statistics. We demonstrate two analyses of asset-level transition risk: a counter-factual demand shock scenario consistent with the IEA Sustainable Development Scenario; and supply shock scenarios developed by interdicting regionally-aggregated coal, oil, and gas supplies. Our contribution lies in the global scope of our asset-level supply chain and the novelty of our minimum-cost flow method. We conclude with a discussion of further research directions and make the graph database and supporting code publicly available. [1]

[1] Code and data are available here: https://github.com/Lkruitwagen/global-fossil-fuel-supply-chain

B.1 Motivation

Climate change risks have been qualitatively compared to shock events like the global financial crisis (Carney 2015) and the COVID-19 pandemic (Salas et al. 2020). The resilience of complex networks (e.g. of financial institutions, of human populations) and the diffusion of shock events are phenomena emergent from the network's topology (i.e. the nodes and edge connections of the network)(Gai & Kapadia 2010). A complex network arrangement of the large fixed-capital infrastructure assets that comprise the global energy sector enables analyses of the resilience of the energy sector to climate change risks, and the diffusion of climate change shocks. The energy sector represents a store of $22tn in invested capital (Dan Tong & Davis 2019); is the largest source of greenhouse gas emissions (IPCC 2014); and (literally) powers the real economy. It is the most salient industrial sector for a high-resolution analysis of climate change risk.

A clear typology of climate risks has emerged following the wide stakeholder consultation of the Bank of England (2015) and the Taskforce on Climate-Related Finance Disclosures (TCFD) (2016). Asset-level analyses of climate risks are increasingly available, (e.g. Caldecott et al. 2016, Pfeiffer et al. 2016, Caldecott, Dericks, Bouveret, Schumacher, Pfeiffer, Tulloch, Kruitwagen & McCarten 2018*a*, Dan Tong & Davis 2019) however risk is typically assessed in the first order only, e.g.. the direct impact of physical climate change or a policy intervention on carbon emissions. Progress is being made to better understand climate change risk through complex supply chains. Research to date has focused on physical climate change risks and transition risk in the form of direct exposure to greenhouse gas mitigation policies (Ghadge et al. 2020).

In this work, we conceive transition climate risk as also being an emergent property of the topography of a supply chain network, where the network arrangement is described both geospatially and relationally between supply chain constituents. The drivers of transition risk, namely policy, markets, technology change, and reputational risks (TCFD 2016), occur at various spatial scales, and have a reflexive relationship with the incumbent arrangement of the infrastructure

network. Climate transition risk in the energy sector can then be characterised as a complex adaptive system (Holland 1992), resistant to the strict boundaries of geographies, markets, and industrial (sub-)sectors.

Understanding the complex dynamics of transition risk in the energy sector can help policy makers target interventions to deliver climate change mitigation. We seek both to describe the impacts of policy decisions on indivual supply chain assets as well as normatively identify *Sensitive Intervention Points*, where a relatively small impulse might trigger irreversibilities and nonlinear feedbacks, leading to an outsized contribution to climate change mitigation (Farmer et al. 2019*b*). We posit that some sensitive intervention points might be geospatial in nature, and might be identified in a complex network of the physical layout of the world's energy infrastructure.

The purpose of our research is thus twofold: first to spatially localise transition risk in the real economy, and second to identify sensitive interventions points for the transition to a post-carbon economy in the layout of real infrastructure. This paper presents significant progress towards these two ends by implementing the coal, oil, and gas supply chains as a complex network. We obtain asset-level data for onshore and offshore oil and gas fields, coal mines, oil and gas refining and processing facilities, pipelines, railways, liquified natural gas (LNG) terminals, ports, shipping routes, urban centres, and power stations. We connect these real-economy assets into a graph, using geospatial intersection and proximity to form the network edges between hierarchical asset types. We solve for energy flow between energy 'sources' (i.e. oil and gas fields and coal mines) and energy 'sinks' (i.e. urban centres and power stations), where the impedance of the network edges is proportional to their geodesic length. We tune our supply chain model to aggregate production and trade data obtained from the IEA(2020) and the UN Comtrade database(2021). The network arrangement and flow simulation provide a starting point for a wide suite of complexity and optimisation studies with wide-reaching implications for the diverse stakeholders of climate change risks and impacts.

We use our coal, oil, and gas supply chain models to interrogate asset-level impacts of a bundle of policies associated with the IEA's Sustainable Development

Scenario. Treating the change in energy demand as a demand shock, we solve for asset-level energy flow in 2040. We identify the energy supply and transmission assets most affected by a rapid transition to a sustainable energy system. Finally, we interdict oil and gas fields, oil and gas wells, and coalmines to identify sensitive intervention points where supply-side energy policies would have the largest impact on energy system costs.

Further work is now required to make our energy system more representative: to identify and fill data gaps; to better obtain marginal costs for energy flow through infrastructure; and to bring our modelled results to real data.

B.2 Related Work

This work builds on emerging studies of climate change risk and seeks to make bottom-up asset-level risk analysis coherent with integrated assessment models of the climate-economy-energy system. Even prior to the establishment of the typology of 'physical' and 'transitional' climate risks (The Bank of England 2015), scholarly work identified, as examples, the impacts that carbon pricing might have on power generation assets (Green 2008), and the impact that policy consistent with certain carbon budgets might have on the stranding of coal, oil, and gas assets (McGlade & Ekins 2015). Physical risks emerge from environmental conditions which are inherently spatial, so they are naturally assessed at the asset level (e.g. water risk at power stations (Ganguli et al. 2017), heat stress on livestock operations (Key et al. 2014), etc.). Assessing *transition* risk at the asset level, however, has been elusive until now. Top-down assessments of the impacts to the global energy system under various policy and technology interventions are available from integrated assessment models (Hamilton et al. 2015) and various government (e.g. The Bank of England 2015), NGO (e.g. CTI 2011, 2013), and corporate authors (e.g. Paltsev et al. 2018). Asset-level data has been used to assess risks of exceeding global carbon budgets (e.g. Pfeiffer et al. 2016, Dan Tong & Davis 2019), alignment with transition pathways in the power sector (e.g. Alova 2020), and to quantify 'unlockable' carbon budget from technology retrofits (namely

carbon capture and storage) (e.g. Caldecott, Dericks, Bouveret, Schumacher, Pfeiffer, Tulloch, Kruitwagen & McCarten 2018*b*). Asset-level energy system models are available for individual sectors and geographies (e.g. Konda et al. 2011, Dunnett et al. 2008, Elahi et al. 2014), but global studies are unavailable.

We provide a global asset-level energy model for the three main energy carriers: coal, oil, and natural gas. A key contribution of this work is that the model is built entirely with open-source data and code, making our work reproducible and a starting point for further improvements and analysis by the research community. Until this work, global energy system models with asset-level resolution have largely been the purview of private companies - such as the proprietary asset-level coal model of Deloitte (Trüby 2013) and the oil model of Rystad (Wold 2015). To our knowledge, this is the first global and publicly-available asset-explicit model of the world's energy infrastructure. This work is made possible by the substantial efforts of public sector organisations to develop and publish open datasets of energy system infrastructure and activity, particularly the work of the Global Energy Monitor (Shearer et al. 2017*b*), the Global Oil and Gas Features Database (Rose, Bauer, Baker, Barkhurst, Bean, DiGiulio, Jones, Jones, Justman, Miller, Sabbatino & Tong 2018), the European Commission Joint Research Council (Janssens-Maenhout et al. 2019, Schiavina et al. 2019) and OpenStreetMap (OpenStreetMap contributors 2017).

B.3 Data Sources and Preparation

The global energy system is modelled as a flow network of sources, sinks, and transmission infrastructure. The nodes and edges of the network correspond to real-economy infrastructure assets, see Table B.1. All the data are geospatially localised, i.e. have a geospatial shape attribute with latitude and longitude coordinates. Figure B.1 shows all the data arranged on a world map. Table B.1 also shows the number of sources from which the energy data was originally collected. We use only open-source data to enable our research to be used by others as widely as possible.

Energy Demand

We develop a high-resolution geospatial dataset of global energy demand by

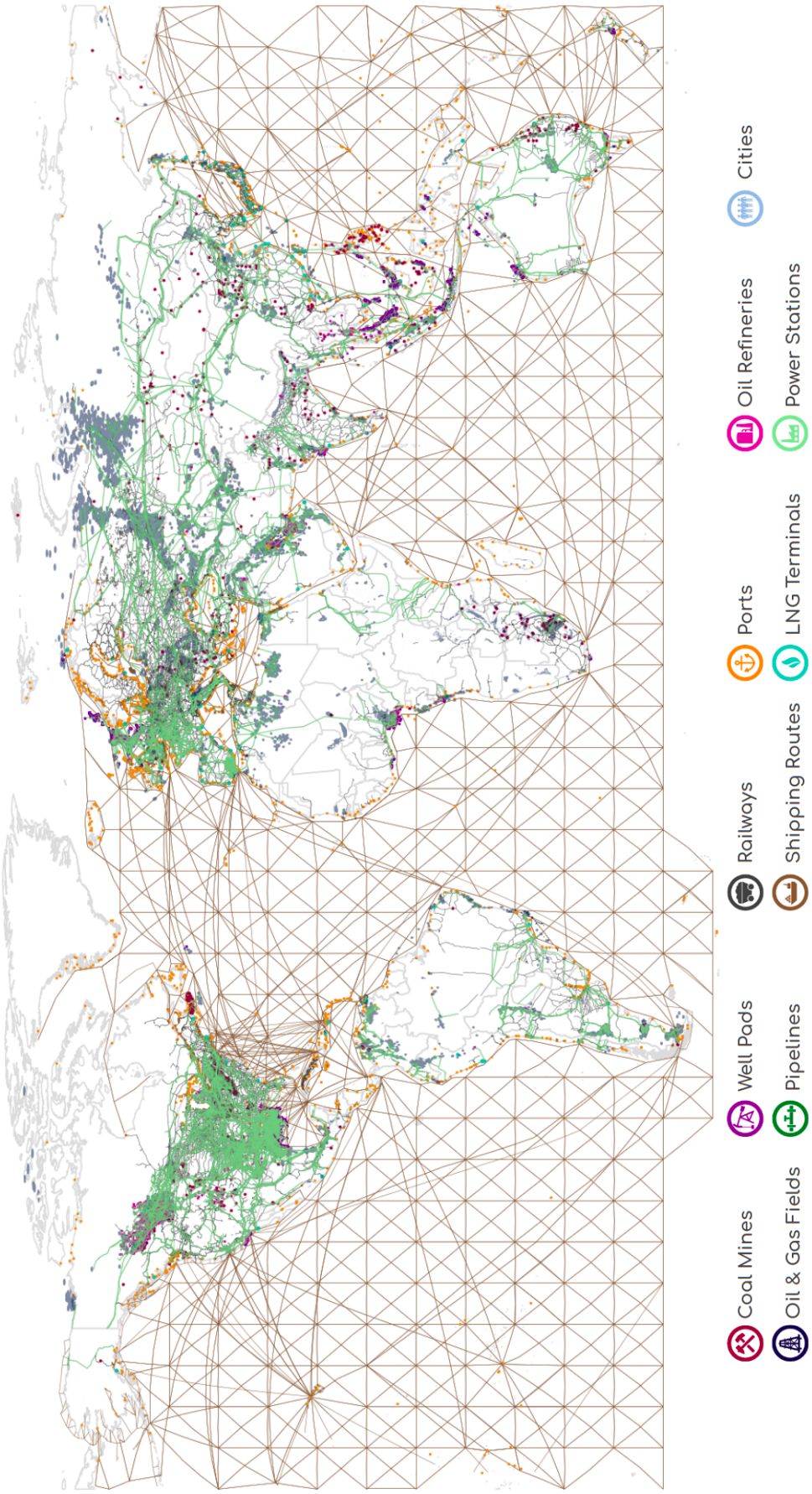


Figure B.1: The global fossil fuel supply chain.

Table B.1: Data Description

Name	$N_{Installations}$	Sources[1]	Quality	Icon
Well Pads	9,845	Rose et al. $(2018)_{24}$	Good	
Oil Fields	25,236	Rose et al. $(2018)_{63}$	Good	
Coal Mines	3,099	Rose et al. $(2018)_{6}$,[2]	Adequate	
Processing & Refining Facilities	2,501	Rose et al. $(2018)_{55}$	Good	
LNG Liquification & Regasification	329	Rose et al. $(2018)_{15}$	Excellent	
Pipelines	94,448	Rose et al. $(2018)_{82}$	Good	
Ports	3,702	Rose et al. $(2018)_{10}$	Excellent	
Shipping Routes	8,273	*Global Shipping Lane Network, World, 2000* (2015)	Excellent	
Railways	496,808	Rose et al. $(2018)_{51}$, (OpenStreetMap contributors 2019)	Excellent	
Power Stations	28,664	Byers et al. (2018)	Good	
Population Centres	13,229	Schiavina et al. (2019), International Energy Agency (2018)	Excellent	

1: Rose et al. (2018) provides secondary reference to a large number of primary data sources. The number of primary sources for each asset type is indicated in subscript.
2: Additional sources for coal mines include *Project Payments Database* (2019), *Tweet: Major coal mines and coal-fired power plants in China* (2015), *Distribution of Coal in India: Gondwana Coalfields & Tertiary Coalfields* (2019), *Indonesia's coal mines and forests* (2015), *Major Coal Projects* (2014), Yuk (2011), *Coal mining areas in North and Central Vietnam* (2015), Nakbanpote et al. (2018), *EIA-7A: Coal Production and Preparation Report* (2018), Ranada (2013), *Pakistan and its Mineral Resources* (2015), Wilson (2015), Dias et al. (2018), *Euracoal: Turkey* (2020), *Euracoal: Ukraine* (2020), *Euracoal: Serbia* (2020), Pantelia (2018), *Industrial sites in Kosovo* (2007), *Current Developments in the Russian Mining Industry* (2007), Kalmykov & Malikova (2017), Vorsler (2002), Kirshner & Power (2015), Baruya & Kessels (2013), *Mining* (1979), Oei & Mendelevitch (2019), Fontana (2019), *Carbon* (2012)

combining the European Commission Joint Research Centre's Emissions Database for Global Atmospheric Research (EDGAR) (Crippa et al. 2020) and Global Human Settlement Urban Centre Database (UCDB) (Schiavina et al. 2019) with country-level energy system data obtained from the International Energy Agency (IEA) (2020). The UCDB is comprised of dense population centres around the planet derived from satellite and census data, which we use as loci for a euclidean allocation of the planet's land surface area. We then spatially aggregate EDGAR emissions maps by industrial sector for each urban center and it's surrounding spatial allocation. For each industrial sector we back-calculate annual final energy consumption from emissions using standard emissions factors (Department for Business, Energy, & Industrial Strategy 2019) and energy balances for 181 countries obtained from the IEA (2020). The result is a spatial dataset of 13,229 named urban centres and their surrounding areas with corresponding coal, oil, and gas consumption.

We account for power sector final energy consumption separately to the other industrial sectors. We use data from the World Resource's Institute Global Power Plant Database (2018), which is spatially localised and is exhaustively complete for the world's large fossil fuel power stations. We derive estimates for annual energy demand per plant by using the plant's nameplate capacity and basic assumptions around the plant's operating characteristics. As a preliminary assumption, we use 60% as the capacity factor for all of the world's coal, oil, and gas power stations.

Energy Supply and Transmission Infrastructure

We obtain asset-level oil and gas data supply and transmission data from the Global Oil and Gas Infrastructure Database produced by the National Energy Technology Laboratory (NETL) and the Environmental Defense Fund (EDF) (2018). We supplement this data with further data from the Oak Ridge National Laboratory (2015) for shipping routes, and OpenStreetMaps OpenStreetMap contributors (2019) for railways. In addition to the data from NETL and EDF, we add data from the Natural Resources Governance Institute (2019), and systemic desk research

georeferencing coal mine maps for all coal producing countries reported in the IEA Coal Market Report.[2]

There are critical gaps in the data, which limit our analysis. While the data is quite feature rich, with a geospatial geometry object and a number of properties which include beneficial corporate ownership and the asset name, most assets do not include capacity, commodity, or production data. Our assumptions concerning asset operating costs and efficiencies are therefore quite coarse, see Table B.3. The richness of our analysis emerges from the complexity of the network arrangement.

We prepare the data for network arrangement with a series of geospatial operations. Linear assets (railways, shipping routes, and pipelines) are spatially self-joined and dissolved to form a consistent set of points and connecting edges. Point asset data (well pads, coal mines, processing and refining facilities, LNG liquification and regasification facilities, ports, and powerstations) are spatially joined with linear assets using a 10km spatial buffer around each point asset. We proceed with the assumption that point assets spatially joined to the linear assets represent connection points between the real infrastructure assets. Population centres and oil fields are provided as polygon area data. We likewise spatially join population centres and oil fields to linear assets using a 10km spatial buffer, and reduce population centres and oil fields to representative points for subsequent analysis. Where no transmission infrastructure is in sufficiently close proximity to energy sources, urban centres, powerstations, and ports, we assume the existence of missing 'first mile' and 'last mile' connections with the supply network. See Methods for further details.

[2]Coal mine map references: *Project Payments Database* (2019), International Energy Agency (2019*a*), *Tweet: Major coal mines and coal-fired power plants in China* (2015), *Distribution of Coal in India: Gondwana Coalfields & Tertiary Coalfields* (2019), *Indonesia's coal mines and forests* (2015), *Major Coal Projects* (2014), Yuk (2011), *Coal mining areas in North and Central Vietnam* (2015), Nakbanpote et al. (2018), *EIA-7A: Coal Production and Preparation Report* (2018), Ranada (2013), *Pakistan and its Mineral Resources* (2015), Wilson (2015), Dias et al. (2018), *Euracoal: Turkey* (2020), *Euracoal: Ukraine* (2020), *Euracoal: Serbia* (2020), Pantelia (2018), *Industrial sites in Kosovo* (2007), *Current Developments in the Russian Mining Industry* (2007), Kalmykov & Malikova (2017), Vorsler (2002), Kirshner & Power (2015), Baruya & Kessels (2013), *Mining* (1979), Oei & Mendelevitch (2019), Fontana (2019), *Carbon* (2012)

B.4 Network Topography

We prepare a disconnected multi-level network arrangement, with a level each for coal, oil, and gas energy carriers, each drawing on the same asset-level data. We treat the levels independently, each a directional graph representing the global supply, demand, and transmission infrastructure of the energy carrier. Schematic arrangements are shown in Figure B.2. The oil graph is constrained such that oil must flow through a refinery prior to proceeding to final use in power stations and urban centres. We do this by duplicating the energy transmission infrastructure in the graph on either side of the refineries and processing plants. This divides the oil supply chain into 'crude oil' and 'oil products', analogous to the energy balances of the IEA (2020). Coal, crude oil, and oil products may be transferred from land transport to seaborne transport via a port connected to both the seaborne shipping routes and the railway or pipeline. For natural gas, we only allow modal shift to ocean freight via a liquified natural gas (LNG) liquification terminal.

Finally, we simplify the graph structure to facilitate faster computation. We identify nodes in each energy carrier graph with only one entering and one leaving edge each. We obtain reduceable connected component subgraphs composed of these nodes and identify their entering and leaving nodes. We simplify each connected component, joining edge impedances in series. This simplification reduces network size by an order of magnitude without any loss of network topological information.

The full data preparation and analysis pipeline is implemented in a highly extensible Python library, employing cutting edge parallelisation of dataframe and geospatial operations. On a 16-core cloud computation instance, the full pipeline for network assembly takes approximately 16 hours to run. The library also includes scripts for implementing the full network in a specialised graph analysis database for faster querying.

Table B.2: Freight and Modal Transfer Costs (Seljom 2019, Banks 1977, Kruitwagen 2020)

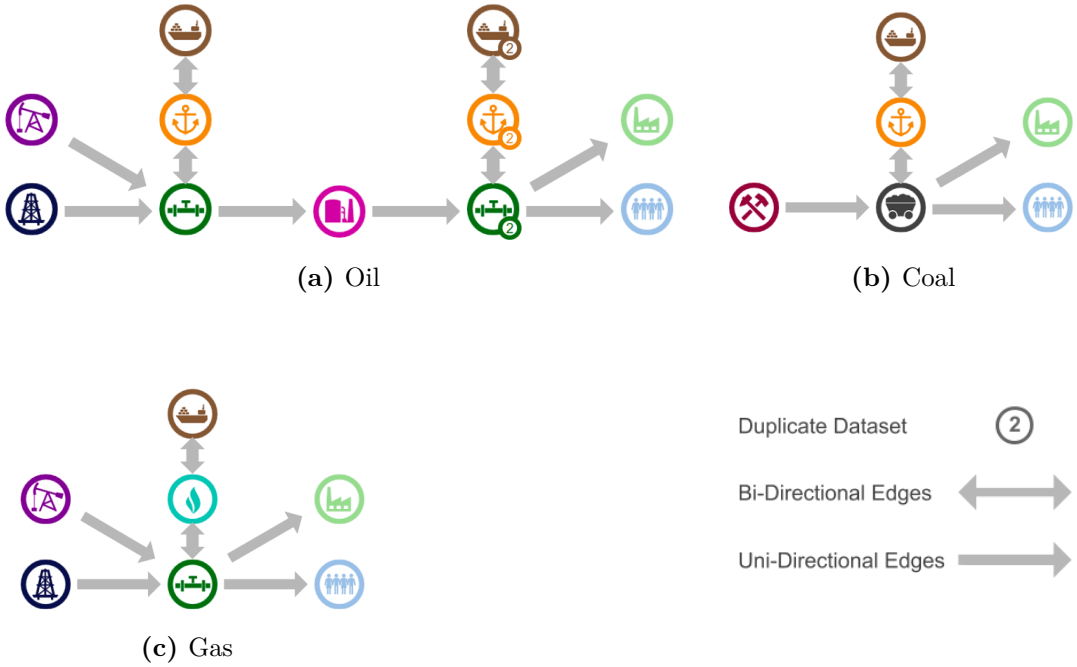
Component	Energy Carrier	Levelised Capital Costs	Parasitic, Operating, and Maintence Costs	Total Cost
Oil Pipeline	Oil	0.00486	0.00801	0.0129 $\$/t/km$
Gas Pipeline	Gas	0.00486	0.00193	0.00678 $\$/t/km$
Ocean Freight	All	-	-	0.00317 $\$/t/km$
Rail Freight	Coal	-	-	0.00254 $\$/t/km$
Ocean Transfer	All	-	-	1.429 $\$/t$
LNG Transfer	Gas	100.15	59.15	159.30 $\$/t$

Table B.3: Energy Technology Properties (Department for Business, Energy, & Industrial Strategy 2019)

Energy Carrier	Energy Density $[t/TJ]$	Power Generation Efficiency	Power Generation Utilisation Rate
Coal	34.1	35%	60%
Oil	23.9	40%	60%
Gas	19.2	55%	60%

Table B.4: Network Summaries

Energy Carrier	N_{Nodes}	N_{Edges}	Average Degree
Coal	1.22mn	2.89mn	4.733
Oil	3.23mn	8.50mn	5.260
Gas	1.64mn	4.31mn	5.271



(a) Oil **(b)** Coal **(c)** Gas

Figure B.2: Network arrangements for coal, oil, and gas supply chains.

B.5 Simulated Energy Flow

We model the global energy system by simulating energy flow through the supply chain network. We ascribe impedances to our network edges corresponding to the unit cost of energy tranmission through the corresponding real infrastructure. For pipelines, railways, and shipping routes, network impedance is assumed to be proportional to the geodesic distance of energy transmission. We obtain illustrative costs per tonne of goods transmitted per km for each linear asset type. We also impose a fixed volumetric impedance for modal shifts to and from linear asset types, i.e. for transfer to and from ocean freight. See Table B.2 for a summary of edge impedances.

We solve for energy flow as a minimum-cost flow problem, where demand at energy 'sinks' must be met at the minimum possible cost from a set of energy 'sources' through a network of known impedances. Because edge capacities are unknown, we assume that they are infinite but introduce convex costs at supply nodes to prevent a trivial solution of minimum-cost paths. To accomodate that the supply at our network's energy sources is unknown, we add a 'supersource' to

our network, and zero-impedance edges between the supersource and source nodes (i.e. oil fields and well pads). The supersource is given a supply equal to the sum of our known energy demands at cities and power stations for each energy carrier. We then solve for minimum cost flow using a cost scaling algorithm (Goldberg & Tarjan 1990) which can accomodate convex marginal costs. The solved flow is shown in Figure B.3, B.4, and B.5.

Our flow solutions are promisingly representative of actual energy flows. In coal, for example, we show major flow between Australia, Indonesia, and the Philippines to China and other Southeast Asia nations. A strong North Atlantic trade emerges, and South Africa plays a key role as a transfer point between East and Western hemispheres. Our simulations of gas flow is less convincing. Major LNG markets are missing, perhaps suggesting that the modal transfer impedance we obtained from literature is too large. Our simulation of oil flow shows a good distribution of flow and production with well-travelled ocean routes. Certain data gaps are obvious, such as a lack of transmission infrastructure data in Russia.

To interrogate whether our network topology and impedances provide an accurate representation of the global energy system, we cross-validate simulated aggregate production and trade data against historical data from 2018. Figure B.6, B.7, and B.8 shows simulated and actual volumes of energy production and international trade. The flow and cross-validation results shown are entirely emergent from the network topology and prelimary assumptions about supply marginal cost curves, uniform infrastructure impedances, and power station capacity factors. With only these naive assumptions the flow simulation provide an order-of-magnitude approximation to actual production and transmission of energy. It is promising that with proper parameterisation our model could provide a realistic representation of the global energy system.

The flow computational graph is considerably more expensive than the network assembly pipeline. The methods to obtain the minimum cost and minimum-cost path between all source and sink nodes, in particular, took approximately one week on a cluster with 96 cores and 672gb of memory. More details concerning

the network implementation and flow simulation are available in the Methods. We also discuss some improvements which would drastically reduce the computational overhead of flow calculation. The improvements are required in order to reduce the expense of paramter tuning.

B.6 Transition Risk Analysis

To identify transition risks in the energy supply chain, we introduce two types of shocks and demonstrate how our flow model might be used to assess their impact on energy flow through the supply chain. First, we introduce shocks in the demand for oil, coal, and gas brought on by a transition to a net-zero emissions economy. Second, we consider supply shocks to specific oil, coal, and gas supply regions, which may be a result of supply-side energy policies, such as those motivated by conventional pollution mitigation, water resource conservation, biodiversity and species conservation, land protection for indigenous and community use, or other reasons. As our flow model still requires further tuning to accurately reflect real energy production and transmission, the results shown here are illustrative of the asset-level insights that might be obtained but are not projections.

Demand Shocks

We model the transition to a sustainable energy system as demand shocks. We reduce population centre and power stations energy demand using projections from the IEA sustainable development scenario (International Energy Agency 2019*a*), which is a scenario aligned with the aspirations of the Paris Agreement and Sustainable Development Goals. We use projected final energy demand and energy demand for power stations for 2040. Using the same marginal cost curves as developed in our flow calculation, we then re-calculate energy flow under the first-order effects of the demand shock. The result is an augmented energy flow through the asset-level supply chain. The resulting flows are shown in Figure B.9, B.10, and B.11. The supply and transmission assets with the greatest increase or decrease in energy flow are shown in Table B.5.

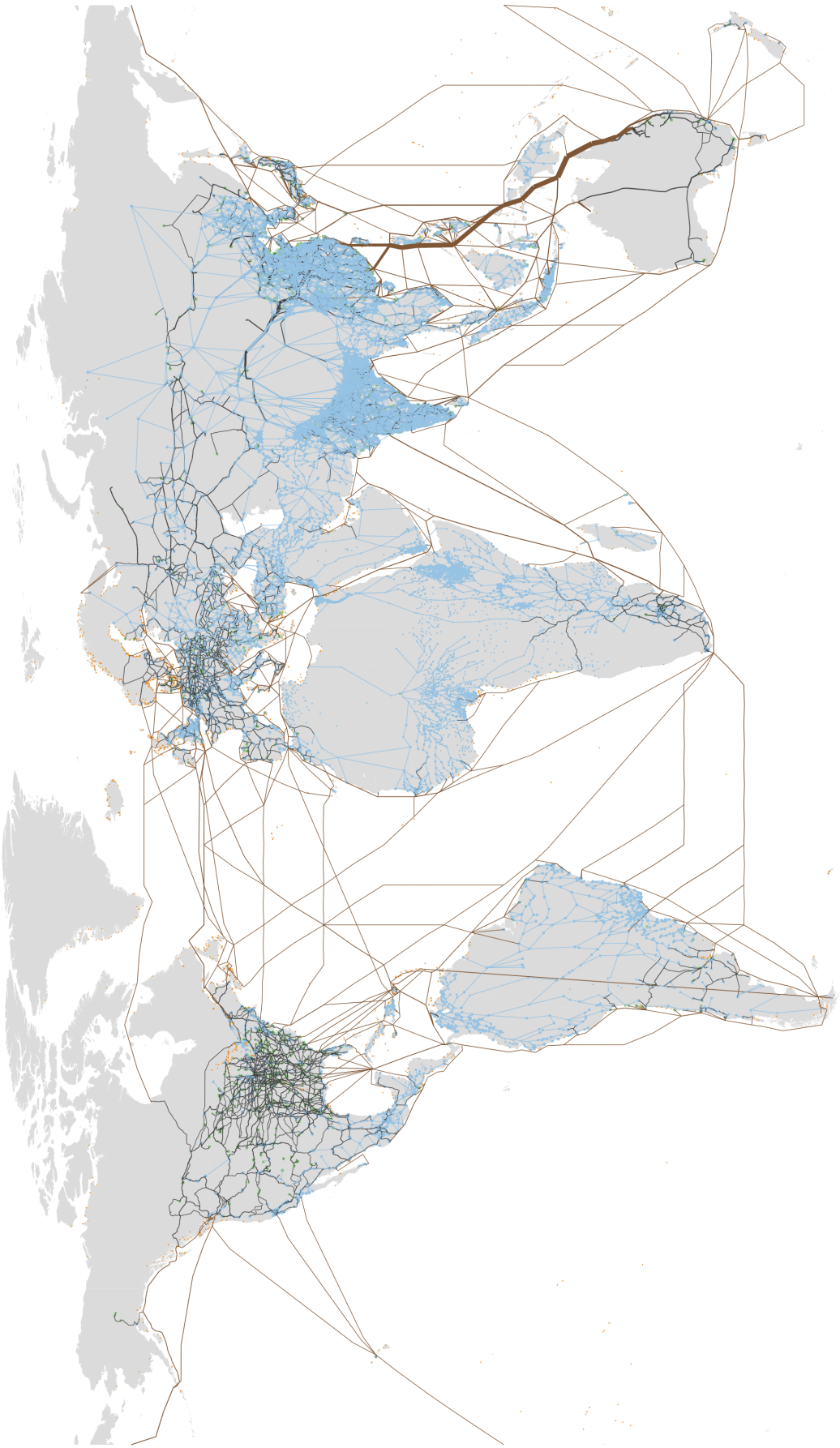


Figure B.3: Global Coal Flow.

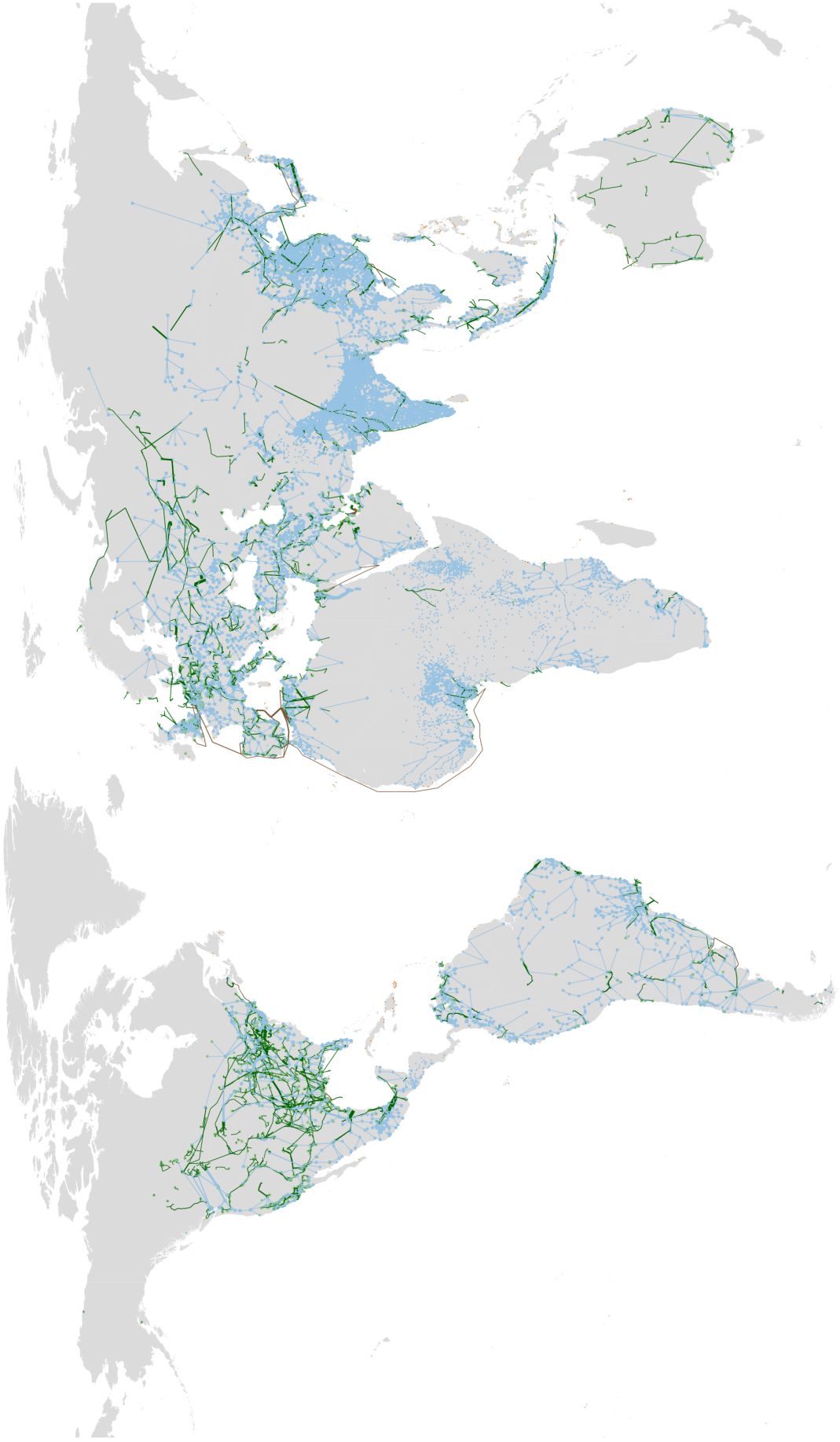


Figure B.4: Global Gas Flow.

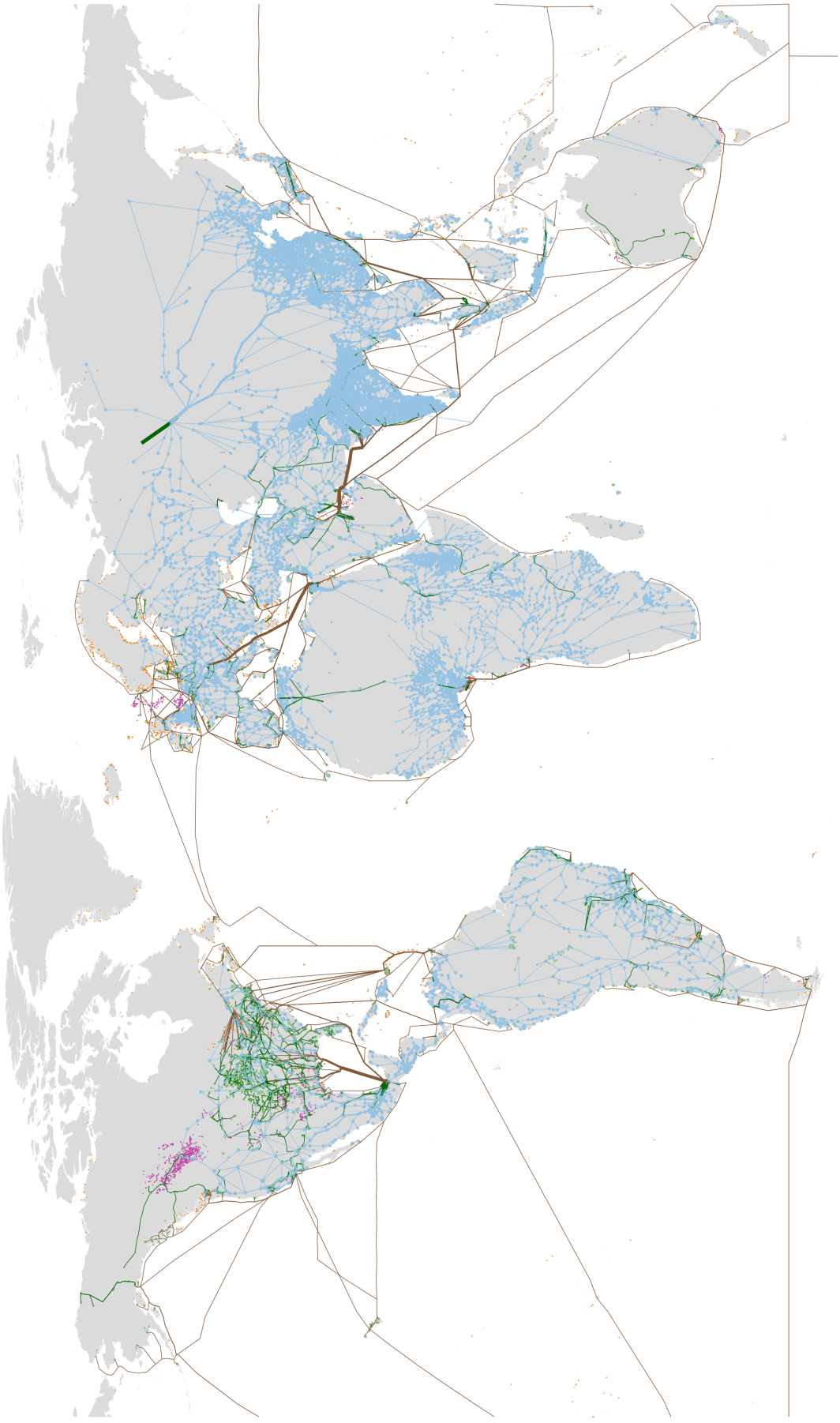


Figure B.5: Global Oil Flow.

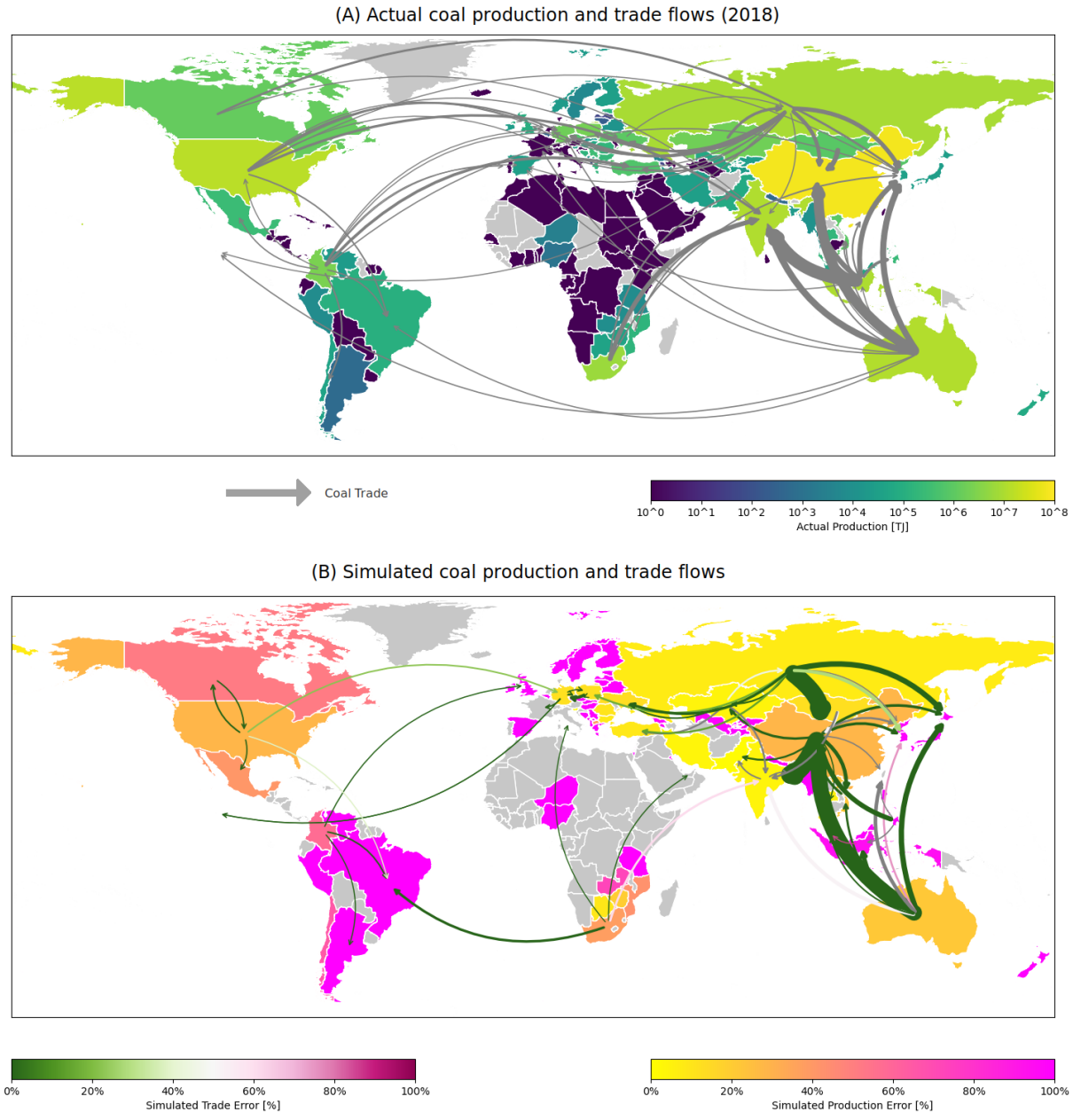


Figure B.6: Coal flow cross-validation with production and trade data.

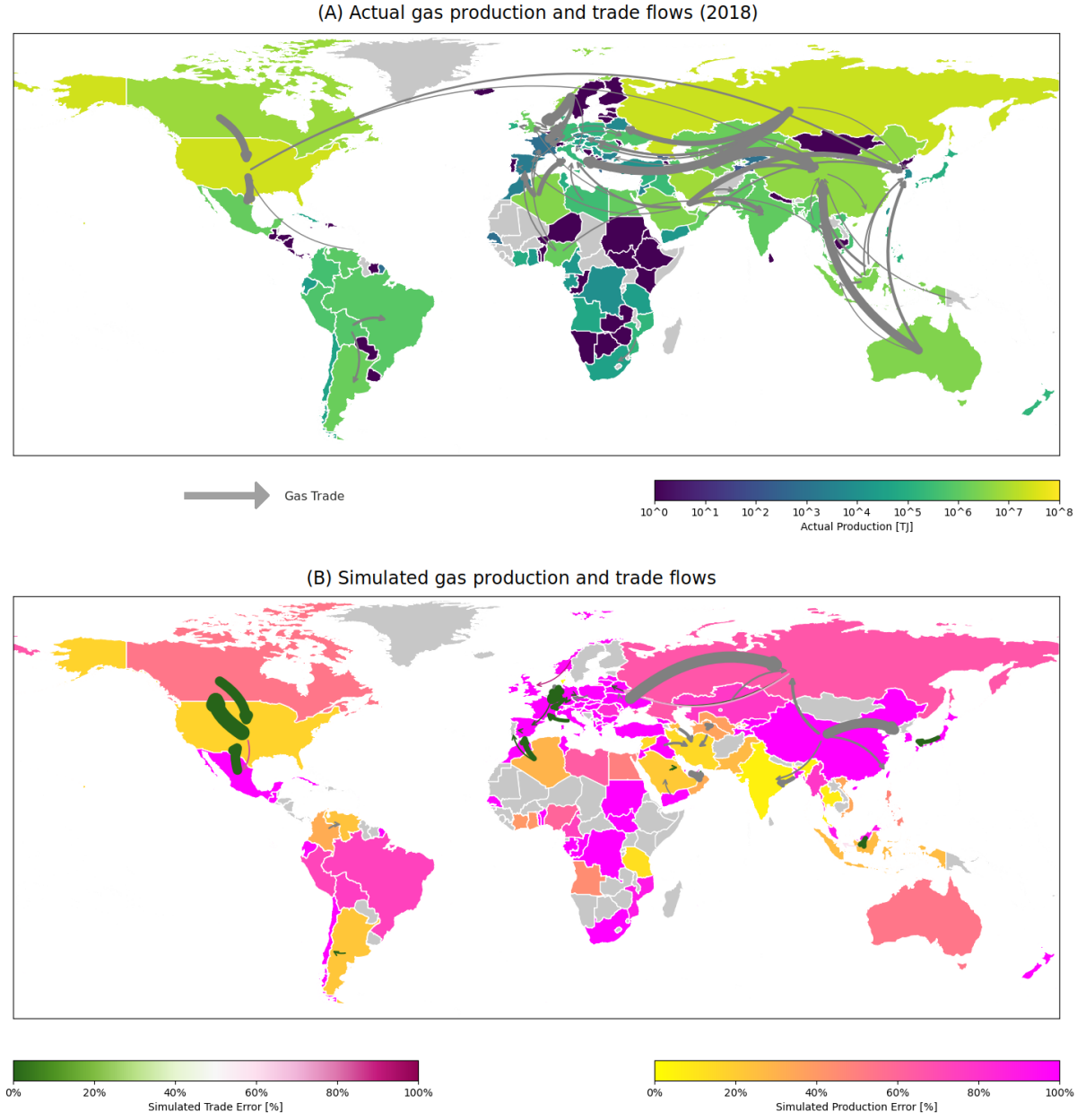


Figure B.7: Gas flow cross-validation with production and trade data.

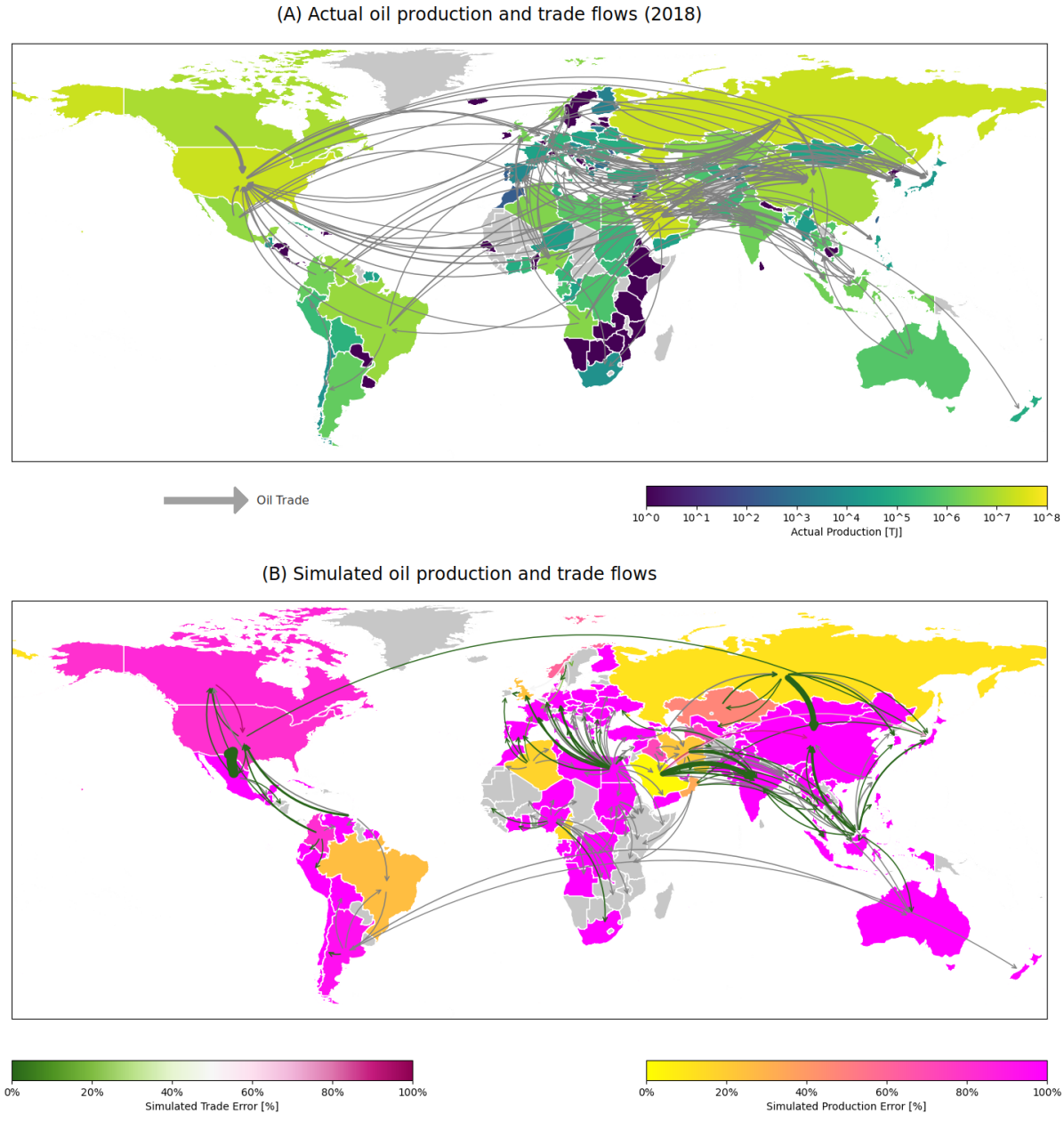


Figure B.8: Oil flow cross-validation with production and trade data.

Table B.5: Demand shock as a result of a transition to a sustainable development scenario. Shown are the most impacted supply nodes, as well as the most impacted transmission edges for each carrier. The nodes are referenced by their regional administrative names and country codes. These results are illustrative only - the model is not yet sufficiently tuned to real data.

Carrier	Largest Flow Increases (% vs baseline, 2040)	Largest Flow Decreases (% vs baseline, 2040)
Coal		
Supply	Bicol Region, PH (262TJ, %690); Queensland, AU (234TJ, %222); Meghalaya, IN (118TJ, %121); Jammu and Kashmir, IN (113TJ, %150); Western Visayas, PH (101TJ, %337)	Shandong, CN (1.19EJ, %51); Inner Mongolia, CN (1.18EJ, %51); Anhui, CN (1.16EJ, %50); Shanxi, CN (1.14EJ, %49); Shaanxi, CN (1.13EJ, %48)
Trans.	Intercity route: Gujranwala-Tamboli-Muridke-Lahore, PK (2.24EJ, %210); Intercity route: Gujrat-Mohla Khurd-Wazirabad, PK (2.17EJ, %226)	Shipping route: Banda Sea (5.35JEJ, %72); Intercity route: Xingyang-Zhengzhou, CN (5.42EJ, %56); Railway route in Shaanxi, CN (6.85EJ, %70)
Oil		
Supply	British Columbia, CA (286TJ, %119); Offshore, CM (11TJ, %106); North Dakota, US (6TJ, %100); Yunlin County, TW (234GJ, %100)	El Menia, DZ (1.59EJ, %54); Eastern Province, SA (2.69EJ, %88); Khanty-Mansi Autonomous Okrug, RU (5.21EJ, %74); Veracruz, MX (8.6EJ, %74); Port Said, EG (11.8EJ, %55)
Trans.	Intercity route: Harrisburg-State College-Buffalo-Niagara Falls (US)-Niagara Falls (CA)-St. Catharines-Toronto, US,CA (1.12EJ - 1.35EJ, %122 - %225)	Port connection: Port Said, EG (10.0EJ, %48); Shipping route: Mediterranean Sea (11.6EJ, %72); Port connection: Pajaritos, MX (12.0EJ, %57)
Gas		
Supply	Veracruz, MX (2.06EJ, %291); Gujarat, IN (1.25EJ, %475); Guangxi, CN (870TJ, %168); Xinjiang, CN (863TJ, %538); Shanghai, CN (858TJ, %151)	Lower Saxony, DE (550TJ, %52); Rhineland-Palatinate, DE (498TJ, %49); Baden-Wurttemberg, DE (396TJ, %39); Modena, IT (595TJ, %59); Mascara, DZ (477TJ, %47)
Trans.	Intercity route: Erie-Buffalo-Niagara Falls (US)-Niagara Falls (CA)-St. Catharines-Toronto, US,CA (2.18EJ - 2.31EJ,%139 - %280); Pipeline connection: Erie, US (2.07EJ, %280)	Intercity route: Alkmaar-Heemskerk-Haarlem, NL (682TJ, %63); LNG Bethioua, DZ (681TJ, %67); Leiden-Rotterdam-Roosendaal(NL)-Antwerp(BE)-Ghent, NL,BE (849TJ - 1.03EJ, %49-%55)

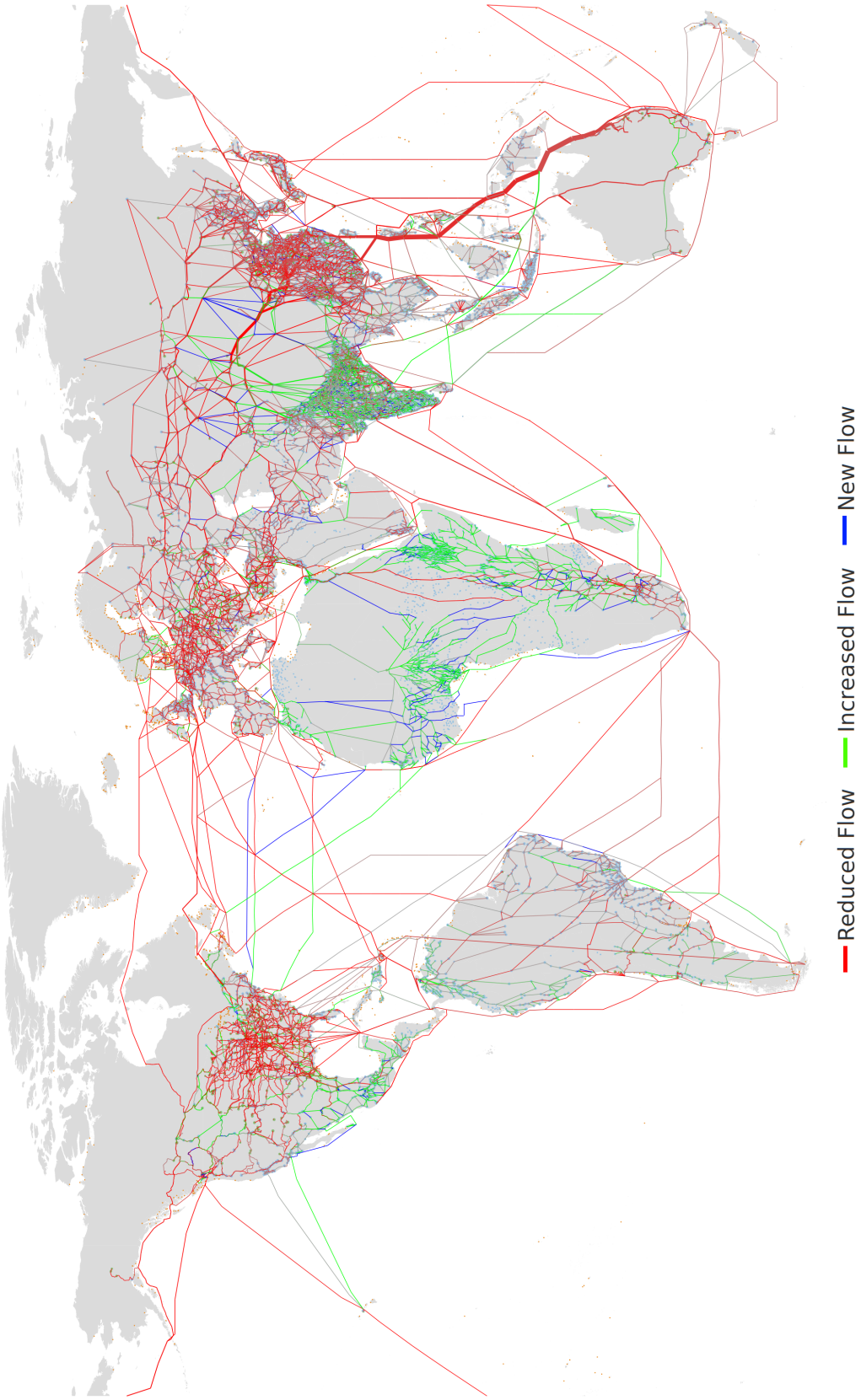


Figure B.9: Coal energy flow under a demand shock to the IEA Sustainable Development Scenario.

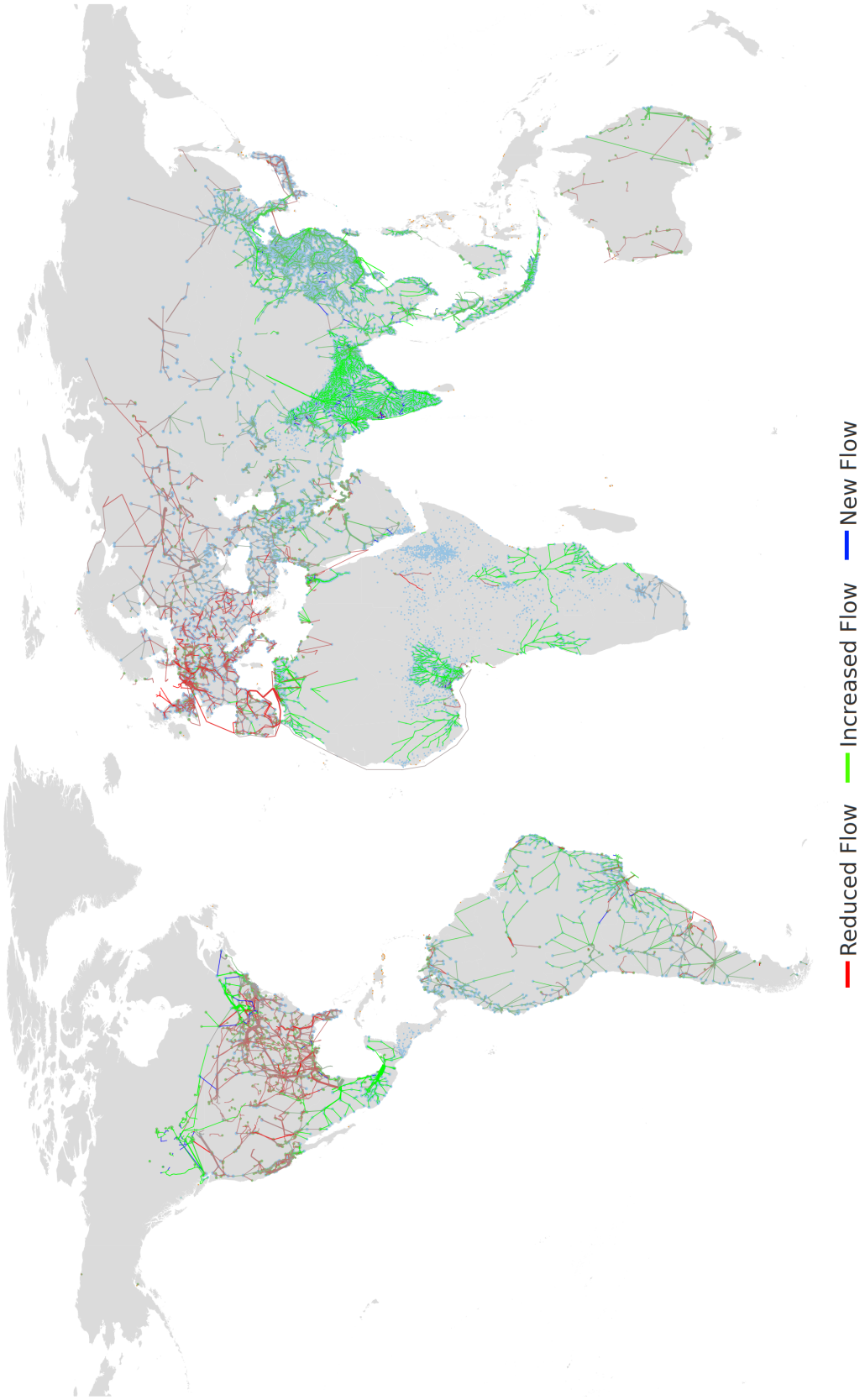


Figure B.10: Gas energy flow under a demand shock to the IEA Sustainable Development Scenario.

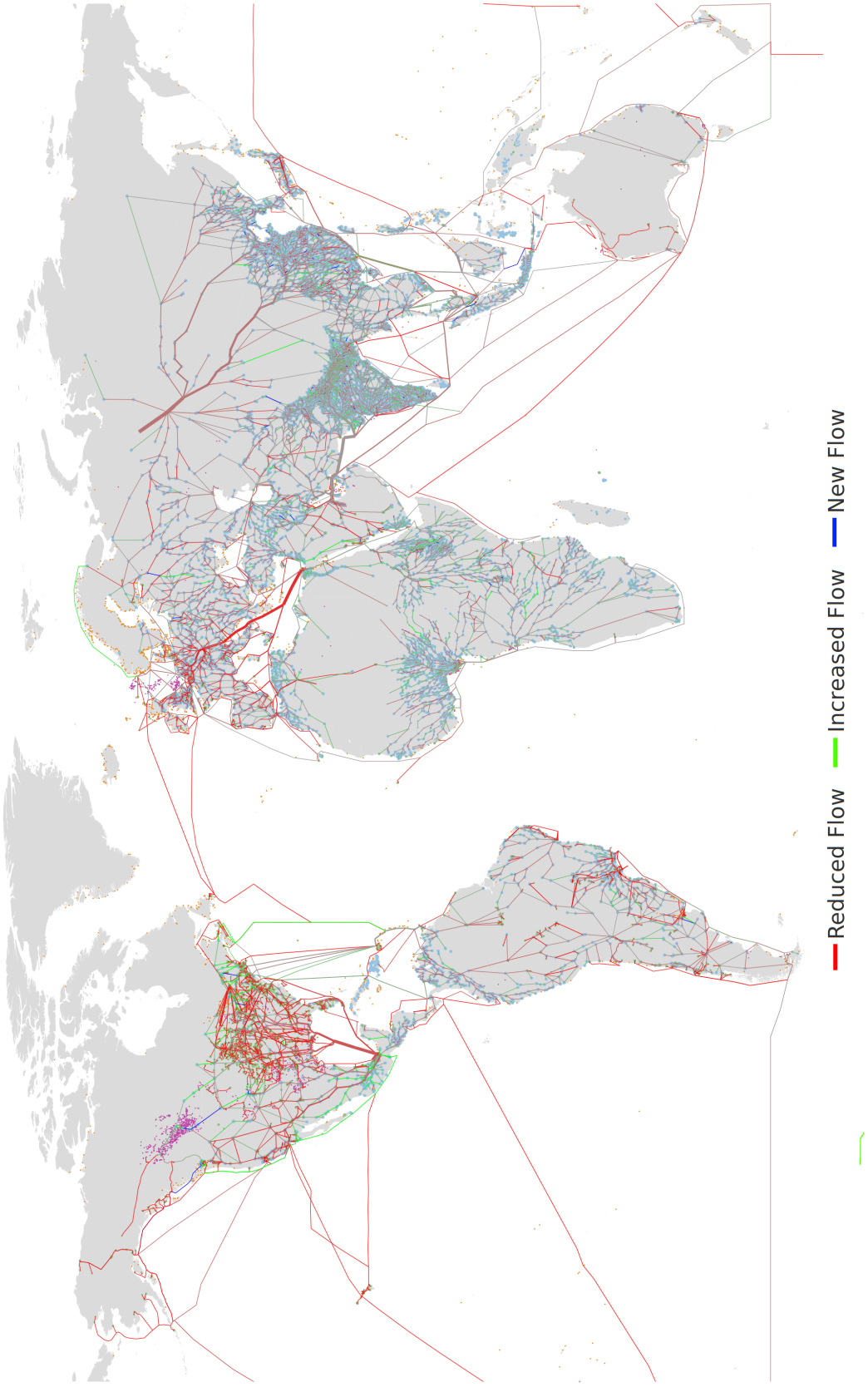


Figure B.11: Oil energy flow under a demand shock to the IEA Sustainable Development Scenario.

While these results are still only illustrative, Figures B.9, B.10, and B.11 show the considerable impact that a transition to a sustainable economy will have on the world's energy transmission infrastructure and the geographies of energy supply, transmission, and demand. Coal supply from Australia, the Philippines, and Indonesia to China and Southeast Asia countries will be significantly impaired, however small opportunities may open with transmission to India. Seaborn coal exports via South Africa may also be significantly impaired, although new overland routes may develop through subsaharan Africa. Coal production, demand, and transmission across the Global North must fall, as well as in China and South America.

For oil and gas routes, our illustrative model shows supply, demand, and transmission reductions across the Global North. Gas transmission across the Global South increases significantly, in line with the projections of the IEA Sustainable Development Scenario. Oil supply and transmission changes are varied across the Global South with the exception of South America.

Further work is required to develop these results from being merely 'illustrative' to being projections of how the world's energy asset network will change with a transition to a sustainable global economy.

Supply Shocks

We also consider shocks to energy supply. These shocks might occur for a number of policy, technology, or economic reasons to which we are agnostic. In identifying where in the supply chain topology a shock has the largest impact, we find *sensitive intervention points* in the transition to a net-zero emissions energy system. While our flow model is reflects only our preliminary assumptions, these results are again illustrative of the use of our model, rather than projections.

In solving for energy flow, we assume supply nodes have convex costs, i.e. marginal costs increase with increased production. We simulate energy supply shocks as the sudden doubling of the quadratic term of an energy supply asset's production cost. We then calculate flow using the new marginal cost curve, and

consider the changes in the new flow versus the baseline. For each energy carrier, we apply this supply shock to each energy source. We sensibly cluster energy source nodes to limit computational requirements. The total increase in supply chain costs, as well as the specific increase for the cities and powerstations most affected, is shown in Table B.6.

While most interdictions only have a small impact on total energy system costs, some cause a considerable change to the energy costs in specific cities or power stations. While climate change mitigation has usually been posed as a problem for demand-side policy making, this shows a method for identifying the broader downstream impacts of supply-side greenhouse gas mitigation policies. Because of the considerable and localised impacts of energy extraction, these policies might also be tied to other policy priorities, such as conventional pollution mitigation, water resource conservation, biodiversity and species conservation, land protection for indigenous and community use, or indeed climate change mitigation. Because these results are merely illustrative, analysis of the specific impacts is unwarranted at this time.

B.7 Future Research Directions

This work is highly extensible to a wide range of energy and climate-related research questions. Future directions for this work should include both improvements to and cross-validation of the energy model as well as further applications for policy insights.

The energy model can be improved by tuning supply curves, capacity factors, and edge impedances to minimise error against aggregate trade and production data. Underlying asset-level data should be improved, especially in geographies where data coverage is poor. Edge impedances and capacities can be tailored to additional ground-truth data rather than estimated. Energy demand can be better allocated to cities and powerstations drawing on true data and measured utilisation rates. Marginal cost curves of energy supply can be either taken from true data or obtained from (better) optimisation of minimum-cost flow. Heirachical improvements to Dijkstra's method's (e.g. Bauer et al. 2008) would facilitate faster minimum-cost

Table B.6: Illustrative supply shock impacts as a result of local policies. Shown are the supply nodes which, if their marginal costs are doubled, cause the largest increase in total network costs. The demand nodes most impacted by each increase are also shown.

Supply Node	Total Cost Increase	Most Impacted Demand Nodes (% cost increase)
Coal		
Barcolod, PH	%0.029	Fuzhou, CN (%0.08); Taizhong Taichung Powerstation, TW (%0.04); Pingnanzhen, CN (%0.05); Jieyang, CN (%0.02); Kaohsiung, TW (%0.03)
Punjab, PK	%0.034	Mailsi, PK (%29.99); Mangowal, PK (%20.35); Katas Raj Dham, PK (%6.26); Peshawar, PK (%7.21); Wan Bhachran, PK (%7.62)
Davao City, PH	%0.039	Taichung, TW (%0.19); Pingnanzhen, CN (%0.07); Liuheng Coal Powerstation, CN (%0.19); Haimen Coal Powerstation, CN (%0.13); Chun'an, CN (%0.20)
Sardegna, IT	%0.06	Los Cristianos, ES (%81.71); ALCUDIA 1 GRUPO N 1 Powerstation, ES (%5.04); BRINDISI NORD Powerstation, IT (%4.11); Torrevaldaliga Nord Powerstation, IT (%1.12); BRINDISI SUD CE Powerstation, IT (%0.83)
Lanao del Norte, PH	%0.073	Taizhong Taichung Powerstation, TW (%0.10); Pingnanzhen, CN (%0.13); Hangzhou, CN (%0.09); Jieyang, CN (%0.05); Shanghai, CN (%0.01)
Oil		
Offshore, TH	%.27	Tam Quan, VN (%17); Cao Lanh, VN (%21); Can Tho Powerstation, VN (%23); Saraburi, TH (%26); Bangkok, TH (%5)
Veracruz, MX	%.32	Duluth, US (%1.4); Spokane, US (%1.0); Miami, US (%2.5); Fargo, US (%1.3); Bismarck, US (%1.3)
Port Said, EG	%.35	Paris, FR (%2.6); Dortmund, DE (%3.0); London, UK (%2.5); Cairo, EG (%14); Benghazi, LI (%5.1)
Offshore, MY	%.53	Sora-myeon, KR (%11); Takasaki, JP (%6.4); Nagoya, JP (%4.7); Osaka, JP (%3.1); Tokyo, JP (%1.8)
Offshore, PH	%.72	Brisbane, AU (%11); Toowoomba, AU (%11); Kaohsiung, TW (%11); Quezon City, PH (%29); New Taipei, TW (%19)
Gas		
Tennessee, US	%0.00	Redcliffe, AU (%0.00); Evansville, US (%0.12); Louisville US (%0.38); Miami, US (%0.05); Atlanta, US (%0.24)
Ohio, US	%0.01	Aurora Powerstation, US (%0.52); Cincinnati, US (%0.26); Indianapolis, US (%0.26); Columbus, US (%0.35); Chicago, US (%0.25)
Pennsylvania, US	%0.01	Saint-Jean-sur-Richelieu, CA (%0.21); Detroit, US (%0.10); Sault Ste. Marie, CA (%0.13); Pittsburgh, US (%0.40); Toronto, CA (%0.54)
Pennsylvania, US	%0.02	Toledo, US (%0.81); Cleveland, US (%0.94); Sault Ste. Marie, CA (%0.39); Toronto, CA (%0.51); Pittsburgh, US (%1.64)
Pennsylvania, US	%0.02	Allentown, US (%2.54); New Bedford, US (%0.71); Philadelphia, US (%0.68); Portland, US (%0.56); New York, US (%0.21)

path discovery in the supply chain, whereby mode transfer between rail/pipeline to seaborn transmission could be treated analogously to mode transfer between town roads and highways in route-finding applications. Faster calculation of least-cost paths would allow iteration on supply curve constants and transmission costs, unlocking non-linear optimisation methods for reducing error between modelled and actual energy production and transmission. As a final test of model quality, it's predictive power should be tested, for example by natural experiments resulting from supply chain disruptions or pricing prescience.

A number of policy and industry questions might then be considered. Treating the whole network as a complex adaptive system, we suggest several avenues for further novel research:

- **Robust net-zero pathway analysis:** What might a robust path to a net-zero energy system look like, considering incumbent infrastructure? Such analysis might necessarily include additional data layers for renewable energy, and modelling dynamic infrastructure pathways (e.g. by selecting nodes and edges for removal according to a set of criteria).

- **Technology transitions to minimise stranded assets:** What opportunities might exist to minimise asset stranding under various transition pathways? Might there be opportunities to switch to other net-zero liquid fuels, for example, that re-use existing infrastructure? Such analysis would include retooling our analysis into a mixed-integer problem.

- **Robust pathways against interdiction:** In the future, the topology of the fossil fuel supply chain might reflect a two-player game between policy makers seeking to mitigate greenhouse gas emissions, and an industry seeking to grow and expand. This is the structure of a network interdiction game, which might give additional insight into robust pathways for contracting the fossil fuel energy system

- **Pricing prescience:** As a test of the prescience of our model and the quality of our data, we can bring pricing data to our data. We would expect metrics from our network topology (e.g. equivalent impedance, betweenness, centrality) to be reflected in commodity pricing, e.g. arbitrage between different oil markets, coal markets, gas, and LNG markets.

A wide range of public, private, and policy outreach possibilities are also enabled by this work.

Data Availability

All data used in this study is open source and freely available. Source datasets can be downloaded here.

Code Availability

All code to reproduce our work is available in repository at https://github.com/Lkruitwagen/global-fossil-fuel-supply-chain.

Contributions

L.K. conceived the project, collected the data, built the energy demand pipeline, designed and built the network preparation pipeline, the flow pipeline, and the analysis pipeline, and drafted the paper. J.K., A.B.L., and JF. joined the project on a 3-week pro-bono engagement from QuantumBlack. J.K. and A.B.L. built the network preparation pipeline. J.F. helped with the paper draft. The authors would also like to thank Aaron Vitarana who joined the project on a 6-week internship from the Ogden Foundation.

B.8 Supplemental Methods

B.8.1 Network Assembly

The computational pipeline for the network assembly is shown in Figure B.12 and takes raw geolocalised asset-level data as input and produces node and edge

dataframes for the complex supply chain as outputs. Here we describe, in detail, the assembly of the energy supply chain graph using asset-level infrastructure data.

The first step, for all infrastructure types, is generic *preprocessing*. This step harmonises data formats, assigning unique identifiers, dropping superfluous data fields, and providing consistent dataframe columns. The second step, *preparation*, provides three key geospatial geometry operations on the harmonised data. Linear assets, i.e. pipelines and railways, need to be broken into individual line segements to ensure a planar arrangement with no self-intersections. These geometries are self-intersected, then 'dissolved' into individual line segments. Line segments are the 'exploded' such that every segment is comprised of two points. The shipping route data is already formatted in this way. Point and polygon assets need to be buffered prior to spatial joining. They are buffered to a distance of 10km, involving a roundtrip reprojection from ellipsoidal latitude-longitude coordinates to local universal transverse mercator (UTM) coordinates where euclidean geometric buffering can be applied. A crucial step at this point is also managing the international dateline. Buffered geometries are split along the international dateline so that supply chains are able to traverse the Pacific ocean.

The *spatial join* step implements this common geospatial analysis method, where linear and buffered datasets are joined according to their geospatial intersection. The spatial joins are performed according to the network arrangement shown conceptually in Figure B.2. The challenge of such large datasets required the implementation of a novel method to allow parallelised reading of geospatial objects in shared memory, achieving scalable performance without the excessive tooling, indexing, and data storage that a database implementation would require. Where multiple linear assets intersect a buffered asset, only the closest is retained. The resulting joined datasets provide the edges between the point and polygon assets and the linear transmission assets.

The *first & last mile* step compensates for missing data. For example, where there is a coalmine in the data but no nearby railways, we assume that this mine must be getting its product to market, and that there must exist an edge between

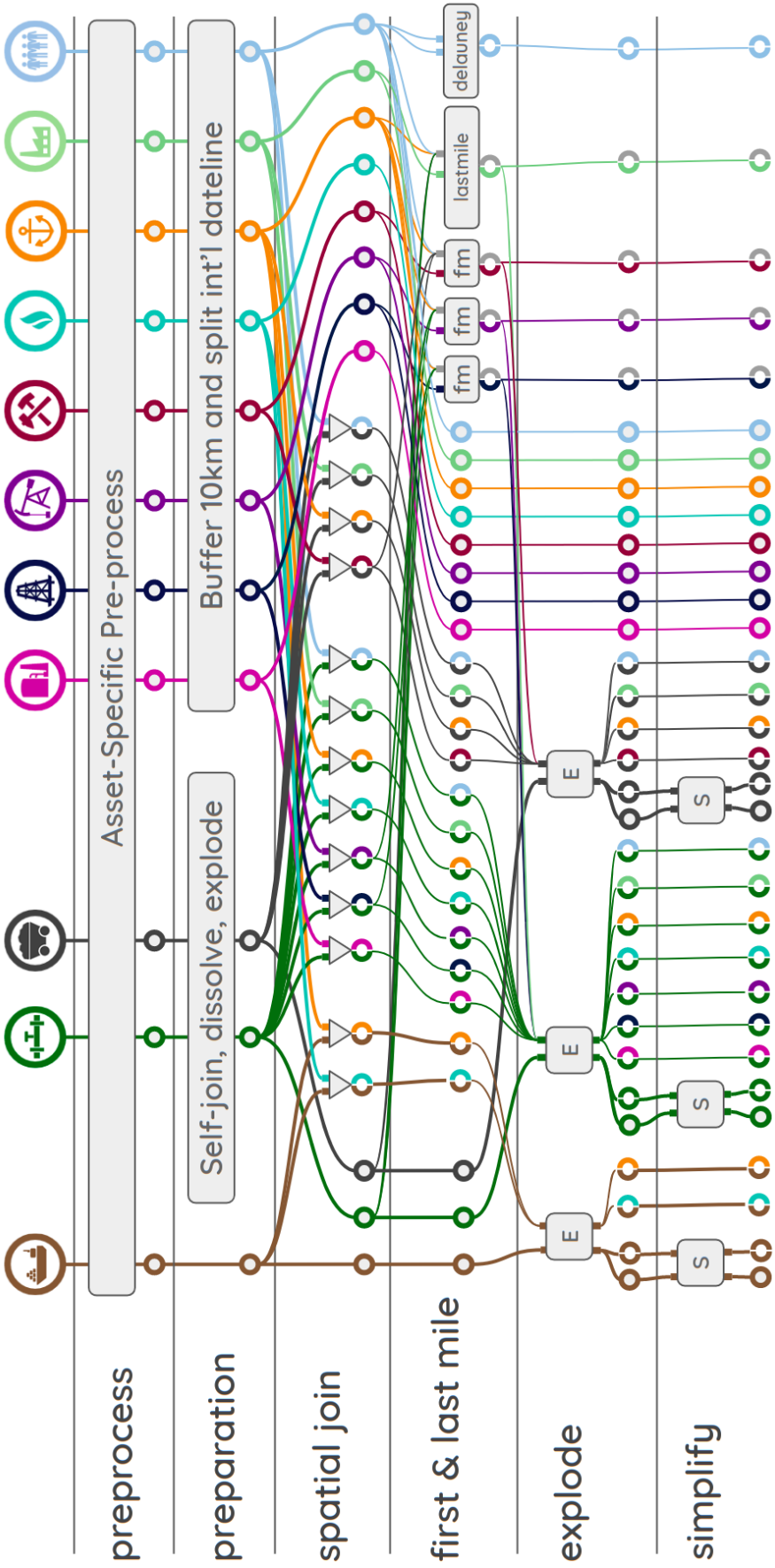


Figure B.12: Directed acyclic computational graph for the network assembly.

the mine and the closest railway, port, or city which is missing from the data. We determine these 'first mile' edges for all coal mines, oil wells, and oil fields which have otherwise not been connected to the supply chain. We also determine 'last mile' connections for unconnected power stations and cities. 'Last mile' edges are made between power stations and ports, cities, and railways or pipelines for coal or oil/gas power stations respectively. We use Delauney triangulation (1934) to produce a 'last mile' web of connections between human settlements. Finally, we also connect any unconnected ports to the nearest shipping route to reflect that the existence of a port necessarily implies a connection to the seaborn supply chain.

The *explode* and *simplify* steps simplify the linear assets to economise downstream calculations. Linear assets linestrings are first 'exploded' around nearest-point intersections with connection edges to other assets. Then graph components with in-degree and out-degree equal to one (i.e. chains of single-input-single-output nodes) are aggregated, summing total geodesic distance along the path. The resulting savings are substanial: 33% of nodes in the coal graph, 27% of nodes in the gas graph, and 72% of nodes in the oil graph.

B.8.2 Flow Calculation

This work models the transmission of energy through the physical infrastructure supply chain by calculating minimum-cost flow through the complex network. Minimum cost flow problems are common in operations research and a number of methods and software packages exist for their solution (Kovács 2015). Because our data lacks cost curve or capacity data, we implement our network with an assumption of convex costs at supply nodes, reflecting the increasing marginal costs of energy supply and Hotelling's rule regarding rent seeking for non-renewable resources (Hotelling 1931). We tune the quadratic coefficient of each supply node's cost taking the total production of that country and dividing it by the number of energy sources in that country. In the absence of known capacity constrains, this prevents, as an example, coal mines in Germany from being the source of

supply for all of Europe, as the marginal costs increase too steeply and make it uncompetitive against lower marginal cost imports.

Because we assume convex marginal costs, our problem formulation is non-linear, and cannot be solved with either conventional mixed-integer linear programming or network simplex methods (Orlin 1997). Network simplex methods for convex costs are known (Végh 2016) but do not have a widely available implementation. We instead use a cost-scaling method (Goldberg & Tarjan 1990) whereby differential quantities are added to the shortest path between each of the demand nodes and the supply 'supersource', with marginal costs updated after each iteration.

Because transmission marginal costs are constant, and edges have infinite capacity, we only need to find the minimum cost path between every supply and demand node once. For this we use Dijkstra's algorithm (Dijkstra et al. 1959). Substantial improvements are available for improving this step, see Future Research Directions.

We use our cost-scaling method to provide develop the policy and counterfactual analysis. For the policy interventions we model as a demand shock, we change solve for minimum-cost flow with augmented node demands. For the supply interdictions, we double the quadratic coefficient of the marginal cost of the supply node being interdicted.

C

Towards *DeepSentinel*

An extensible corpus of labelled Sentinel-1 and -2 imagery and a general purpose sensor-fusion semantic embedding model

Kruitwagen, L.,[1,*]

[1] University of Oxford, Smith School of Enterprise and the Environment
[*] Corresponding Author

Contents

C.1 Earth observation for climate change mitigation 147
C.2 Sensor fusion 148
C.3 Data preparation 152
C.4 Experimentation 153
C.5 Results 158
C.6 Application 161
C.7 Discussion 161
C.8 Conclusions 163

Abstract

Earth observation offers new insight into anthropogenic changes to nature, and how these changes are effecting (and are effected by) the built environment and the real economy. With the global availability of medium-resolution (10-30m) synthetic aperture radar (SAR) Sentinel-1 and multispectral Sentinel-2 imagery, machine learning can be employed to offer these insights at scale, unbiased to the reporting of companies and countries. In this paper, I introduce *DeepSentinel*, a data pipeline and experimentation framework for producing general-purpose semantic embeddings of paired Sentinel-1 and Sentinel-2 imagery. I document the development of an extensible corpus of labelled and unlabelled imagery for the purposes of sensor fusion research. With this new dataset I develop a set of experiments applying popular self-supervision methods and encoder architectures to a land cover classification problem. Tile2vec spatial encoding with a self-attention enabled ResNet model outperforms deeper ResNet variants as well as pretraining with variational autoencoding and contrastive loss. As a final test of generaliseability, the top-performing encoder is further fine-tuned on a small dataset of the world's mining sites and their commodities, obtaining a classifier to detect coalmines. This classifier is deployed on imagery obtained for global mining dataset to produce a dataset of the world's coal mines. All supporting and derived data and code are made publicly available.

C.1 Earth observation for climate change mitigation

Satellite-based earth observation plays a central role in measuring climate change impacts and risks (Caldecott, Kruitwagen, McCarten, Zhou, Lunsford, Marchand, Hadjikyriakou, Bickel, Sachs & Bohn 2018). Medium-resolution satellites (10-30m spatial resolution), despite being initially designed for environmental monitoring, are being increasingly used in applications focusing on the interface between the environment and the real economy for the purposes of financial risk measurement: estimating carbon dioxide emissions (CTI 2018), methane emissions (Varon et al. 2020), and localising large fixed-capital assets (Kruitwagen et al. 2019). For the purposes of assessing financial risk due to climate change, these satellites provide the globally exhaustive and unbiased view required by financial decision makers. Deploying analysis at this global scale can only be accomplished with the use of machine learning.

These climate change risk applications are impaired by two perennial challenges with satellite-based earth observation: the presence of atmospheric interference, and a shortage of training labels. Atmospheric interference is not equally distributed around our planet. Excessive cloud cover makes surface retrievals using multispectral imagery very challenging in certain geographies, negating its otherwise 'exhaustive' coverage. Many cloudy geographies are in the global south, precisely where populations are the most vulnerable to climate change and where conventional reported data is the most sparse. These same geographies are where financial institutions in the global north concentrate their risk exposure to generate outsize returns, and are where civil society groups must be most vigilant to identify and respond to neocolonial practises by these same institutions.

The shortage of training labels for machine learning with earth observation data is well documented. Land use and land cover data is available with moderate spatial and temporal resolution in Europe (e.g. the Copernicus Corine Land Cover 2020a) and the US (e.g. the USDA Cropland Data Layer 2020b), however data for the rest of the planet are sparse (e.g. OpenStreetMaps 2020). These datasets include

only broad categories of land use and land cover, and are not fit for the purpose of localising specific categories of industrial infrastructure. Financial institutions just beginning to reckon with geospatial data, leaving a large gap in existence and availability of spatially-localised asset-level data (Burks 2020).

I develop a general-purpose sensor fusion semantic embedding model to overcome these dual challenges of atmospheric interference and label availability. *DeepSentinel* uses sensor fusion of SAR Sentinel-1 data and multispectral Sentinel-2 data to provide latent space embeddings of surface conditions even in excessively cloudy conditions. Self-supervised pretraining followed by fine tuning on land-use and land-cover data creates a general purpose embedding model suitable for a wide range of downstream applications using transfer learning, see Figure C.1. This paper provides a background on the current state-of-the-art in sensor fusion, documents a data corpus for training purposes, and describes experimentation and initial results towards *DeepSentinel*, a general-purpose embedding model.

C.2 Sensor fusion

SAR imagery carries complementary information to multispectral imagery and so is useful beyond its penetration of atmospheric conditions. Sentinel-1 SAR C-band backscatter is sensitive to moisture and surface types, and has applications in both natural environments (e.g. with classification of forests (Rüetschi et al. 2018) and crop types (Nelson et al. 2014, Setiyono et al. 2017, Ndikumana et al. 2018)), and the built human environment (e.g. road classification (Zhang, Kong, Zhang, You, Wei, Sun & Li 2019)). These properties are complimentary to the multispectral data provided by Sentinel-2, which has sensors designed to detect aerosols, water vapour, chlorophyll, and visual spectra.

Sensor fusion describes the complementary combination of data from multiple sources to improve inference quality beyond what would be possible from either sensor individually. Fusion of Sentinel-1 and Sentinel-2 imagery has been studied for cloud removal (e.g. Eckardt et al. 2013, Grohnfeldt et al. 2018, Bermudez et al. 2018, Meraner et al. 2020), synthetic imagery generation (e.g. He & Yokoya 2018,

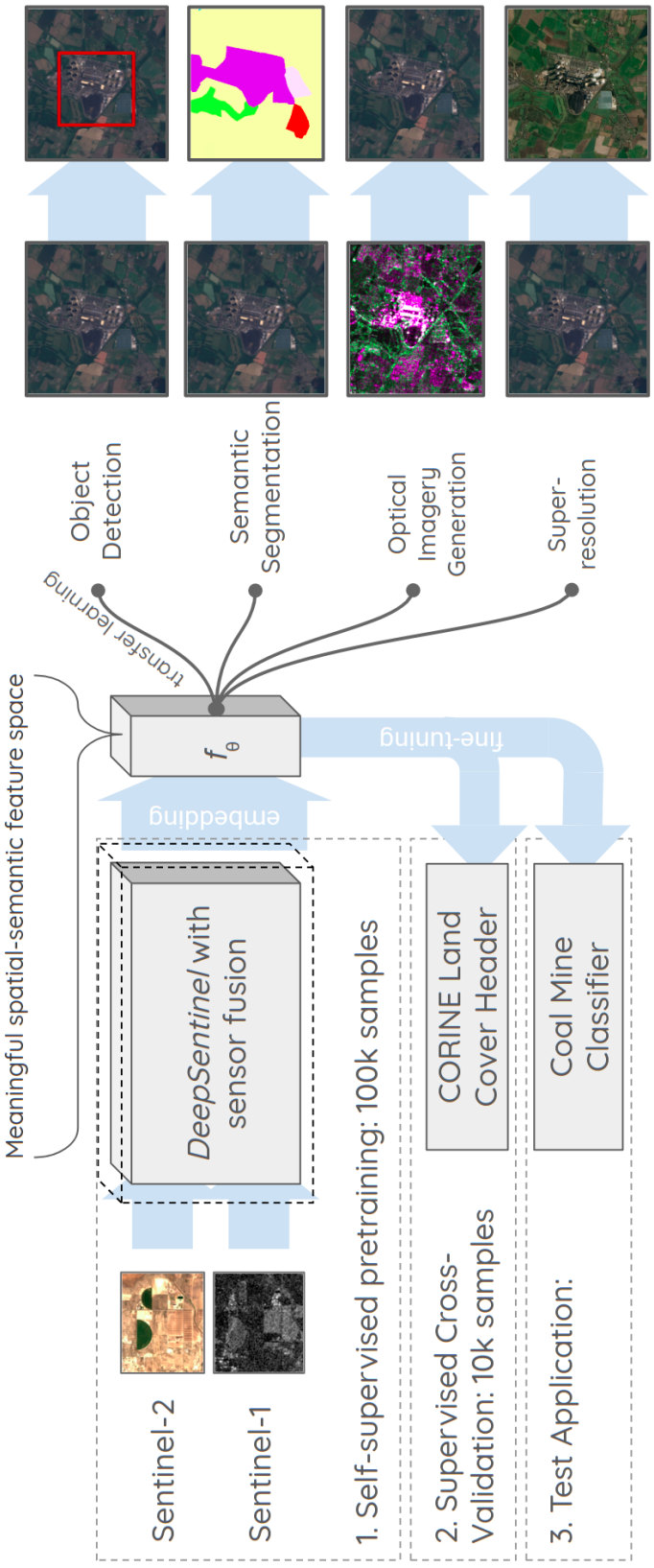


Figure C.1: *DeepSentinel* summary, showing pretraining and fine-tuning curriculum and the variety of use case applications.

Bermudez et al. 2019, Fuentes Reyes et al. 2019), and land cover classification (e.g. Van Tricht et al. 2018, Ferrant et al. 2017, Torbick et al. 2018).

With the exception of Meraner et al. (2020), most of these studies obtain a limited corpus for a specific area of interest. With *DeepSentinel*, we propose a general-purpose encoder than can be used for any land-surface area-of-interest on the planet, and fine-tuned for any of the applications above. We prepare an unlabelled training dataset of similar size to Meraner et al. (2020), see Table C.1. Our dataset differs in that it is provided corrected for atmosphere conditions (but not clouds), and so shows the best-available surface reflectance retrievals. It has not been generated with any bias against cloud cover and has been uniformly sampled over the planet's land surface area, so it represents preprocessed analysis-ready data from the natural distributions of the two sensing platforms. The data have been generated twice, sampling from both Google Earth Engine and DescartesLabs imagery access pipelines, making our results as universally accessible as possible. Our goal is to unlock increasingly niche applications where limited training data is available, allowing a proliferation of earth observation use cases with impact analogous to the release of pre-trained conventional imagery convolutional neural networks of the mid-2010s (e.g. ResNet, VGG-16/19, etc).

DeepSentinel includes a dataset with labels obtained from two sources: land cover labels from Copernicus CORINE Land Cover (2020*a*), and polygon, line, and point data from OpenStreetMaps (OSM). (2020) Copernicus CORINE Land Cover data was rasterised from geodatabases with less than 100m spatial resolution. This dataset was geographically restricted to the EU27+GB, where the land cover data was available and where OSM data is of significantly higher quality. While the maximum resolution of the geodatabases is coarser than the spatial resolution of imagery, rasterising directly from the geodatabases rather than use derived 100m raster products means the actual resolution of the land cover samples is considerably higher (although not consistent across the EU27+GB countries). In this sense these are 'weak' labels which do not confidently match the true land cover at each location.

A sample of the labelled data is shown in Figure C.2.

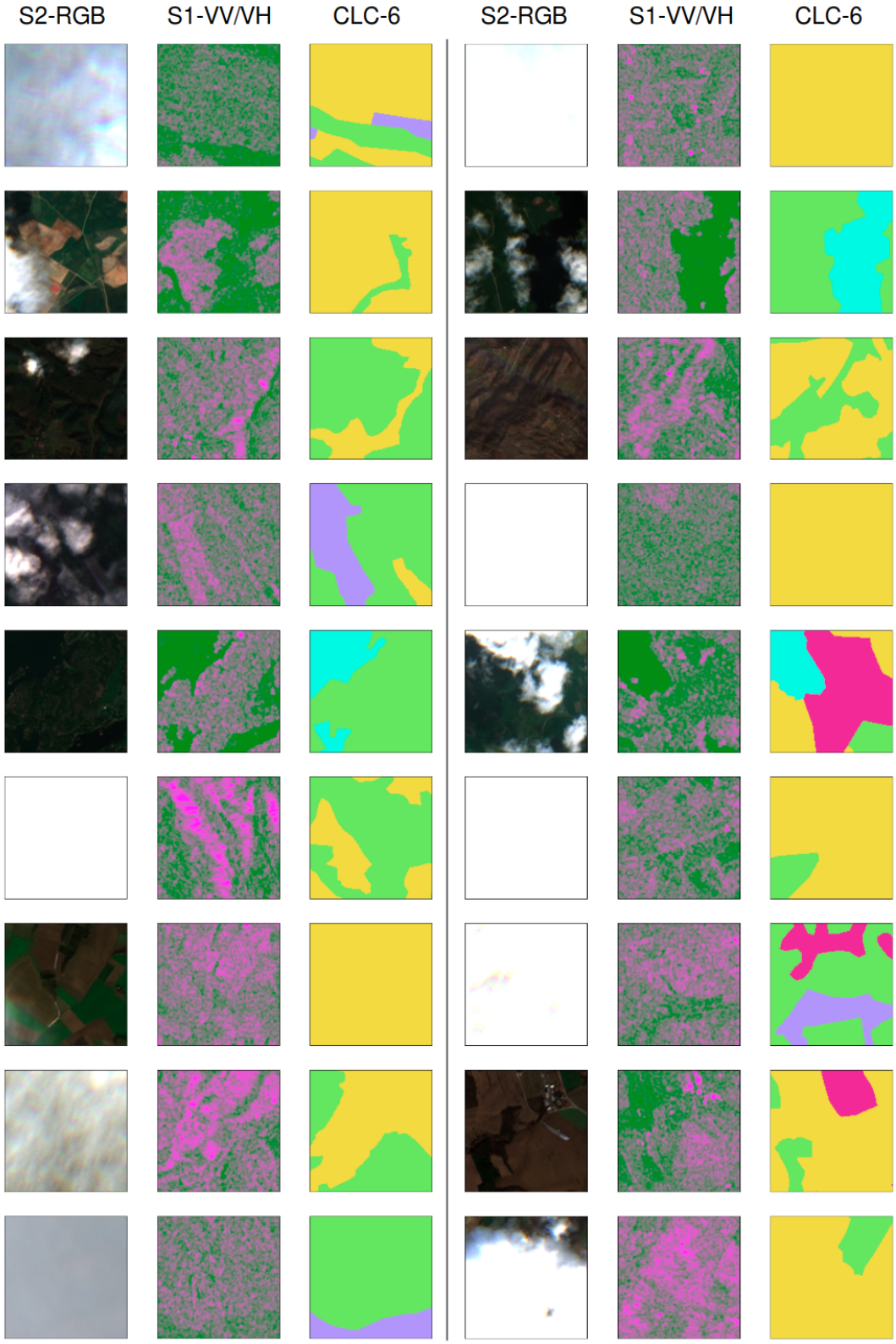


Figure C.2: Samples from the training corpus. S2-RGB: Sentinel-2 red, green, and blue channels; S1-VV/VH: Sentinel-1 synthetic RGB composed of VV and VH channels; CLC-6: 6-class Copernicus CORINE land cover classification. Blank Sentinel-2 RGB images area not missing data - they are cloudy images where the brightness saturates all pixel values.

Table C.1: Sensor Fusion Datasets

Dataset (Study)	Components*	Pixels	Resolution	Geography
DeepSentinel labelled (ours)	S1 (VV+VH) + S2 (L2A - all bands) + CCLC + OSM	655 Mpx	10m	EU27+GB
SEN12MS-CR (Meraner et al. 2020)	S1 (VV+VH) + S2 (L1C - all bands)	10,323 Mpx	10m	169 Global ROIs
DeepSentinel unlabelled (ours)	S1 (VV+VH) + S2 (L2A - all bands)	6,554 Mpx * 2	10m	Global Uniform

* S1: Sentinel-1; S2: Sentinel-2; CCLC: Copernicus CORINE Land Cover; OSM: OpenStreetMaps; L1C: Level 1-C; L2A: Level 2-A

C.3 Data preparation

Two novel datasets are prepared for the development of *DeepSentinel*. The first dataset is prepared without labels for the purpose of self-supervised pretraining. Random sample patches are obtained from the planet's land surface area. The square patches are sampled at a pixel resolution of 10m with 256-pixel side length. Patches are obtained for both Sentinel-1 and -2, only where the image acquisition dates are within 3 days of each other. For Sentinel-2, all 12 multispectral bands are sampled. For Sentinel-1, VV and VH polarisation bands of interferometric wide swath (IW) retrievals are sampled, the retrieval mode used over land. For maximum reproduceability and impact, samples for all patches are obtained from both Google Earth Engine (2017) and Descartes Labs computation platform (2017). Sentinel-2 acquisitions from Google Earth Engine are provided at 'Level 2A' surface reflectance level, from Descartes Labs they are obtained at 'surface-level' using Descartes Labs proprietary atmosphere correction algorithm. Sentinel-1 acquisitions from Google Earth Engine have been thermal noise corrected, radiometrically calibrated, and terrain corrected. Sentinel-1 acquisitions from Descartes Labs have been similarly terrain corrected.

The second dataset is prepared with labels for the purposes of fine-tuning and cross-validation. Land use and land cover labels are obtained from the 2018 Copernicus CORINE Land Cover inventory, rasterised at 10m, as well as OSM Point, Line, and Polygon datasets. For the second dataset, samples are only obtained

for European Union (plus the United Kingdom) where Copernicus CORINE Land Cover data are available and OSM data is of substantially higher quality. The two datasets are made publicly available via Google Cloud Storage and Microsoft Azure Storage. Four datasets have been prepared: a demonstration 1,000 sample dataset with CORINE and OSM land use and land cover labels, a 10,000 sample dataset without labels, a 10,000 dataset with CORINE and OSM land use and land cover labels, and a 100,000 sample dataset without labels.[1] The code for sampling the earth observation, land cover, and OSM data is available via Github.[2] With 100,000 unlabelled samples, our dataset is similar in size to the largest dataset available for sensor fusion research (Meraner et al. 2020).

C.4 Experimentation

The potential of DeepSentinel is demonstrated with a set of experiments producing semantically-meaningful features embeddings from Sentinel-1 and -2 imagery. The ultimate goal of DeepSentinel is to be able to take advantage of the abundant unlabelled imagery and limited or weak imagery labels to produce a general-purpose embedding machine prior to fine-tuning on specific end-use applications. Such a system features two primary design decisions: the choice of self-supervision training curriculum, and a choice of embedding architecture. This paper presents experimentation and results for three self-supervision algorithms and three embedding architectures to see which produces the best image representations for land cover classification and coal mine classification.

The three self-supervision approaches used in this paper are variational autoencoding (Kingma & Welling 2013), tile2vec (Jean et al. 2019) spatial encoding, and contrastive learning (Oord et al. 2018, Tian et al. 2019). These three methods follow different intuitions in the design of self-supervision criteria. Variational autoencoding combines the information compression of a standard autoencoder with generative

[1] Datasets are accessible at `https://console.cloud.google.com/storage/browser/deepsentinel`

[2] *DeepSentinel* code is available at `https://github.com/Lkruitwagen/deepsentinel.git` and the OSM server at `https://github.com/Lkruitwagen/deepsentinel-osm.git`

sampling of a latent space to ensure continuous and semantically-meaningful embeddings. Both tile2vec and contrastive learning seek to minimise euclidean distances of latent space representations of 'alike' samples, while maximising the distance between the representations of 'distant' samples. With tile2vec 'alike-ness' is dictated by geospatial proximity. With contrastive learning, 'alike-ness' is determined by the generation of different randomly-augmented views of the same sample. Experimenting with all three of these methods will show which is most efficient at encoding high-level feature information in fused Sentinel-1 and -2 imagery.

A variational autoencoder (VAE) is prepared which encodes samples from the unlabelled dataset and generates a reconstructed image. The autoencoder is trained using reconstruction loss – the mean-squared error between the 14-channel generated image and the input image. Several variations are made to the architecture which draws on Ha & Schmidhuber's (2018) implementation. Rather than flattening the entire embedding space, dimension reduction is achieved by first using 1x1 convolutional kernels. This reduces embedding space dimensions while maintaining spatial relationships. Deconvolution of the generated latent vector uses transposed convolution followed by two standard convolution-batch-normalisation-rectified-linear-unit blocks (Table C.3).

Tile2vec (Jean et al. 2019) (T2V) is a spatial analog to word2vec (Mikolov et al. 2013) and functions on the (Tobler-ian)[3] assumption that objects that are nearby are more alike than objects that are distant. Within the unlabelled dataset, for each point, the geodesic distance to the nearest neighbours that are within 1^o is obtained. The average distance between points ('anchors') and their closest neighbour is 18.2km. Training uses triplet loss, where for each anchor point, the euclidean distance between the latent space representation of the anchor and neighbour points is minimised, and distance between the anchor and distant point representations is maximised. Neighbour points are chosen from the neighbours within 1^o, with probability equal to the softmax of their distances to the anchor point. Distant

[3]Tobler's First Law of Geography (Tobler 1970): "everything is related to everything else, but near things are more related than distant things."

points are chosen from the remaining non-neighbour population. T2V uses the embedding model directly and has no other learned parameters.

The final self-supervision approach uses contrastive learning. In contrastive learning, the same underlying latent space representation may have multiple possible views in the data space. Adapting Contrastive Sensor Fusion (CSF) from Swope et al. (2020), paired image views are prepared where each view has been augmented with band dropout, random cropping, and contrast, brightness, saturation, and hue jitter. A loss function is defined as the euclidean distance between the embedded representations of these augmented views. To avoid a null solution, the distance between other randomly selected samples should also be maximised, and the ensemble is trained using triplet loss as with tile2vec. As in Swope et al. (2020), a learning curriculum is employed whereby augmentation intensifies over a period of 10 epochs following a 'warm-up' phase of 5 epochs.

All pretraining methods are trained on the 100,000 sample unlabelled dataset of matched Sentinel-1 and Sentinel-2 imagery. The imagery bands are concatenated at model input and are normalised using the band-wise means and standard deviations. Pretraining for all methods is run for 20 epochs with a maximum batch size of 1024 on a cluster of eight NVidia k80 GPUs. Adam optimisers are used for training, with an initial learning rate of $5e{-}4$.

The challenge with self-supervision methods is the development of an evaluation criteria suitable for a target use case. This is why *DeepSentinel* is made available with a second dataset which includes land cover labels for the EU27+GB. Pretrained encoders are finetuned using the labelled data in a semantic segmentation land cover classification task. Pixels are classified into the categories shown in Table C.10. Cross-validation performance on the fine-tuning task is also used to optimise the self-supervised pretraining method and to develop insight into the representations the encoders are learning.

Three well-studied variants of the ResNet architecture (He et al. 2015) are used to develop insight into favourable encoder properties. ResNet18 and ResNet34 are topologically similar and differ only in their depth and number of parameters. The

Table C.2: Convolutional Block Architecture (DownBlock) (He et al. 2015)

Layer	Input	Operation	Output Shape
Conv1	INPUT [B,C,X,Y]	Convolution2D	[B,C*2,X/2,Y/2]
BatchNorm1	Conv1	Batch Normalisation	[B,C*2,X/2,Y/2]
ReLu1	BatchNorm1	Rectified Linear Activation	[B,C*2,X/2,Y/2]
Conv2	ReLu1	Convolution2D	[B,C*2,X/2,Y/2]
BatchNorm2	Conv2	Batch Normalisation	[B,C*2,X/2,Y/2]
ReLu2	BatchNorm2	Rectified Linear Activation	[B,C*2,X/2,Y/2]
Out	[INPUT,ReLu2]	Sum	[B,C*2,X/2,Y/2]

Table C.3: Deconvolutional Block Architecture (UpBlock) (He et al. 2015)

Layer	Input	Operation	Output Shape
DeConv1	INPUT [B,C,X,Y]	Transpose Convolution2D	[B,C/2,X*2,Y*2]
Conv1	DeConv1	Convolution2D	[B,C/2,X*2,Y*2]
BatchNorm1	Conv1	Batch Normalisation	[B,C/2,X*2,Y*2]
ReLu1	BatchNorm1	Rectified Linear Activation	[B,C/2,X*2,Y*2]
Conv2	ReLu1	Convolution2D	[B,C/2,X*2,Y*2]
BatchNorm2	Conv2	Batch Normalisation	[B,C/2,X*2,Y*2]
ReLu2	BatchNorm2	Rectified Linear Activation	[B,C/2,X*2,Y*2]

difference in performance between the two models gives insight into the parameter spaces required to encode deep semantic information. The third embedding architecture is a ResNet variant incorporating context-sensitive attention (Vaswani et al. 2017*b*). Attention used in this way transports broader contextual information into the convolutional perceptive field. The three embedding architectures are shown in Tables C.5, C.6, C.7, with reference to Tables C.2, C.3, C.4, and C.8.

Table C.4: Attention Block Architecture (AttnBlock) (Zhang, Goodfellow, Metaxas & Odena 2019)

Layer	Input	Operation	Output Shape
QueryConv	INPUT [B,C,X,Y]	Convolution2D$_{1x1}$	[B,C/8,X,Y]
KeyConv	INPUT [B,C,X,Y]	Convolution2D$_{1x1}$	[B,C/8,X,Y]
ValConv	INPUT [B,C,X,Y]	Convolution2D$_{1x1}$	[B,C,X,Y]
Energy	[QueryConv^T, KeyConv]	Matrix Multiplication	[B,C/8,X,Y]
AttentionMap	Energy	Softmax	[B,C,X,Y]
PreOut	[ValConv, AttentionMap]	Matrix Multiplication	[B,C,X,Y]
Out	[PreOut, INPUT]	Sum	[B,C,X,Y]

Table C.5: ResNet18 (He et al. 2015)

Layer	Input	Operation	Output Shape	Parameters
Conv1	INPUT [B, 14, 128, 128]	Convolution2D	[B, 64, 64, 64]	43,904
BatchNorm1	Conv1	Batch Normalisation	[B, 64, 64, 64]	128
ReLu1	BatchNorm1	Rectified Linear Activation	[B, 64, 64, 64]	0
Layer1	ReLu1	DownBlockx2 (Table C.2)	[B, 64, 64, 64]	147,960
Layer2	Layer1	DownBlockx2 (Table C.2)	[B, 128, 32, 32]	525,568
Layer3	Layer2	DownBlockx2 (Table C.2)	[B, 256, 16, 16]	2,099,712
Layer4	Layer3	DownBlockx2 (Table C.2)	[B, 512, 8, 8]	8,393,728
Out	Layer4	Softmax	[B, 512, 8, 8]	0
			Trainable Parameters:	**11,211,008**

Table C.6: ResNet34 (He et al. 2015)

Layer	Input	Operation	Output Shape	Parameters
Conv1	INPUT [B, 14, 128, 128]	Convolution2D	[B, 64, 64, 64]	43,904
BatchNorm1	Conv1	Batch Normalisation	[B, 64, 64, 64]	128
ReLu1	BatchNorm1	Rectified Linear Activation	[B, 64, 64, 64]	0
Layer1	ReLu1	DownBlockx3 (Table C.2)	[B, 64, 64, 64]	221,952
Layer2	Layer1	DownBlockx4 (Table C.2)	[B, 128, 32, 32]	1,116,416
Layer3	Layer2	DownBlockx6 (Table C.2)	[B, 256, 16, 16]	6,822,400
Layer4	Layer3	DownBlockx3 (Table C.2)	[B, 512, 8, 8]	13,114,368
Out	Layer4	Softmax	[B, 512, 8, 8]	0
			Trainable Parameters:	**21,319,168**

Table C.7: ResNet18 with Attention (He et al. 2015, Zhang, Goodfellow, Metaxas & Odena 2019)

Layer	Input	Operation	Output Shape	Parameters
Conv1	INPUT [B, 14, 128, 128]	Convolution2D	[B, 64, 64, 64]	43,904
BatchNorm1	Conv1	Batch Normalisation	[B, 64, 64, 64]	128
ReLu1	BatchNorm1	Rectified Linear Activation	[B, 64, 64, 64]	0
Layer1	ReLu1	DownBlockx2 (Table C.2)	[B, 64, 64, 64]	221,952
Layer2	Layer1	DownBlockx2 (Table C.2)	[B, 128, 32, 32]	1,116,416
Layer3	Layer2	DownBlockx2 (Table C.2)	[B, 256, 16, 16]	6,822,400
Attn1	Layer2	AttnBlock (Table C.4)	[B, 256, 16, 16]	82,241
Sum1	[Attn1,Layer3]	Sum	[B, 256, 16, 16]	0
Layer4	Layer3	DownBlockx2 (Table C.2)	[B, 512, 8, 8]	13,114,368
Attn2	Layer4	AttnBlock (Table C.4)	[B, 512, 8, 8]	328,321
Sum2	[Attn2, Layer4]	Sum	[B, 512, 8, 8]	0
Out	Sum2	Softmax	[B, 512, 8, 8]	0
			Trainable Parameters:	**11,621,570**

Table C.8: Deconvolutional Header

Layer	Input	Operation	Output Shape	Parameters
Layer1	INPUT [B, 512, 8, 8]	UpBlockx1 (Table C.3)	[B, 256, 32, 32]	1,705,728
Layer2	Layer1	UpBlockx1 (Table C.3)	[B, 128, 64, 64]	426,880
Layer3	Layer2	UpBlockx1 (Table C.3)	[B, 53, 128, 128]	106,944
Layer4	Layer3	UpBlockx1 (Table C.3)	[B, 32, 256, 256]	26,848
Conv1	Layer4	Convolution2D$_{1x1}$	[B, 6, 128, 128]	231
			Trainable Parameters:	**2,266,631**

C.5 Results

Encoders and deconvolutional headers are trained for all combinations of encoders and pretaining algorithms. Select results are shown in Figure C.3. The test performance of landcover classfication is shown in Table C.9 and Figure C.4. The Jaccard index (intersection-over-union, IoU) for each class is used as an evaluation metric. Table C.9 shows IoU performance weighted by the size of each class. For all pretraining methods, the ResNet18 architecture with attention performed the best on the out-of-sample test set. This suggests that parameter (and training data) size is not a key constraint in the development of useful embeddings, and that more can be done with novel architecture search.

Among the pretraining methods, variational autoencoding and tile2vec perform similarly well, while contrastive sensor fusion performs even worse than no pretraining with a ResNet34. It is remarkable that the humble variational autoencoder was so competitive with the other pretraining methods. Perhaps it is because the autoencoder must actively learn to encode levels of atmospheric interference in order to reconstruct it.

Figure C.4 shows more detail into the performance of each pretraining and encoder combination for the 5 different land classes. Many of the models struggled to classify urban areas, and none of them detected wetlands. The models that were effective in classifying urban areas were three different combinations of encoder architecture and pretraining method, so a clear lesson is difficult to draw. Perhaps the greater parameter space of ResNet34 allows for effective autoencoding of more

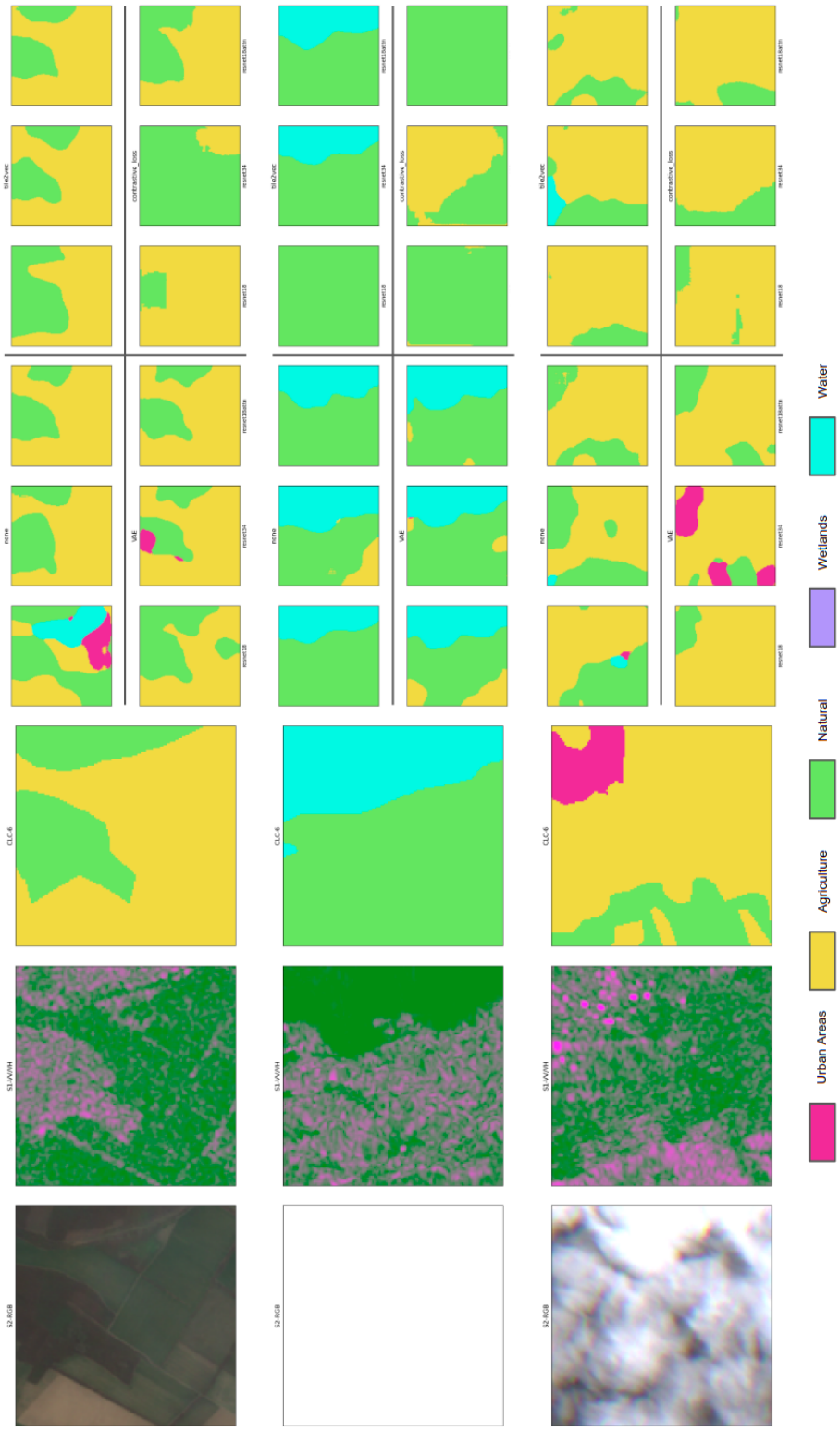


Figure C.3: Out-of-sample land cover classification inference samples

Table C.9: Land Cover Classification, CLC level 0[1], out-of-sample class-weighted mean IoU[2]

Encoder Parameters[3]	ResNet18 13.5mn	ResNet34 23.5mn	ResNet18-Attn 13.9mn
None	38.5	48.2	**55.7**
VAE	49.4	54.8	**59.4**
T2V	50.2	51.3	**59.7**
CSF	46.3	37.6	**48.8**

1: Copernicus CORINE level 0 land cover classes, see Table C.10

2: **bold** indicates top performing encoder by pretraining, underline shows top performing pretraining by encoder.

3: Includes encoder plus deconvolutional header

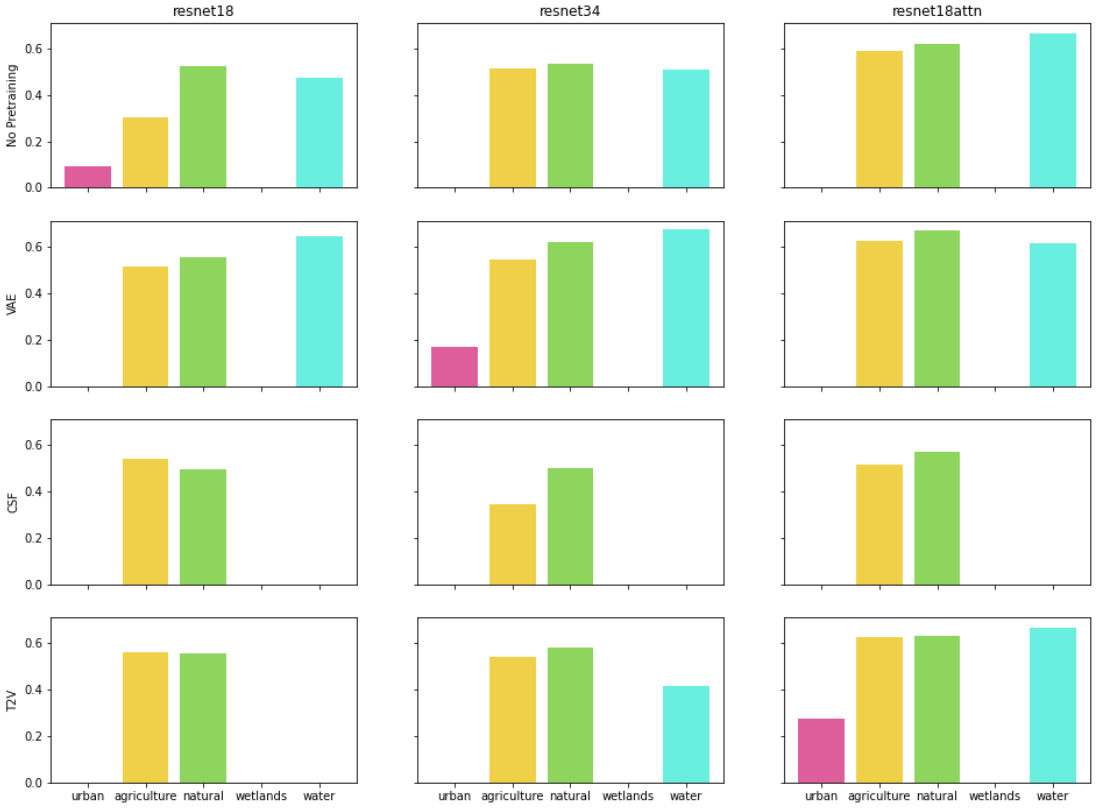


Figure C.4: Test Jaccard index (intersection-over-union) for 5 level-0 CORINE land cover classes

sophisticated semantic classes. Urban areas, with their distinct shapes, textures, and colors, may benefit more from attentions and spatially-explicit training, leading to the outperformance of tile2vec with ResNet18 with attention.

Table C.10: Copernicus CORINE Land Cover Classes

Class Level 0	Description
1XX	Urban and built-up areas
2XX	Agriculture and mixed land use
3XX	Natural land uses
4XX	Wetlands
5XX	Permanent water

Table C.11: Coal Mine classifier, test AUC-ROC*

Encoder Parameters	ResNet18 13.5mn	ResNet34 23.5mn	ResNet18-Attn 13.9mn
None	0.600	0.539	**0.629**
VAE	0.649	**0.684**	0.681
T2V	0.660	0.580	**0.734**
CSF	0.621	0.635	**0.690**

*: **bold** indicates top performing encoder by pretraining, underline shows top performing pretraining by encoder.

C.6 Application

As a demonstration of the generaliseability of the pretraining and finetuning methods developed in this paper, a secondary fine-tuned model is prepared to classify whether mining sites produce coal or not. Mines sites are obtained from Maus et al. (2020*b*) and corresponding imagery is prepared for all mine sites. A small fine-tuning set of 2684 Sentinel-1 and Sentinel-2 images corresponding to 608 mines is obtained with hand labels provided by the Natural Resources Governance Institute (2019). The mine classifier is evaluated against a held-out test set, using area under the receiving operator curve (AUC-ROC) as the evaluation metric. Finetuning results are shown in Table C.11.[4]

C.7 Discussion

While these results are novel, the poor performance suggests that more refinement is needed before this end-to-end self-supervised-plus-supervised method can be put into production systems. This machine learning task is difficult; atmospheric conditions obscure ground conditions, affect the brightness of the entire scene, and can cause

[4]A visualisation of the mines and the estimated probability of their being coal mines in available at `https://lucaskruitwagen.users.earthengine.app/view/deepsentinelmines`

artefacts like shadow. Figure C.2 shows how much atmospheric interference is present in this 'raw' data, with a very small portion of the scenes being cloud-free, and many scenes being entirely saturated by cloud cover. Atmosphere conditions also affect SAR imagery, albeit to a lesser extent. Target distributions are dynamic, changing (marginally) through time and also cyclically with the seasons and natural growing cycles. Training labels are up-scaled from a nominally coarser resolution and so are 'noisy', further compounding the challenge of the learning task.

It is also quite likely that the ResNet computer vision models employed in this paper are not fit for remote sensing tasks. Remote sensing imagery has several properties which make it distinctly different from the general imagery (as might be captured by a phone or dashboard camera) that these models were developed for. Some intuitions are accessible:

- Remote sensing imagery has no horizon and fixed focal distance. Unlike general-purpose encoders, those for remote sensing do not need to codify perspective, angle, and scale.

- Objects of interest are often fewer pixels in size than in conventional computer vision problems. This means information does not need to traverse large distances in the pixel space to inform a classification or regression. Conventional computer vision architectures trade off pixel-space for feature space and use pooling compress representations. This might not be necessary with remote sensing imagery.

- Remote sensing imagery is more often confounded by atmospheric conditions. This means surface representations are more variant and models need to be designed to accomodate this dynamicism and uncertainty.

- Often more spectra are available than the red-green-blue optical spectra of conventional imagery. This means there is combinationally more complexity accessible in the lower latent space encodings which these models have not been designed to capture.

Some ideas for future research might include exploring the trade-off between contextual information filtered with attention and local information. Grouping shallow convolutions might help draw out relations between a high number of input bands. Multi-look training, especially with SAR data, might help normalise data within a single scene. Shallower architectures might be tried to more fully encode the shallow latent space. New methods for modelling uncertainty might help isolate the effect of atmospheric conditions on model predictions. *DeepSentinel* provides a platform for this future research.

C.8 Conclusions

This paper presents *DeepSentinel*, a new codebase and dataset for Sentinel-1, Sentinel-2 sensor fusion research and applications. A preliminary experiment with *DeepSentinel* is developed, testing a number of encoder architectures and pretraining methods for accuracy on a landcover classification task. Of the encoder architectures and pretraining methods trialed, tile2vec pretraining with a shallow ResNet variant with self-attention is found to produce the best classifier in intersection-over-union tests against held-out data. The mixed results help give intuition for future research of this type, particularly the novelties of earth observation data. *DeepSentinel* is an open-source tool with accompanying datasets which can facilitate these future efforts.

Acknowledgements

I am immensely grateful for the ongoing support of Descartes Labs who have provided platform access for this project as part of their Impact Science Program. This work is also being prepared with the generous support of Microsoft AI for Earth and Google Cloud Platform research credits. Without the support of these organisations this work would not be possible; they have my immense gratitude.

Bibliography

2 Degrees Investing Initiative (2015), Developing sustainable energy investment (sei) metrics, benchmarks, and assessment tools for the financial sector, Technical report. Available at `http://2degrees-investing.org/IMG/pdf/sei_metrics_summary-3.pdf` [accessed 2017.03.07].

Ackerlof, G. A. (1970), 'The market for 'lemons': Quality uncertainty and the market mechanism', *The Quarterly Journal of Economics* **84**(3), 488–500.

Adams, J. (2014), 'Pra response to defra on climate change adaptation reporting'.

Adewumi, A., Misra, S. & Omoregbe, N. (2015), Evaluating open source software quality models against iso 25010, *in* '2015 IEEE International Conference on Computer and Information Technology', pp. 872–877.

Agency, I. E. (2020), 'Balances & statistics'.
URL: *https://www.iea.org/data-and-statistics*

Airbus Intelligence" (2013), Spot 6 | Spot 7 Imagery, Technical report, Leiden, NL.
URL: *https://www.intelligence-airbusds.com/files/pmedia/public/r49229*$_{9}$*s*$_{}$*pot*$_{6}$*7.pdf*

Allen, F. (1990), 'The market for information and the origin of financial intermediation', *Journal of Financial Intermediation* **1**(1), 3–30.

Alova, G. (2020), 'A global analysis of the progress and failure of electric utilities to adapt their portfolios of power-generation assets to the energy transition', *Nature Energy* **5**(11), 920–927.

Ansar, A. (2013), 'Location decisions of large firms: analyzing the procurement of infrastructure services', *Journal of Economic Geography* **13**(5), 823–844.

Armour, J., Awrey, D., Daives, P., Enriques, L., Gordon, J. N., Mayer, C. & Payne, J. (2016), *Principles of Financial Regulation*, Oxford University Press, Oxford.

Arrhenius, S. (1896), 'Xxxi. on the influence of carbonic acid in the air upon the temperature of the ground', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **41**(251), 237–276.

Arrow, K. (1962), *Economic Welfare and the Allocation of Resources for Investion*, National Bureau of Economic Research, Inc, Cambridge, pp. 609–626.

Audebert, N., Le Saux, B. & Lefèvre, S. (2018), 'Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks', *ISPRS Journal of Photogrammetry and Remote Sensing* **140**, 20–32.

Badrinarayanan, V., Kendall, A. & Cipolla, R. (2015), 'Segnet: A deep convolutional encoder-decoder architecture for image segmentation', *CoRR* **abs/1511.00561**.
URL: *http://arxiv.org/abs/1511.00561*

Balkissoon, K. & Heaps, T. (2014), 'Performance and impact: Can low carbon equity portfolios generate healthier financial returns?', *SSRN* .

Banks, W. F. (1977), Energy consumption in the pipeline industry, Technical report.

Barbose, G. & Darghouth, N. (2019), 'Tracking the sun'.
URL: *https://openpv.nrel.gov/*

Baruya, P. & Kessels, J. (2013), 'Coal prospects in botswana, mozambique, zambia, zimbabwe and namibia'.

Bauer, B., Koedijk, K. & Otten, R. (2005), 'International evidence on ethical mutual fund performance and investment style', *Journal of Banking & Finance* **29**(7), 1751–1767.

Bauer, R., Delling, D., Sanders, P., Schieferdecker, D., Schultes, D. & Wagner, D. (2008), Combining hierarchical and goal-directed speed-up techniques for dijkstra's algorithm, *in* 'International Workshop on Experimental and Efficient Algorithms', Springer, pp. 303–318.

Baumol, W. J. (1970), *A New Rationale for Corporate Social Policy*, Committee for Economic Development, chapter Enlightened Self-Interest and Corporate Philanthropy.

Belkaoui, A. (1976), 'The impact of the disclosure of the environmental effects of organisational behaviour on the market', *Financial Management* **5**(4), 26–31.

Bello, I., Zoph, B., Vaswani, A., Shlens, J. & Le, Q. V. (2019), 'Attention augmented convolutional networks'.

Benabou, S., Moussu, N. & Müller, B. (2017), The business voice at cop21, *in* S. C. Aykut, J. Foyer & E. Morena, eds, 'Globalising the CLimate: COP21 and the climatisation of global debates'.

Bender, E. M. & Koller, A. (2020), Climbing towards nlu: On meaning, form, and understanding in the age of data, *in* 'Proc. of ACL'.

Beneke, C. M., Skillman, S., Warren, M. S., Kelton, T., Brumby, S. P., Chartrand, R. & Mathis, M. (2017), A Platform for Scalable Satellite and Geospatial Data Analysis, *in* 'AGU Fall Meeting Abstracts', Vol. 2017, pp. IN32C–04.

Benkler, Y. (2002), 'Coase's penguin, or, linux and the nature of the firm', *Yale Law Journal* **112**, 369–446.

Benkler, Y. (2006), *The Wealth of Networks*, Yale University Press.

Bennear, L. S. & Olmstead, S. M. (2008), 'The impacts of the "right to know": Information disclosure and the violation of drinking water standards', *Journal of Environmental Economics and Management* **56**(2), 117–130.

Bentes, C., Velotto, D. & Tings, B. (2018), 'Ship classification in terrasar-x images with convolutional neural networks', *IEEE Journal of Oceanic Engineering* **43**(1), 258–266.

Berardi, G., Esuli, A., Fagni, T. & Sebastiani, F. (2015), Classifying websites by industry sector: A study in feature design, *in* 'Proceedings of the 30th Annual ACM Symposium on Applied Computing', SAC '15, Association for Computing Machinery, New York, NY, USA, p. 1053–1059.
URL: *https://doi.org/10.1145/2695664.2695722*

Berle, A. & Means, G. (1932), *The Modern Corporation and Private Property*, Transaction Publishers.

Bermudez, J. D., Happ, P. N., Feitosa, R. Q. & Oliveira, D. A. (2019), 'Synthesis of multispectral optical images from sar/optical multitemporal data using conditional generative adversarial networks', *IEEE Geoscience and Remote Sensing Letters* **16**(8), 1220–1224.

Bermudez, J., Happ, P., Oliveira, D. & Feitosa, R. (2018), 'Sar to optical image synthesis for cloud removal with generative adversarial networks.', *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences* **4**(1).

Birch, T., Brumby, S. P., Brown, C. F. & Hanson, C. (2020), Dynamic world, *in* 'GeoforGood 2020', Virtual Conference.

Blodgett, S. L., Barocas, S., Daumé III, H. & Wallach, H. (2020), 'Language (technology) is power: A critical survey of" bias" in nlp', *arXiv preprint arXiv:2005.14050* .

BloombergNEF (2018), New Energy Outlook 2018, Technical report.

Bogle, J. C. (2005), *The battle for the soul of capitalism*, Yale University Press.

Bolinger, M., Seel, J. & Robson, D. (2019), 'Utility-scale solar: Empirical trends in project technology, cost, performance, and ppa pricing in the united states–2019 edition'.

Bolton, P. & Dewatripont, M. (2005), *Contract Theory*, MIT Press, Cambridge.

Bonner, S. E. & Sprinkle, G. B. (2002), 'The effects of monetary incentives on effort and task performance: theories, evidence, and a framework for research', *Accounting, Organizations and Society* **27**(4–5), 303–345.

Bostrom, N. (2014), *Superintelligence*, Oxford University Press.

Bower, J. L. & Christensen, C. M. (1995), 'Disruptive technologies: Catching the wave', *Harvard Business Review* **Jan–Feb**, 43–53.

BP plc (2018), Statistical Review of World Energy 2018, Technical report, London, UK.

Branco, M. C. & Rodrigues, L. L. (2007), 'Postioning stakeholder theory within the debate on corporate social responsibility', *Electronic Journal of Business Ethics and Organization Studies* **12**(1), 5–15.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(5), 5–32.

Britannia, E. (2017), 'Title', Available at `https://www.britannica.com/`. [accessed 2017.03.02].

Brown, J. A. & Foster, W. R. (2012), 'Csr and stakeholder theory: A tale of adam smith', *Journal of Business Ethics* **112**(2), 301–312.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020), 'Language models are few-shot learners', *arXiv preprint arXiv:2005.14165* .

Bryan, R. D. (2012), 'Company valuation: Oil and gas', *Society of Petroleum Engineers: The Way Ahead* **8**(3), 10–13.

Buhr, B. (2017), 'Assessing the sources of stranded asset risk: a proposed framework', *Journal of Sustainable Finance & Investment* **7**(1), 37–53.

Bukhary, S., Ahmad, S. & Batista, J. (2018), 'Analyzing land and water requirements for solar deployment in the southwestern united states', *Renewable and Sustainable Energy Reviews* **82**, 3288 – 3305.
URL: *http://www.sciencedirect.com/science/article/pii/S136403211731393X*

Burks, B. (2020), 'Space, The Next Frontier: Spatial Finance And Environmental Sustainability', `https://www.spglobal.com/ratings/en/research/articles/200122-space-the-next-frontier-spatial-finance-and-environmental-sustainability-11317146`.

Burlea, A. S. & Popa, I. (2013), *Legitimacy Theory*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1579–1584.

Byers, L., Friedrich, J., Hennig, R., Kressig, A., Li, X., McCormick, C. & Valeri, L. M. (2018), A Global Database of Power Plants, Technical report, Washington DC, USA.

Caldecott, B. (2016), The future of climate-related disclosure, Technical report, Economist Intelligence Unit: Perspectives. Available at: `https://eiuperspectives.economist.com/sustainability/future-climate-related-disclosure` [Accessed 2020.11.30].

Caldecott, B., Dericks, G., Bouveret, G., Schumacher, K., Pfeiffer, A., Tulloch, D. J., Kruitwagen, L. & McCarten, M. (2018*a*), Asset-level data and the energy transition, Technical report, Smith School of Enterprise and the Environment.

Caldecott, B., Dericks, G., Bouveret, G., Schumacher, K., Pfeiffer, A., Tulloch, D. J., Kruitwagen, L. & McCarten, M. (2018*b*), 'Asset-level data and the energy transition: Findings from et risk work package 2', *University of Oxford, Sustainable Finance Programme< http://www. smithschool. ox. ac. uk/research/sustainable-finance/publications/Asset-level-dataand-the-Energy-Transition-Findings-from-ET-Risk-Work-Package2. pdf> accessed* **19**.

Caldecott, B. & Elders, G. (2013), Bloomberg Carbon Risk Valuation Tool, Technical report, Bloomberg New Energy Finance.

Caldecott, B., Howarth, N. & McSharry, P. (2013), 'Stranded assets in agriculture: Protecting value from environment-related risks'.

Caldecott, B. & Kruitwagen, L. (2016), Guest opinion: How asset level data can improve the assessment of environmental risk in credit analsyis, Technical report, Standard & Poor's MArket Intelligence. Available at: `https://www.globalcreditportal.com/ratingsdirect/renderArticle.do?articleId=1728982&SctArtId=402968&from=CM&nsl_code=LIME&sourceObjectId=9796605&sourceRevId=1&fee_ind=N&exp_date=20261003-21:55:05` [Accessed 2020.11.30].

Caldecott, B., Kruitwagen, L., Dericks, G., Tulloch, D. J., Kok, I. & Mitchell, J. (2016), Stranded assets and thermal coal: An analysis of environment-related risk, Technical report, Smith School of Enterprise and the Environment.

Caldecott, B., Kruitwagen, L., McCarten, M., Zhou, X., Lunsford, D., Marchand, O., Hadjikyriakou, P., Bickel, V., Sachs, T. & Bohn, N. (2018), 'Climate risk analysis from space: remote sensing, machine learning, and the future of measuring climate-related risk'.

Callon, M. (1994), 'Is science a public good? fifth mullins lecture, virginia polytechnic institute, 23 march 1993', *Science, Technology, & Human Values* **19**(4), 395–424.

Cambria, E., Poria, S., Gelbukh, A. & Thelwall, M. (2017), 'Sentiment analysis is a big suitcase', *IEEE Intelligent Systems* **32**(6), 74–80.

Camilo, J. A., Wang, R., Collins, L. M., Bradbury, K. & Malof, J. M. (2018), 'Application of a semantic segmentation convolutional neural network for accurate automatic detection and mapping of solar photovoltaic arrays in aerial imagery', *CoRR* **abs/1801.04018**.
URL: *http://arxiv.org/abs/1801.04018*

Campbell, T. S. & Kracaw, W. A. (1980), 'Information production, market signalling, and the theory of financial intermediation', *The Journal of Finance* **35**(4), 863–882.

Carbon (2012).
URL: *https://web.archive.org/web/20120526014231/http://www.cne.cl/energias/hidrocarburos/tipos-de-energia/365-carbon*

Carlisle, J. E., Kane, S. L., Solan, D., Bowman, M. & Joe, J. C. (2015), 'Public attitudes regarding large-scale solar energy development in the u.s.', *Renewable and Sustainable Energy Reviews* **48**, 835–847.

Carney, M. (2014), 'Letter from governor of bank of england on carbon bubble risk'. Available at `https://www.parliament.uk/globalassets/documents/commons-committees/environmental-audit/Letter-from-Governor-of-Bank-of-England-on-Carbon-Bubble-risk.pdf` [accessed 2020.11.30].

Carney, M. (2015), 'Breaking the tragedy of the horizon - climate change and financial stability'. Speech given at Lloyd's of London, 2015.09.29.

Carney, M. (2016), 'Resolving the climate paradox'. Speech given at Arthur Burns Memorial Lecture, Berlin, 2016.09.22.

Carroll, A. B. (1979), 'A three-dimsensional model of corporate performance', *Academy of Management Review* **4**(4), 479–505.

Carroll, M. W. (2013), 'Creative commons and the openness of open access', *The New England Journal of Medecine* **368**(9), 789–791.

Castells, M. (1996), *The information age*, Vol. 98, Oxford Blackwell Publishers.

Chen, L., Zhao, X., Tang, O., Price, L., Zhang, S. & Zhu, W. (2017), 'Supply chain collaboration for sustainability: A literature review and future research agenda', *International Journal of Production Economics* **194**, 73–87.

Chen, T. & Guestrin, C. (2016), XGBoost: A scalable tree boosting system, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '16, ACM, New York, NY, USA, pp. 785–794. **URL:** *http://doi.acm.org/10.1145/2939672.2939785*

Cheng, B., Ioannou, I. & Serafeim, G. (2014), 'Corporate social responsibility and access to finance', *Strategic management journal* **35**(1), 1–23.

Cherlet, M., Hutchinson, C., Reynolds, J., Hill, J., Sommer, S. & Von Maltitz, G. (2018), *World atlas of desertification: Rethinking land degradation and sustainable land management*, Publications Office of the European Union.

Chomsky, N. (1999), *Profit over PEople: Neoliberalism and Global Order*, Seven Stories Press, New York.

Chubirka, M. (2014), 'Open-source vs. commercial software: A false dilemma', Available at `http://www.informationweek.com/strategic-cio/it-strategy/open-source-vs-commercial-software-a-false-dilemma/d/d-id/1252665`. [accessed 2017.03.02].

Clark, G. L. (2017), *The New Oxford Handbook of Economic Geography*, Oxford University Press, chapter Behaviour in Context. Forthcoming.

Clark, G. L., Feiner, A. & Viehs, M. (2015), 'From the stockholder to the stakeholder: How sustainability can drive financial outperformance', *SSRN* .

Clark, G. L., Feldman, M. P., Gertler, M. S. & Wójcik, D. (2017), *The New Oxford Handbook of Economic Geography*, Oxford University Press, chapter Introduction: Economic Geography in the 21^st Century. Forthcoming.

Clark, P. (2015), 'Shell bows to shareholder demands on climate change', *Financial Times* **29**.

Coal mining areas in North and Central Vietnam (2015). **URL:** *https://www.vd-office.org/en/trainmine-2/*

Coase, R. H. (1960), 'The problem of social cost', *Journal of Law and Economics* **3**(1), 1–44.

Cohen, B. D. & Dean, T. J. (2005), 'Information asymmetry and investor valuation of ipos: Top management team lefitimace as a capital market signal', *Strategic Management Journal* **26**(7), 683–690.

Cohen, K. (2014*a*), 'Some thoughts on divestment'. Available at `http://www.exxonmobilperspectives.com/2014/10/10/some-thoughts-on-divestment/` [accessed 2015.4.3].

Cohen, K. (2014*b*), Some thoughts on divestment, Technical report, ExxonMobil Perspectives. Available at `http://www.exxonmobilperspectives.com/2014/10/10/some-thoughts-on-divestment/` [accessed 2017.03.02].

Cohen, M. A. (2001), 'Information as a policy intrument in protecting the environment: What have we learned?', *Environmental Law Reporter* **31**, 10425–10431.

Cohen, M. A. & Santhakumar, V. (2007), 'Information disclosure as environmental regulation: A theoretical analysis', *Environmental and Resource Economics* **37**(3), 599–620.

Committee for Economic Development (1971), Social responsibilities of business corporations, Technical report.

Comtrade, U. (2021), 'Un comtrade database', *UN Comtrade Online* .

Conine, T. E. & Madden, G. P. (1987), 'Corporate social responsibility and investment value', *Accounting, Organizations and Society* **12**(2), 111–122.

Connelly, B. L., Certo, S. T., Ireland, R. D. & Reutzel, C. R. (2011), 'Signalling theory: A review and assessment', *Journal of Management* **37**(1), 39–67.

Couture, H. D., O'Connor, J., Mitchell, G., Söldner-Rembold, I., D'souza, D., Karra, K., Zhang, K., Kargar, A. R., Kassel, T., Goldman, B., Tyrrell, D., Czerwinski, W., Talekar, A. & McCormick, C. (2020), Towards tracking the emissions of every power plant on the planet, *in* 'NeurIPS Climate Change AI Workshop 2020', NeurIPS Climate Change AI.

Couvet, D., Jiguet, F., Julliard, R., Levrel, H. & Teyssedre, A. (2008), 'Enhancing citizen contributions to biodiversity science and public policy', *Interdisciplinary Science Reviews* **33**(1), 95–103.

Cowen, S. S., Ferreri, L. B. & Parker, L. D. (1987), 'The impact of corporate characteristc on social responsibility disclosure: A typology and frequency-based analysis', *Accounting, Organizations and Society* **12**(2), 111–122.

Crawford, K. (2017), The trouble with bias, *in* 'Conference on Neural Information Processing Systems, invited speaker'.

Crippa, M., Solazzo, E., Huang, G., Guizzardi, D., Koffi, E., Muntean, M., Schieberle, C., Friedrich, R. & Janssens-Maenhout, G. (2020), 'High resolution temporal profiles in the emissions database for global atmospheric research', *Scientific Data* **7**(1), 1–17.

CTI (2011), Unburnable carbon: Are the world's financial markets carrying a carbon bubble?, Technical report, Carbon Tracker Initiative.

CTI (2013), Unburnable carbon 2013: Wasted capital and stranded assets, Technical report, Carbon Tracker Initiative.

CTI (2014), Carbon supply cost curves: oil capital expenditures, Technical report, Carbon Tracker Initiative.

CTI (2015), Lost in transition: How the energy sector is missing potential demand destruction, Technical report.

CTI (2018), Nowhere to hide: Using satellite imagery to estimate the utilisation of fossil fuel power plants, Technical report.

Cummins, J. D. & Tennyson, S. (1996), 'Moral hazard in insurance claiming: Evidence from automobile insurance', *Journal of Risk and Uncertainty* **12**(1), 29–50.

Current Developments in the Russian Mining Industry (2007).
URL: *https://mining-report.de/english/current-developments-in-the-russian-mining-industry/mining*$_r ussia_2$*/*

Dan Tong, Qiang Zhang, Y. Z. K. C. C. S. C. H. Y. Q. & Davis, S. J. (2019), 'Committed emissions from existing energy infrastructure jeopardize 1.5°c climate target', *Nature* **572**(4), 373–377.

Davidson, S. (2014), 'A critique of the coal divestment campaign', *Minerals Council of Australia* .

Deephouse, D. L. (2000), 'Media reputation as a strategic resource: An integration of mass communication and resource-based theories', *Journal of Management* **26**(6), 1091–1112.

Delaunay, B. et al. (1934), 'Sur la sphere vide', *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk* **7**(793-800), 1–2.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009), ImageNet: A Large-Scale Hierarchical Image Database, *in* 'CVPR09'.

Department for Business, Energy, & Industrial Strategy (2019), Greenhouse gas reporting: conversion factors 2019, Technical report.
URL: *https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2019*

Der Kiureghian, A. & Ditlevsen, O. (2009), 'Aleatory or epistemic? does it matter?', *Structural safety* **31**(2), 105–112.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018), 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805* .

Dewar, J. A. (1998), 'The information age and the printing press'. Research Paper P-8014.

Dias, L., Gouveia, J. P., Lourenço, P. & Seixas, J. (2019), 'Interplay between the potential of photovoltaic systems and agricultural land use', *Land Use Policy* **81**, pp725–735.

Dias, P., Kanellopoulos, K., Medarac, H., Kapetaki, Z., Miranda-Barbosa, E., Shortall, R., Veronica, C., Telsnig, T., Cristina, V.-H., Lacal Arantegui, R., Nijs, W., González-Aparicio, I., Trombetti, M., Mandras, G., Peteves, S. & Tzimas, E. (2018), 'Eu coal regions: opportunities and challenges ahead'.

Dijkstra, E. W. et al. (1959), 'A note on two problems in connexion with graphs', *Numerische mathematik* **1**(1), 269–271.

Distribution of Coal in India: Gondwana Coalfields & Tertiary Coalfields (2019).
URL: *https://www.pmfias.com/coal-in-india-gondwana-coal-tertiary-coal-coking-coal-non-coking-coal/*

Dorogush, A. V., Gulin, A., Gusev, G., Kazeev, N., Prokhorenkova, L. O. & Vorobev, A. (2017), 'Fighting biases with dynamic boosting', *CoRR* **abs/1706.09516**.
URL: *http://arxiv.org/abs/1706.09516*

Dosi, G. (1982), 'Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change', *Research Policy* **11**(3), 147–162.

Dreisbach, C., Koleck, T. A., Bourne, P. E. & Bakken, S. (2019), 'A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data', *International journal of medical informatics* **125**, 37–46.

Dunnett, A. J., Adjiman, C. S. & Shah, N. (2008), 'A spatially explicit whole-system model of the lignocellulosic bioethanol supply chain: an assessment of decentralised processing potential', *Biotechnology for biofuels* **1**(1), 13.

Dye, R. A. (1985), 'Disclosure of nonproprietary information', *Journal of Accounting Research* **23**(1), 123–145.

Eccles, R. G., Ioannou, I. & Serafeim, G. (2014), 'The impact of corporate sustainability on organizational processes and performance', *Management Science* **60**(11), 2835–2857.

Eckardt, R., Berger, C., Thiel, C. & Schmullius, C. (2013), 'Removal of optically thick clouds from multi-spectral satellite images using multi-frequency sar data', *Remote Sensing* **5**(6), 2973–3006.

Economist, T. (2017), 'Regulating the internet giants', *The Economist* .

Edmans, A. (2020), 'Response to the european commission study on sustainable corporate governance'. Available at: `https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12548-Sustainable-corporate-governance/F556360` [Accessed 2020.12.03].

EIA-7A: Coal Production and Preparation Report (2018).
URL: *https://www.eia.gov/maps/layer*$_{i}nfo - m.php$

Eisenhardt, K. M. (1989), 'Agency theory: An assessment and review', *Academy of Management Review* **14**(1), 57–74.

Elahi, N., Shah, N., Korre, A. & Durucan, S. (2014), 'Multi-period least cost optimisation model of an integrated carbon dioxide capture transportation and storage infrastructure in the uk', *Energy Procedia* **63**, 2655–2662.

Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. (1996), A density-based algorithm for discovering clusters in large spatial databases with noise, AAAI Press, pp. 226–231.

Esty, D. C. (2004), 'Environmental protection in the information age', *New York University Law Review* **79**(1), 115–211.

Euracoal: Serbia (2020).
URL: *https://euracoal.eu/info/country-profiles/serbia/*

Euracoal: Turkey (2020).
URL: *https://euracoal.eu/library/archive/turkey/*

Euracoal: Ukraine (2020).
URL: *https://euracoal.eu/info/country-profiles/ukraine/*

European Commission, Joint Research Centre; Columbia University, C. f. I. E. S. I. N. C. (2015), 'Ghs population grid, derived from gpw4, multitemporal (1975, 1990, 2000, 2015)', =http://data.europa.eu/89h/jrc-ghsl-ghs$_p op_g pw4_g lobe_r 2015a$.

Evans, L., Owda, M., Crockett, K. & Vilas, A. F. (2019), 'A methodology for the resolution of cashtag collisions on twitter–a natural language processing & data fusion approach', *Expert Systems with Applications* **127**, 353–369.

ExxonMobil Inc (2014), Energy and Carbon - Managing the Risks, Technical report, ExxonMobil.

Fama, E. F. (1970), 'Efficient capital markets: A review of theory and empirical work', *The Journal of Finance* **25**(2), 383–417.

Fama, E. F. (1980), 'Agency problems and the theory of the firm', *The Journal of Political Economy* **88**(2), 288–307.

Faria, P., Labutong, N., Wasmuth, K., Pineda, A. C., Delgado, P., Tornay, C., Huusko, H., Aden, N. & Russel, S. (2015), Science based targets: Mind the science, Technical report, CDP, WWF, UN Global Compact, WRI. Available at `http://sciencebasedtargets.org/mindthescience/MindTheScience.pdf` [accessed 2017.03.07].

Farmer, J., Hepburn, C., Ives, M., Hale, T., Wetzer, T., Mealy, P., Rafaty, R., Srivastav, S. & Way, R. (2019*a*), 'Sensitive intervention points in the post-carbon transition', *Science* **364**(6436), 132–134.

Farmer, J., Hepburn, C., Ives, M., Hale, T., Wetzer, T., Mealy, P., Rafaty, R., Srivastav, S. & Way, R. (2019*b*), 'Sensitive intervention points in the post-carbon transition', *Science* **364**(6436), 132–134.

Farrell, J. (1987), 'Information and the coase theorem', *Economic Perspectives* **1**(2), 113–129.

Faust, D. (2013), Fossil Fuel Divestment Statement, Technical report, Harvard University Office of the President. Available at `http://www.harvard.edu/president/news/2013/fossil-fuel-divestment-statement` [accessed 2017.03.02].

Ferrant, S., Selles, A., Le Page, M., Herrault, P.-A., Pelletier, C., Al-Bitar, A., Mermoz, S., Gascoin, S., Bouvet, A., Saqalli, M. et al. (2017), 'Detection of irrigated crops from sentinel-1 and sentinel-2 data to estimate seasonal groundwater use in south india', *Remote Sensing* **9**(11), 1119.

Filatotchev, I. & Bishop, K. (2002), 'Boad composition, share ownership, and 'underpricing' of uk ipo firms', *Strategic Management Journal* **23**(10), 941–955.

Financial Stability Board (2020), 2020 status report: Task force on climate-related financial disclosures, Technical report, Financial Stability Board. Available at `https://www.fsb.org/2020/10/2020-status-report-task-force-on-climate-related-financial-disclosures/` [accessed 2020.11.30].

Fink, L. (2020), 'A fundamental reshaping of finance. blackrock'.

Fishburn, P. C., Odlyzko, A. M. & Siders, R. C. (2000), *The Economics of Digital Information and Intellectual Property*, MIT Press, Cambridge, chapter Fixed-Fee versus Unit Pricing for Information Goods: Competition, Equilibria, and Price Wars.

Fontana, V. (2019), 'Open pits in the south'.
URL: *https://brazilian.report/business/2019/10/08/coal-mining-open-pits-south-rio-grande-do-sul/*

Forum Pour L'Investissement Responable (2016), Article 173-vi understanding the french regulation on investor climate reporting, Technical report. Available at `https://www.frenchsif.org/isr-esg/wp-content/uploads/Understanding_article173-French_SIF_Handbook.pdf` [accessed 2020.11.30].

Foulon, J., Lanoie, P. & Laplante, B. (2002), 'Incentives for pollution control: Regulation or information?', *Journal of Environmental Economics and Management* **44**(1), 169–187.

Freedman, M. & Jaggi, B. (1982), 'Pollution disclosures, pollution performance, and economic performance', *Omega: The International Journal of Management Science* **10**(2), 167–176.

Freeman, R. E. (1984), *Strategic management: a stakeholder approach*, Pitman, Boston.

Freeman, R. E. & Reed, D. L. (1983), 'Stockholders and stakeholders: A new perspective on corporate governance', *California Management Review* **25**(3), 88–106.

Frey, C. B. & Osborne, M. A. (2013), The future of employment: How susceptible are jobs to computerisation?, Technical report, Oxford Martin Programme on Technology and Employment.

Friedman, M. (1962), *Capitalism and Freedom*, The University of Chicago Press, Chicago.

Friedman, M. (1970), 'The social resposibility of business is to increase its profits'.

F.R.S., K. P. (1901), 'Liii. on lines and planes of closest fit to systems of points in space', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**(11), 559–572.

FSB (2015), 'Fsb to establish task force on climate-related financial disclosures'. Press release, 2015.12.04.

Fuentes Reyes, M., Auer, S., Merkle, N., Henry, C. & Schmitt, M. (2019), 'Sar-to-optical image translation based on conditional generative adversarial networks—optimization, opportunities and limits', *Remote Sensing* **11**(17), 2067.

Funtowicz, S. O. & Ravetz, J. R. (1993), 'Science for the post-normal age', *Futures* **25**(7), 739–755.

G20 Finance Ministers and Central Bank Governors (2015), 'Communiqué', *G20 Information Centre* . Available at `http://www.g20.utoronto.ca/2015/150417-finance.html` [accessed 2020.11.30].

Gai, P. & Kapadia, S. (2010), 'Contagion in financial networks', *Proceedings of the Royal Society of Mathematical, Physical, and Engineering Sciences* **466**(1), 2401–2423.

Ganguli, P., Kumar, D. & Ganguly, A. R. (2017), 'Us power production at risk from water stress in a changing climate', *Scientific reports* **7**(1), 1–13.

Gao, Y., Gao, X. & Zhang, X. (2017), 'The 2 c global temperature target and the evolution of the long-term goal of addressing climate change—from the united nations framework convention on climate change to the paris agreement', *Engineering* **3**(2), 272–278.

Geddes, P., Goldberg, L., Tymoczko, R. & Branch, M. (2016), Building a carbon-free equity portfolio, Technical report, Aperio Group. Available at `https://www.aperiogroup.com/resource/138/node/download` [accessed 2017.03.07].

Ghadge, A., Wurtmann, H. & Seuring, S. (2020), 'Managing climate change risks in global supply chains: a review and research agenda', *International Journal of Production Research* **58**(1), 44–64.
URL: *https://doi.org/10.1080/00207543.2019.1629670*

Global Shipping Lane Network, World, 2000 (2015).
URL: *https://github.com/geoiq/gc*$_d$*ata/blob/master/datasets/*25*.geojson*

gofossilfree.org (2020), 'Divestment commitments'. Available at `https://gofossilfree.org/commitments/` [accessed 2020.11.30].

Goldberg, A. V. & Tarjan, R. E. (1990), 'Finding minimum-cost circulations by successive approximation', *Mathematics of Operations Research* **15**(3), 430–466.

Goldin, I. & Kutarna, C. (2016), *Age of Discovery: Navigating the Risks and Rewards of Our New Renaissance*, St. Martin's Press, New York.

Gompers, P., Ishii, J. & Metrick, A. (2003), 'Corporate governance and equity prices', *The quarterly journal of economics* **118**(1), 107–156.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014), 'Generative adversarial networks'.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. & Moore, R. (2017), 'Google earth engine: Planetary-scale geospatial analysis for everyone', *Remote Sensing of Environment* .
URL: *https://doi.org/10.1016/j.rse.2017.06.031*

Gouveia, C. & Fonseca, A. (2008), 'New approaches to environmental monitoring: the use of ict to explore volunteered geographic information', *GeoJournal* **72**(3-4), 185–197.

Graham, P. (2009), 'G8 summit to set 2 degrees Celcius global warming goal', Available at `http://www.reuters.com/article/us-g8-summit-climate-factbox-sb-idUSTRE56746M20090708`. [accessed 2017.03.02].

Graves, A. & Schmidhuber, J. (2009), Offline handwriting recognition with multidimensional recurrent neural networks, *in* D. Koller, D. Schuurmans, Y. Bengio & L. Bottou, eds, 'Advances in Neural Information Processing Systems 21', Curran Associates, Inc., pp. 545–552.
URL: *http://papers.nips.cc/paper/3449-offline-handwriting-recognition-with-multidimensional-recurrent-neural-networks.pdf*

Gray, M., Watson, L., Ljungwaldh, S. & Morris, E. (2018), Nowhere to hide: Using satellite imagery to estimate the utilisation of fossil fuel power plants, Technical report, London, UK.
URL: *https://www.carbontracker.org/reports/nowhere-to-hide/*

Gray, R., Kouhy, R. & Lavers, S. (1995), 'Corporate social and environmental reporting', *Accounting, Auditing, & Accountability Journal* **8**(2), 47–77.

Green, R. (2008), 'Carbon tax or carbon permits: The impact on generators risks', *The Energy Journal* **29**(3).

Grodsky, S. M. & Hernandez, R. R. (2020), 'Reduced ecosystem services of desert plants from ground-mounted solar energy development', *Nature Sustainability* pp. 1–8.

Grohnfeldt, C., Schmitt, M. & Zhu, X. (2018), A conditional generative adversarial network to fuse sar and multispectral optical data for cloud removal from sentinel-2 images, *in* 'IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium', IEEE, pp. 1726–1729.

Grossman, G. M. & Helpman, E. (1994), 'Protection for sale', *The American Economic Review* **84**(4), 833–850.

Grossman, S. J. (1981), 'The role of warranties and private disclosure about product quality', *Journal of Law and Economics* **24**(3), 461–483.

Grossman, S. J. & Hart, O. D. (1980), 'Disclosure laws and takeover bids', *The Journal of Finance* **35**(2), 323–334.

Grossman, S. J. & Hart, O. D. (1986), 'The costs and benefits of ownership: A theory of verical and lateral integration', *Journal of Political Economy* **94**(14), 691–719.

Grossman, S. J. & Stiglitz, J. E. (1980), 'On the impossibility of informationally efficient markets', *The American Economic Review* **70**(3), 393–408.

Guha, S., Rastogi, R. & Shim, K. (2001), 'Cure: an efficient clustering algorithm for large databases', *Information Systems* **26**(1), 35 – 58.
URL: *http://www.sciencedirect.com/science/article/pii/S0306437901000084*

Gunningham, N., Kagan, R. A. & Thornton, D. (2004), 'Social license and environmental protection: Why businesses go beyond compliance', *Law & Social Inquiry* **29**(2), 307–341.

Gupta, A. & Mason, M. (2014), *Transparency in Global Environmental Governance: Critical Perspectives*, MIT Press, Cambridge.

Ha, D. & Schmidhuber, J. (2018), 'World models', *arXiv preprint arXiv:1803.10122* .

Hall, K. (1981), 'The economic nature of information', *The Information Society* **1**(2), 143–166.

Hamilton, S. H., ElSawah, S., Guillaume, J. H., Jakeman, A. J. & Pierce, S. A. (2015), 'Integrated assessment and modelling: overview and synthesis of salient dimensions', *Environmental Modelling & Software* **64**, 215–229.

Hand, E. (2010), 'Citizen science: People power', *Nature* **466**, 685–687.

Harari, Y. N. (2016), *Homo Deus: A Brief History of Tomorrow*, Harvill Secker, London.

Hardy, Q. (2016), 'Artificial intelligency software is booming. but why now?', *The New York Times* . Available at `https://www.nytimes.com/2016/09/19/technology/artificial-intelligence-software-is-booming-but-why-now.html` [accessed 2017.03.02].

Hart, O. D. (1983), 'Optimal labour contracts under asymmetric information: An introduction', *The Review of Economic Studies* **50**(1), 3–35.

Hart, O. D. & Moore, J. (1990), 'Property rights and the nature of the firm', *Journal of Political Economy* **98**(6), 1119–1158.

Hawley, J. (2017), 'Esg ratings and rankings'.

Hayek, F. A. (1945), 'The use of knowledge in society', *The American economic review* **35**(4), 519–530.

He, K., Zhang, X., Ren, S. & Sun, J. (2015), 'Deep residual learning for image recognition', *CoRR* **abs/1512.03385**.
URL: *http://arxiv.org/abs/1512.03385*

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 770–778.

He, W. & Yokoya, N. (2018), 'Multi-temporal sentinel-1 and-2 data fusion for optical image simulation', *ISPRS International Journal of Geo-Information* **7**(10), 389.

Healy, P. M. & Palepu, K. G. (2001), 'Information assymetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature', *Journal of Accounting and Economics* **31**(1–3), 405–440.

Hernandez, R., Easter, S., Murphy-Mariscal, M., Maestre, F., Tavassoli, M., Allen, E., Barrows, C., Belnap, J., Ochoa-Hueso, R., Ravi, S. & Allen, M. (2014), 'Environmental impacts of utility-scale solar energy', *Renewable and Sustainable Energy Reviews* **29**, 766 – 779.
URL: *http://www.sciencedirect.com/science/article/pii/S1364032113005819*

Hernandez, R. R., Hoffacker, M. K. & Field, C. B. (2014), 'Land-use efficiency of big solar', *Environmental science & technology* **48**(2), 1315–1323.

Hernandez, R. R., Hoffacker, M. K., Murphy-Mariscal, M. L., Wu, G. C. & Allen, M. F. (2015), 'Solar energy development impacts on land cover change and protected areas', *Proceedings of the National Academy of Sciences* **112**(44), 13579–13584.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D. et al. (2020), 'The era5 global reanalysis', *Quarterly Journal of the Royal Meteorological Society* **146**(730), 1999–2049.

Hillis, D. (2016), 'The new world order for open-source and commercial software', Available at `https://techcrunch.com/2016/06/13/the-new-world-order-for-open-source-and-commercial-software/`. [accessed 2017.08.02].

Hirshleifer, J. (1971), 'The private and social value of information and the reward to inventive activity', *The American Economic Review* **61**(4), 561–574.

HM Treasury (2020), A roadmap towards mandatory climate-related disclosures, Technical report, HM Teasury. Available at `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/933783/FINAL_TCFD_ROADMAP.pdf` [accessed 2020.11.30].

Hobbes, T. (1651), *Leviathan*.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural Computation* **9**(8), 1735–1780.
URL: *https://doi.org/10.1162/neco.1997.9.8.1735*

Hofmokl, J. (2010), 'The internet commons: towards an eclectic theoretical framework', *International Journal of the Commons* **4**(1), 226–250.

Holland, J. H. (1992), 'Complex adaptive systems', *Daedalus* **121**(1), 17–30.

Holström, B. (1979), 'Moral hazard and observability', *The Bell Journal of Economics* **10**(1), 74–91.

Hotelling, H. (1931), 'The economics of exhaustible resources', *Journal of political Economy* **39**(2), 137–175.

Hou, X., Wang, B., Hu, W., Yin, L., Huang, A. & Wu, H. (2020), Solarnet: A deep learning framework to map solar plants in china from satellite imagery, *in* 'Climate Change AI Workshop, ICLR2020', ICLR.

Howard, A. W. & Harp, A. B. (2009), 'Oil and gas company valuations', *Business Valuation Review* **29**(1), 30–35.

Hubel, D. H. & Wiesel, T. N. (1962), 'Receptive fields, binocular interaction and functional architecture in the cat's visual cortex', *The Journal of physiology* **160**(1), 106.

Hummel, R., Michalak, R., Sroka, R. & Olczyk, M. (2016), Cdp climate change report, Technical report, CDP.

Hunt, C., Weber, O. & Dordi, T. (2017), 'A comparative analysis of the anti-apartheid and fossil fuel divestment campaigns', *Journal of Sustainable Finance & Investment* **7**(1), 64–81.

Hunt, V., Simpson, B. & Yamada, Y. (2020), The case for stakeholder capitalism, Technical report, McKinsey & Company. Available at `https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-case-for-stakeholder-capitalism` [accessed 2020.12.01].

IHSMarkit (2020), 'Electric plants'.
URL: *https://catalogue.datalake.ihsmarkit.com/*

Imamoglu, N., Kimura, M., Miyamoto, H., Fujita, A. & Nakamura, R. (2017), 'Solar power plant detection on multi-spectral satellite imagery using weakly-supervised cnn with feedback features and m-pcnn fusion', *arXiv preprint arXiv:1704.06410* .

IMPAX Asset Management (2013), Beyond fossil fuels: The investment case for fossil fuel divestment, Technical report. Available at `http://www.impaxam.com/media-centre/white-papers/beyond-fossil-fuels-investment-case-fossil-fuel-divestment` [accessed 2017.03.07].

Indonesia's coal mines and forests (2015).
URL: *http://www.coalforest.org/maps.php?id=indonesia*

Industrial sites in Kosovo (2007).
URL: *https://www.grida.no/resources/6867*

Ingram, R. W. (1978), 'An investigation of the information content of (certain) social responsiblity disclosures', *Journal of Accounting Research* **16**(2), 270–285.

International Energy Agency (2018), World Energy Outlook 2018, Technical report, Paris, France.

International Energy Agency (2019*a*), Coal Market Report 2019, Technical report, Paris, France.
URL: *https://webstore.iea.org/market-report-series-coal-2019*

International Energy Agency (2019*b*), Renewables 2019, Technical report, Paris, France.
URL: *https://www.iea.org/reports/renewables-2019*

International Energy Agency (2019*c*), World Energy Outlook 2019, Technical report, Paris, France.

International Renewable Energy Agency (2019), Renewable capacity statistics 2019, Technical report, Abu Dhabi.

IPCC (2014), *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press.

Ishii, T., Simo-Serra, E., Iizuka, S., Mochizuki, Y., Sugimoto, A., Ishikawa, H. & Nakamura, R. (2016), Detection by classification of buildings in multispectral satellite imagery, *in* '2016 23rd International Conference on Pattern Recognition (ICPR)', IEEE, pp. 3344–3349.

Jackson, G. & Apostolakpou, A. (2010), 'Corporate social responsibility in western europe: An institutional mirror or substitute?', *Journal of Business Ethics* **94**(3), 371–394.

Janssens-Maenhout, G., Crippa, M., Guizzardi, D., Muntean, M., Schaaf, E., Dentener, F., Bergamaschi, P., Pagliari, V., Olivier, J. G., Peters, J. A. et al. (2019), 'Edgar v4. 3.2 global atlas of the three major greenhouse gas emissions for the period 1970–2012', *Earth System Science Data* **11**(3), 959–1002.

Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D. & Ermon, S. (2019), Tile2vec: Unsupervised representation learning for spatially distributed data, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 33, pp. 3967–3974.

Jensen, M. C. & Meckling, W. H. (1976), 'Theory of the firm: Managerial behaviour, agency costs, and ownership structure', *Journal of Financial Economics* **3**(4), 305–360.

Jovanovic, B. (1982), 'Turthful disclosure of information', *The Bell Journal of Economics* **13**(1), 36–44.

Joyce, S. & Thomson, I. (2000), 'Earning a social licence to operate: Social acceptability and resource development in latin america', *The Canadian Mining and MEtallurgical Bulletin* **93**(1037), 49–53.

Kaiser, M. J. & Yu, Y. (2012), 'Part 1: Oil and gas company valuation, reserves, and production', *Oil and Gas Financial Journal* **9**. Available at `http://www.reuters.com/article/us-g8-summit-climate-factbox-sb-idUSTRE56746M20090708` [accessed 2017.03.02].

Kalmykov, D. & Malikova, A. (2017), 'Doomed to coal'.
URL: *https://bankwatch.org/wp-content/uploads/2018/01/KZ$_C$oal$_E$NG.pdf*

Karkkainen, B. C. (2000), 'Information as environmental regulation: Tri and performance benchmarking, percursor to a new paradigm?', *The Georgetown Law Journal* **89**, 257–370.

Karkkainen, B. C. (2006), 'Information-forcing environmental regulation', *Florida State University Law Review* **33**, 861–902.

Keim, G. D. (1978), 'Corporate social responsibility: An assessment of the enlightened self-interest model', *The Academy of Management Review* **3**(1), 32–39.

Kelly, H. J. (1960), 'Gradient theory of optimal flight paths', *American Rocket Society Review* **30**, 947–954.

Kelly, M., Ferranto, S., Lei, S., ichi Ueda, K. & Huntsinger, L. (2012), 'Expanding the table: The web as a tool for participatory adaptive management in california forests', *Journal of Environmental Management* **12**(1), 1–11.

Kemp, R. (2010), 'Open source software (oss) governance in the organisation', *Computer Law & Security Review* **26**(3), 309–316.

Kessler, D. J. & Cour-Palais, B. G. (1978), 'Collision frequency of artificial satellites: The creation of a debris belt', *Journal of Geophysical Research: Space Physics* **83**(A6), 2637–2646.

Key, N., Sneeringer, S. & Marquardt, D. (2014), 'Climate change, heat stress, and us dairy production', *USDA-ERS Economic Research Report* (175).

Keynes, J. M. (1921), *A Treatise on Probability*, Macmillan & Co., London.

Khan, M., Serafeim, G. & Yoon, A. (2015), 'Corporate sustainability: First evidence on materiality', *Harvard Business School Working Paper* . No. 15-073.

Khan, M., Serafeim, G. & Yoon, A. (2016), 'Corporate sustainability: First evidence on materiality', *The accounting review* **91**(6), 1697–1724.

Kim, S., Robson, C., Zimmerman, T., Pierce, J. & Haber, E. M. (2011), Creek watch: pairing usefullness and usability for successful citizen science, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada, May 07-12, 2011', pp. 2125–2134.

Kingma, D. P. & Ba, J. (2014), 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980* .

Kingma, D. P. & Welling, M. (2013), 'Auto-encoding variational bayes', *arXiv preprint arXiv:1312.6114* .

Kirshner, J. & Power, M. G. (2015), Mining and extractive urbanism: Postdevelopment in a mozambican boomtown.

Knight, F. H. (1921), *Risk, Uncertainty, and Profit*, Houghton Mifflin Company, Boston.

Knox-Hayes, J. & Levy, D. (2011), 'The politics of carbon disclsure as climate governance', *Strategic Organization* **9**(1), 1–9.

Koks, E., Jongman, B., Husby, T. & Botzen, W. (2015), 'Combining hazard, exposure and social vulnerability to provide lessons for flood risk management', *Environmental Science Policy* **47**, 42 – 52.
URL: *http://www.sciencedirect.com/science/article/pii/S1462901114002056*

Konadu, D., Mourão, Z. S., Allwood, J. M., Richards, K. S., Kopec, G., McMahon, R. & Fenner, R. (2015), 'Land use implications of future energy system trajectories—the case of the uk 2050 carbon plan', *Energy Policy* **86**, 328 – 337.
URL: *http://www.sciencedirect.com/science/article/pii/S0301421515300197*

Konar, S. & Cohen, M. A. (1997), 'Information as regulation: The effect of community right to know laws on toxic emissions', *Journal of Environmental Economics and Management* **12**(1), 109–124.

Konda, N. M., Shah, N. & Brandon, N. P. (2011), 'Optimal transition towards a large-scale hydrogen infrastructure for the transport sector: The case for the netherlands', *International Journal of Hydrogen Energy* **36**(8), 4619–4635.

Kotsantonis, S., Pinney, C. & Serafeim, G. (2016), 'Esg integration in investment management: Myths and realities', *Journal of Applied Corporate Finance* **28**(2), 10–16.

Kovács, P. (2015), 'Minimum-cost flow algorithms: an experimental evaluation', *Optimization Methods and Software* **30**(1), 94–127.

Kozacek, C. (2014), 'Rio tinto's michigan nickel mine introduces citizen water quality testing program', Available at `http://www.circleofblue.org/2014/world/rio-tintos-michigan-nickel-mine-introduces-citizen-water-quality-testing-program/`. [accessed 2017.03.02].

Krause, F., Bach, W. & Koomey, J. (1989), *From warming fate to warming limit: benchmarks for a global climate convention*, Vol. 1, International Project for Sustainable Energy Paths.

Krishnamurthy, R. & Awazu, Y. (2016), 'Liberating data for public value: The case of data.gov', *International Journal of Information Management* **36**(4), 668–672.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012), Imagenet classification with deep convolution neural networks, *in* 'Advances in Neural Information Processing Systems 25', pp. 1106–1114.

Kruitwagen, G. (2020), personal communication.

Kruitwagen, L., Harnett, E. & Caldecott, B. (2016), Ultra high-net-worth individuals, private banks, and the state of sustainable investment, Technical report, Smith School of Enterprise and the Environment.

Kruitwagen, L., Stoy, K., Friedrich, J., Skillman, S. & Hepburn, C. (2019), A global census of solar facilities using deep learning and remote sensing, *in* 'NeurIPS Climate Change AI Workshop 2019', NeurIPS Climate Change AI.

Kuhn, T. S. (1962), *The Structure of Scientific Revolutions*, University of Chicago Press.

La Porta, R., Lopez-De-Silanes, F., Shleifer, A. & Vishny, R. W. (2000), 'Agency problems and dividend policies around the world', *Journal of Finance* **55**(1), 1–33.

Laffont, J.-J. & Martimort, D. (2001), *The Theory of Incentives: The Principal-Agent Model*, Princeton University Press, Princeton.

Laffont, J.-J. & Tirole, J. (1988), 'The dynamics of incentive contracts', *Econometrica* **56**(5), 1153–1175.

Lamarche, C., Santoro, M., Bontemps, S., d'Andrimont, R., Radoux, J., Giustarini, L., Brockmann, C., Wevers, J., Defourny, P. & Arino, O. (2017), 'Compilation and validation of sar and optical data products for a complete and global map of inland/ocean water tailored to the climate modeling community', *Remote Sensing* **9**(1), 36.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. & Jackel, L. D. (1989), 'Backpropagation applied to handwritten zip code recognition', *Neural Computation* **1**, 541–551.

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998), 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE* **86**(11), 2278–2324.

Lee, E. (2010), 'Information disclosure and environmental regulation: Green lights and gray areas', *Regulation & Governance* **4**(3), 303–328.

Leland, H. E. & Pyle, D. H. (1977), 'Informational asymetries, financial structure, and financial intermediation', *The Journal of Finance* **77**(2), 371–387.

Lewis, M. C. & Voisin, S. (2014), Oil & Carbon Revisited: Value at risk from "unburnable" reserves, Technical report, Kepler Cheuvreux.

Li, X. & Wu, X. (2014), 'Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition', *CoRR* **abs/1410.4281**.
URL: *http://arxiv.org/abs/1410.4281*

Liao, S.-H., Hu, D.-C. & Ding, L.-W. (2017), 'Assessing the influence of supply chain collaboration value innovation, supply chain capability and competitive advantage in taiwan's networking communication industry', *International Journal of Production Economics* **191**, 143–153.

Liddle, J. D., Holt, A. P., Jason, S. J., O'Donnell, K. A. & Stevens, E. J. (2020), 'Space science with cubesats and nanosatellites', *Nature Astronomy* **4**(11), 1026–1030.

Linux Foundation (2014), Collaborative development trends report, Technical report, Linux Foundation.

Liu, Y., Zeng, Q., Ordieres Meré, J. & Yang, H. (2019), 'Anticipating stock market of the renowned companies: A knowledge graph approach', *Complexity* **2019**.

Lord, R., Bullock, S. & Birt, M. (2019), Understanding climate risk at the asset level: The interplay of transition and physical risks, Technical report, SP Global. Available at: `https://www.spglobal.com/en/research-insights/featured/understanding-climate-risk-at-the-asset-level-the-interplay-of-transition-and-physical-risks` [Accessed 2020.11.30].

Luan, Y., He, L., Ostendorf, M. & Hajishirzi, H. (2018), 'Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction', *arXiv preprint arXiv:1808.09602* .

Lukyanenko, R., Parsons, J. & Wiersma, Y. F. (2016), 'Emerging problems of data quality in citizen science', *Conservation Biology* **30**(3), 447–449.

MacQueen, J. B. (1967), Some methods for classification and analysis of multivariate observations, *in* 'Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, Berkeley, California.

Magee, C. L. & Devezas, T. C. (2016), 'A simple extension of dematerialization theory: Incorporation of technical progress and the rebound effect', *Working Paper* .

Major Coal Projects (2014).
URL: *https://web.archive.org/web/20190322130543/https://minerals.org.au/files/major-coal-projects-ga-map-13-7856-1jpg*

Malof, J. M., Bradbury, K., Collins, L. M. & Newell, R. G. (2016), 'Automatic detection of solar photovoltaic arrays in high resolution aerial imagery', *Applied Energy* **183**, 229 – 240.
URL: *http://www.sciencedirect.com/science/article/pii/S0306261916313009*

Marmot, M. G. (2004), 'Evidence based policy or policy based evidence?', *BMJ* **328**, 906–907.

Martimort, D. & Sand-Zantman, W. (2013), 'Solving the global warming problem: beyond markets, simple mechanisms may help!', *Canadian Journal of Economics* **46**(2), 361–378.

Mattauch, M. (2015), '2015 - the fossil fuel endgame begins', *The Ecologist* . Available at `http://www.theecologist.org/blogs_and_comments/commentators/2709491/2015_the_fossil_fuel_endgame_begins.html` [accessed 2017.03.02].

Matten, D. & Moon, J. (2008), "'implicit" and "explicit csr": A conceptual framework for a comparative understanding of corporate social responsiblity', *Academy of Management Review* **33**(2), 404–424.

Maus, V., Giljum, S., Gutschlhofer, J., da Silva, D. M., Probst, M., Gass, S. L., Luckeneder, S., Lieber, M. & McCallum, I. (2020*a*), 'A global-scale data set of mining areas', *Scientific Data* **7**(1), 1–13.

Maus, V., Giljum, S., Gutschlhofer, J., da Silva, D. M., Probst, M., Gass, S. L., Luckeneder, S., Lieber, M. & McCallum, I. (2020*b*), 'A global-scale data set of mining areas', *Scientific Data* **7**(1), 1–13.

Mazzucato, M. (2018), 'Mission-oriented innovation policies: challenges and opportunities', *Industrial and Corporate Change* **27**(5), 803–815.

McCain, R. A. (1988), 'Information as property and as a public good: Perspectives from the economic theory of property rights', *The Library Quarterly: Information, Community, Policy* **58**(3), 265–282.

McClelland, J. D. & Horowitz, J. K. (1999), 'The costs of water pollution regulation in the us pulp and paper industry', *Land Economics* **75**(2), 220–232.

McCoy, R. T., Pavlick, E. & Linzen, T. (2019), 'Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference', *arXiv preprint arXiv:1902.01007* .

McGlade, C. & Ekins, P. (2015), 'The geographyical distribution of fossil fuels unused when limiting global warming to 2^oc', *Nature* **517**(7533), 187–190.

McKibben, B. (2012), 'Global warming's terrifying new math', *The Rolling Stone* . Available at `http://www.rollingstone.com/politics/news/global-warmings-terrifying-new-math-20120719` [accessed 2017.03.02].

McKie, R. (2020), 'The moon, mars and beyond... the space race in 2020', *The Guardian* . **URL:** *https://www.theguardian.com/science/2020/jan/05/space-race-moon-mars-asteroids-commercial-launches*

Meinshausen, M., Meinshausen, N., Hare, W., Raper, S. C. B., Frieler, K., Knutti, R., Frame, D. J. & Allen, M. R. (2009), 'Greenhouse-gas emission targets for limiting global warming to 2^oc', *Nature* **458**(7242), 1158–1162.

Melas, D. (2019), 'The future of emerging markets'. **URL:** *https://www.msci.com/documents/10199/239004/Research-Insight-The-Future-of-Emerging-Markets/13e0765f-9506-4011-2d6d-704346c1cc30*

Menou, V. & Nishikawa, L. (2015), Toward sustainable impact through public markets, Technical report, MSCI ESG Research Inc. Available at `http://2degrees-investing.org/IMG/pdf/sei_metrics_summary-3.pdf` [accessed 2017.03.07].

Meraner, A., Ebel, P., Zhu, X. X. & Schmitt, M. (2020), 'Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion', *ISPRS Journal of Photogrammetry and Remote Sensing* **166**, 333–346.

Mercer LLC (2015), Investing in a time of climate change, Technical report, Mercer LLC, International Finance Corporation, UK Department for International Development. Available at `http://2degrees-investing.org/IMG/pdf/sei_metrics_summary-3.pdf` [accessed 2017.03.07].

Meyers, S. C. & Majluf, N. S. (1984), 'Corporate financing and investment decisions when firms have information that investors do not have', *The National Bureau of Economic Research* **13**(2), 187–221.

Microway Inc. (2017), 'Nvidia tesla p100 price analysis', Available at `https://www.microway.com/hpc-tech-tips/nvidia-tesla-p100-price-analysis/`. [accessed 2017.03.02].

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013), 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781* .

Milgrom, P. R. (1981), 'Good news and bad news: Representation theorems and applications', *The Bell Journal of Economics* **12**(2), 380–391.

Milner, M. J. (2002), 'Positive accounting theory, political costs and social disclosure analyses: A critical look', *Critical Perspectives on Accounting* **13**(3), 369–395.

Mining (1979). **URL:** *https://legacy.lib.utexas.edu/maps/africa/south$_r$hodesia$_m$ine$_1$979.jpg*

Mirrlees, J. A. (1971), 'An exploration in the theory of optimum income taxation', *The Review of Economic Studies* **38**(2), 175–208.

Mishra, R. (1979), 'Technology and social structure', *Science & Society* **43**(2), 132–157.

Mitroff, I. (1983), *Stakeholders of the organizational mind*, Jossey-Bass, San Francisco.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G. et al. (2015), 'Human-level control through deep reinforcement learning', *Nature* **518**(7540), 529.

Mol, A. P. (2006), 'Environmental governance in the information age: the emergence of informational governance', *Environment and Planning C: Government and Policy* **24**(4), 497–514.

Moore, G. E. (1965), 'Cramming more components onto integrated circuits', *Electronics* pp. 114–117.

Moore, J. (2001), *AAAS Science and Technology Policy Yearbook*, American Association for the Advancement of Science, Washington, D.C., chapter Frankenfood or Doubly Green Revolution: Europe vs. America on the GMO debate.

MSCI ESG Research (2013), Responding to the call for fossil-fuel free portfolios, Technical report. Available at `http://fossilfreeindexes.com/wp-content/uploads/bp-attachments/356/MSCI-ESG-Research_FAQ-on-Fossil-Free-Investing_June-2013.pdf` [accessed 2017.03.07].

Mulvaney, D. (2017), 'Identifying the roots of green civil war over utility-scale solar energy projects on public lands across the american southwest', *Journal of Land Use Science* **12**(6), 493–515.
URL: *https://doi.org/10.1080/1747423X.2017.1379566*

Murphy, D. F. & Bendell, J. (1999), Partners in time? business, ngos, and sustainable development, Technical report, United Nations Research Institute for Social Development. UNRISD Discussion Paper No. 109.

Nace, T., Plant, L. & Browning, J. (2019), 'The new gas boom: tracking global lng infrastructure', *Global Energy Monitor* .

Nagar, V. (1999), 'The role of the managers' human capital in discretionary disclosure', *Journal of Accounting Research* **37**(1), 167–181.

Nakbanpote, W., Prasad, M. N., Mongkhonsin, B., Panitlertumpai, N., Munjit, R. & Rattanapolsan, L. (2018), Chapter 33 - strategies for rehabilitation of mine waste/leachate in thailand, *in* M. N. V. Prasad, P. J. de Campos Favas & S. K. Maiti, eds, 'Bio-Geotechnologies for Mine Site Rehabilitation', Elsevier, pp. 617 – 643.
URL: *http://www.sciencedirect.com/science/article/pii/B9780128129869000336*

Ndikumana, E., Ho Tong Minh, D., Baghdadi, N., Courault, D. & Hossard, L. (2018), 'Deep recurrent neural network for agricultural classification using multitemporal sar sentinel-1 for camargue, france', *Remote Sensing* **10**(8), 1217.

Neale, A. (1997), 'Organizational learning in contested environments: Lessons from brent spar', *Business Strategy and Environment* **6**(2), 93–103.

Nelson, A., Setiyono, T., Rala, A. B., Quicho, E. D., Raviz, J. V., Abonete, P. J., Maunahan, A. A., Garcia, C. A., Bhatti, H. Z. M., Villano, L. S. et al. (2014), 'Towards an operational sar-based rice monitoring system in asia: Examples from 13 demonstration sites across asia in the riice project', *Remote Sensing* **6**(11), 10773–10812.

Neu, D., Warsame, H. & Pedwell, K. (1998), 'Managing public impressions: Environmental disclosures in annual reports', *Accounting, Organisations and Society* **23**(3), 265–282.

Newman, G., Wiggins, A., Crall, A., Graham, E. & Newamn, S. (2012), 'The future of citizen science: emerging technologies and shifting paradigms', *Frontiers in Ecology and the Environment* **10**(6), 298–304.

Nordhaus, W. (2015), 'Climate clubs: Overcoming free-riding in international climate policy', *American Economic Review* **105**(4), 1339–1370.

Oei, P.-Y. & Mendelevitch, R. (2019), 'Colombian coal'.
URL: *https://thecoalhub.com/articles/colombian-coal*

Ong, S., Campbell, C., Denholm, P., Margolis, R. & Heath, G. (2013), Land-use requirements for solar power plants in the united states, Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States).

Oord, A. v. d., Li, Y. & Vinyals, O. (2018), 'Representation learning with contrastive predictive coding', *arXiv preprint arXiv:1807.03748* .

OpenStreetMap contributors (2017), 'Planet dump retrieved from https://planet.osm.org ', `https://www.openstreetmap.org`.

OpenStreetMap contributors (2019), 'Planet dump retrieved from https://planet.osm.org ', `https://www.openstreetmap.org`.

OpenStreetMap contributors (2020), 'Planet dump retrieved from https://planet.osm.org ', `https://www.openstreetmap.org`.

Orlin, J. B. (1997), 'A polynomial time primal network simplex algorithm for minimum cost flows', *Mathematical Programming* **78**(2), 109–129.

Orlitzky, M., Schmidt, F. L. & Rynes, S. L. (2003), 'Corporate social and finance performance: A meta-analysis', *Organization Studies* **24**(3), 403–441.

Ostrom, E. (1990), *Governing the Commons*, Cambridge University Press.

Pachauri, R. K. & Reisinger, A. (2007), 'Ipcc fourth assessment report', *IPCC, Geneva* **2007**.

Pakistan and its Mineral Resources (2015).
URL: *http://pakistanminerals.blogspot.com/2015/*

Paltsev, S., Sokolov, A. P., Gao, X. & Haigh, M. (2018), 'Meeting the goals of the paris agreement: Temperature implications of the shell sky scenario'.

Panahifar, F., Byrne, P. J., Salam, M. A. & Heavey, C. (2018), 'Supply chain collaboration and firm's performance', *Journal of Enterprise Information Management* .

Pantelia, A. (2018), 'Lignite mining: Greece's dirty secret - in pictures'.
URL: *https://www.theguardian.com/environment/2018/mar/23/lignite-mining-greeces-dirty-secret-in-pictures*

Park, C. (2000), 'Monitoring and structure of debt contracts', *The Journal of Finance* **55**(5), 2157–2195.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N. & Ku, A. (2018), 'Image transformer', *CoRR* **abs/1802.05751**.
URL: *http://arxiv.org/abs/1802.05751*

Patel, N. V. (2020), 'Are we in the middle of a new space race for this century?', *MIT Technology Review* .
URL: *https://www.technologyreview.com/2020/02/07/349033/are-we-in-the-middle-of-a-new-space-race-for-this-century/*

Pearl, J. (1985), Bayesian networks: A model of self-activated memory for evidential reasoning, *in* 'Seventh Annual Conference of the Cognitive Science Society', pp. 1–20.

Pearson, K. (1894), 'Contributions to the mathematical theory of evolution', *Philosophical Transactions of the Royal Society of London. A* **185**, 71–110.

Perez, C. (1985), 'Microelectronics, long waves and world structural change: new perspectives for developing countries', *World Development* **13**(3), 357–375.

Pfeiffer, A., Millar, R., Hepburn, C. & Beinhocker, E. (2016), 'The 2c capital stock for electricity generation: Committed cumulative carbon emissions from the electricity generation sector and the transition to a green economy', *Applied Energy* **179**(1), 1395–1408.

Platts, S. G. (2018), 'World electric power plant database'.
URL: *https://www.spglobal.com/platts/en/products-services/electric-power/world-electric-power-plants-database*

Polanyi, K. (1944), *The Great Transformation*, Farrar & Rinehart, New York.

Pratt, L. Y. (1993), Discriminability-based transfer between neural networks, *in* 'Advances in neural information processing systems', pp. 204–211.

Prize Committee of the Royal Swedish Academy of Sciences (2007), Mechanism design theory, Technical report, The Royal Swedish Academy of Sciences.

Project Payments Database (2019).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019), 'Language models are unsupervised multitask learners', *OpenAI Blog* **1**(8).

Ramanathan, K. V. (1976), 'Towards a theory of corporate social accounting', *The Accounting Review* **51**(3), 516–528.

Ranada, P. (2013), 'Green groups to doe: What happened to renewable energy?'.
URL: *https://www.rappler.com/nation/43029-pmcj-doe-renewable-energy-act*

Revelli, C. & Viviani, J.-L. (2015), 'Financial performance of socially responsible investing (sri): What have we learned? a meta-analysis', *Business Ethics: A European Review* **24**(2), 158–185.

Ricardo, D. (1817), *On the Principles of Political Economy and Taxation*, John Murray, London.

Rogers, J. L., Skinner, D. J. & Zechman, S. L. (2017), 'Run edgar run: Sec dissemination in a high-frequency world', *Journal of Accounting Research* **55**(2), 459–505.

Ronneberger, O., Fischer, P. & Brox, T. (2015*a*), 'U-net: Convolutional networks for biomedical image segmentation', *CoRR* **abs/1505.04597**.
URL: *http://arxiv.org/abs/1505.04597*

Ronneberger, O., Fischer, P. & Brox, T. (2015*b*), 'U-net: Convolutional networks for biomedical image segmentation', *CoRR* **abs/1505.04597**.
URL: *http://arxiv.org/abs/1505.04597*

Rose, F. (1999), *The economics, concept, and design of iformation intermediaries*, Heidelberg, New York.

Rose, K., Bauer, J., Baker, V., Barkhurst, A., Bean, A., DiGiulio, J., Jones, K., Jones, T., Justman, D., Miller, R., Sabbatino, M. & Tong, A. (2018), 'Global oil gas features database'.

Rose, K., Bauer, J., Baker, V., Barkhurst, A., Bean, A., DiGiulio, J., Jones, K., Jones, T., Justman, D., Miller, R. et al. (2018*a*), Global oil & gas features database, Technical report, National Energy Technology Laboratory (NETL), Pittsburgh, PA, Morgantown, WV ....

Rose, K., Bauer, J., Baker, V., Barkhurst, A., Bean, A., DiGiulio, J., Jones, K., Jones, T., Justman, D., Miller, R. et al. (2018*b*), Global oil & gas features database, Technical report, National Energy Technology Laboratory (NETL), Pittsburgh, PA, Morgantown, WV ....

Rosenblatt, F. (1958), 'The perceptron: A probabilistic model for information storage and organization in the brain', *Psychological Review* **65**, 386–408.

Ross, S. (1973), 'The economic theory of agency: The principal's problem', *American Economic Review* **63**(2), 134–139.

Ross, S. A. (1977), 'The determination of financial structure: the incentive-signalling approach', *The Bell Journal of Economics* **8**(1), 23–40.

Rosset, C. (2019), 'Turing-nlg: A 17-billion-parameter language model by microsoft', *Microsoft Blog* .

Rothman, K. J. & Poole, C. (1985), 'Science and policy making', *American Journal of Public Health* **75**(4), 340–341.

Roundtable, B. (2019), 'Statement on the purpose of a corporation'.

Roweis, S. T. & Saul, L. K. (2000), 'Nonlinear dimensionality reduction by locally linear embedding', *science* **290**(5500), 2323–2326.

Rüetschi, M., Schaepman, M. E. & Small, D. (2018), 'Using multitemporal sentinel-1 c-band backscatter to monitor phenology and classify deciduous and coniferous forests in northern switzerland', *Remote Sensing* **10**(1), 55.

Rumelhart, D., Hinton, G. & Williams, R. (1986), 'Learning representations by back-propagating errors', *Nature* **323**, 533–536.

Rusbridger, A. (2015), 'Keep it in the ground', *The Guardian* . Available at `https://www.theguardian.com/environment/2015/mar/16/argument-divesting-fossil-fuels-overwhelming-climate-change?CMP=share_btn_tw` [accessed 2017.03.02].

Sadiq, S. & Indulska, M. (2017), 'Open data: Quality over quantity', *International Journal of Information Management* **37**, 150–154.

Salas, R. N., Shultz, J. M. & Solomon, C. G. (2020), 'The climate crisis and covid-19—a major threat to the pandemic response', *New England Journal of Medicine* **383**(11), e70.

SAMS Admin (2017), 'Coalition joins in defense of stream protection rule', Available at `http://www.samsva.org/`. [accessed 2017.03.02].

Samuelson, P. A. (1954), 'The pure theory of public expenditure', *The Review of Economics and Statistics* **36**(4), 387–389.

Sand, P. H. (2002), The right to know: Environmental information disclosure by government and industry, *in* 'Conference on" Human Dimensions of Global Environmental Change: Knowledge for the Sustainability Transition"(Berlin, 7 December 2002)'.

SASB (2018), 'Sasb materiality map'.

Schere, A. G. & Palazzo, G. (2011), 'The new political role of business in a globalized world: A review of new persepctives on csr and its implications for the firm, governance, and democracy', *Journal of Management Studies* **48**(4), 899–931.

Schiavina, M., Freire, S. & MacManus, K. (2019), 'Ghs population grid multitemporal (1975, 1990, 2000, 2015)'.
URL: *http://data.europa.eu/89h/0c6b9751-a71f-4062-830b-43c9f432370f*

Schumpeter, J. (1949), *Change and the Entrepreneur: Postulates and Patters for Entrepreneurial History*, Harvard University Press, Cambridge, chapter Economic Theory and Entrepreneurial History.

Schwab, K. (2019), 'Davos manifesto 2020: The universal purpose of a company in the fourth industrial revolution', *WEF website* .

Sehra, S. S., Singh, J. & Rai, H. S. (2017), 'Assessing openstreetmap data using intrinsic quality indicators: an extension to the qgis processing toolbox', *Future Internet* **9**(2), 15.

Seljom, P. (2019), Oil and gas logistics, Technical report.

Service, C. L. M. (2020*a*), 'Corine land cover'.
URL: *https://land.copernicus.eu/pan-european/corine-land-cover*

Service, N. A. S. (2020*b*), 'Cropland data layer'.
URL: *https://www.nass.usda.gov/Research*$_a$*nd*$_S$*cience/Cropland/SARS*1*a.php*

Setiyono, T., Holecz, F., Khan, N., Barbieri, M., Quicho, E., Collivignarelli, F., Maunahan, A., Gatti, L. & Romuga, G. (2017), Synthetic aperture radar (sar)-based paddy rice monitoring system: Development and application in key rice producing areas in tropical asia, *in* 'IOP Conference Series: Earth and Environmental Science', Vol. 54, IOP Publishing, p. 012015.

Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Hall, W. & Schraefel, M. C. (2012), 'Linked open government data: Lessons from data.gov.uk', *IEEE Intelligent Systems* **27**(3), 16–24.

Shearer, C., Ghio, N., Myllyvirta, L., Yu, A. & Nace, T. (2017*a*), *Boom and bust 2017: Tracking the global coal plant pipeline*, CoalSwarm.

Shearer, C., Ghio, N., Myllyvirta, L., Yu, A. & Nace, T. (2017*b*), *Boom and bust 2017: Tracking the global coal plant pipeline*, CoalSwarm.

Sheng, H., Irvin, J., Munukutla, S., Zhang, S., Cross, C., Story, K., Rustowicz, R., Elsworth, C., Yang, Z., Omara, M. et al. (2020), 'Ognet: Towards a global oil and gas infrastructure database using deep learning on remotely sensed imagery', *arXiv preprint arXiv:2011.07227* .

Shleifer, A. & Vishny, R. W. (1997), 'A survey of corporate governance', *The Journal of Finance* **52**(2), 737–783.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T. et al. (2018), 'A general reinforcement learning algorithm that masters chess, shogi, and go through self-play', *Science* **362**(6419), 1140–1144.

Silvertown, J. (2009), 'A new dawn for citizen science', *Trends in Ecology & Evolution* **24**(9), 467–471.

Simerly, R. L. (1994), 'Corporate social performance and firms' financial performance: An alternative perspective', *Psychological Reports* **75**(3), 1091–1103.

Simonyan, K. & Zisserman, A. (2015*a*), Very deep convolutional networks for large-scale image recognition, *in* 'International Conference on Learning Representations'.

Simonyan, K. & Zisserman, A. (2015*b*), Very deep convolutional networks for large-scale image recognition, *in* 'International Conference on Learning Representations'.

Skancke, M., Dimson, E., Hoel, M., Kettis, M., Nystuen, G. & Starks, L. (2014*a*), 'Fossil-fuel investments in the norwegian government pension fund global: Addressing climate issues through exclusion and active ownership', *Report, Expert Group appointed by the Norwegian Ministry of Finance, Oslo* .

Skancke, M., Dimson, E., Hoel, M., Kettis, M., Nystuen, G. & Starks, L. (2014*b*), Fossil-Fuels Investments in the Norwegian Government Pension Fund Global: Addressing Climate Issues through Exclusion and Active Ownership, Technical report, Expert Group Appointed by the Norwegian Ministry of Finance.

Smith, A. (1776), *The Wealth of Nations*, W. Strahan and T. Cadell, London.

Smith, H. J. (2003), 'The shareholders vs. stakeholders debate', *MIT Sloan Management Review* **44**(4), 85–90.

SolarGIS (2016), 'Global solar atlas'.
URL: *https://globalsolaratlas.info/downloads/world*

Solow, R. M. (1957), 'Technical change and the aggregate production function', *The Review of Economics and Statistics* **39**(3), 312–320.

Spedding, P., Mehta, K. & Robins, N. (2013), Oil & Carbon Revisited: Value at risk from "unburnable" reserves, Technical report, HSBC.

Spence, M. (1973), 'Job market signaling', *The Quarterly Joural of Economics* **87**(3), 355–374.

Stallman, R. (2016), 'Why "open source" misses the point of free software', *Communications of the ACM* **52**(6).

Standard & Poor's Ratings Services (2014), Corporate ratings methodology, Technical report, Standard & Poor's.

Stern, N. (2015), *Why are we waiting?: The logic, urgency, and promise of tackling climate change*, Mit Press.

Stigler, G. J. (1961), 'The economics of information', *The Journal of Political Economy* **69**(3), 213–225.

Stiglitz, J. E. (1999), Knowledge as a global public good, *in* 'Global Public Goods: International Cooperation in the 21st Century', pp. 308–325.

Stiglitz, J. E. (2000), 'The contributions of the economics of information to twentieth century economics', *The Quarterly Journal of Economics* **115**(4), 1441–1478.

Stiglitz, J. E. (2002), 'Information and the change in the paradigm in economics', *The American Economic Review* **92**(3), 460–501.

Stoughton, N. M. (1993), 'Moral hazard and the portfolio management problem', *The Journal of Finance* **48**(5), 2009–2028.

Subramanian, K., Lemaire, J., Hershey, J. C., Pauly, M. V., Armstrong, K. & Asch, D. A. (1999), 'Estimating adverse selection costs from genetic testing for breast and ovarian cancer: The case of life insurance', *The Journal of Risk and Insurance* **66**(4), 531–550.

Sutskever, I., Vinyals, O. & Le, Q. V. (2014), 'Sequence to sequence learning with neural networks', *CoRR* **abs/1409.3215**.
URL: *http://arxiv.org/abs/1409.3215*

Sutton, R. (2019), 'The bitter lesson', Available at `http://www.incompleteideas.net/IncIdeas/BitterLesson.html`. [accessed 2020.12.03].

Sutton, R. S. & Barto, A. G. (199), *Reinforcement Learning: An introduction*, MIT Press.

Sverdlik, Y. (2014), 'Linux foundation: Open source is eating the software world', Available at `http://www.datacenterknowledge.com/archives/2014/11/03/linux-foundation-open-source-eating-enterprise-software-world/`. [accessed 2017.08.02].

Swope, A. M., Rudelis, X. H. & Story, K. T. (2020), 'Representation learning for remote sensing: An unsupervised sensor fusion approach'.
URL: *https://openreview.net/forum?id=SJlVn6NKPB*

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2015), Going deeper with convolutions, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 1–9.

TCFD (2016), Recommendations of the task force on climate-related financial disclosures, Technical report, Financial Stability Board.

Tegmark, M. (2017), *Life 3.0*, Knopf.

Teoh, S. H. & Hwang, C. Y. (1991), 'Nondisclosure and adverse disclosure as signals of firm value', *Review of Financial Studies* **4**(2), 283–313.

The Bank of England (2015), The impact of climate change on the uk insurance sector, Technical report, Prudential Regulatory Authority.

The History of Computing Project (2013), 'Cray 1', Available at `https://www.thocp.net/hardware/cray_1.htm`. [accessed 2017.03.02].

Thomä, J., Dupré, S., Hayne, M., Weber, C., Hassan, F. & Fulton, M. (2015), Assessing the alignment of portfolios with climate goals, Technical report, 2^o Investing Initiative. Available at `http://2degrees-investing.org/IMG/pdf/2dportfolio_v0_small.pdf` [accessed 2017.03.07].

Tian, Y., Krishnan, D. & Isola, P. (2019), 'Contrastive multiview coding', *arXiv preprint arXiv:1906.05849* .

Tin Kam Ho (1995), Random decision forests, *in* 'Proceedings of 3rd International Conference on Document Analysis and Recognition', Vol. 1, pp. 278–282 vol.1.

Tobler, W. R. (1970), 'A computer movie simulating urban growth in the detroit region', *Economic geography* **46**(sup1), 234–240.

Tong, D., Zhang, Q., Zheng, Y., Caldeira, K., Shearer, C., Hong, C., Qin, Y. & Davis, S. J. (2019), 'Committed emissions from existing energy infrastructure jeopardize 1.5 c climate target', *Nature* **572**(7769), 373–377.

Toonders, J. (2014), 'Data is the new oil of the information economy', *WIRED* .

Torbick, N., Huang, X., Ziniti, B., Johnson, D., Masek, J. & Reba, M. (2018), 'Fusion of moderate resolution earth observations for operational crop type mapping', *Remote Sensing* **10**(7), 1058.

Trüby, J. (2013), 'Strategic behaviour in international metallurgical coal markets', *Energy Economics* **36**, 147–157.

Turconi, R., Boldrin, A. & Astrup, T. (2013), 'Life cycle assessment (lca) of electricity generation technologies: Overview, comparability and limitations', *Renewable and Sustainable Energy Reviews* **28**, 555 – 565.
URL: *http://www.sciencedirect.com/science/article/pii/S1364032113005534*

Tweet: Major coal mines and coal-fired power plants in China (2015).
URL: *https://twitter.com/IEA/status/580120032539230208?s=20*

UN General Assembly (2015), Transforming our world: the 2030 agenda for sustainable development, Technical report, Decentralized Autonomous Organization Integral Platform for Climate Initiative. A/RES/70/1, Available at `https://www.refworld.org/docid/57b6e3e44.html` [accessed 2020.11.30].

UNEP Financial Inquiry (2015), 'The financial system we need', *Aligning the Financial System with Sustainable Development. Geneva: UNEP Inquiry into the Design of a Sustainable Financial System* .

UNEP-WCMC, I. (2016), 'World database on protected areas'.

UNFCCC (2009), Copenhagen accord, Technical report, United Nations Framework Convention on Climate Change Fifteenth Session of the Conference of Parties.

UNFCCC (2010), Cancun agreements, Technical report, United Nations Framework Convention on Climate Change Fifteenth Session of the Conference of Parties.

UNFCCC (2015), The paris agreement, Technical report, United Nations Framework Convention on Climate Change Twenty-First Session of the Conference of Parties.

Union of Concerned Scientists (2020), 'Ucs satellite database'.

US Energy Information Administration (2020), EIA-860, Technical report, Washington DC, USA.
URL: *https://www.eia.gov/maps/layer*$_i nfo - m.php$

van Rooij, S. W. (2007), 'Shahron williams van rooij', *Journal of Research on Technology in Education* **39**(4), 433–453.

Van Tricht, K., Gobin, A., Gilliams, S. & Piccard, I. (2018), 'Synergistic use of radar sentinel-1 and optical sentinel-2 imagery for crop mapping: a case study for belgium', *Remote Sensing* **10**(10), 1642.

Vapnik, V. & Lerner, A. (1963), 'Pattern recognition using generalized portrait method', *Automation and Remote Control* **24**, 774–780.

Varon, D. J., Jacob, D. J., Jervis, D. & McKeever, J. (2020), 'Quantifying time-averaged methane emissions from individual coal mine vents with ghgsat-d satellite observations', *Environmental Science & Technology* **54**(16), 10246–10253.

Vasco, D. W., Farr, T. G., Jeanne, P., Doughty, C. & Nico, P. (2019), 'Satellite-based monitoring of groundwater depletion in california's central valley', *Scientific reports* **9**(1), 1–14.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017*a*), Attention is all you need, *in* 'Advances in neural information processing systems', pp. 5998–6008.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017*b*), 'Attention is all you need', *arXiv preprint arXiv:1706.03762* .

Végh, L. A. (2016), 'A strongly polynomial algorithm for a class of minimum-cost flow problems with separable convex objectives', *SIAM Journal on Computing* **45**(5), 1729–1761.

Verrecchia, R. E. (1983), 'Discretionary disclosure', *Journal of Accounting and Economics* **5**(1), 179–194.

Verrecchia, R. E. (2001), 'Essays on disclosure', *Journal of Accounting and Economics* **32**(1–3), 97–180.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P. et al. (2019), 'Grandmaster level in starcraft ii using multi-agent reinforcement learning', *Nature* **575**(7782), 350–354.

von Hippel, E. (2012), 'Lecture notes: The paradigm shift from producer to user innovation'.

Vorsler, C. (2002), 'Simplified geology and activate coal mines south africa, lesotho, and swaziland'.
URL: *http://www.geoscience.org.za/images/Maps/coalactive.gif*

Wang, A. X., Tran, C., Desai, N., Lobell, D. & Ermon, S. (2018), Deep transfer learning for crop yield prediction with remote sensing data, *in* 'Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies', ACM, p. 50.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X. & Tang, X. (2017), Residual attention network for image classification, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3156–3164.

Wang, H., Li, G., Wang, G., Peng, J., Jiang, H. & Liu, Y. (2017), 'Deep learning based ensemble approach for probabilistic wind power forecasting', *Applied Energy* **188**, 56 – 70.
URL: *http://www.sciencedirect.com/science/article/pii/S0306261916317421*

Wang, Z., Zhang, J., Feng, J. & Chen, Z. (2014), Knowledge graph and text jointly embedding, *in* 'Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)', pp. 1591–1601.

Watts, R. (1973), 'The information content of dividends', *The Journal of Business* **46**(2), 191–211.

Watts, R. L. & Zimmerman, J. L. (1978), 'Towards a positive theory of the determination of accounting standards', *The Accounting Review* **53**(1), 112–134.

Watts, R. L. & Zimmerman, J. L. (1990), 'Postitive accounting theory: A ten year perspective', *The Accounting Review* **65**(1), 131–156.

Weber, C., Dupre, S., Thomae, J. & Braschi, T. (2017), Asset-level data and climate-related financial analysis, Technical report. Available at `https://www.transitionmonitor.com/wp-content/uploads/2017/04/assetdata_v0.pdf` [accessed 2020.11.30].

Weber, S. (2000), The political economy of open source software, *in* 'BRIE Working Paper 140'.

WFE Sustainability Working Group (2015), Exchange guidance & recommendation, Technical report, World Federation of Exchanges. Available at `https://www.world-exchanges.org/home/index.php/news/world-exchange-news/world-exchanges-agree-enhanced-sustainability-guidance` [accessed 2017.03.07].

Wiersma, Y., Parsons, J. & Lukyanenko, R. (2016), Data quality in citizen science - a research study, *in* 'They Walk Among Us: The Rise of Citizen Science', Journal of the Institution of Environmental Sciences, pp. 75–78.

Wildsmith, H. (2012), 'Why asset owners and mutuals are "aiming for a"', *Responsible Investor* . Available at `https://www.responsible-investor.com/articles/why-were-aiming-for-a/` [accessed 2020.11.30].

Williams, C. K. & Rasmussen, C. E. (1996), Gaussian processes for regression, *in* 'Advances in neural information processing systems', pp. 514–520.

Wilson, C. (2015), 'Mexico resources mining'.
URL: *https://www.tes.com/lessons/imAOK-WpTHGL1g/comparison-of-mexico-and-venezuela*

Wold, E. (2015), 'Bottom-up modelling of future global oil supply and associated spend', *Oil Age* **1**(4), 19–39.

Woods, M. (2003), 'The global reporting initiative', *The CPA journal* **73**(6), 60.

WRI, L. (2017), 'Landmark: Global platform of indigenous and community lands'.
URL: *http://www.landmarkmap.org/*

Wu, G. C., Leslie, E., Allen, D., Sawyerr, O., Cameron, D. R., Brand, E., Cohen, B., Ochoa, M. & Olson, A. (2019), Power of Place: Land Conservation and Clean Energy Pathways for California, Technical report.

Yu, J., Wang, Z., Majumdar, A. & Rajagopal, R. (2018), 'Deepsolar: A machine learning framework to efficiently construct a solar deployment database in the united states', *Joule* **2**(12), 2605 – 2617.
URL: *http://www.sciencedirect.com/science/article/pii/S2542435118305701*

Yuk, P. K. (2011), 'Us, china, russia: winners in the gobi'.
URL: *https://www.ft.com/content/12879151-bfec-3695-bfc3-1771d4fff0c8*

Zador, A. M. (2000), 'The basic unit of computation', *Nature Neuroscience* **3**(11), 1167–1167.

Zhang, H., Goodfellow, I., Metaxas, D. & Odena, A. (2019), Self-attention generative adversarial networks, *in* 'International conference on machine learning', PMLR, pp. 7354–7363.

Zhang, H. & Malczewski, J. (2019), Quality evaluation of volunteered geographic information: The case of openstreetmap, *in* 'Crowdsourcing: Concepts, Methodologies, Tools, and Applications', IGI Global, pp. 1173–1201.

Zhang, Q., Kong, Q., Zhang, C., You, S., Wei, H., Sun, R. & Li, L. (2019), 'A new road extraction method using sentinel-1 sar images based on the deep fully convolutional neural network', *European Journal of Remote Sensing* **52**(1), 572–582.

Zhang, Z., Liu, Q. & Wang, Y. (2017), 'Road extraction by deep residual u-net', *CoRR* **abs/1711.10684**.
URL: *http://arxiv.org/abs/1711.10684*

Zhang, Z., Liu, Q. & Wang, Y. (2018), 'Road extraction by deep residual u-net', *IEEE Geoscience and Remote Sensing Letters* **15**(5), 749–753.

Zuo, J., Xu, G., Fu, K., Sun, X. & Sun, H. (2018), 'Aircraft type recognition based on segmentation with deep convolutional neural networks', *IEEE Geoscience and Remote Sensing Letters* **15**(2), 282–286.