# What Actually Happened? Novel Econometric Methods to Improve Estimates of Climate Impacts and Policies


Moritz P. Schwarz

Brasenose College

University of Oxford

A thesis submitted for the degree of

*Doctor of Philosophy*

Hilary 2023

# Acknowledgements

This four year journey of putting this doctoral thesis together would not have been possible without the immense support of a long list of incredible people.

First of all, I want to thank my longest academic supporter, most important co-author, and friend Prof Felix Pretis. Without Felix, I have no doubt that I would not have managed to finish this thesis – and indeed it was a short conversation with him at Nuffield College in 2017 that has paved the way for my academic career. Felix has continuously been extremely supportive and has always elevated people around him to make sure that they get the recognition that they deserve. I am incredibly glad to have been lucky enough to work with Felix for many years and am excited for all the amazing projects that lie ahead.

I am immensely grateful for the invaluable support of my two fantastic DPhil supervisors, Prof Sir David F. Hendry and Prof Cameron Hepburn.

David has given me the feeling that I could always turn to him with questions and queries, that he would always find the time should I need his advice, and that he values my thoughts and contributions greatly. I am especially grateful to him for creating an environment in Oxford that has become my academic home over most of my academic career. To be able to learn from someone of his academic stature and with his kind, thoughtful, and compassionate manner has been an incredible experience to work with him for which I am very grateful.

There were of course moments in this DPhil journey when I questioned whether I had what it takes to continue; whether investing so much time and effort into this document was actually worth it. There were moments when I envied the success of others and doubted I could ever achieve something similar – in other words, at times I lived out the imposter syndrome to its fullest. However, whenever such moments arrived, I knew that I could always turn to Cameron. Cameron has the unique gift, aside from his obvious professional skills and acing life in general, of being able to lift my spirits and pick my confidence back up in just a thirty minute conversation. After each interaction with Cameron, I was more determined to reach for the stars and to strive for the next challenge than before. Pair this with his incredible intuition for attention-grabbing research topics, his razor-sharp mind, and his high degree of empathy, I am incredibly grateful to have had Cameron's support throughout the years.

The community of Climate Econometrics, both at Oxford and all across the Climate Econometrics Network, has allowed me to find an academic home where I could strive, challenge myself, eat fantastic lunches, and find close friends for life. To Angela who would always give me the time of day when complaining about Brexit or Trump, to Lisa and Jonas who I am lucky to call friends and collaborators at the same time, to Xiyu who has surprised me with small acts of kindness so many times (and who I will someday watch playing Basketball again) and to Ryan, Susana, and Sam who have inspired me countless times to find new academic endeavours and who made sure that we celebrated every success that came along. Jennie has been one of the most outstanding teachers that I have been lucky enough to come across and I am convinced that Climate Econometrics is in fantastic hands with her at the helm. I am also grateful to have had the opportunity to meet a number of impressive MSc students and am excited by the incredible work that Antonina and Ebba have created, despite my lousy co-supervisions for their dissertations. Climate Econometrics has been my home in the DPhil journey and no matter how often I complained to each one of them about my inability to stop procrastinating, they have been the most incredible support throughout.

I am grateful to the dedicated and supportive colleagues at the Austrian Ministry of Finance and my former colleagues at the Austrian Ministry for Climate Action. Throughout the years, I have been inspired by countless colleagues and friends within the Budget DG, the Climate Team, the Austrian COP delegation, and many others. I am especially grateful to Elfi More, whose professionalism, empathy, and friendship I cherish greatly and who I will always consider to be somewhat of a mentor for me. I am also glad to finally be able to tell my friend and colleague Jesus that I have finally managed to finish this doctorate and that now I won't have to laugh off his questions about my progress with him anymore. I am grateful to José for always having my back and supporting me in my academic endeavours. I am also extremely glad to work with fantastic friends and colleagues like Anna who has been a close and honest friend for a long time - and now is an even more inspiring colleague who always astounds me with her sharp mind and the rationality that she sometimes pushes to the extreme. To Henrik, Fabian, Kerstin, Eva, and Philipp who continue to inspire and motivate me to push ahead with attempting to drive climate policy alongside amazing colleagues like Valentin, Kristina, Josef, Lisa, Hendrik and so many others. To the two Christians who, as great office neighbours, make navigating Austrian bureaucracy life all the more enjoyable and especially to the people that I have learnt many valuable lessons with like Daniel, Björn, and Julia – learning those lessons provided essential support to me finally finishing this huge document. I would also like to thank Dietmar for his support in allowing me to complete this thesis while contributing to Austria's budget and climate policy and for supporting us in advancing Green Budgeting in Austria considerably.

I know that I could not have made it through this journey without the love and support from my closest friends, so I want to thank them from the bottom of my heart. I have always known that I can count on friends who will always have a special place in my life like Jakob and Lisa who I know will always be incredibly important to me; especially with their sunshine Anton that I am so glad to be able to accompany in such a special way. To Marissa who has the unfortunate job of listening to every detail of my complicated personal life, to Benny who has played such a unique role in my life for a long time and whose opinions and thoughts have shaped me and my character greatly together. To Yanick, who I know will always be a kind and generous person that I cherish to have around me and who continues to provide joy to my life every time I see him. I'm incredibly thankful to the friends that have supported me in Vienna throughout this time like a pack of wolves, to Hanna, Laurenz, Nina, Teresa, Antonia, Max, Anna, Alex, and Cordi. Those friendships, which at times span multiple decades by now, mean the world to me.

I am also thankful to all the close and amazing friends that I have made in Oxford, to Linda, who will always call me out of nothing for a simple chat, still owes me an ACL, rubs it in every time the Swiss beat the Austrians in skiing, and who has made two weeks of unexpected COVID-19 quarantine an absolute blast, for Lena who I have never once seen in a bad mood and who I know will always join me for Kaiserschmarrn mit Zwetschkenröster and Schneewitchentorte in a Cotswold Cottage in a heartbeat, to Simi who has supported me now for a long time and who I know will go on to create incredible things everywhere she goes and inspire the people around her (but who will hopefully lose that bet to me in 2040), to Olya who was the best Oxford housemate that I could have hoped for, to Anna who has always been completely true to herself and will always offer her honest thoughts on any issue I bring to her and to Rebecca who is one of the kindest and most genuine people that I think I have ever met, and finally to Giles who has challenged me intellectually like few people ever could and would not hesitate to join if I called him up to climb a mountain for seven days again. You all have been incredible friends and supporters that have made me laugh since I was lucky enough to meet you. And of course a special shout-out to Tim and Sugandha for who I will always come out of my Bollywood-Dance retirement, if called upon and to Leo, who has been my earliest proof-reader and ping-pong adversary.

I would also like to thank many of the people mentioned here already for providing me with useful comments on various aspects of this thesis. I also thank two former supervisors, Sefi Roth and Linus Mattauch. Without Sefi, I might have never picked up statistical methods and Linus has always been a close adviser on all aspects of my academic career for which I am very grateful. I would also like to thank many amazing co-authors that I have been lucky to have worked with

such as Alex Clark, Philippe Benoit, Matt Ives, Nico Koch, Nolan Ritter, Lennard Naumann, Francois Cohen, Yangsiu Lu, Anant Jani, Sihan Li, Arjuna Dibley, Ben Caldecott, Debbie Hopkins, and many more.

I would like to thank Prof Sam Fankhauser and Prof Thomas Sterner for their time and for their valuable comments on this thesis in the context of an incredibly enjoyable DPhil viva.

Lastly, I owe so much to the unconditional love that my family has given me in any situation in life. I dedicate this thesis to them, my mother Johanna, my father Philipp, the best sister I could have wished for, Sarah, with her amazing husband Hannes and my dearest niece Rosa, who has lit up my life in a whole new light.

Moritz P. Schwarz
Brasenose College, Oxford
19 February 2023

# Abstract

Climate change is one of the most crucial societal challenges of the 21st century, affecting a wide range of social, economic, and environmental aspects of modern society. To design and implement policies that can deal with the climate challenge requires an accurate and robust understanding of the physical impacts of climate change as well as understanding the potential impacts of different policy instruments, their limitations, as well as their successes and failures in the past. Despite this necessity, there remains substantial empirical uncertainty around the effectiveness of policy approaches both in the context of mitigation and adaptation. In this doctoral thesis, I present a total of five papers that advance the field of impact estimation in the context of climate change. By using and developing novel econometric methods, I show how existing gaps in this literature can be addressed. In Paper 1, I develop a novel statistical test to illustrate the impact that outlying observations have on regression coefficients of econometric climate impact estimates. In Paper 2, I use these methods and advance climate impact estimates further by presenting a first set of economic climate damage projections that incorporate the effects of extreme weather events and adaptation. While current climate and weather impact data collection approaches focus on manual bottom-up database records, Paper 3 uses machine learning algorithms to predict the occurrence of weather impact events reliably without manual input or on-the-ground knowledge. In Paper 4, I operationalise an alternative way to empirically evaluate policy, which is used in Paper 5 to identify the effects of 10 distinct road transport mitigation policies in the EU15. Overall, I argue that when econometric methods are specified correctly, are applied to the most pressing research questions, and make use of appropriate data then using these methods can allows us to direct adaptation funding more efficiently, track Loss and Damage events around the world, and allow policy-makers to focus on those policy packages that have the largest chance of making a difference.

# Contents

# Part I: Introduction

Climate change is one of the most crucial challenges of the 21st century, affecting a wide range of social, economic, and environmental aspects of modern society. Dealing with climate change requires policy-makers to devise and implement strategies to both reduce greenhouse gas emissions and ensure that society is capable of dealing with the impacts of climate change. While these two objectives, climate mitigation and adaptation, have been central principles in our collective global understanding of climate change for decades, progress to date on both has been slow.

Continuously high global greenhouse gas emissions, despite a global pandemic (Davis et al. 2022), have led to recent analyses concluding that the famous 1.5°C target of Article 2.1a of the Paris Agreement is currently "not plausible" (Engels et al. 2023, p. 5). While the methodology and the conclusions of this particular analysis can be questioned, alongside the political implications of framing a goal as "not plausible", I regard it as a further indication that progress on climate action has been too slow. The challenges of deep decarbonisation, paired with the highly controversial discussions on adaptation and Loss and Damage[1], as has been demonstrated at COP27 in Egypt in November 2022 (CarbonBrief 2022), underscore the necessity for well-informed and effective policy-making.

Devising and implementing effective climate action is complicated by a vast range of technological, physical, social, and political factors. It is these factors that determine whether policy-makers take climate change seriously, whether appropriate policies are agreed upon and implemented and that determine the ultimate effect of these policy. To assess the highest potential of policies for impact therefore requires

[1]The definition of Loss and Damage is controversial and not yet universally agreed (Boyd et al. 2017) but can understood to refer to the impacts of climate change where adaptation is highly difficult or impossible for this thesis.

researchers to consider each of these factors in detail to identify potential windows of opportunity for climate action – especially given the inherent uncertainties that are associated with attempting to devise decadal strategies for society (see IPCC 2013, Chapter 11).

To identify such windows of opportunity for climate policy, however, requires an accurate and robust understanding of the status quo – it requires an understanding of the impact potential of different policy instruments, their limitations, as well as their successes and failures in the past. Despite this necessity, there remains substantial empirical uncertainty around the effectiveness of policy approaches both in the context of mitigation and adaptation. Challenges like multi-causal relationships, long forecasting periods, high internal variability systems, and high levels of heterogeneity in impacts and responses combine to make any kind of impact estimation in the context of climate change inherently difficult.

This empirical uncertainty can be exemplified by considering that extreme climate and weather events are much less likely to be recorded when they are occurring in developing countries (Harrington and Otto 2020). This simple fact can have immense consequences for Loss and Damage discussions, can lead to a misallocation of adaptation finance, will influence expectations of future damages, and plays a key role in shaping the insurance market in these countries. Similarly, a limited understanding of the effectiveness of climate mitigation policies, e.g. by ignoring relevant rebound effects, not considering the most influential policies, overrelying on theoretical ex-ante policy estimates, or failing to evaluate policies ex-post can seriously hamper climate action in practice.

It is this understanding of the impacts of the physical climate and of climate policy in the past that I improve in this thesis. I argue, that it is essential to understand the extent to which we actually know what has happened in the past in order to use this knowledge as a guide to the future. To improve this understanding, I present a range of novel methods and approaches that deal with many of the gaps that the existing literature still contains. In this existing literature,

I identify a number of crucial gaps that the papers of this thesis, and the methods contained within them, address.

To do so, I place the focus of this thesis on *empirical* methods, which analyse real-world observations and data. This is in contrast to many of the theoretical approaches that heavily rely on mathematical descriptions of relationships. Nonetheless of course, theoretical knowledge often inspires a particular empirical model formulation and is therefore crucial in empirical analysis as well. More concretely, I focus on a literature embedded in the field of econometric methods, which are methods that use statistical analysis to test and validate economic theories (Stock and Watson 2015).

In the following pages of this doctoral thesis, I present novel methods and approaches that demonstrate the potential of advanced econometric techniques to improve the estimation of climate impacts and policies considerably. I apply these methods to more accurately estimate and project the macro-economic consequences of physical climate impacts, to detect previously unrecorded extreme weather events, and to identify a concrete set of climate policies that have reduced carbon emissions in the road transport sector of the European Union.

The work presented here complements the research that I have conducted over the past few years, such as a study published in the *Philosophical Transactions of the Royal Society* in 2018 on the effects of 1.5°C and 2°C of warming on economic growth, which has sparked much of the work leading up to Paper 1 and 2 of this thesis (Pretis et al. 2018). Similarly, a study recently published in *Global Environmental Change* evaluates and quantifies the influence that weather and conflict has on displacement in Somalia using similar methods to Papers 1 and 2 (Thalheimer, Schwarz, and Pretis 2023). Further work has focused on the effects of climate policy on labour dynamics in order to gain a more thorough understanding of how policies would need to be designed to ensure a Just Transition. This line of inquiry has led to two pieces of work, with one paper that evaluates the employment impacts of oil price shocks on the Albertan labour market being published in

*Climate Policy* (Scheer et al. 2022) and a Working Paper estimating the impacts of coal mine closures on local US labour markets (Mark, Rafaty, and Schwarz 2022). Alongside more qualitative work on designing ideal climate policy for state-owned power companies, which was published in the *Journal of Cleaner Production* in 2022 and a paper on the difficulty of relating COVID-19 cases to temperatures and precipitation published in *Environmental and Resource Economics: Perspectives* in 2020 (Cohen et al. 2020), I have contributed to several reports and policy papers and have released and contributed to open-source estimation software such as packages for the programming language R.

This thesis overall concludes, that to combat climate change and to deal with the climate crisis in a sustainable, equitable, and efficient manner, we must question the current focus of research critically. Much of the efforts in the macro-economic climate impact community have focused on trying to identify the 'efficient' or 'optimal' level of warming given economic preferences; a consideration that in my mind has only little additional value in the context of clearly defined temperature goals within the Paris Agreement. However, comparatively little attention of economists has been diverted towards identifying where climate and weather impacts have actually occurred, what their effect has been, especially on developing countries, and how such impacts could challenge societies in the next few decades. Similarly, efforts to study deep decarbonisation policies have largely focused on theoretical ex-ante assessments, which in practice have often shown little accuracy.

The individual pieces of work presented in this thesis illustrate that with the support of novel and innovative econometric methods, the estimates of climate impacts and policies can be improved considerably. Advanced and novel econometric techniques can deal with issues that have plagued the climate debate for decades; these techniques can quantify and alleviate misspecification concerns, can allow a more specific and insightful debate surrounding adaptation options and likely future damage, while providing a more holistic and objective data environment for climate policy and weather impact events. Implementing and utilising these

techniques more broadly can overall contribute to distributing adaptation efforts and finance more equitably and efficiently, can allow policy-makers to focus their political capital on more effective policy packages, and will hopefully contribute to a more resilient and low-carbon society.

There remains of course a significant potential to improve such methods further, yet this thesis demonstrates that existing estimation approaches can be improved by using novel misspecification tests, the use of more specialised data, as well as more advanced machine learning type model selection algorithms, and also that asking new questions in the field of policy evaluation and weather impact detection can offer promising avenues for further research. The methods used and developed in the five papers presented in this doctoral thesis illustrate distinct opportunities to advance our understanding in this field – and it is this knowledge that will be crucial in our collective endeavour to create a low-carbon equitable economic system for the future.

# Thesis Outline

In Part I, I outline the overall motivation of my doctoral thesis. I discuss why improving the estimation of impacts is so crucial to understanding and combatting climate change and for designing effective policy. After this outline, a set of identified gaps in the academic literature is put forward alongside the relevant evidence in the literature. For each gap, a short description is presented of how these gaps are addressed in the individual papers in Part II.

In Part II, I present a total of five individual pieces of research that tackle one or more aspects of these challenges; with four of them produced as co-authored pieces of work and one single-author paper. Out of the five papers presented, three have been submitted to peer-reviewed journals, complying with the relevant thesis requirements. One of these papers, Paper 5, has been published in Nature Energy, while Paper 1 is in the second Revise & Resubmit (R&R) stage. Paper 2 has been submitted to a peer-reviewed journal and alongside with Paper 4 has been published as a Working Paper. Part II is structured as follows.

In Paper 1, I present a co-authored paper that evaluates the influence that misspecification, in particular the presence of outliers, has on regression estimates. The paper presents Asymptotic Theory that defines and establishes a new Outlier Distortion Test (the Jiao-Pretis-Schwarz Test). This test is then evaluated using Asymptotic and Bootstrap Testing to evaluate its performance under different conditions. Finally, the test is applied to a climate impact estimation and the effect of using an outlier-robust estimation method is presented. This paper has been submitted to the Journal of Econometrics and is currently in its second R&R stage.

In Paper 2, I present a co-authored paper that builds on the preceding paper by utilizing an outlier-robust estimation algorithm as well as advanced econometric model selection methods to explore the robustness of existing econometric climate impact estimates. The resulting estimates of GDP per capita impacts are then projected to 2100 using the full CMIP5 climate model variability range, which shows that current existing damage estimates significantly underestimate the likely socio-economic climate impacts in the future. This paper has been submitted to Nature Climate Change.

While Papers 1 and 2 represent an attempt to improve existing climate impact estimation approaches, the following Papers aim to advance more emerging fields of inquiry.

In Paper 3, a novel approach to identifying and recording weather impacts and disasters is presented. While current climate and weather impact data collection approaches focus on manual bottom-up database records, Paper 3 uses machine learning algorithms to predict the occurrence of weather impact events reliably without manual input or on-the-ground knowledge. The results of this proof-of-concept Paper exhibit very high levels of predictive accuracy. This suggest that there is immense potential for such methods to provide more accurate disaster estimates in the future – a crucial prerequisite for adaptation policy-makers and international Loss and Damage considerations.

Papers 4 and 5 consider a slightly different type of climate impact as they are motivated not by quantifying the effect of physical climate impacts, but rather

they attempt to enable researchers to quantify the effect of government policies on socio-economic processes.

To this end, in Paper 4, I present an alternative way to phrase policy evaluation questions, which is operationalised in the paper. Rather than asking forward causal questions, i.e. 'Does X cause Y?', Paper 4 explores and operationalises reverse causal questions that concern themselves with asking the question 'What causes Y?'. In this paper, therefore, a novel interpretation of existing machine learning algorithms is presented in a way that operationalises this conceptual approach and applies it to a case study. While the case study in Paper 4 is not climate related, in Paper 5 I apply the algorithm developed in Paper 4 to climate policies in the transport sector in 15 EU countries and agnostically detect the impact of 10 climate policies that have reduced $CO_2$ road transport emissions by up to 26%. This paper has been published in Nature Energy in August 2022.

Finally in Part IV, a brief Conclusion is presented where I summarise the main insights that the collective work of this thesis has yielded. Overall, the individual pieces of work combined in this thesis illustrate that there remains significant potential to improve various aspects of climate impact estimation. Advanced and novel econometric techniques can deal with omnipresent issues that have plagued the academic climate debate for decades; these techniques can quantify and alleviate misspecification concerns, can allow for more specific and insightful debates surrounding adaptation policy and can give policy-makers an idea about likely future damage scenarios. Simultaneously, the efforts of this thesis could provide the basis for a more holistic and objective data environment for climate policy and weather impact event data collection. This thesis has demonstrated that existing estimation approaches can be improved by using novel misspecification tests and the use of more specialised data, as well as more advanced machine learning type model selection algorithms. It further shows that that asking new questions in the field of policy evaluation and weather impact detection can offer promising avenues for further research.

*Thesis Outline*

While Parts I, II, and III capture the essence of each research effort, the Appendix of this thesis in Part IV contains a Contribution Statement and also contains the relevant appendices for the individual papers. Finally, a reference list for the citations referred to in Parts I and III is presented.

# Identifying Gaps in the Literature

Having analysed the vast body of literature on impact estimation over the past years in the context of this doctoral thesis – both within the climate context and beyond – I present three key areas below where I argue that the impact literature has significant gaps – many of which I am able to address in the papers in Part II.

Firstly, I identify three distinct gaps in this literature relevant to climate impact estimation. Firstly, I argue that the research questions that are being set are inadequate both in terms of research objective and effectiveness of the insight gained given the challenges that climate change is presenting society with. Secondly, I shed light on significant aspects of impact estimation models that could be improved in order to maximise the insight relevant for policy. Thirdly, I argue that robust impact estimation in the field of climate change is faced with significant hurdles to acquire and use the data that would be required. In each section below, I outline how I have addressed these gaps in the papers in Part II of this thesis.

## Asking inadequate questions

The first major gap in the literature refers to the research questions considered. I argue, that the majority of the impact estimation literature in this field focuses on aspects of the climate crisis that have either already been answered, are comparatively well understood, or, most significantly, do not provide us with the most useful information on how to overcome the challenges of climate change.

In the context of physical climate impacts and the estimation of their macro-economic consequences, much of the academic debate focuses on the question which level of warming is optimal to global society, given a set of economic preferences. Here, I argue, however, that this question has been answered by making concious political decisions in the Paris Agreement and that therefore much more research should focus on implementing effective climate adaptation strategies.

On these physical climate impacts, the Intergovernmental Panel on Climate Change (IPCC) in its Sixth Assessment Report (AR6) has once more established that the impacts of a changed climate are a key challenge for society as "Climate change has caused substantial damages, and increasingly irreversible losses, in terrestrial, freshwater and coastal and open ocean marine ecosystems" (IPCC 2022b, B.1.2 on p. 9). The IPCC further shows that climate change increases the frequency and intensity of extreme events alongside significant slow onset events. The literature on physical climate impacts is vast, with countless papers analysing dynamics such as heatwaves, meteorological droughts, and the effects of various climate dynamics on biological systems and biodiversity that document such impacts in all regions and ecosystems, see Figure 1.

However, translating these physical measures to socio-economic impacts to understand the societal consequences of a warmer world – especially in a coherent and holistic manner – has troubled social scientists for decades. This manifests itself in the difficulty of providing well accepted academic and political definitions of emerging concepts such as Loss and Damage, which despite its rising political significance in the context of COP27 in 2022 has no fully agreed upon definition (Boyd et al. 2017).

Socio-economic impact modelling of the consequences of a changing climate - and the mitigation and adaptation policies that go along with it - have especially until the 2010s largely followed a mathematical process-based modelling approach. This means that relationships between physical and social systems are not estimated using real-world data but are mostly grounded in theory derived using economics and physics. In this context, the most famous concept is the so-called damage function. These damage functions specify a continuous and simple relationship between global mean temperature change and changes in gross economic output. The exact formulation or specification of a damage function, however, remains extremely controversial.

These damage functions have been subject to numerous meta-analyses (Newbold and Marten 2014; Nordhaus 2018; Tol 2009), yet most estimates of global damage functions continue to be fairly hard to distinguish from 'back of the envelope'

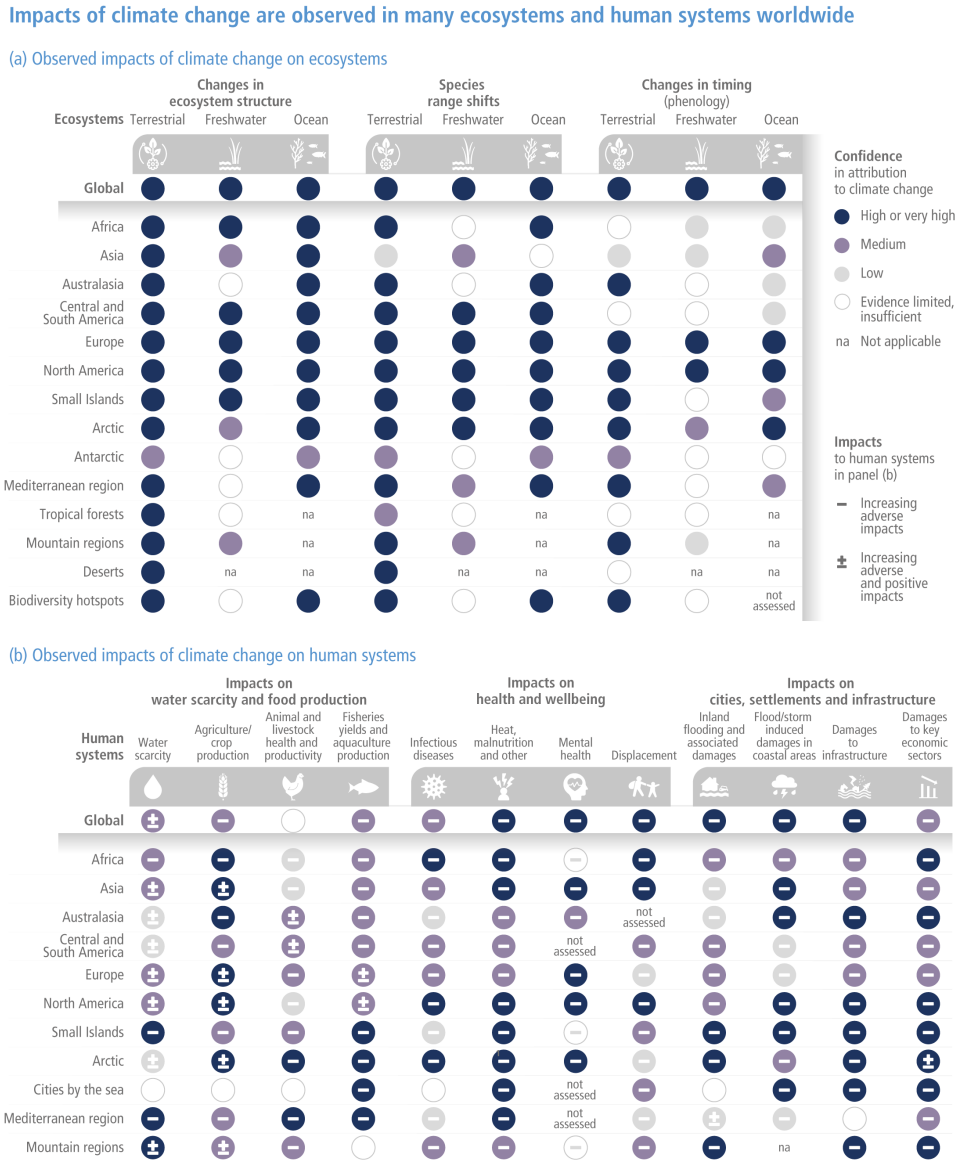


**Figure 1:** Figure SPM.2 reproduced from the IPCC Summary for Policymakers WGII.

calculations, as Howard and Sterner (2017) frame it. While back-of-the envelope calculations certainly have their use, it is the significance of damage functions in Cost-Benefit Integrated Assessment Models (IAMs) that highlight the necessity to estimate such functions accurately.[2]

Cost-Benefit IAMs, such as DICE, FUND, and PAGE (Pindyck 2013) essentially attempt to shed light on 'optimal' levels of mitigation, i.e. they attempt to derive the ideal level of warming. The (IPCC 2014, see Section 6.2.1, p. 422), however, notes that these IAMs essentially "use economics as the basis for decision making [and] typically assume fully functioning markets and competitive market behaviour, meaning that factors such as non-market transactions, information asymmetries, and market power influencing decisions are not effectively represented". This poses a number of challenges when aiming to inform the public policy debate.

The work by William Nordhaus (Nordhaus 1992; Nordhaus 2018) is most famous in this field, with him receiving the Nobel Memorial Prize in Economic Sciences in 2018 for his work on the DICE model. In the DICE model, the specification of the damage function is key to estimating the Social Cost of Carbon that tries to illustrate the societal cost of emitting one tonne of $CO_2$, which then informs an 'optimal' level of warming for global society. The DICE model in its current form continues to rely on a controversial literature survey by Tol (2009) that is based largely on process-based estimates of climate impacts and then adds a relatively arbitrary adjustment of 25% for non-economic impacts (Nordhaus and Sztorc 2013).

Given the significance of this method and DICE's popularity, with thousands of papers adapting and using the model, a fair number of studies have sought to more formally use empirical methods to derive such damage functions and impact estimates (Hsiang 2016). These approaches represent a distinct departure from early estimates of macro-economic climate impacts that was mainly dominated by the use of process-based climate impact representations.

[2]They are also sometimes relevant for process-based Cost-Effectiveness IAMs, such as IMAGE, MESSAGE, REMIND, and others but in a much more nuanced manner (Wilson et al. 2017).

While the physical climate modelling community has been embracing such empirical approaches for some time now, i.e. using statistical tools to describe certain dynamics with historical data to overcome the existing gaps in the physical and mathematical understanding of natural systems (Majda and Gershgorin 2010), empirical data had not been used to the same effect in modelling economic impacts up until about the late 2000s (Hsiang 2016). Such scientific contributions have been published in Economics journals (e.g. Dell, Jones, and Olken 2012, in American Economic Journal: Macroeconomics; as well as Kalkuhl and Wenz 2020; and Newell, Prest, and Sexton 2021, in the Journal of Environmental Economics and Management), and general-interest journals (e.g. Burke, Hsiang, and Miguel 2015; Burke, Davis, and Diffenbaugh 2018; Kotz, Levermann, and Wenz 2022, in Nature; or Pretis et al. 2018, in Phil. Trans. R. Soc.) alike, emphasizing the interdisciplinary nature of this research agenda.

While our understanding in this area has considerably improved based on these studies, the resulting climate damage estimates continue to exhibit known shortcomings. The range of potential damages remains extremely large as Figure 2 shows. Furthermore, more recognition has emerged that these continuous, smooth, and differentiable global damage functions are incompatible with our increasing understanding of tipping points and regional climate impacts (Sterner 2015). Substantial efforts have been undertaken to incorporate these dynamics, such as Dietz et al. (2021) who consider the economic implications of such tipping points and find, unsurprisingly, that they substantially increase economic costs. Work by Hänsel et al. (2020) improves the estimation by including a more accurate carbon cycle model and update damage estimates into DICE and find that such a specification can suggest that the Paris Agreement targets are 'optimal'. Similar conclusions are reached when questioning the mitigation assumptions of the DICE model (Grubb, Wieners, and Yang 2021).

Yet, despite these advances and efforts, considerable uncertainties about future climate impacts on economic quantities, such as productivity, GDP per capita growth, social equity, or output in key industries remain. These, and other aspects,

have motivated numerous academics to challenge the usefulness of these concepts in their entirety (Pindyck 2013; Farmer et al. 2015). While the fundamental question of an 'optimal' level of warming has been useful in the 1990s and 2000s, the Paris Agreement, and the Copenhagen Accord before it, have clearly changed the relevance of such assessments. While these assessments of course still have significant use in the public discourse, their estimates are always highly dependent on social and political preferences – and I would argue that the political consensus that has led to the Paris Agreement has determined these social and political preferences. In other words, the underlying assumed preferences in every IAM to illustrate the trade-offs that are fundamental to every political and economic decisions have in a way been set clearly by the Paris Agreement.

While it is self-evident that quantifying climate impacts in terms of economic outcomes is crucial to inform mitigation and adaptation policy decisions in a context of finite (fiscal and economic) resources (Roughgarden and Schneider 1999; Stern 2008), finding the 'optimal' level of global warming is now not really a political question anymore. Therefore, I would argue that academic efforts in this field should now shift towards accelerating mitigation policy and maximising the efficiency of adaptation policies. And arguably, the damage function literature does little to contribute to these two goals.

Asking the right question is, however, not only a key factor in deriving reliable estimates of future physical climate impacts – it is also an essential factor in estimating the effects of climate policies. While it is clear that robust policy-making must be based on an understanding of the likely impact of a policy, most of the causal inference literature has concerned itself with questions that consider the 'effects of causes' rather than the 'causes of effects'; a circumstance that I argue has limited the potential of creating the most effective policies for tackling the climate crisis.

These questions that concern themselves with the 'effects of causes' have been termed 'forward causal questions' by Gelman and Imbens (2013) and are widely considered to be standard approaches in the wider literature (Gelman 2011; Mill

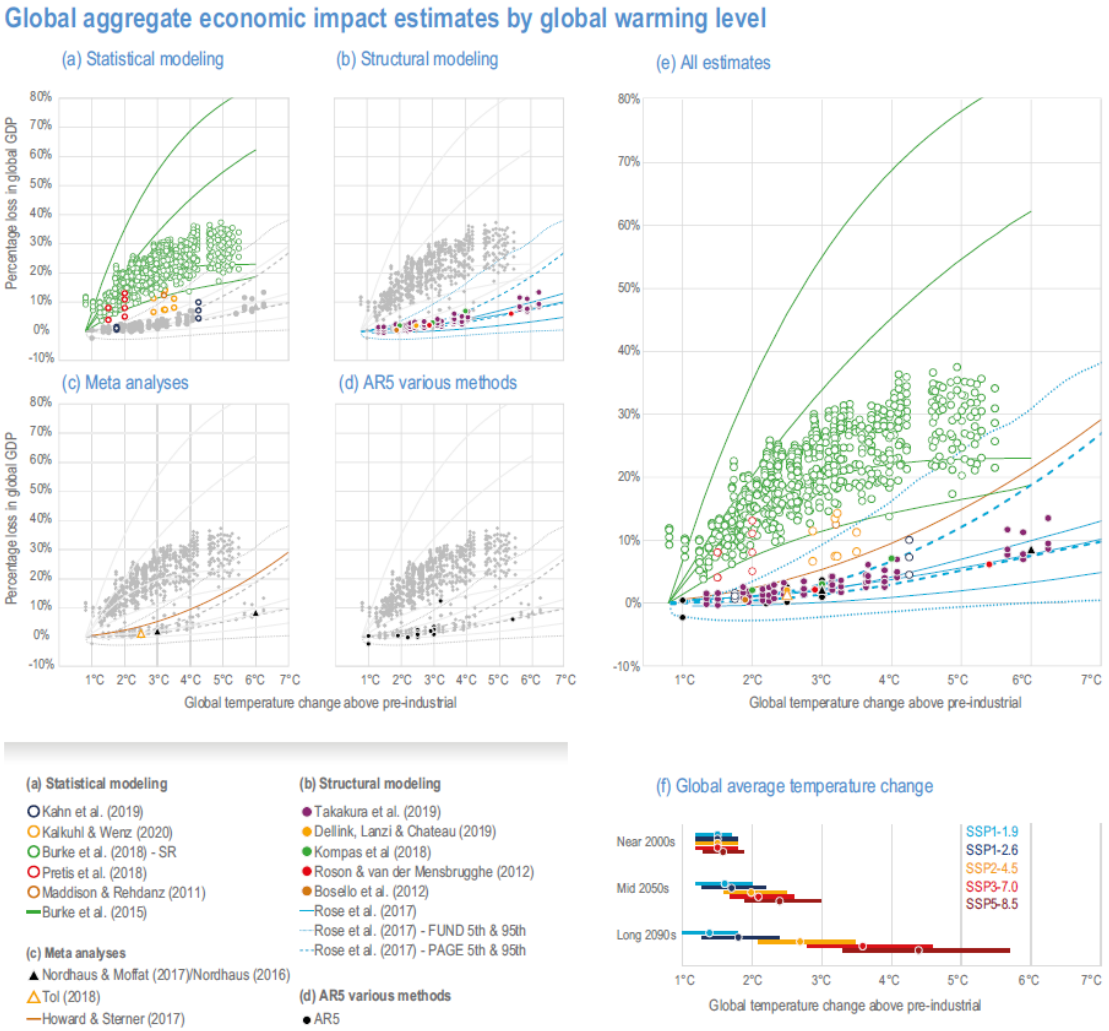


**Figure 2:** Figure Cross-Working Group Box ECONOMIC.1 reproduced from the IPCC WGII Report.

1843). These questions are used to answer inquires like 'did this particular car engine regulation policy reduce $CO_2$ emissions in road cars'. In other words, they are focused on a single policy and attempt to find the specific effect of that policy. Such an approach is for example the standard set-up for methods like Differences-in-Difference studies.

However, the decision-making process as to which policy should be analysed is itself already a major research design question – potential policies that are much more effective might not be identified for such an analysis a-priori and given the fact that most policy making in recent years includes policy packages with more than one policy being implemented, just focusing on the effectiveness of a single

measure might understate the effectiveness of political decisions.

'Reverse causal questions'[3] on the other hand attempt to estimate the 'causes of effects' – so rather consider the road $CO_2$ emissions and look at all potential policies, trends, and shocks that might have had an effect on them. According to Gelman and Imbens (2013, p. 3) "Reverse causal reasoning is different; it involves asking questions and searching for new variables that might not yet even be in our model". The motivation of this way of inquiry closely resembles advances from the break detection literature that is mostly used in the time series literature where e.g. Perron (1989) used break detection to show the effects of the Great Depression and oil price shocks in a GNP time series or Hendry (2020) who uses a similar technique to identify the effect of the UK climate policy in per capita $CO_2$ emissions. Such an approach could shed an entirely new light on the comparative effectiveness of policies by allowing analyses to really hone in on those policies that have shown to be effective in real life.

This is in stark demand, as most impact estimates of policies are done ex-ante using economic models like Dynamic Stochastic General Equilibrium (DSGE) or Computable General Equilibrium models and only few policies are actually evaluated using ex-post data (Farmer et al. 2015). The study by Rafaty, Dolphin, and Pretis (2022), while also having to make an a-priori selection of policies, uses a wide set of implemented carbon pricing policies to evaluate their effects on carbon emissions. While the study finds relatively low effects overall, their conclusions do seem to suggest that the prices that have been in place over the past decades have simply been too low to have a significant effect on carbon emissions. Similarly, a study by Eskander and Fankhauser (2020) considers over 1,800 climate laws worldwide and presents the effects of an additional climate law on $CO_2$ emissions.

[3]Note that the term 'reverse causal questions' is distinct and not directly related to the concept of 'reverse causality', see Gelman and Imbens (2013).

**Adressing this gap**

While there are naturally always insights that can be gained from the questions I discussed above, adapting the focus of inquiry more towards the dynamics that can really help us deal with the climate crisis will be essential.

This means, as I have argued, abandoning huge academic efforts to find an 'optimal' level of warming. Despite the obvious and significant academic benefits and successes that can be gained from adapting models like the DICE model, I am convinced that publishing yet another study on how important climate mitigation actually is will not change the societal consensus to considerably change our climate targets. Given that the physical realities and the urgency of climate change has been established beyond reasonable doubt, I would argue that remaining climate denialism should be investigated using political and social sciences rather than attempting to convince these people using an economic argument derived using an IAM.

While I do present an econometric study on climate impacts in this thesis in Paper 1 and 2, I based the formulation of this damage function on empirical methods rather than relying on process-based damage estimates. Furthermore, I purposefully did not incorporate those damage estimates into an Integrated Assessment Model as the question that such a model answers is in my mind not relevant anymore. What is relevant, however, is to identify regions most at risk of suffering from climate impacts, which is what Papers 1, 2, and 3 do. It is for those regions that we must develop robust adaptation and development policies capable of dealing with a warmer world.

In Paper 4, I have operationalised the 'reverse causal modelling' approach that now enables us to implement the approach introduced by Gelman and Imbens (2013). This approach is then used in Paper 5 to demonstrate how much we can learn about emission mitigation policies when we do not a-priori decide which policies to analyse but rather allow the data to reveal the policies that have actually been successful in reducing emissions.

## Running inadequate models

The second major gap in the existing academic literature that I have identified in my DPhil research, relates to the way that empirical estimation is carried out – in other words how statistical estimation of impacts is being done and how it can be significantly improved by employing robust and novel estimation methods. Below, I argue that although significant progress has been made in particular aspects of climate impact estimation, approaches that consider these inadequacies holistically are still required.

Incidentally, much of the progress in the empirical estimation of macro-economic climate impacts was driven by changes to the functional form and specification of the models used (Hsiang 2016). While Dell, Jones, and Olken (2012) specified their model of GDP per capita growth in a quasi-linear set-up with linear regressors of temperature and precipitation, but with a dummy for poorer countries, Burke, Hsiang, and Miguel (2015) presented a highly successful study using linear regressors in a polynomial form. The study by Pretis et al. (2018) then added further regressors that not only consider the mean of a variable but also the variance of temperature and precipitation across a year. However, the impact and effect of extreme climate variables has not yet been assessed in such a set-up and therefore remains a major gap in the macro-economic literature of climate impacts.

While there have been many further studies experimenting with the exact specification of such models (Kalkuhl and Wenz 2020; Kotz, Levermann, and Wenz 2022), it seems clear that the appropriate specification of economic climate impact models is non-trivial and can have significant consequences for its conclusions. This was highlighted in great detail by Newell, Prest, and Sexton (2021) who assess the plausibility of 800 different specifications, both implying growth and level effects on GDP per capita. Their results show clearly that there is vast uncertainty with models that consider climate growth effects on GDP per capita but that the much more stable results of GDP level effect models suggest that more conventional

estimates of the magnitude of climate impacts that have traditionally been used in Integrated Assessment Models could be more realistic.

These advances and other meta-studies (Hsiang et al. 2017) highlight that numerous aspects of econometric model misspecification have not been considered systematically and much more work on this topic is needed. After all it seems that the existing estimates of climate impacts from an economic perspective continue to be incompatible with the dire warnings of the physical effects of climate change, best summarised in the IPCC (2022a) reports. This is highly implausible especially when considering the stability of future climate systems with high warming scenarios and the risks of crossing significant tipping points (Lenton et al. 2008). Simultaneously, existing macro-economic impact studies continue to assume that global economic output will be several times higher than today, albeit slightly smaller than in a no-further-climate-change-baseline (Burke, Hsiang, and Miguel 2015; Pretis et al. 2018; Burke, Davis, and Diffenbaugh 2018). This is in part driven by the assumptions contained in the Shared Socio-economic Pathway scenarios (SSP, Riahi et al. 2017), but nevertheless considering the physical realities of a world 4°C to 6°C warmer than pre-industrial levels and then claiming that macro-economic output in terms of wealth per capita might be similar to today (see Figure 2) seems hardly plausible.

A further example of such a misspecification could arise from the coefficient distortion of unmodelled outlying observations. This is a common concern in applied modelling, as a small set of outliers can affect and bias regression coefficients of estimates – in this instance, outliers could therefore considerably change the expectation of climate impacts. Generally, such concerns can be alleviated by using an outlier-robust estimation method, such as using a so called Huber-skip estimator like Trimmed Least Squares (Ruppert and Carroll 1980) and Impulse Indicator Saturation (Hendry, Johansen, and Santos 2008). But of course there is a robustness-efficiency trade-off when using such a method – if outliers affect the regression estimates, a robust estimator should be used. However, when outliers do not affect regression estimates, an OLS estimation would be more efficient. While

the resulting estimates can be compared to one another, there is no formal test that has been established to check the distortion of regressors in this case.

In the existing climate impact literature there has been a lack of comprehensive attention to the distortion of climate impacts by unmodelled outlying observations. Apart from Pretis et al. (2018), I am not aware of any other study explicitly using a robust estimator, while a formal discussion surrounding the distortion of regression coefficients is missing At the same time, in studies attempting to identify extreme weather events, there is a risk that (unmodelled) outlying events can be attributed to climate and extreme weather and end up being identified as an extreme weather event.

Lastly, I argue that the extent and speed to which society will be able to adapt to the effects of climate change is a major remaining source of uncertainty that is currently not incorporated into most macro-economic climate impact studies. While a number of studies attempt to test for the stability of their estimated climate damage functions (Burke, Hsiang, and Miguel 2015), formal discussions of adaptation in this context have often focused on the agricultural sector (Schlenker and Roberts 2009; Burke and Emerick 2016) or have not directly considered GDP per capita impacts (Carleton et al. 2020; Barreca et al. 2016). Other studies, such as Dell, Jones, and Olken (2012) have considered the role of wealth in their estimation by incorporating a rich/poor dummy in their estimation but then do not use this insight further in projecting future climate impacts. Outside of the macro-economic literature, adaptation estimates are of course being considered in many different ways (Hinkel et al. 2014), highlighting the fact that macro-economic studies must bridge this gap in future research.

The fact that adaptation has not been tackled systematically and is hardly ever considered in future projections of macro-economic impacts raises two key challenges. Firstly, this information is crucial to understand where regions should invest in adaptation solutions and could inform policy-makers where potential limits to adaptation might lie (Dow et al. 2013; Tol, Fankhauser, and Smith

1998). Secondly, the lack of holistic and robust estimates of future adaptation rates has allowed critics of climate impact estimation and of the overall significance of climate impacts in a macro-economic sense, such as Bjorn Lomborg, to claim that existing damage estimates are widely exaggerated and that action against climate change is not really necessary.

**Adressing this gap**

It is for these reasons, I argue, that critically assessing the specification of the impact models being run is so crucial. When running models of physical climate impacts, researchers should consider a wider potential set of covariates that can incorporate the variability of climate and weather, while more closely representing the full distribution of these indicators and therefore also representing extreme climate events. Similarly, more formal consideration of the impact of the functional form of the estimation and of the impact of outlying observations is necessary to ensure that the estimates being used are robust and useful in a policy context. Lastly, the role of adaptation and development will be a key factor in future societal welfare, so should be assessed from a much more rigorous and sophisticated macro-economic perspective.

In this thesis, I address a number of these gaps. In Paper 1, I present a novel statistical test that can transparently document whether an estimate has been distorted by outlying observations – and as the paper shows, traditional econometric climate impact estimates are significantly affected by outliers. In Paper 2, I incorporate the insight from Paper 1 by using an outlier-robust estimator and consider a much wider set of regressors to estimate climate impacts by considering the effects that extreme climate indicators have. In Papers 1 and 2, I also consider the role of adaptation and of income levels in the estimation of climate impacts while Paper 2 then includes a set of projections of climate impacts that take adaptation into account that illustrate the major role that adaptation is likely to have in the future.

## Using inadequate data

The third and last major gap that I have identified in the climate impact estimation literature is the inadequate data that is being used to answer relevant policy questions.

The vast majority of published studies in the field of macro-economic climate impact estimation currently relies on a small number of highly aggregated measures of climate. Often the only variable considered is a country-aggregated average annual temperature series, occasionally accompanied with an annual average of monthly total precipitation (see e.g. Dell, Jones, and Olken 2012; Burke, Hsiang, and Miguel 2015; Burke, Davis, and Diffenbaugh 2018; Newell, Prest, and Sexton 2021). This approach is likely to mask a number of crucial dynamics relevant to climate impacts, such as extreme climate events like torrential precipitation and floods, heatwaves, tropical cyclones, and storms. All of these events might not be represented in annually-constructed aggregate measures due to their localised nature and short lifespan – but of course the impacts from such events are often immense (Otto et al. 2016).

To address some of these gaps, Pretis et al. (2018) considered the monthly variability alongside annual averages to account for the changes that climate change is exerting on weather patterns and Kalkuhl and Wenz (2020) consider sub-national economic output for their study on climate impacts. Most of these studies also rely on observational climate data, such as the data from the GHCN-D network (Menne et al. 2012) or processed interpolated products, such as the one from Matsuura and Willmott (2018). This can be sufficient to analyse annual average temperatures over the past few decades, but is completely inadequate for analysing extreme climate events especially in developing countries, as this type of data is simply missing (see e.g. Donat et al. 2013).

Identifying the true occurrence of extreme impact events is incredibly difficult. Existing databases of such extreme weather events are of a fairly poor quality

when it comes to reporting and analysing extreme events, especially in developing countries. Based on one of these databases, EM-DAT, heatwaves have have resulted in more than 140,000 deaths in the past 40 years in Europe – but have only killed 71 people in Sub-Saharan Africa from 1900 to 2019 (Harrington and Otto 2020). This lack of structured and complete data has substantial consequences for disaster response efforts as well as long-term adaptation financing. This is despite extreme weather events and disasters being the most visible and present type of climate impact that social scientists for decades.

The short-term impact of individual disasters has been studied extensively – including using empirical econometric methods – such as in the case of Hurricane studies by Martinez (2020b) and Martinez (2020a) who uses various indicator saturation and super saturation methods to investigate the beneficial effects of improved hurricane forecasts and normalized damage estimates as well as Strobl (2011) and Strobl (2012) who uses econometric methods to study hurricanes in the Caribbean and in the coastal areas of the United States. Fomby, Ikeda, and Loayza (2013) consider the effects of floods, droughts, earthquakes, and storms using an empirical vector autoregression models that contain both endogenous variables and exogenous shocks (VARX model) while Felbermayr and Gröschl (2014, p. 93) consider the effect of geophysical and meteorological events and find that a "disaster in the top 1-percentile of the disaster index distribution reduces GDP per capita by at least 6.83%, while the top 5-percentile disasters cause per capita income to drop at least by 0.46%", although studies such as Loayza et al. (2012) point out using panel econometric methods of moments estimators that these effects can be heterogeneous across sectors and are not always negative.

But also the long-term effect of individual disasters have been studied extensively, with Hsiang and Jina (2014) econometrically analysing 6,700 Cyclones and finding that for both rich and poor countries impacts from cyclones are long-lasting, while the authors reject the hypothesis that there are easily avoidable losses through migration or wealth transfers or that indeed these disasters stimulate short-term

growth. Nonetheless, they concur with Berlemann and Wenzel (2018) that these effects are most pronounced in developing countries.

When considering the available bottom-up databases of natural disasters that most of these studies are based on, however, countless studies highlight issues and challenges with them (Panwar and Sen 2020). Edmonds and Noy (2018) for example finds that the dataset available to them leads the authors to underestimate disaster risk and contains significant inconsistencies. Gall, Borden and Cutter (2009, p. 807) identify a total of six general biases with these databases, which are "1) hazard bias, which produces an uneven representation and distribution of losses between hazard types; 2) temporal bias, which makes it difficult to compare losses across time due to less reliable loss data in past decades; 3) threshold bias, which results in an underrepresentation of minor and chronic events; 4) accounting bias, which underreports indirect, uninsured, and others losses; 5) geographic bias, which generates a spatially distorted picture of losses by over- or underrepresented certain locales; and 6) systemic bias, which makes it difficult to compare losses between databases due to different estimation and reporting techniques." Felbermayr and Gröschl (2014) show that that the likelihood of an event appearing in a disaster database is affected by a country's GDP per capita; a fact that is deeply concerning given the significant vulnerabilities in developing countries which increase the demand for such accurate data in those regions even more.

**Adressing this gap**

The lack of reliable and available data is always going to make robust impact estimation more challenging. Given this data shortage, especially in terms of high frequency spatially explicit economic data as well as accurate extreme weather event data in developing countries, current estimates of climate impacts are very likely to be underestimated and highly imprecise.

To address these gaps, given the data that is available, I add 27 regressors of extreme climate to the standard set of temperature and precipitation variables to represent the full weather distribution more accurately in Paper 2. To ensure that

my estimation sample is not biased by a lack of observational extreme weather data, I also make use of reanalysis data, i.e. output data from a climate model rather than observational data directly, which is currently not standard in many other climate impact studies. While this data is less precise in areas where reliable observations are available, reanalysis data is crucial in areas where these observations are missing, such as developing countries, and also provide a consistent set of indicators that can be used in impact estimation.

Lastly in Paper 3, I present a novel method to validate and identify extreme weather events independent of local administrative resources using machine-learning methods. This paper is supposed to act as a first proof-of-concept paper. Given the success of the method presented in Paper 3, however, I argue that such an approach could be expanded further and could eventually make use of the vast data that is being produced by satellites, reanalysis models, search-engines, and digital payment providers and could further be paired with extensive textual analysis to considerably improve datasets on extreme weather events in developing countries.

# Part II: Individual Thesis Papers

# Paper 1: Testing for Coefficient Distortion due to Outliers with an Application to the Economic Impacts of Climate Change

# Testing for Coefficient Distortion due to Outliers with an Application to the Economic Impacts of Climate Change

Xiyu Jiao[1,4], Felix Pretis[2,4]*, and Moritz Schwarz[3,4]

[1]Department of Economics, University of Oxford
[2]Department of Economics, University of Victoria
[3]School of Geography and the Environment, University of Oxford
[4]Nuffield College, University of Oxford

**Abstract**

Outlying observations can bias regression estimates, requiring the use of robust estimators. Comparing robust estimates to those obtained using OLS is a common robustness check, however, such comparisons have been mostly informal due to the lack of available tests. Here we introduce a formal test for coefficient distortion due to outliers in regression models. Our proposed test is based on the difference between OLS and robust estimates obtained using a class of Huber-skip M-type estimators (such as Impulse Indicator Saturation or Robustified Least Squares). Establishing asymptotics of the corresponding Huber-skip M-estimators using an empirical process CLT recently developed in the literature, we show that our distortion test has an asymptotic chi-squared distribution. The test is valid for cross-sectional, as well as panel, and stationary or deterministically-trending time series models. To improve finite sample performance and to alleviate concerns on distributional assumptions, we explore several bootstrap testing schemes. We apply our outlier distortion test to estimates of the macro-economic impacts of climate change allowing for adaptation.

**JEL Classification:** C12, C52, Q54.
**Keywords:** outlier robustness, robust estimation, iterated 1-step Huber-skip M-estimator, indicator saturation, climate econometrics, climate change, adaptation.

*Corresponding author: fpretis@uvic.ca

## 1 Introduction

A common concern in empirical modelling centres around whether estimated regression coefficients are affected by a small set of outlying observations. Comparing OLS estimates to those obtained using an outlier-robust estimator is a common robustness check in the applied literature. However, such comparisons are mostly done heuristically, due to a lack of formal tests. Here we propose a formal test for outlier distortion assessing whether robust estimates obtained using popular outlier-robust Huber-skip estimators are different from those obtained using OLS. We apply our test to estimates of the macro-economic impacts of climate change allowing for adaptation.

Distorted estimates due to outlying observations are of particular concern in empirical analyses of the economic impacts of climate change. The standard approach in the empirical macro-economic climate-impact literature is to estimate country or region-level panel models, modelling GDP per capita growth as a function of population-weighted climate observations, see for example Dell et al. (2012); Burke et al. (2015); Pretis, Schwarz, et al. (2018); Kalkuhl & Wenz (2020); Newell et al. (2021). Control variables in these impact models are conventionally limited to unit and time fixed effects, together with unit-specific non-linear time trends. With such limited controls and the wide range of determinants of economic growth, there is the potential for many un-modelled idiosyncratic shocks (i.e. outliers) to bias the estimated coefficients on climate variables. Existing estimates in the climate impact literature have relied on heuristic comparisons of OLS and outlier-robust estimates due to a lack of formal tests. For instance, comparing robust estimates to their main OLS specification, Dell et al. (2012) write in their seminal paper modelling GDP growth as a function of temperatures: "*When we use median regressions to reduce the impact of outliers, the estimated impact [of temperatures] for poor countries becomes slightly larger and substantially more statistically significant*". However, it is unclear whether 'slightly larger' implies that the estimates are statistically different from those obtained using OLS. Similarly, Pretis, Schwarz, et al. (2018) estimate an empirical impact model robust to outliers using Impulse Indicator Saturation (IIS - see Hendry et al. 2008), finding that "*controlling for these outlying observations reduces the slope of the estimated temperature curve [...] (suggesting reduced impact of temperatures on countries with low average annual temperatures)*". Nevertheless, this comparison also remains informal as no test was available to compare OLS to robust IIS estimates. Such potentially-distorted coefficients may result in biases in the projections of the economic impacts of climate change and subsequently distort cost-benefit analyses of climate policy.

Outlier-robust estimators as in the above two cases pose a robustness-efficiency trade-off. When outliers are present, we should rely on robust estimators. When no outliers are present, OLS is more efficient. Unfortunately in practise the presence of outliers and any resulting distortion in coefficients is unknown. To assess whether the gain in robustness offsets the loss of efficiency when using robust estimators, we propose an outlier distortion test. Specifically, we introduce a test for outlier distortion comparing the difference between OLS and robust estimates derived from Huber-skip estimators. The robust procedure underlying Huber-skip estimators is to use least squares estimators on "clean" data obtained by removing observations with extreme residuals from an initial least squares regression fit. This two-step procedure originates from the robust statistics literature, where it has been referred to as Trimmed Least Squares (Ruppert & Carroll 1980); the Data-analytic Strategy (Welsh & Ronchetti 2002); and the 1-step Huber-skip M-estimator (Johansen & Nielsen 2009). This procedure may start with any initial estimator, including robust

ones. Two special cases are Robustified Least Squares (RLS) and IIS, which respectively use the full sample and split sample least squares as initial estimators. Huber-skip type estimators have been widely applied in empirical studies to conduct outlier robustness checks, with RLS being used to estimate the social returns to equipment investment (De Long & Summers 1991, 1994, Auerbach et al. 1994), the institutional impact on economic growth (Acemoglu et al. 2001, 2012, Albouy 2012, Acemoglu et al. 2019); and IIS being used to model wages (Castle & Hendry 2009), food expenditure and demand (Hendry & Mizon 2011), money demand (Dreger & Wolters 2014), housing markets (Anundsen 2015), unemployment (Nymoen & Sparrman 2015), exchange rates (Stillwagon 2016), debt forecasts (Ericsson 2017); with climate applications ranging from assessing climate model performances (Pretis et al. 2015) to detecting volcanic eruptions in temperature reconstructions (Schneider et al. 2017) to hurricane damages (Martinez 2020). However, since no statistical test was available to compare OLS to RLS/IIS all of these studies had to rely on heuristic comparisons of OLS and outlier robust estimators (RLS/IIS) without being able to formally assess whether outliers distort the coefficients on the variables of interest.

The main contribution of our paper is thus to construct a formal test for outlier-robustness of regression coefficients. When no outliers are present, we expect the difference between OLS and robust estimators to be small. In turn, when outliers are present and distort regression estimates, the difference between the two estimators will be large.[1] We show that our test based on the difference between OLS and RLS/IIS has an asymptotic $\chi^2$ distribution and mirrors the test statistics in Durbin (1954)-Hausman (1978)-Wu (1973). We arrive at this result by analytically deriving the limiting distribution of the robust Huber-skip estimators RLS/IIS in cross-sectional and time series regressions, where regressors can be either stationary or deterministically-trending. These results enable us to evaluate whether the gain in robustness is more valuable than the corresponding loss in efficiency when using RLS/IIS. These results also allows to test whether the parameter of interest changes its value significantly before and after removing outliers, providing a formal analysis of outlier sensitivity. Our simulations show that the test has a size close to nominal levels once sample sizes are moderately large ($n \geq 200$) and exhibits high power under a range of alternatives for both vertical outliers as well as bad leverage points. To improve finite sample performance and to alleviate concerns around target reference distributions, we further introduce and explore several bootstrap testing schemes of our distortion test.

The literature on outlier robustness has not commonly focused on testing distortion of coefficients. Two papers, however, are notable exceptions and closely related to our approach. First, Kaji (2018) asymptotically studies outlier-robust estimators, such as RLS or winsorized estimators, all of which can be represented as L-statistics (integrals of transformations of empirical quantile functions with respect to corresponding random sample selection measures). His argument can be extended to instrumental variables regressions, but requires iid observations. Compared to Kaji (2018), we characterise the RLS/IIS and other Huber-skip type estimators as a new class of weighted and marked empirical processes. Our argument is constructed specifically for RLS/IIS allowing us to explore these algorithms in depth. For example, our theory can explore the asymptotics of the variance estimator, iteration of the algorithms, variations of the robustification parameter, and different initial estimators, such as least trimmed squares. In addition, our analysis does not require

[1] For example, Hendry & Mizon (2011) show extreme distortion of regression estimates due to outliers when investigating the effect of relative prices on food expenditure. The sign of the coefficient changes from positive to negative after using the robust Huber-skip IIS estimator.

iid data and is also valid for time series regressions.

Second, Dehon et al. (2012) propose a Hausman-type test for comparing OLS to robust S-estimators. Asymptotic theory of the S-estimators has been well established in the robust statistics literature, thus Dehon et al. (2012) do not need to derive the asymptotics when constructing the outlier distortion test for S-estimators. In contrast, we have to first study the statistical properties of RLS/IIS using the empirical process theory recently developed in the literature in order to explore the difference between OLS and RLS/IIS. In addition, Dehon et al. (2012) restrict their study in the iid setup, whereas time series regressions are allowed in our framework.

In order to develop our outlier distortion test, we have to study the asymptotic behaviour of RLS/IIS and two-step outlier-robust procedures in general. The Huber-skip M-type robust estimators involve weighted and marked empirical processes. These have been studied by Berenguer-Rico et al. (2019) using a martingale decomposition argument with the main results summarized in their Theorem 4.4. To emphasize the theoretical connection to Berenguer-Rico et al. (2019), we present the tightness and first order expansions of their one-sided empirical processes in our Lemma A.1 and extend it to Lemma A.2. We consider two-sided processes which can be applied directly to analyze Huber-skip M-estimators. Some closely related work includes Hendry et al. (2008), Johansen & Nielsen (2009, 2013, 2016a,b, 2019), and Jiao & Nielsen (2015) on analyzing M- and L-type estimators.

Our theory suggests two improvements to existing simple two-step procedures for outlier detection. First, the ordinary variance estimator needs to be bias-corrected due to the fact that some observations may be wrongly classified as outliers and subsequently removed under the null of no outliers. Second, to further gain robustness, the two-step procedure can be iterated until convergence to a fixed point which is shown to have the same first order asymptotics as the Huber-skip M-estimator. Our paper then establishes tightness, a stochastic expansion, and weak convergence of the robust estimators produced by the improved iterated procedure in cross-sectional or time series regressions with stationary and deterministically-trending regressors.

In the wider literature on Huber-skip estimators, our work is related to Jiao & Pretis (2022) who study whether the proportion or number of outliers is different from their expected values when no outliers are present. Jiao (2019) and Jiao & Kurle (2021) further extend asymptotic theory of the RLS/IIS algorithms and their false outlier detection rate to instrumental variables regressions. Berenguer-Rico & Nielsen (2018) and Berenguer-Rico & Wilms (2021) consider diagnostic testing on residuals for normality and heteroskedasticity after outlier removal by robust Huber-skip regressions.

We apply our proposed test of outlier distortion to estimates of the macro-economic impacts of climate change using the robust IIS estimator. Specifically, we estimate a macro-economic climate impact model in line with the growing panel-econometric literature modelling GDP per capita growth as a function of (non-linear) climate variables. We make two contributions to the existing climate-econometric literature. First, to address un-modelled (and *a-priori* unknown) idiosyncratic shocks, we apply the robust IIS estimator and re-estimate the macro-economic impact model of Burke, Hsiang, and Miguel (2015). We then test for outlier distortion of the estimated coefficients to check whether conventional OLS estimates are likely biased due to outliers. Second, as countries become richer, they may be better able to mitigate weather/climate shocks. Beyond existing estimates, we thus consider income-based adaptation. Specifically, we allow the impact of year-

on-year changes in temperatures to vary by time-varying country-specific income levels (similar to Carleton et al. (2020) for mortality).

Our results show that conventional OLS panel estimates of the growth-temperature relationship are significantly different from those obtained using the robust estimator. Once we control for detected outliers, the estimated impacts of temperatures on economic growth are attenuated both for our base model and our adaptation model at all income levels compared to conventional OLS estimates. We also find significant evidence of income-driven adaptation to temperatures. Climate effects are dampened as incomes increase, suggesting that richer countries are likely to have a greater capacity to deal with the consequences of continued climate change, in turn exacerbating existing cross-country inequality.

Our paper proceeds as follows: Section 2 presents our main results with all proofs shown in the online Appendix A. In particular, Section 2.1 and Section 2.2 introduces the regression model, assumptions required for analysis, and a class of outlier-robust algorithms including RLS and IIS. Section 2.3 establishes asymptotic theory of RLS and IIS, whilst Section 2.4 proposes the outlier distortion test. Then, Section 3 conducts Monte Carlo studies with additional simulation results in the online Appendix. Finally, Section 4 applies the outlier distortion test to the macro-economic impacts of climate change using the robust IIS estimator.

## 2 Outlier Distortion Test

We consider a linear regression model

$$y_i = x_i'\beta + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.1}$$

for the data $\{(y_i, x_i)\}_{i=1}^n$, where $y_i$ is univariate and $x_i$ is multivariate with the dimension $d_x$. This setting can represent both classical, time series, and panel regression models. Moreover, in our analysis, regressors $x_i$ can be either stationary or deterministically-trending. Innovations $\varepsilon_i/\sigma$ are independent of the filtration $\mathcal{F}_{i-1} = \sigma(x_1, \ldots, x_i, \varepsilon_1, \ldots, \varepsilon_{i-1})$ with the common density $\mathsf{f}$ and distribution function $\mathsf{F}(c) = \mathsf{P}(\varepsilon_i/\sigma \leq c)$. Denote $\mathsf{g}$ as the density of the absolute error $|\varepsilon_i|/\sigma$ and its distribution function by $\mathsf{G}(c) = \mathsf{P}(|\varepsilon_i|/\sigma \leq c)$ for $c > 0$. Assuming symmetry of $\mathsf{f}$, $\mathsf{G}(c) = 2\mathsf{F}(c) - 1$ and $\mathsf{g}(c) = 2\mathsf{f}(c)$. Define $\psi_c = \mathsf{G}(c)$ so the probability of exceeding the cut-off $c$ is $\gamma_c = 1 - \psi_c$. Suppose the $k$-th (truncated) moment of the density $\mathsf{f}$ exists so that they can be defined as

$$\tau_k = \int_{-\infty}^{\infty} u^k \mathsf{f}(u) du, \qquad \tau_k^c = \int_{-c}^{c} u^k \mathsf{f}(u) du. \tag{2.2}$$

Thus, $\tau_0^c = \psi_c$, $\tau_2 = 1$ while $\tau_k = \tau_k^c = 0$ for odd $k$ under symmetry. We define the conditional variance of $\varepsilon_i/\sigma$ given $(|\varepsilon_i|/\sigma \leq c)$ as

$$\varsigma_c^2 = \frac{\tau_2^c}{\psi_c} = \frac{\int_{-c}^{c} u^2 \mathsf{f}(u) du}{\mathsf{P}(|\varepsilon_i| \leq \sigma c)}. \tag{2.3}$$

This is the bias correction factor for the variance estimate computed from the selected non-outlying sample. For a Normal reference $\varepsilon_i/\sigma \sim \mathsf{N}(0, 1)$, then $\tau_2^c = \psi_c - 2c\mathsf{f}(c)$, $\tau_4^c = 3\psi_c - 2c(c^2 + 3)\mathsf{f}(c)$ and $\tau_4 = 3$.

Outliers are pairs of observations that do not conform with the model (2.1) or with the reference density $\mathsf{f}$. We are interested in the presence of outliers where the errors $\varepsilon_i/\sigma$ are drawn from the reference distribution $\mathsf{f}$ but potentially contaminated by an arbitrary (possibly fatter tail) unknown distribution $\mathsf{f}^{\mathrm{c}}$ under $\epsilon$-contamination framework as in Huber (1964)

$$(1-\epsilon)\mathsf{f}+\epsilon\mathsf{f}^{\mathrm{c}}. \tag{2.4}$$

Compared to not imposing a parametric distribution on errors, the mixture model of $\mathsf{f}$ and $\mathsf{f}^{\mathrm{c}}$ in (2.4) is an alternative way to relax the assumption on $\varepsilon_i/\sigma \sim \mathsf{f}$. The following section describes the iterated 1-step Huber-skip M-estimator where we subsequently derive its asymptotic properties and present a new Durbin-Hausman-Wu type test to formalize outlier robustness checks.

## 2.1 Algorithms Robust to Outliers

We consider algorithms that remove outliers following an initial estimator. The simplest approach is to use least squares estimators on "clean" data obtained by removing observations with extreme residuals from an initial least squares regression fit. There are, however, two potential improvements to this simple two-step procedure (also referred to as trimmed least squares). First, Johansen & Nielsen (2009) suggest that the updated variance estimator should be corrected by the factor $\varsigma^{-2}$ introduced in (2.3), since the simple procedure underestimates $\sigma^2$ in the case where observations are identified by chance and falsely removed as outliers. Second, robustness of the estimator could be improved by iterating the procedure. Considering these two improvements, we study the so called *iterated 1-step Huber-skip M-estimators* in Algorithm 2.1:

**Algorithm 2.1.** *Choose a cut-off $c>0$.*
*1. Choose initial estimators $\widehat{\beta}_c^{(0)}$, $(\widehat{\sigma}_c^{(0)})^2$ and let $m=0$.*
*2. Define indicator variables for selecting non-outlying observations*

$$v_{i,c}^{(m)} = 1_{(|y_i - x_i'\widehat{\beta}_c^{(m)}| \le \widehat{\sigma}_c^{(m)} c)}. \tag{2.5}$$

*3. Compute least squares estimators*

$$\widehat{\beta}_c^{(m+1)} = (\sum_{i=1}^n x_i x_i' v_{i,c}^{(m)})^{-1}(\sum_{i=1}^n x_i y_i v_{i,c}^{(m)}), \tag{2.6}$$

$$(\widehat{\sigma}_c^{(m+1)})^2 = \varsigma_c^{-2}(\sum_{i=1}^n v_{i,c}^{(m)})^{-1}\{\sum_{i=1}^n (y_i - x_i'\widehat{\beta}_c^{(m+1)})^2 v_{i,c}^{(m)}\}. \tag{2.7}$$

*4. Let $m=m+1$ and repeat 2 and 3.*

Having defined the robust algorithm 2.1, we need to specify initial estimators $\widehat{\beta}_c^{(0)}$, $(\widehat{\sigma}_c^{(0)})^2$. It is common to choose the full sample OLS as the initial estimator (see for example Acemoglu et al. (2019) and other references listed in the introduction).

We refer to using OLS as the initial estimator as *Robustified Least Squares* (RLS). While this estimator is popular in practise (and thus part of our analysis), it is itself is not an ideal robust estimator as the initial estimator is not robust to outliers. Therefore we primarily focus our analysis on a robust initial estimator, such as the split sample least squares approach described in Algorithm

2.2 and referred to as *Impulse Indicator Saturation* (IIS – see Hendry, Johansen, and Santos 2008). In our climate application, we rely on the robust estimates using IIS to draw inference on macroeconomic impacts of climate change. IIS starts with an initial estimator which divides the full sample into two sub-samples and uses regression estimates calculated from each sub-sample to detect outliers in the other sub-sample:

**Algorithm 2.2.** ***Stylized Impulse Indicator Saturation****. Choose a cut-off $c > 0$.*
*1.1. Split full sample into two sets $\mathcal{I}_j$, $j = 1, 2$ of $n_j$ observations where $\sum_{j=1}^{2} n_j = n$.*
*1.2. Calculate least squares estimators based upon each sub-sample $\mathcal{I}_j$ for $j = 1, 2$*

$$\widehat{\beta}_j = (\sum_{i\in\mathcal{I}_j} x_i x_i')^{-1}(\sum_{i\in\mathcal{I}_j} x_i y_i), \qquad \widehat{\sigma}_j^2 = \frac{1}{n_j}\sum_{i\in\mathcal{I}_j}(y_i - x_i'\widehat{\beta}_j)^2. \tag{2.8}$$

*1.3. Define the initial indicator variables for selecting non-outlying observations*

$$v_{i,c}^{(0)} = 1_{(i\in\mathcal{I}_1)}1_{(|y_i - x_i'\widehat{\beta}_2|\leq\widehat{\sigma}_2 c)} + 1_{(i\in\mathcal{I}_2)}1_{(|y_i - x_i'\widehat{\beta}_1|\leq\widehat{\sigma}_1 c)}. \tag{2.9}$$

*1.4. Compute $\widehat{\beta}_c^{(1)}$, $(\widehat{\sigma}_c^{(1)})^2$ using (2.6), (2.7) with $m = 0$, and then let $m = 1$.*
*2. Follow the step 2,3,4 in Algorithm 2.1.*

The initial sets $\mathcal{I}_1$ and $\mathcal{I}_2$ should be iterated as the location of of contaminated observations is generally unknown. The iterated 1-step Huber-skip M-estimator mimics the Huber (1964) skip[2] estimator with the criterion function

$$\rho(t) = \begin{cases} \frac{t^2}{2}, & \text{if } |t| \leq c, \\ \frac{c^2}{2}, & \text{otherwise,} \end{cases} \tag{2.10}$$

which is immune to either outliers or a fatter tail distribution (defined relative to the reference $\mathsf{f}$) under the $\epsilon$-contamination $(1-\epsilon)\mathsf{f} + \epsilon\mathsf{f}^{\mathrm{c}}$. In practise, instead of applying the simple two-step approach, we recommend iterating Algorithm 2.1 until it converges to a fixed point. We show later in our theory that the fixed point estimator has the same first order asymptotics as the Huber-skip estimator. Thus, Algorithm 2.1 can be understood as an approximation to a the Huber-skip regression. Even with a non-robust OLS starting point, iterating the algorithm increases the robustness of the resulting estimators and results in a fixed point which behaves similarly to the robust Huber-skip regression. Jiao & Pretis (2022) provide guidance on selecting the robustification parameter $c$ and address the testing problem for overall presence of outliers by theoretically analyzing the false outlier detection rate (referred to as the gauge). The main purpose of our paper here is to formalize outlier robustness checks of comparing OLS estimates of $\beta$ to the robust estimates produced by Algorithm 2.1 as the Durbin-Hausman-Wu type test by establishing weak convergence of the iterated 1-step Huber-skip M-estimator with a drifting cut-off $c$.

[2]See Hampel et al. (1986) (p. 104) for the Huber-skip type estimator as opposed to the Huber estimator with the criterion function

$$\rho(t) = \begin{cases} \frac{t^2}{2}, & \text{if } |t| \leq c, \\ c|t| - c^2/2, & \text{otherwise.} \end{cases}$$

.

### 2.2 Assumptions for Asymptotic Theory

We briefly discuss the assumptions required to derive the weak convergence of the robust estimators and formalise our robustness test. Innovations $\varepsilon_i$ and regressors $x_i$ must satisfy moment conditions as outlined in Assumption 2.1 for our asymptotic analysis. Regressors $x_i$ can be temporally dependent and deterministically trending. We therefore require a normalisation matrix $N$ that allows for different behaviour of the components of the regressor vector $x_i$. In the case of a stationary regressor we need a standard $n^{-1/2}$ normalisation so that $N$ must be proportional to the identity matrix of the same dimension as $x_i$, that is $N = n^{-1/2}I_{d_x}$. If the regressors are unbalanced as in $x_i = (1, i)'$ we can choose $N = \text{diag}(n^{-1/2}, n^{-3/2})$. Thus, denote the normalized regressors as $x_{in} = N'x_i$. Explosive bubble processes and regressions with co-trending regressors are not analyzed in the paper.

**Assumption 2.1.** *Let $\mathcal{F}_i$ be an increasing sequence of $\sigma$-fields so $\varepsilon_{i-1}$ and $x_i$ are $\mathcal{F}_{i-1}$ measurable and $\varepsilon_i$ is independent of $\mathcal{F}_{i-1}$. Let $\varepsilon_i/\sigma$ have a symmetric, continuously differentiable density $\mathsf{f}$ which is positive on the real line $\mathbb{R}$. For some values of $\eta$ such that $0 < \eta \leq 1/4$, choose an integer $r \geq 2$ so*

$$2^{r-1} > 1 + (1/4 - \eta)(1 + d_x). \tag{2.11}$$

*Let $q = 1 + 2^{r+1}$. Suppose*
*(i) the density $\mathsf{f}$ satisfies*
*(a) $|u|^q\mathsf{f}(u)$, $|u^{q+1}\dot{\mathsf{f}}(u)|$ are decreasing for large $u$;*
*(ii) the regressors $x_i$ satisfy*
*(a) $\Sigma_n = \sum_{i=1}^n x_{in}x_{in}' \xrightarrow{\mathsf{D}} \Sigma \overset{a.s.}{>} 0$;*
*(b) $n^{-1}\mathsf{E}\sum_{i=1}^n |n^{1/2}x_{in}|^q = \mathrm{O}(1)$;*
*(iii) the initial estimator $(\widetilde{\beta}, \widetilde{\sigma}^2)$ satisfies*
*(a) $N^{-1}(\widetilde{\beta} - \beta) = \mathrm{O}_{\mathsf{P}}(n^{1/4-\eta})$;*
*(b) $n^{1/2}(\widetilde{\sigma}^2 - \sigma^2) = \mathrm{O}_{\mathsf{P}}(n^{1/4-\eta})$.*

While these assumptions may appear abstract, conditions $(i)$, $(ii)$ are satisfied in a range of situations. In particular, the condition $(i)$ is satisfied by the Normal and t distributions; see discussions on the assumptions on $\mathsf{f}$ in Berenguer-Rico & Nielsen (2018), while the condition $(ii)$ is satisfied by stationary and deterministically trending regressors; see assumptions and examples regarding time series regressors $x_i$ in Johansen & Nielsen (2016a), as well as by explosive processes; see a remark in Berenguer-Rico et al. (2019). Condition $(iii)$ allows the standardised estimation errors to diverge at a rate of $n^{1/4-\eta}$ rather than being bounded in probability. In particular, $\eta = 1/4$ can be chosen for estimators with standard convergence rates.

There is a trade-off in (2.11) between $\eta$, the divergence rate of initial estimators $\widetilde{\beta}$, $\widetilde{\sigma}^2$, and $r$, the required number of moments for innovations $\varepsilon_i$ and regressors $x_i$. If we have standard initial estimators which are bounded in probability after normalization, such as RLS and IIS, then $\eta$ becomes $1/4$ so we can choose $r = 2$ regardless of dimension of regressors, implying we only require the lower number of moments. Whereas for non-standard diverging estimators, i.e. $0 < \eta < 1/4$ and $1/4 - \eta > 0$, then the required number $r$ of moments grows linearly with the dimension of the regressor. This would be relevant for the $n^{1/3}$-consistent least median of squares regression estimator proposed by Rousseeuw (1984).

It is also feasible to extend the asymptotic analysis to the case of an asymmetric error distribution. The assumption of symmetry of $\mathsf{f}$ could be relaxed if the cut-off values $\underline{c} < \overline{c}$ are chosen accordingly. Notably, this applies under the null hypothesis of no outlier contamination as our proposed approach will have power under the alternative when contamination is asymmetric. The machinery of deriving asymptotics under the null in our paper holds when $\mathsf{f}$ is asymmetric, however some notation has to be adapted to accommodate this scenario. For example, the truncated moments have to be re-defined as $\tau_k^c = \int_{\underline{c}}^{\overline{c}} u^k \mathsf{f}(u)du$. We now still have $\tau_0 = \tau_2 = 1$, $\tau_1 = 0$, and $\tau_0^c = \psi_c$, where $\psi_c = \mathsf{P}(\sigma\underline{c} \leq \varepsilon_i \leq \sigma\overline{c})$ and $\gamma_c = 1 - \psi_c$. In Algorithm 2.1, we then select two critical values $\underline{c}$ and $\overline{c}$ such that $\tau_1^c = 0$. The indicator variables (2.5) for selecting non-outlying observations then become $v_{i,c}^{(m)} = 1_{(\widehat{\sigma}_c^{(m)}\underline{c} \leq y_i - x_i'\widehat{\beta}_c^{(m)} \leq \widehat{\sigma}_c^{(m)}\overline{c})}$. The rest of the algorithm is then well defined to accommodate the asymmetric $\mathsf{f}$ with two chosen cut-offs $\underline{c}$ and $\overline{c}$. Asymptotic analysis then remains identical to our main specification even under asymmetry of $\mathsf{f}$, however a theory is required for adjusting to the new type of empirical processes. In Appendix A we present a LLN and CLT for the necessary type of empirical processes.

When $\mathsf{f}$ is symmetric and two symmetric cut-off values $-c$ and $c$ are chosen, the robust estimator $\widehat{\beta}_c^{(m)}$ is consistent though less efficient under the null of no outliers while $\widehat{\sigma}_c^{(m)}$ needs to be bias-corrected in order to be consistent. For an asymmetric $\mathsf{f}$, two critical values $\underline{c}$ and $\overline{c}$ can be selected similarly in the sense that $\tau_1^c = 0$, then $\widehat{\beta}_c^{(m)}$ and $\widehat{\sigma}_c^{(m)}$ have the same asymptotic behaviour as analyzed under symmetry. Otherwise, if $\underline{c}$ and $\overline{c}$ are chosen in a way such that $\tau_1^c \neq 0$, then $\widehat{\beta}_c^{(m)}$ is biased as well, thus it is essential to control such additional bias terms through asymptotic studies (see Johansen & Nielsen (2009) for a similar argument and detailed discussion).

### 2.3 Weak Convergence

To formalise the comparison of OLS to outlier-robust estimators like RLS/IIS as a statistical test, we first need to study the asymptotic behaviours of such robust estimators. We focus on a class of iterated one-step Huber-skip estimators, whose analysis relies on the empirical process theory recently developed by Berenguer-Rico et al. (2019). Therefore, in this section we provide asymptotic theory such as tightness, stochastic expansions, fixed points of the iterated estimators computed by Algorithm 2.1, and establish weak convergence of RLS/IIS. Equipped with weak convergence results, we then introduce our outlier distortion test in Section 2.4. Our arguments hold uniformly in cut-off values $c$ and can be extended to develop outlier distortion tests using other types of robust estimators as long as their limiting distributions are well established.

Theorem 2.1 shows that the iterated estimator produced by Algorithm 2.1 is tight in the iteration $m \in [0, \infty)$ and in the cut-off value $c \in [c_+, \infty)$. Note that $c_+ > 0$ is a small positive number.

**Theorem 2.1.** *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 2.1. Suppose Assumption 2.1 holds with $\eta = 1/4$. Then, as $n \to \infty$*

$$\sup_{0 \leq m < \infty} \sup_{c_+ \leq c < \infty} |N^{-1}(\widehat{\beta}_c^{(m)} - \beta)| + |n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)| = \mathrm{O}_{\mathsf{P}}(1).$$

First, Assumption 2.1(*iii*) with $\eta = 1/4$ corresponds to a standard convergence rate for the initial estimator. With the 1-step relationship between the updated and the original estimator provided by Lemma A.4 (see also Corollary 2.3), the tightness can then be demonstrated by a

geometric argument and mathematical induction.

The above tightness theorem implies uniform consistency of the iterated estimators computed by Algorithm 2.1, that is $\widehat{\beta}_c^{(m)} \xrightarrow{\mathsf{P}} \beta$ and $\widehat{\sigma}_c^{(m)} \xrightarrow{\mathsf{P}} \sigma$ uniformly in the cut-off value $c$ and iteration step $m$ as $n \to \infty$. Secondly, tightness will also be used to establish fixed points $\widehat{\beta}_c^{(*)}$ and $\widehat{\sigma}_c^{(*)}$ of Algorithm 2.1 upon infinite iterations when $m \to \infty$ and to demonstrate weak convergence theory of RLS and IIS.

Next, theorem 2.2 shows the stochastic expansions of any iterated step estimators of Algorithm 2.1 in terms of the initial estimators, kernels, and small remainder terms.

**Theorem 2.2.** *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 2.1. Suppose Assumption 2.1 holds with $\eta = 1/4$. Then, as $n \to \infty$ and uniformly in $c \in [c_+, \infty)$, we have for any $m \in [0, \infty)$*

$$N^{-1}(\widehat{\beta}_c^{(m+1)} - \beta) = \varrho_{\beta,c}^{(m+1)} N^{-1}(\widehat{\beta}_c^{(0)} - \beta) + \varrho_{x\varepsilon,c}^{(m+1)} \Sigma_n^{-1} \sum_{i=1}^{n} x_{in}\varepsilon_i 1_{(|\varepsilon_i| \le \sigma c)} + \mathsf{o}_{\mathsf{P}}(1),$$
$$n^{1/2}(\widehat{\sigma}_c^{(m+1)} - \sigma) = \varrho_{\sigma,c}^{(m+1)} n^{1/2}(\widehat{\sigma}_c^{(0)} - \sigma) + \varrho_{\varepsilon\varepsilon,c}^{(m+1)} n^{-1/2} \sum_{i=1}^{n} (\frac{\varepsilon_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|\varepsilon_i| \le \sigma c)} + \mathsf{o}_{\mathsf{P}}(1),$$

*where coefficients have expressions*

$$\varrho_{\beta,c}^{(m+1)} = \{\frac{2c\mathsf{f}(c)}{\psi_c}\}^{m+1}, \qquad \varrho_{x\varepsilon,c}^{(m+1)} = \frac{\psi_c^{m+1} - \{2c\mathsf{f}(c)\}^{m+1}}{\psi_c^{m+1}\{\psi_c - 2c\mathsf{f}(c)\}},$$
$$\varrho_{\sigma,c}^{(m+1)} = \{\frac{c(c^2 - \varsigma_c^2)\mathsf{f}(c)}{\tau_2^c}\}^{m+1}, \qquad \varrho_{\varepsilon\varepsilon,c}^{(m+1)} = \sigma \frac{(\tau_2^c)^{m+1} - \{c(c^2 - \varsigma_c^2)\mathsf{f}(c)\}^{m+1}}{2(\tau_2^c)^{m+1}\{\tau_2^c - c(c^2 - \varsigma_c^2)\mathsf{f}(c)\}}.$$

The above theorem generalises Lemma A.4 in the sense that its proof is to recursively apply the one-step expansion of the updated estimator in terms of the original ones. Let $m = 0$ such that $\varrho_{\beta,c}^{(1)} = 2c\mathsf{f}(c)/\psi_c$, $\varrho_{x\varepsilon,c}^{(1)} = \psi_c^{-1}$, $\varrho_{\sigma,c}^{(1)} = c(c^2 - \varsigma_c^2)\mathsf{f}(c)/\tau_2^c$, and $\varrho_{\varepsilon\varepsilon,c}^{(1)} = \sigma/(2\tau_2^c)$, then the expansion immediately reduces to the one-step case. Here, we provide the following corollary to re-express Lemma A.4 as a stochastic expansion of the fist-step estimators in terms of the initial ones and the kernel terms.

**Corollary 2.3.** *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 2.1. Suppose Assumption 2.1 holds with $\eta = 1/4$. Then, as $n \to \infty$ and uniformly in $c \in [c_+, \infty)$, we have*

$$N^{-1}(\widehat{\beta}_c^{(1)} - \beta) = \frac{2c\mathsf{f}(c)}{\psi_c} N^{-1}(\widehat{\beta}_c^{(0)} - \beta) + (\psi_c \Sigma_n)^{-1} \sum_{i=1}^{n} x_{in}\varepsilon_i 1_{(|\varepsilon_i| \le \sigma c)} + \mathsf{o}_{\mathsf{P}}(1),$$
$$n^{1/2}(\widehat{\sigma}_c^{(1)} - \sigma) = \frac{c(c^2 - \varsigma_c^2)\mathsf{f}(c)}{\tau_2^c} n^{1/2}(\widehat{\sigma}_c^{(0)} - \sigma)$$
$$+ \frac{\sigma}{2\tau_2^c} n^{-1/2} \sum_{i=1}^{n} (\frac{\varepsilon_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|\varepsilon_i| \le \sigma c)} + \mathsf{o}_{\mathsf{P}}(1).$$

Initially the tight estimator is assumed to be available and it is subsequently iterated through the above one-step expansion. Assumption 2.1$(i)$ implies that autoregressive coefficients $2c\mathsf{f}(c)/\psi_c$ and $c(c^2 - \varsigma_c^2)\mathsf{f}(c)/\tau_2^c$ are strictly bounded by one holding uniformly in $c$, suggesting that the above

equation is a contraction mapping. Thus, Algorithm 2.1 will converge to a fixed point as the iteration step increases to be sufficiently large. Let $m \to \infty$ in Theorem 2.2 such that $\varrho_{\beta,c}^{(\infty)} = 0$, $\varrho_{x\varepsilon,c}^{(\infty)} = 1/\{\psi_c - 2c\mathsf{f}(c)\}$, $\varrho_{\sigma,c}^{(\infty)} = 0$, and $\varrho_{\varepsilon\varepsilon,c}^{(\infty)} = \sigma/[2\{\tau_2^c - c(c^2 - \varsigma_c^2)\mathsf{f}(c)\}]$, then the below theorem finds the fixed point $N^{-1}(\widehat{\beta}_c^{(*)} - \beta) = N^{-1}(\widehat{\beta}_c^{(\infty)} - \beta)$, $n^{1/2}(\widehat{\sigma}_c^{(*)} - \sigma) = n^{1/2}(\widehat{\sigma}_c^{(\infty)} - \sigma)$.

**Theorem 2.4.** *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 2.1. Suppose Assumption 2.1 holds with $\eta = 1/4$. Then, for all $\epsilon, \delta > 0$ a pair $m_0, n_0 > 0$ exists, so for all $m > m_0$ and $n > n_0$*

$$\mathsf{P}\{\sup_{c_+ \leq c < \infty} |N^{-1}(\widehat{\beta}_c^{(m)} - \widehat{\beta}_c^{(*)})| + |n^{1/2}(\widehat{\sigma}_c^{(m)} - \widehat{\sigma}_c^{(*)})| > \delta\} < \epsilon,$$

*where*

$$N^{-1}(\widehat{\beta}_c^{(*)} - \beta) = \frac{1}{\psi_c - 2c\mathsf{f}(c)} \Sigma_n^{-1} \sum_{i=1}^n x_{in}\varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)},$$
$$n^{1/2}(\widehat{\sigma}_c^{(*)} - \sigma) = \frac{\sigma}{2\{\tau_2^c - c(c^2 - \varsigma_c^2)\mathsf{f}(c)\}} n^{-1/2} \sum_{i=1}^n (\frac{\varepsilon_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|\varepsilon_i| \leq \sigma c)}.$$

The proof is conducted as follows. According to Theorem 2.1, if the initial estimator is bounded in a large compact set with a large probability, then any iterated estimators of Algorithm 2.1 take values in the same compact set no matter what value of the cut-off $c$ is chosen. Next, the argument is to further demonstrate that the deviation between the $m$-fold iterated estimator and the fixed point is the sum of two terms vanishing exponentially and in probability respectively as $m$ and $n$ go to infinity.

Theorem 2.4 shows the first order asymptotics of the fixed point of Algorithm 2.1, and it is in fact the same as that of the Huber (1964) skip estimator, thus the iterated one-step Huber-skip M-estimator mimics the Huber-skip estimator and Algorithm 2.1 can be understood as an implementation of the Huber-skip estimation as a non-linear optimisation problem.

The choice of the initial estimator does not affect Algorithm 2.1 in terms of finding the fixed point, as long as it has the tightness property. Thus, the fixed point theorem also applies in the case of RLS and IIS. It also applies to the theorems regarding to tightness and stochastic expansions, since the full sample or split sample least squares as their starting estimators are tight. We now show the stochastic expansions of RLS and IIS in order to establish weak convergence theory.

**Theorem 2.5.** *Consider Robustified Least Squares (RLS) or split half Impulse Indicator Saturation where $n_1 = \text{int}[n/2]$ and $n_2 = n - n_1$ (IIS). Suppose Assumption 2.1$(i, ii)$ holds for each sub-sample set $\mathcal{I}_1$, $\mathcal{I}_2$. Then, as $n \to \infty$ and uniformly in $c \in [c_+, \infty)$ we have for any $m \in [0, \infty)$*

$$N^{-1}(\widehat{\beta}_c^{(m+1)} - \beta) = \varrho_{\beta,c}^{(m+1)} \Sigma_n^{-1} \sum_{i=1}^n x_{in}\varepsilon_i + \varrho_{x\varepsilon,c}^{(m+1)} \Sigma_n^{-1} \sum_{i=1}^n x_{in}\varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)} + \mathrm{o_P}(1),$$
$$n^{1/2}(\widehat{\sigma}_c^{(m+1)} - \sigma) = \varrho_{\sigma,c}^{(m+1)} \frac{\sigma}{2} n^{-1/2} \sum_{i=1}^n (\frac{\varepsilon_i^2}{\sigma^2} - 1)$$
$$+ \varrho_{\varepsilon\varepsilon,c}^{(m+1)} n^{-1/2} \sum_{i=1}^n (\frac{\varepsilon_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|\varepsilon_i| \leq \sigma c)} + \mathrm{o_P}(1),$$

*where the coefficients* $\varrho_{\beta,c}^{(m+1)}, \varrho_{x\varepsilon,c}^{(m+1)}, \varrho_{\sigma,c}^{(m+1)}, \varrho_{\varepsilon\varepsilon,c}^{(m+1)}$ *are defined in Theorem 2.2.*

Substituting the expansion of the full sample least squares as the initial estimator in Theorem 2.2, we immediately prove the above theorem for RLS. Then, we find that the split half version of IIS has the identical expansion as RLS. The first step updated estimator and the fixed point of RLS and IIS are two special cases of our main result, so their expansions can be obtained by letting $m = 0$ and $m \to \infty$. With Theorem 2.5, we are ready to establish the weak convergence theory for RLS and IIS in cases where we primarily focus on stationary regressions with deterministic trends.

Theorem 2.1 demonstrates uniform consistency of the iterated estimators of $\beta, \sigma^2$. Next, we concentrate on distributional analysis of the iterated estimator of $\beta$. Using the stochastic expansion of the $\beta$ estimator in Theorem 2.5, it suffices to analyse the distribution of the kernel vector $\sum_{i=1}^{n}(x_{in}'\varepsilon_i, x_{in}'\varepsilon_i 1_{(|\varepsilon_i|\leq\sigma c)})'$. With this purpose in mind, we first need to discuss the choice of the normalisation matrix $N$ in $x_{in} = N'x_i$ and the limiting behaviour of the covariance matrix of the normalised regressors $\Sigma_n = \sum_{i=1}^{n} x_{in}x_{in}'$.

*Stationary case.* Suppose regressors are cross-sectional iid or arise from a stationary time series model. Thus, we choose $N = n^{-1/2}I_{d_x}$ for normalising the regressors such that $x_{in} = N'x_i = n^{-1/2}x_i$. Then, $\Sigma_n = \sum_{i=1}^{n} x_{in}x_{in}' = n^{-1}\sum_{i=1}^{n} x_i x_i'$ converges in probability to a deterministic term $\Sigma = \mathsf{E}x_i x_i'$ by LLN. To investigate the asymptotic behaviour of the kernel vector, we apply martingale CLT so that for any $c \in [c_+, \infty)$

$$\sum_{i=1}^{n}\begin{pmatrix} x_{in}\varepsilon_i \\ x_{in}\varepsilon_i 1_{(|\varepsilon_i|\leq\sigma c)}\end{pmatrix} = n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} x_i\varepsilon_i \\ x_i\varepsilon_i 1_{(|\varepsilon_i|\leq\sigma c)}\end{pmatrix} \xrightarrow{\mathsf{D}} \mathsf{N}\left\{\begin{pmatrix}0_{d_x}\\0_{d_x}\end{pmatrix}, \sigma^2\tau_2^c\begin{pmatrix}\frac{1}{\tau_2^c}\Sigma & \Sigma \\ \Sigma & \Sigma\end{pmatrix}\right\}.$$

Drifting cut-off values $c$ in the interval $[c_+, \infty)$, we obtain a sequence of processes $\mathbb{G}_n^{(m+1)}(c) = N^{-1}(\widehat{\beta}_c^{(m+1)} - \beta) = n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta)$ for any $m \in [0, \infty)$. We can now establish a weak convergence theory for $\mathbb{G}_n^{(m+1)}$, which follows from a finite dimensional convergence and tightness; see Billingsley (1968). The below theorem then shows that RLS and IIS are asymptotically approximated by a Gaussian process.

**Theorem 2.6.** *Consider RLS or IIS. Suppose Assumption 2.1*$(i, ii)$ *holds. For any* $m \in [0, \infty)$*, denote the processes* $\mathbb{G}_n^{(m+1)}(c) = n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta)$ *with the argument* $c \in [c_+, \infty)$*. Then as* $n \to \infty$*,* $\mathbb{G}_n^{(m+1)}$ *weakly converges to a zero mean Gaussian process* $\mathbb{G}^{(m+1)}$ *with variance*

$$\mathsf{Var}\{\mathbb{G}^{(m+1)}(c)\} = \{(\varrho_{\beta,c}^{(m+1)})^2 + 2\tau_2^c\varrho_{\beta,c}^{(m+1)}\varrho_{x\varepsilon,c}^{(m+1)} + \tau_2^c(\varrho_{x\varepsilon,c}^{(m+1)})^2\}\sigma^2\Sigma^{-1},$$

*where* $\varrho_{\beta,c}^{(m+1)}$*,* $\varrho_{x\varepsilon,c}^{(m+1)}$ *are defined in Theorem 2.2.*

Again, the one-step updated estimator and fixed point are of particular interest in RLS or IIS. To explore their weak convergence, let $m = 0$ and $m \to \infty$ in Theorem 2.6 such that $\varrho_{\beta,c}^{(1)} = 2c\mathsf{f}(c)/\psi_c$, $\varrho_{x\varepsilon,c}^{(1)} = \psi_c^{-1}$ and $\varrho_{\beta,c}^{(\infty)} = 0$, $\varrho_{x\varepsilon,c}^{(\infty)} = 1/\{\psi_c - 2c\mathsf{f}(c)\}$, then we have the below corollary.

**Corollary 2.7.** *Consider RLS or IIS. Suppose Assumption 2.1*$(i, ii)$ *holds. The cut-off* $c$ *drifts in the interval* $[c_+, \infty)$*. Then as* $n \to \infty$*, for* $m = 0$ *a sequence of processes* $\mathbb{G}_n^{(1)}$ *of the initial estimator weakly converges to a zero mean Gaussian process* $\mathbb{G}^{(1)}$ *with variance*

$$\mathsf{Var}\{\mathbb{G}^{(1)}(c)\} = \frac{4c^2\mathsf{f}^2(c) + 4\tau_2^c c\mathsf{f}(c) + \tau_2^c}{\psi_c^2}\sigma^2\Sigma^{-1}.$$

*In addition, for $m \to \infty$ a sequence of processes $\mathbb{G}_n^{(*)}$ of the fixed point estimator weakly converges to a zero mean Gaussian process $\mathbb{G}^{(*)}$ with variance*

$$\mathsf{Var}\{\mathbb{G}^{(*)}(c)\} = \frac{\tau_2^c}{\{\psi_c - 2c\mathsf{f}(c)\}^2}\sigma^2\Sigma^{-1}.$$

*Deterministic trends.* To keep notations simple and analysis clear, here we consider an example which follows the regression

$$y_i = \beta_0 + \beta_1 i + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

For the regressors $x_i = (1, i)'$, we choose the normalisation matrix $N = \begin{pmatrix} n^{-1/2} & 0 \\ 0 & n^{-3/2} \end{pmatrix}$ so that $x_{in} = N'x_i = (n^{-1/2}, n^{-3/2}i)'$. Then, it follows

$$\Sigma_n = \sum_{i=1}^{n} x_{in}x_{in}' = \sum_{i=1}^{n} \begin{pmatrix} n^{-1} & n^{-2}i \\ n^{-2}i & n^{-3}i^2 \end{pmatrix} \to \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{pmatrix} = \Sigma.$$

Notice that we use $\sum_{i=1}^n i = n(n+1)/2$ and $\sum_{i=1}^n i^2 = n(n+1)(2n+1)/6$ to obtain the above deterministic limit. The kernel vector $\sum_{i=1}^n (x_{in}'\varepsilon_i, x_{in}'\varepsilon_i 1_{(|\varepsilon_i| \le \sigma c)})'$ has a limiting normal distribution with the same form of mean and variance as given in the stationary case where instead $d_x = 2$ and $\Sigma$ is derived immediately above. For any $m \in [0, \infty)$, denote a sequence of processes of the iterated estimators computed by RLS or IIS as $\mathbb{G}_n^{(m+1)}(c) = N^{-1}(\widehat{\beta}_c^{(m+1)} - \beta) = \begin{Bmatrix} n^{1/2}(\widehat{\beta}_{0,c}^{(m+1)} - \beta_0) \\ n^{3/2}(\widehat{\beta}_{1,c}^{(m+1)} - \beta_1) \end{Bmatrix}$ with $c \in [c_+, \infty)$. Thus as $n \to \infty$, $\mathbb{G}_n^{(m+1)}$ weakly converges to a zero mean Gaussian process with the same form of variance as given in Theorem 2.6 and Corollary 2.7 where again $\Sigma$ needs to be changed to the one shown above.

Other cases of deterministic trends can be studied using a similar analysis to the above example. For instance, the argument applies to trend stationary autoregressions but involves a notationally tedious detrending derivation; see Section 1.5.1 in Johansen & Nielsen (2009) for a related and more detailed description.

With the above weak convergence results of RLS and IIS, the next step is to construct statistical tests for coefficient distortion due to outliers. The proposed tests formalise the common practice for assessing outlier robustness by comparing IIS (or RLS) to OLS estimates.

### 2.4 Testing for Outlier Distortion in Regression Coefficients

A frequent concern in empirical economics is whether a small set of outliers may have invalidated empirical results. A common practice to check for outlier-robustness is to compare full sample OLS estimates to those obtained from a sample having trimmed all outliers detected by Algorithm 2.1. Instead of heuristically checking the difference between trimmed LS and OLS, we formalize such an outlier robustness comparison as a new type of Durbin (1954)-Hausman (1978)-Wu (1973) test.

The test is based on the trade-off between robustness and efficiency and enables us to judge whether least squares estimation is appropriate or the robust method should be preferred. The robust estimator produced by Algorithm 2.1 is consistent both under the null and alternatives

(although less efficient under the null[3]), whereas OLS is efficient (and consistent) under the null, but inconsistent otherwise. Our proposed test statistics is to test for a significant difference between coefficients estimated using RLS/IIS and OLS. If the model is correctly specified, the Hausman type test statistics should be small under the null of no outliers, since two consistent methods should produce coefficient estimates that are very close to the true population values. In turn, when outliers have a large influence on least squares estimates, the robust method should be very different from the ordinary estimate. The proposed test thus evaluates whether the gain in robustness offsets the corresponding loss in efficiency.

We compare the full sample OLS estimates $\widetilde{\beta}$ to the robust estimate $\widehat{\beta}_c^{(m+1)}$. Our proposed test can detect whether two estimates are significantly distinct by assessing the L2 norm of the difference between $\widetilde{\beta}$ and $\widehat{\beta}_c^{(m+1)}$. Thus, for any $m \in [0, \infty)$ it is essential first to derive the stochastic expansion and weak limit of a sequence of stochastic processes $\mathbb{H}_n^{(m+1)}(c) = N^{-1}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ with the argument $c \in [c_+, \infty)$. We concentrate on stationary regressions in this section, where $N = n^{-1/2} I_{d_x}$ such that $\mathbb{H}_n^{(m+1)}(c) = n^{1/2}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$, $x_{in} = n^{-1/2} x_i$, and $\Sigma = \mathsf{E} x_i x_i'$.

**Theorem 2.8.** *Consider RLS or IIS. Suppose Assumption 2.1(i, ii) holds. For any $m \in [0, \infty)$, denote the processes $\mathbb{H}_n^{(m+1)}(c) = n^{1/2}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ with the argument $c \in [c_+, \infty)$. Then as $n \to \infty$, we have*

$$\mathbb{H}_n^{(m+1)}(c) = (\varrho_{\beta,c}^{(m+1)} - 1)\Sigma_n^{-1} \sum_{i=1}^{n} x_{in}\varepsilon_i + \varrho_{x\varepsilon,c}^{(m+1)} \Sigma_n^{-1} \sum_{i=1}^{n} x_{in}\varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)} + \mathsf{o_P}(1),$$

*where $\varrho_{\beta,c}^{(m+1)}$, $\varrho_{x\varepsilon,c}^{(m+1)}$ are defined in Theorem 2.2. Furthermore, $\mathbb{H}_n^{(m+1)}$ weakly converges to a zero mean Gaussian process $\mathbb{H}^{(m+1)}$ with variance given as*

$$\mathsf{Var}\{\mathbb{H}^{(m+1)}(c)\} = \{(\varrho_{\beta,c}^{(m+1)} - 1)^2 + 2\tau_2^c(\varrho_{\beta,c}^{(m+1)} - 1)\varrho_{x\varepsilon,c}^{(m+1)} + \tau_2^c(\varrho_{x\varepsilon,c}^{(m+1)})^2\}\sigma^2\Sigma^{-1}.$$

The proof immediately follows from the stochastic expansion and weak convergence of $\widehat{\beta}_c^{(m+1)}$ in Theorem 2.5 and 2.6. Recall the expansion of RLS or IIS from Theorem 2.5

$$N^{-1}(\widehat{\beta}_c^{(m+1)} - \beta) = \varrho_{\beta,c}^{(m+1)} \Sigma_n^{-1} \sum_{i=1}^{n} x_{in}\varepsilon_i + \varrho_{x\varepsilon,c}^{(m+1)} \Sigma_n^{-1} \sum_{i=1}^{n} x_{in}\varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)} + \mathsf{o_P}(1).$$

We then find that the expansion of the test statistic given by the difference between RLS/IIS and OLS is almost identical to that of RLS/IIS except the coefficients on the first kernel term $\Sigma_n^{-1} \sum_{i=1}^{n} x_{in}\varepsilon_i$. The expansion of $N^{-1}(\widehat{\beta}_c^{(m+1)} - \beta)$ has the coefficient $\varrho_{\beta,c}^{(m+1)}$ while $N^{-1}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ has $\varrho_{\beta,c}^{(m+1)} - 1$, therefore the limiting distribution of the testing statistic immediately follows that of RLS/IIS but with the adapted coefficient $\varrho_{\beta,c}^{(m+1)} - 1$ instead of $\varrho_{\beta,c}^{(m+1)}$. This argument thus implies that the asymptotic theory of the outlier distortion test can be attained as soon as the limiting distribution of the robust estimator is derived, regardless of the type of regression. Using the weak convergence result of $\mathbb{H}_n^{(m+1)}$, we next establish the formal outlier distortion test.

[3]The trimmed LS would throw away observations wrongly classified as outliers and thus has a higher asymptotic variance than OLS under the null; See the weak convergence result of RLS and IIS in §2.3; Also see the discussion in Johansen & Nielsen (2009) on the relative efficiency factor (efficiency loss) of IIS.

**Corollary 2.9.** *Consider RLS or IIS. Suppose Assumption 2.1(i, ii) holds. For any $m \in [0, \infty)$, $c \in [c_+, \infty)$ and as $n \to \infty$, we have*

$$n^{1/2}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) \xrightarrow{\mathsf{D}} \mathsf{N}\{0_{d_x}, \mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})\},$$

*where* $\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) = \{(\varrho_{\beta,c}^{(m+1)} - 1)^2 + 2\tau_2^c(\varrho_{\beta,c}^{(m+1)} - 1)\varrho_{x\varepsilon,c}^{(m+1)} + \tau_2^c(\varrho_{x\varepsilon,c}^{(m+1)})^2\}\sigma^2\Sigma^{-1}$. *Then, the proposed test statistics has the weak limit*

$$H_{n,c}^{(m+1)} = n(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})'\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})^{-1}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) \xrightarrow{\mathsf{D}} \chi_{d_x}^2.$$

The robust estimator (RLS/IIS $\widehat{\beta}_c^{(m+1)}$) is always consistent but inefficient under the null of no outliers, while the other (OLS $\widetilde{\beta}$) is efficient under the null but not consistent under alternatives. In this case the asymptotic variance of their difference is given by the difference of their respective asymptotic variances under the null of no outliers. Using a similar argument to Hausman (1978), Lemma A.5 in the appendix demonstrates the above statement in our context. In addition, we provide a different but more direct proof in Remark A.1 to show that $\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) = \mathsf{avar}(\widehat{\beta}_c^{(m+1)}) - \mathsf{avar}(\widetilde{\beta})$ under the null of no outliers. Our argument makes use of the asymptotics derived for $\widehat{\beta}_c^{(m+1)}$ in §2.3 and indicates that the regularity conditions required by Lemma A.5 hold for $\mathsf{f} \overset{\mathsf{D}}{=} \mathsf{N}(0, 1)$.

To ensure power of the outlier distortion test under alternatives, we recommend using $\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ as suggested in Corollary 2.9 rather than $\mathsf{avar}(\widehat{\beta}_c^{(m+1)}) - \mathsf{avar}(\widetilde{\beta})$ when constructing the test statistic. Although the two are equal under the null of no outliers, this may not be the case under alternatives. Furthermore, we need to estimate $\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ in order to conduct the test. Given the chosen iteration step $m$, the cut-off value $c$, and the reference distribution $\mathsf{f}$, the terms $\tau_2^c$, $\varrho_{\beta,c}^{(m+1)}$, $\varrho_{x\varepsilon,c}^{(m+1)}$ appearing in $\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ are known, so parameters that need to be estimated are $\sigma^2$, $\Sigma$. Their estimators should be consistent under the null and robust under alternatives, thus we also recommend to estimate $\sigma^2$, $\Sigma$ using the clean data with all outliers removed, since the full sample estimators are inconsistent under alternatives though efficient under the null. Given any chosen $m \in [0, \infty)$ and $c \in [c_+, \infty)$, we can consistently estimate $\sigma^2$ and $\Sigma = \mathsf{E}x_i x_i'$ under the null using the subsample of all non-outlying observations by

$$(\widehat{\sigma}_c^{(m+1)})^2 = \varsigma_c^{-2}(\sum_{i=1}^n v_{i,c}^{(m)})^{-1}\{\sum_{i=1}^n (y_i - x_i'\widehat{\beta}_c^{(m+1)})^2 v_{i,c}^{(m)}\},$$
$$\widehat{\Sigma}_c^{(m+1)} = (\sum_{i=1}^n v_{i,c}^{(m)})^{-1}(\sum_{i=1}^n x_i x_i' v_{i,c}^{(m)}).$$

Thus, our suggested estimator of $\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ is given by

$$\begin{aligned}&\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) \\ &= \{(\varrho_{\beta,c}^{(m+1)} - 1)^2 + 2\tau_2^c(\varrho_{\beta,c}^{(m+1)} - 1)\varrho_{x\varepsilon,c}^{(m+1)} + \tau_2^c(\varrho_{x\varepsilon,c}^{(m+1)})^2\}(\widehat{\sigma}_c^{(m+1)})^2(\widehat{\Sigma}_c^{(m+1)})^{-1}. \qquad (2.12)\end{aligned}$$

If instead we use $\mathsf{avar}(\widehat{\beta}_c^{(m+1)}) - \mathsf{avar}(\widetilde{\beta})$ to construct the testing statistics and estimate $\sigma^2$ and $\Sigma$ in $\mathsf{avar}(\widetilde{\beta})$ using the full sample, then under alternatives the test would lose power and lead to

incorrect results. More seriously, it is very likely under alternatives to have $\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(m+1)}) \leq \widehat{\mathsf{avar}}(\widetilde{\beta})$ such that $\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(m+1)}) - \widehat{\mathsf{avar}}(\widetilde{\beta}) \leq 0$, thus the Hausman type testing statistics is negative rendering the test meaningless in this case.

**Outlier distortion tests.** We can now establish our proposed outlier distortion test formalising the comparison between OLS and robust estimates. For any chosen $m \in [0, \infty)$ and cut-off $c \in [c_+, \infty)$, we have the test statistic and its limiting distribution

$$n^{1/2}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) \overset{a}{\sim} \mathsf{N}\{0_{d_x}, \widehat{\mathsf{avar}}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})\},$$

and

$$\widehat{H}_{n,c}^{(m+1)} = n(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})' \widehat{\mathsf{avar}}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})^{-1}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) \overset{a}{\sim} \chi^2_{d_x}.$$

Note that $\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ shown in (2.12) is invertible and its inverse is given by

$$\begin{aligned}
&\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})^{-1} \\
&= \{(\varrho_{\beta,c}^{(m+1)} - 1)^2 + 2\tau_2^c(\varrho_{\beta,c}^{(m+1)} - 1)\varrho_{x\varepsilon,c}^{(m+1)} + \tau_2^c(\varrho_{x\varepsilon,c}^{(m+1)})^2\}^{-1}(\widehat{\sigma}_c^{(m+1)})^{-2}\widehat{\Sigma}_c^{(m+1)},
\end{aligned}$$

since $(\varrho_{\beta,c}^{(m+1)}-1)^2+2\tau_2^c(\varrho_{\beta,c}^{(m+1)}-1)\varrho_{x\varepsilon,c}^{(m+1)}+\tau_2^c(\varrho_{x\varepsilon,c}^{(m+1)})^2 \neq 0$. We therefore avoid the rank deficiency problem described and addressed by the generalised inverse in Hausman & Taylor (1981) and Holly (1982). The proposed test can either be performed as the two-sided test with the normal limit or as the one-sided test with the chi-squared limit. When implementing RLS or IIS we are particularly interested in two specific estimators: the trimmed estimator just updated from the full sample OLS and the fixed point estimator iterated upon through infinite steps. The below corollary derives the outlier distortion tests for these two special cases, checking whether $\widetilde{\beta}$ is distinct from $\widehat{\beta}_c^{(1)}$ when $m = 0$ or from $\widehat{\beta}_c^{(*)}$ when $m \to \infty$.

**Corollary 2.10.** *Consider RLS or IIS. Suppose Assumption 2.1$(i, ii)$ holds. For any $c \in [c_+, \infty)$ and large $n$, then when $m = 0$ we have*

$$n^{1/2}(\widehat{\beta}_c^{(1)} - \widetilde{\beta}) \overset{a}{\sim} \mathsf{N}\{0_{d_x}, \widehat{\mathsf{avar}}(\widehat{\beta}_c^{(1)} - \widetilde{\beta})\},$$

*and*

$$\widehat{H}_{n,c}^{(1)} = n(\widehat{\beta}_c^{(1)} - \widetilde{\beta})' \widehat{\mathsf{avar}}(\widehat{\beta}_c^{(1)} - \widetilde{\beta})^{-1}(\widehat{\beta}_c^{(1)} - \widetilde{\beta}) \overset{a}{\sim} \chi^2_{d_x},$$

*where*

$$\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(1)} - \widetilde{\beta}) = \frac{\{2c\mathsf{f}(c) - \psi_c\}^2 + 2\tau_2^c\{2c\mathsf{f}(c) - \psi_c\} + \tau_2^c}{\psi_c^2}(\widehat{\sigma}_c^{(1)})^2(\widehat{\Sigma}_c^{(1)})^{-1}.$$

*In addition, when $m \to \infty$ we have*

$$n^{1/2}(\widehat{\beta}_c^{(*)} - \widetilde{\beta}) \overset{a}{\sim} \mathsf{N}\{0_{d_x}, \widehat{\mathsf{avar}}(\widehat{\beta}_c^{(*)} - \widetilde{\beta})\},$$

*and*

$$\widehat{H}_{n,c}^{(*)} = n(\widehat{\beta}_c^{(*)} - \widetilde{\beta})' \widehat{\mathsf{avar}}(\widehat{\beta}_c^{(*)} - \widetilde{\beta})^{-1}(\widehat{\beta}_c^{(*)} - \widetilde{\beta}) \overset{a}{\sim} \chi^2_{d_x},$$

*where*

$$\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(*)} - \widetilde{\beta}) = \frac{\{2c\mathsf{f}(c) - \psi_c\}^2 + 2\tau_2^c\{2c\mathsf{f}(c) - \psi_c\} + \tau_2^c}{\{2c\mathsf{f}(c) - \psi_c\}^2}(\widehat{\sigma}_c^{(*)})^2(\widehat{\Sigma}_c^{(*)})^{-1}.^4$$

Notice that the density $\mathsf{f}$ of the errors $\varepsilon_i$ enters in the asymptotic variance of the difference between RLS/IIS and OLS, thus our proposed test relies on a reference distribution as the target of outlier-removal. A targeted reference distribution is a common regularity condition in the outlier detection literature and follows from the idea of $\epsilon$-contamination initially proposed by Huber (1964) and recently re-investigated by Johansen & Nielsen (2016a). Under Huber's framework, data come from $(1-\epsilon)\mathsf{f} + \epsilon\mathsf{f}^{\mathsf{c}}$, so outliers have to be defined relative to a reference distribution $\mathsf{f}$. Numerous empirical studies directly assume a form of the density $\mathsf{f}$ to select the cut-off value $c$ for conducting outlier robustness checks. For example, Acemoglu et al. (2019) choose $c = 1.96$ by implicitly imposing the standard normal for the density $\mathsf{f}$ to compute the RLS estimator when assessing the impact of democratisation on economic growth. In addition, nearly all of the IIS applications mentioned in the introduction (from wages to hurricane damages) assume a normal reference distribution of the error term. Further, the assumption of a reference distribution is also made in the majority of theory papers studying IIS, including: Hendry, Johansen and Santos (2008), Johansen and Nielsen (2009, 2016), or Berenguer-Rico and Wilms (2021). Thus, the distributional assumption on $\mathsf{f}$ is commonly-made in practise and should not limit the applicability of our test. The assumption is also potentially testable by a specification test proposed by Berenguer-Rico & Nielsen (2018) and Jiao & Pretis (2022).

*A note on heuristic tests:* In absence of asymptotic theory, some existing studies have used the difference between robust and OLS estimates but incorrectly replaced the standard error of the difference between two estimators $\widehat{\beta}_c^{(m+1)}$ and $\widetilde{\beta}$ with the standard error of the marginal distribution of the baseline estimator $\widetilde{\beta}$ when constructing the proposed test statistic. The heuristic test statistic is then compared to the critical value either drawn from the standard normal for the two-sided test or from the chi-square with $d_x$ degree of freedom for the one-sided test. The size of this 'heuristic' method of assessing outlier robustness does not converge to the nominal level under the null and is also likely to exhibit low statistical power under alternatives. Thus, such a heuristic approach is invalid and likely uninformative in practise. We instead recommend using our above proposed formal test.

### 2.4.1 Bootstrap Outlier Distortion Tests

Our proposed outlier distortion test relies on asymptotic results as well as a reference distribution against which outliers are evaluated. To improve the finite sample performance of the test and to explore robustness against an incorrectly-specified reference distribution, we consider two bootstrap versions using a range of different resampling schemes based on existing approaches to bootstrapping Hausman-type test statistics. The test statistic we are interested in is motivated by our asymptotic test:

$$\widehat{H}_{n,c}^{(1)} = n(\widehat{\beta}_c^{(1)} - \widetilde{\beta})'\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(1)} - \widetilde{\beta})^{-1}(\widehat{\beta}_c^{(1)} - \widetilde{\beta}). \tag{2.13}$$

Since we are interested in the difference between robust and OLS estimates, as a first bootstrap approach we can directly bootstrap the L2 norm of the difference between coefficients $\|\widehat{\beta}_c^{(1)} - \widetilde{\beta}\|_2$

[4] If we assume $\mathsf{f} \overset{\mathsf{D}}{=} \mathsf{N}(0,1)$, then $\tau_2^c = \psi_c - 2c\mathsf{f}(c)$ so that $\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(1)} - \widetilde{\beta})$ and $\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(*)} - \widetilde{\beta})$ can be further simplified.

mirroring the approach for the winsorized estimator in Kaji (2018). We refer to this approach as the L2 bootstrap. Second, the only quantity in the test statistic of interest (2.13) that depends on the reference distribution $\mathsf{f}$ is the variance term. We can thus use the bootstrap to estimate the variance $\mathsf{var}(\widehat{\beta}_c^{(1)} - \widetilde{\beta})$ instead of using the estimate of the asymptotic variance in Corollary 2.10. Replacing the asymptotic variance with the bootstrap estimate, the Hausman-type test statistic for outlier distortion should then still follow the limiting $\chi^2$ distribution. This approach is inspired by the bootstrap method for the conventional Hausman test proposed in Cameron & Trivedi (2005) & (2010) and we refer to it as the 'variance bootstrap'. As a complete analysis of the bootstrap is beyond the scope of this paper, we investigate the performance of the two bootstrap schemes in a range of simulations.

There is a range of different options on how we generate the bootstrap samples for either the L2 or the variance bootstrap. First, we consider non-parametric case re-sampling from the raw data similar to Kaji (2018). When using just the L2 norm, this approach likely suffers from low power due to the risk of the inclusion of outliers in each bootstrap draw - a concern also highlighted by Singh (1998) and Salibián-Barrera et al. (2008). However, when the reference distribution is incorrect this approach has the potential to correct some of the size-distortions. Second, as Singh (1998) and Salibián-Barrera et al. (2008) recommend, we consider re-sampling from the outlier-removed (cleaned) data. Re-sampling from the outlier-removed data is also in-line with the recommendation by Davidson & MacKinnon (1999) to impose the null in bootstrap samples in order to minimize the probability of type I errors. When the reference distribution is correct, sampling from the cleaned, outlier-removed data likely exhibits high power under the alternative. However, when the reference distribution is incorrect this approach is likely to over-reject as the outlier-removed data is made to resemble the specified reference distribution.

We consider both non-parametric case re-sampling as well as a parametric residual bootstrap in our simulations. For the time series setting, we consider both a parametric bootstrap as well as a non-parametric block bootstrap. For the parametric time series bootstrap we resample from the residuals obtained by using the robust estimate and construct the bootstrap sample under the robustly-estimated auto-regressive structure (Davison & Hinkley, 1997). For the non-parametric time series bootstrap we consider block resampling with a fixed block size (for a given sample size) – see e.g. Bühlmann & Künsch (1999). We describe all our bootstrap algorithms in detail in Appendix B.

In summary, we consider two different test statistics to be estimated using the bootstrap: the L2 norm similar to Kaji (2018), and the variance as in Cameron & Trivedi (2010). For each of these two we consider non-parametric case resampling from the raw as well as cleaned data, and a parametric residual bootstrap for both iid and a time series auto-regressive model.

Future work could consider finite-sample improvements by sampling from the reference distribution instead of the residuals, or re-sampling from the truncated residuals as in the bootstrap Hausman-type test for jumps in Dovonon et al. (2019).

## 3 Finite Sample Performance using Simulations

Here we study the performance of the proposed asymptotic outlier distortion test in a series of simulation experiments under the null hypothesis of no distortion, as well as under a range of

alternatives. We also simulate the performance of the bootstrap versions of the test under the null and alternatives. We simulate the DGP in (3.1) under the null of no outliers (and thus no distortion) as

$$y_i = x_i'\beta + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{3.1}$$

with $\beta = 0.5$ varying the sample size $n$ (from $n = 100$ to $n = 500$), the number of regressors whose coefficients are tested for distortion $d_x$ ($d_x = 1$, 5, 10), the degree of persistence in the dependent variable (ranging from iid to a stationary autoregressive process by considering a time series $y_t = \rho y_{t-1} + x_t'\beta + \varepsilon_t$ with $\rho = 0.5$), as well as the underlying reference distribution of the errors $\varepsilon_i$, drawn from either $\varepsilon_i \sim N(0, \sigma^2 = 1)$ or $\varepsilon_i \sim t(3)$, with results for an asymmetric, lognormal, error distribution reported in the Appendix. To simulate the performance under a range of alternatives, we simulate two sets of various outlier-contaminated DGPs. In the first set we look at vertical outliers with the DGP given as

$$y_i = x_i'\beta + u_i, \quad i = 1, 2, \ldots, n, \tag{3.2}$$

where $u_i = \varepsilon_i + \lambda 1_{\{i \in O\}}$ and $O$ is the set of $n_o$ randomly chosen outlier locations (anywhere between $i = 1, ..., n$) where we vary the degree of outlier contamination (the proportion of randomly located outliers $n_o/n$ ranging from 10%-15% of the sample) as well as outlier magnitudes $\lambda$ from $\lambda = 2\sigma$ to $\lambda = 6\sigma$. In the second set, we assess the performance of the test in the presence of bad leverage points (following the setup of Dehon et al. 2012)[5] where $y_i$ is generated using (3.2) with $u_i \sim N(0, 1)$, however, the model is estimated using $x_i^* = x_i + \lambda 1_{\{i \in O\}}$ contaminated with outliers of magnitude $\lambda = 2\sigma$ to $\lambda = 6\sigma$.

For each draw we construct our proposed outlier distortion test statistic and record the rejection at 5% or 1%. Simulations are implemented using the R-package `gets` using IIS with cut-offs $c = 1.96$ and $c = 2.57$, with $M = 10,000$ replications for asymptotic tests and $M = 500$ to $1,000$ replications and 499 bootstrap draws for the bootstrap versions of the test.[6]

### 3.1 Simulation Performance under the Null of No Distortion

Simulation results for the asymptotic test under the null of no outliers are shown in Figure 3.1 (with full results provided in tables in the online Appendix). The asymptotic test appears slightly over-sized for small samples ($n < 200$) but exhibits size close to the nominal level in larger samples ($n > 200$). Size appears unaffected by the threshold $c$ used to classify outliers (left panel in 3.1), the number of coefficients $d_x$ tested for distortion (middle panel in Figure 3.1), or whether the process is iid or stationary autoregressive (right panel in Figure 3.1).

Bootstrap results for the non-parametric bootstrap are shown in Figure 3.2, with parametric bootstrap results provided in the Appendix. For the bootstrap we consider cases where the reference distribution is correctly- as well as incorrectly-specified. For the case of an incorrect reference distribution we assume a normal reference when the DGP errors in fact follow a fatter-tailed t-

[5] Where similar to Dehon et al. (2012) we consider a single regressor $x_i$, ($d_x = 1$) with coefficient $\beta = 1$, contaminated with 5% outliers acting as bad leverage points – i.e. outliers in the x-dimension.

[6] We use 500 Monte Carlo replications for the bootstrap under alternatives and 1000 replications under the null. A smaller number of replications is used for the bootstrap versions of the test due to the computational complexity of the bootstrap.

distribution with three degrees of freedom. We also show the results of the asymptotic test for comparison.

As expected, bootstrapping the L2 norm re-sampling the raw data is drastically under-sized regardless of the reference distribution. Bootstrapping the L2 norm using the cleaned, outlier-removed data yields size improvements relative to the asymptotic test in small samples. However, this approach is not robust to mis-specification of the reference distribution. Similar to the asymptotic test in this setting, bootstrapping the L2 norm using cleaned data is over-sized when the reference distribution is incorrect. Notably, the parametric bootstrap of the L2 norm does not appear to perform better than the asymptotic test (see Appendix).

In turn, the non-parametric variance bootstrap using the raw data performs remarkably well.[7] The size is close to the nominal level, even when the reference distribution is different from the true (albeit symmetric) error distribution, both for time series as well as iid data.[8]

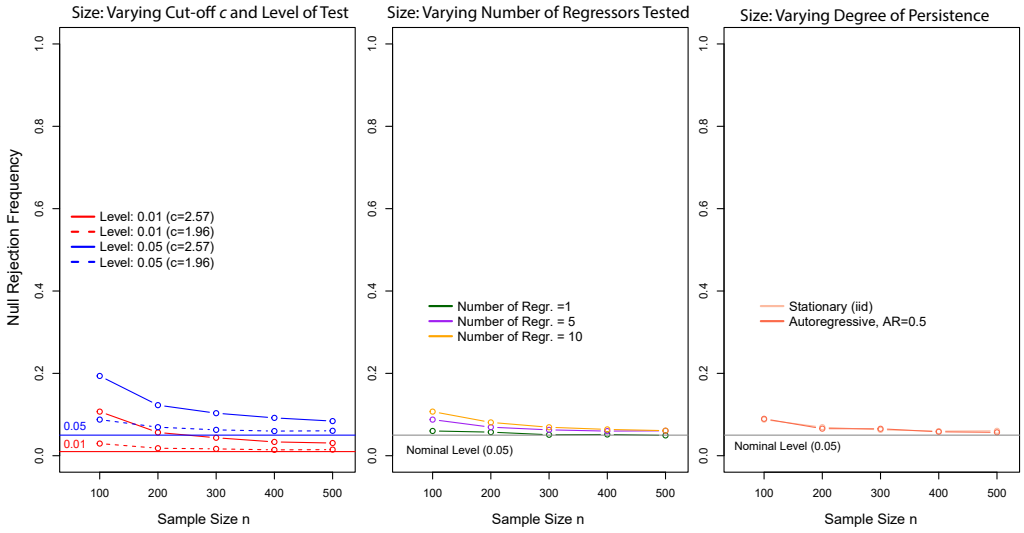


Figure 3.1: Simulation performance of the asymptotic test under the null of no distortion for varying sample sizes, cut-offs, test levels and number of regressors in the DGP for iid and time series regressions.

## 3.2 Simulation Performance in the Presence of Distortion

We assess the performance of the asymptotic test in the presence of outlier distortion by introducing outliers of varying magnitudes to the DGP (both as vertical outliers as well as bad leverage points), varying the sample size and degree of autocorrelation in the dependent variable. We plot the null-rejection frequencies for different simulation specifications using vertical outliers in Figures 3.3 for

[7] The variance bootstrap re-sampling the cleaned data leads to over-rejection potentially due to the variance being under-estimated in the outlier-removed data – see Appendix for full results.

[8] This seems to hold for the case of a mis-specified but symmetric error distribution. For an asymmetric error distribution we consider the case where the reference distribution is assumed to be normal, when in fact the errors are drawn from an asymmetric, lognormal distribution. In this setting, the L2 bootstrap appears to have zero power while the variance bootstrap is over-sized (see simulation Tables in the Appendix).

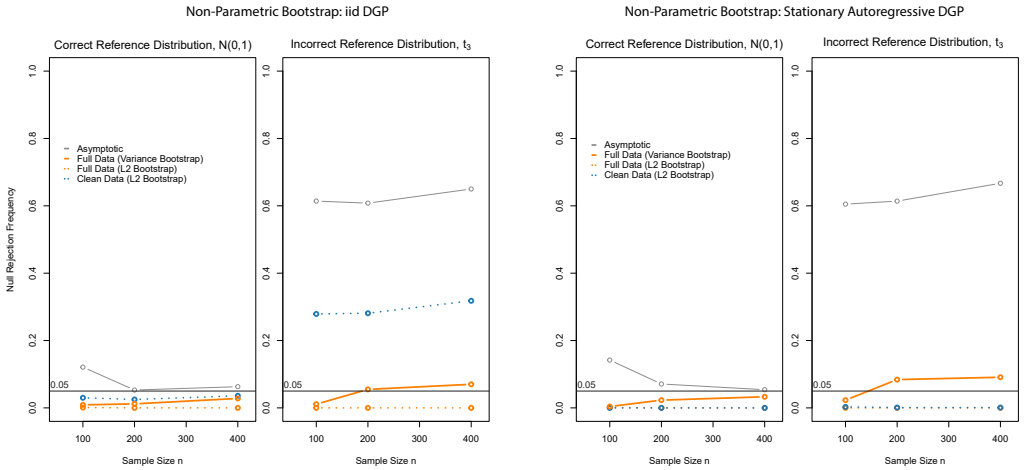


Figure 3.2: Simulation performance of the non-parametric bootstrap tests under the null of no distortion (with $c = 1.96$) for varying sample sizes when the reference distribution does and does not match the error distribution in the DGP. Left panels show an iid process, right panels show a stationary autoregressive process using a non-parametric block bootstrap.

a stationary (iid) DGP and an autoregressive DGP in Figure 3.4. Results for simulations using bad leverage points are plotted in Figure 3.5. To capture the degree of distortion we also plot the null rejection frequency against the euclidian distance of all estimated (OLS) coefficients from their true underlying DGP counterparts, scaled by the maximum distance (right panels in Figures). The power of the test increases with the sample size for both vertical outliers and bad leverage points, regardless of whether $y$ is iid or follows a stationary autoregressive process. Power also generally increases with the degree of distortion – the larger the difference between OLS and robust estimates, the more likely we are to reject the null hypothesis of no distortion. Additional simulation results with varying proportions of outlier contamination and varying degrees of persistence in $y$ are also provided in the appendix.

Bootstrap results under the alternatives for vertical outliers are shown in Figure 3.6, with bad leverage point results provided in the appendix. Beginning with the L2 bootstrap, as expected, bootstrapping using the raw data appears to have zero power as each bootstrap draw has a chance of including some outlying observations, thus rendering the resulting bootstrap distribution uninformative when compared to the original test statistic. In turn, bootstrapping the L2 norm using clean (outlier-removed) data shows desirable power properties: power increases with the sample size as well as with the degree of outlier distortion. The variance bootstrap exhibits good power properties when using the raw data, even when the reference distribution is incorrectly assumed to be normal and the true errors follow a t-distribution.

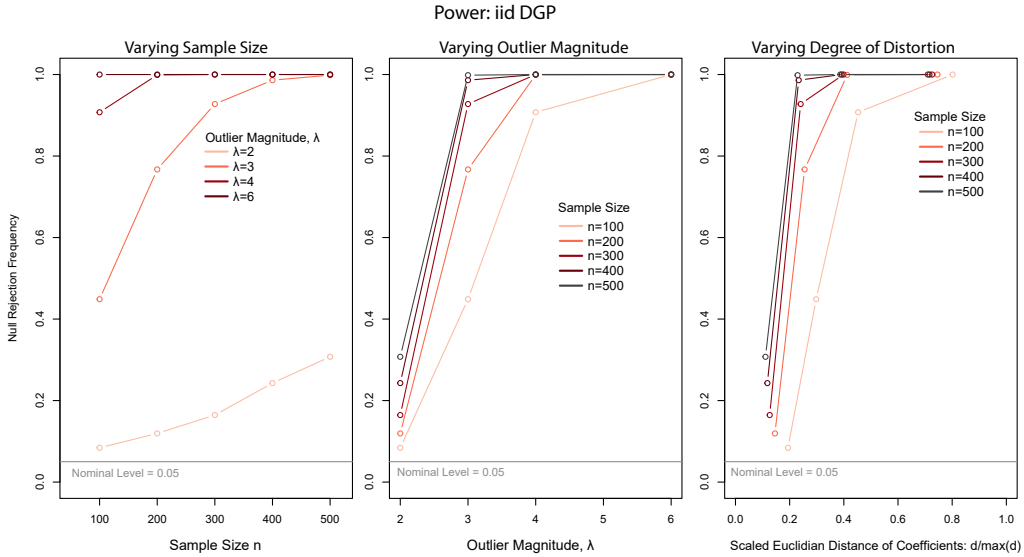


Figure 3.3: Simulation performance of the asymptotic test under alternatives contaminated with vertical outliers for varying sample sizes and outlier magnitudes when the DGP is iid ($\rho = 0$) including five regressors and 10% of the sample is outlier-contaminated.

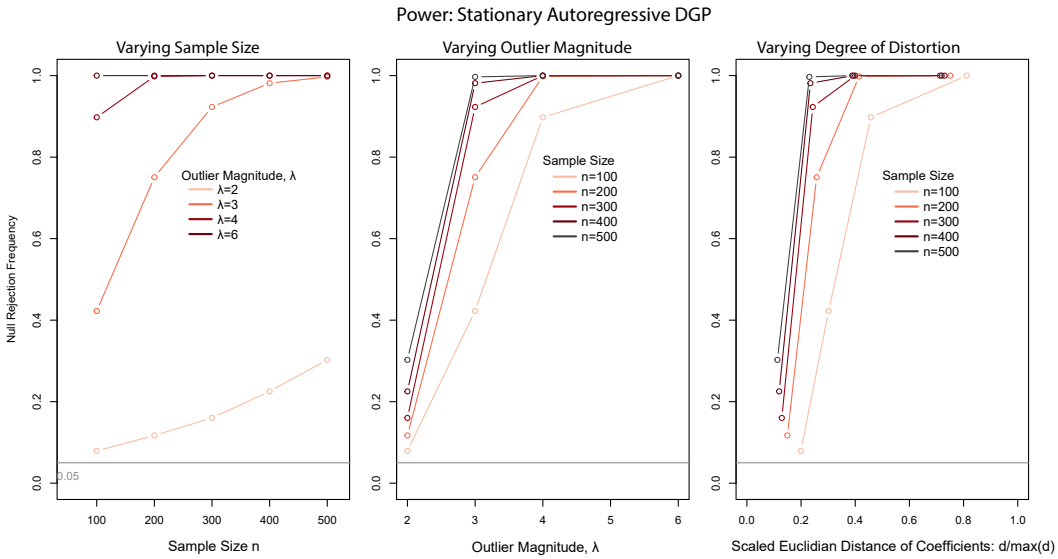


Figure 3.4: Simulation performance of the asymptotic test under alternatives contaminated with vertical outliers for varying sample sizes and outlier magnitudes when the DGP is a stationary autoregressive process, contains five regressors and 10% of the sample is outlier-contaminated.

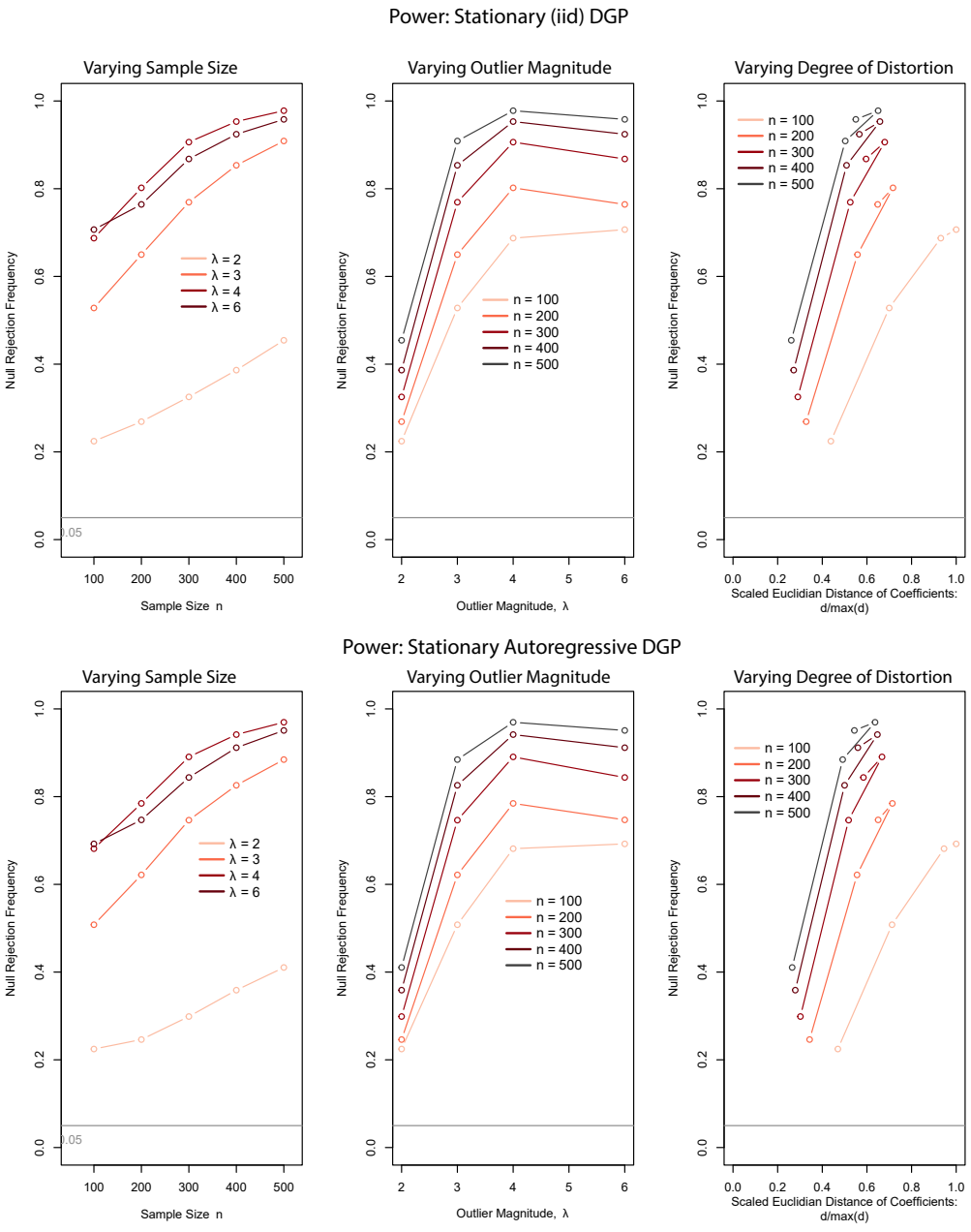


Figure 3.5: Simulation performance of the asymptotic test under alternatives contaminated with bad leverage points for varying sample sizes and outlier magnitudes when the DGP contains five regressors and 10% of the sample is outlier-contaminated.

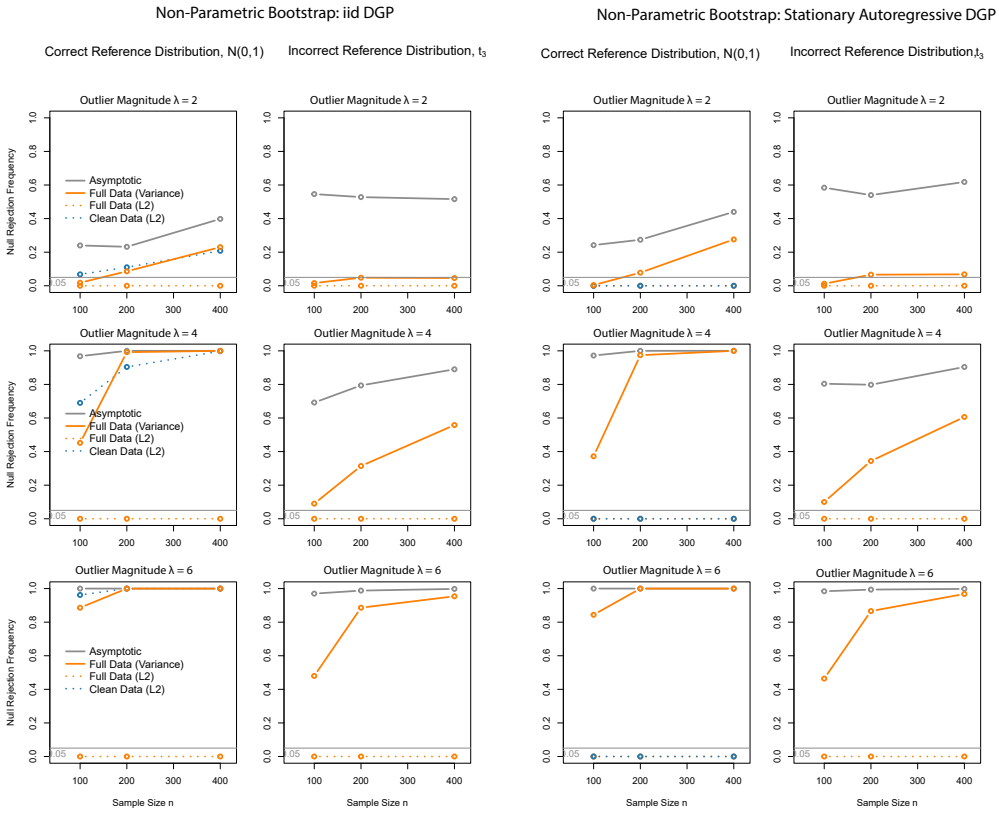


Figure 3.6: Simulation performance of the non-parametric bootstrap tests under the alternatives contamined with vertical outliers for varying sample sizes when the reference distribution does and does not match the error distribution in the DGP (for outlier magnitudes of 2 SD, 4 SD, and 6 SD of the error term and 10% outlier proportion of the sample).

### 3.3 Summary of Simulation Results

The simulation results show well-controlled size and high power of our asymptotic test in large samples (n > 200), with the tests being slightly oversized in small samples (n < 200). In line with our theory, these results hold for both iid as well as stationary autoregressions. The tests exhibit high power for both vertical outliers as well as bad leverage points.

Bootstrapping the L2 norm using the cleaned (outlier-removed) data can improve the size of the test for finite samples, but is not robust to mis-specification of the reference distribution. However, bootstrapping the variance using non-parametric resampling of the raw data appears to yield good size and power even when the reference distribution is incorrectly-specified (albeit symmetric). Thus, if the only concern is over-rejection in finite samples, then the L2 norm can be bootstrapped using a non-parametric bootstrap drawing from the outlier-removed data. If there are concerns around both over-rejection in small samples and the specification of the reference distribution, we recommend using the non-parametric variance bootstrap. Future work will explore these bootstrap tests further.

## 4 Application: Climate Adaptation & the Macro-Economic Impacts of Temperatures

We apply the proposed outlier distortion test to estimates of the economic impacts of climate change. Modelling GDP per capita growth in fixed effects panels as a non-linear (quadratic) function of annual-average temperatures has become common practise in the climate econometric literature. A concern with such global panel-econometric models of climate impacts (e.g. Burke et al., 2015; Dell et al., 2012; Pretis, Schwarz, et al., 2018) is the inability to account for much observed variation in economic growth. Even though conventional panel climate impact models control for time fixed effects, country fixed effects, and country-specific (non-linear) time trends, there are potentially numerous un-modelled idiosyncratic shocks. Such shocks – possibly taking the form of outliers in the model – can erroneously be attributed to climate variation and thus distort the estimated coefficients on climate variables. We therefore use the robust IIS estimator to estimate a panel model of the effects of temperature on GDP per capita growth using a global panel dataset and test for outlier distortion.

First, we estimate a model relating GDP per capita growth to temperatures replicating the specification in Burke et al. (2015) with updated data and compare OLS estimates to those obtained using IIS. Second, there exists substantial evidence of adaptation to climate change globally (Berrang-Ford et al., 2021) as well as regionally (Aguiar et al., 2018; Chen & Gong, 2021) and has been considered empirically e.g. in studies on mortality (Carleton et al., 2020; Deschênes & Greenstone, 2011; Barreca et al., 2016), agriculture (Burke & Emerick, 2016), and electricity consumption (Deschênes & Greenstone, 2011; Auffhammer, 2022). Nonetheless, adaptation to climate change has received considerably less attention in global empirical macro-economic panel analyses of economic growth. Higher incomes might mitigate the impacts of climate change (see e.g. Acevedo et al., 2020), and such adaptation might bias existing historical estimates and subsequent future projections of macro-economic climate impacts, as considered by Schwarz & Pretis (2020) and Kahn et al. (2021). We therefore estimate a panel model of climate impacts while allowing for

adaptation driven by incomes, specified through interaction terms of temperatures and incomes.

### 4.1 Data & Model

Following the macro-econometric literature studying the temperature effects on economic growth, we model the year-on-year change in the log of real GDP per capita (World Development Indicators, World Bank, 2019) from 1961-2017 for 169 countries as a function of population-weighted climate variables. We use historical climate observations on temperatures and precipitation from (Matsuura & Willmott, 2018, version 5) covering global land areas at a 0.5° spatial resolution. To map gridded climate observations to individual countries in each year, we weight climate observations by gridded population data (CIESIN, 2016) at the same spatial resolution. Population-weighting (rather than area-weighting) is more likely to capture the effects of climate onto socio-economic activity (see Tol, 2017).

We first replicate existing results from Burke et al. (2015) and compare these to using a robust estimator. We then expand on these existing models by allowing for climate adaptation at different income levels. Specifically, our model allows for adaptation by potentially attenuating or amplifying the coefficients on temperatures through interaction terms with (lagged) log of GDP per capita. Dell et al. (2012) estimate a simplified version of such a specification by interacting a linear temperature variable with a dummy variable for poor countries, and Burke et al. (2015) explore the interaction of a linear temperature variable with country-level incomes. Here we generalise this to non-linear temperature impacts interacted with the continuum of per capita incomes. A similar approach of estimating income-based adaptation has also been applied for heat-related mortality in Carleton et al. (2020) and Schwarz & Pretis (2020) for climate impacts of extremes. Our base model replicating Burke et al. (2015) is given in equation (4.1), and our adaptation model (we report estimates using lagged incomes in the appendix) is given in equation (4.2):

$$\Delta log(y)_{i,t} = \alpha_i + \lambda_t + \beta_1 T_{i,t} + \beta_2 T_{i,t}^2 + x'_{i,t}\gamma + u_{i,t}, \tag{4.1}$$

$$\Delta log(y)_{i,t} = \alpha_i + \lambda_t + \beta_1 T_{i,t} + \beta_2 T_{i,t}^2 + \beta_3 \left[T_{i,t} \times log(y)_{i,t}\right] + \beta_4 \left[T_{i,t}^2 \times log(y)_{i,t}\right] + x'_{i,t}\gamma + u_{i,t}. \tag{4.2}$$

where $\Delta log(y)_{i,t}$ is the year-on-year change in per capita income in country $i$ in year $t$, $\alpha_i$ denote country fixed effects and $\lambda_t$ are year fixed effects. The coefficients of interest are $\beta_1$ through $\beta_4$. Specifically, $\beta_1$ and $\beta_2$ capture the (potentially non-linear) impacts on GDP per capita growth while coefficients $\beta_3$, $\beta_4$ in (4.2) on income-interactions terms allow for the temperature-growth relationship to vary by country-specific income levels. The vector $x_{i,t}$ denotes a set of control variables including population-weighted annual average precipitation (and its square), and country-specific linear and quadratic trends as in Burke et al. (2015) and Pretis, Schwarz, et al. (2018).

To assess the impact of unknown idiosyncratic shocks, we estimate equations (4.1) and (4.2) using OLS (as is convention in the literature) and then compare the OLS estimates to those obtained when using the robust IIS estimator. Let $\widetilde{\beta}$ denote the vector of coefficients estimated by OLS $\widetilde{\beta} = (\widetilde{\beta}_1, \widetilde{\beta}_2, \widetilde{\beta}_3, \widetilde{\beta}_4)$, and $\widehat{\beta}_c^{(1)}$ denote the vector of coefficients estimated using the robust IIS estimator. The cut-off value $c$ in IIS determines the expected false positive rate of detected outliers. We set the cut-off to remove outlying observations to $\gamma_{c=2.57} = 0.01$ relative to a normal reference distribution `f`. This leads to an expected false-positive rate of 1% of observations to be spuriously labelled as

outlying for a well-specified model which is common in the IIS literature (where often a range from 5% to 0.1% is considered).[9] The hypothesis of interest is $H_0 : \widetilde{\beta} = \widehat{\beta}_c^{(1)}$. In other words, we assess possible outlier distortion in the estimated coefficients linking annual average temperatures and GDP per capita growth.

As a robustness check, we vary the cutoff $c$ to specify a target level of significance (and thus expected false positive rate of outliers) from 5% ($c = 1.96$) to 0.1% ($c = 3.29$) and we also present a model using temperatures interacted with $log(y)_{i,t-1}$ in the Appendix (see equation (D.1)).

### 4.2 Results

Estimating the above models linking GDP per capita growth to temperatures using IIS results in 165 detected outliers for the base model (4.1) and 170 ouliers for the adaptation model (4.2) respectively. The number of detected outliers greatly exceeds the number expected by chance ($0.01 \times 7716 = 77.16$). The observed proportion of outliers $\hat{\gamma}_{c=2.57} = 165/7716 = 0.021$ and $\hat{\gamma}_{c=2.57} = 170/7716 = 0.022$ is statistically different from the expected proportion of $\gamma_{c=2.57} = 0.01$ ($p < 0.001$ using the Jiao & Pretis (2022) proportion test comparing 0.021 and 0.022 to 0.01). The distribution of outliers does not appear to be random over time and space and is indicative of possible model mis-specification for some countries. Few outliers are detected for OECD countries, instead they are concentrated in developing regions and economies in transition (see Figures 4.1 and 4.2).

Table 4.1 and Figure 4.4 show the estimated coefficients on temperature variables using OLS and the robust IIS approach.[10] The robust coefficient estimates are attenuated relative to the OLS estimates for both the base model replicating Burke et al., as well as our adaptation model. In other words, the impacts of temperature on GDP per capita growth are dampened when controlling for potentially-outlying observations. To formally compare robust and OLS estimates we construct our test statistic for outlier distortion in the coefficients of interest as:

$$\widehat{H}_{n,c}^{(1)} = n(\widehat{\beta}_c^{(1)} - \widetilde{\beta})' \widehat{\mathsf{avar}}(\widehat{\beta}_c^{(1)} - \widetilde{\beta})^{-1} (\widehat{\beta}_c^{(1)} - \widetilde{\beta}), \tag{4.3}$$

where $c = 2.57$ for a normal reference distribution. The test statistic for each coefficient and all temperature coefficients of interest are shown in Table 4.1 (third and sixth column). We compare the resulting test statistic $\widehat{H}_{n,c}^{(1)} = 111.27$ for our base model against the critical value of the $\chi^2$ distribution with $df = 2$ and for our adaptation model with $\widehat{H}_{n,c}^{(1)} = 770.69$ against the critical value of the $\chi^2$ distribution with $df = 4$, as we are testing for the distortion of two and four coefficients respectively. The null hypothesis of no distortion is rejected at any conventional significance level for both models, with $p < 0.0001$. Further, all coefficient estimates are also statistically different from their robust counterparts when testing each coefficient individually. These results are highly robust to varying the cut-off used to classify outliers (see Figure 4.4 and the Appendix for full estimation results).

[9] See e.g. Jiao & Pretis (2022) for additional discussion and simulation evidence on the expected false positive rate in IIS.

[10] Note that the coefficient on temperatures and temperatures squared in the adaptation model have to be interpreted taking the interaction into account. The relationship between temperatures and GDP per capita growth does not switch signs in the adaptation model as the interaction terms have to be factored in. The resulting non-linear relationships still follow the familiar inverted-U shape even under adaptation as shown in Figure 4.3.

Beyond demonstrating the presence of outlier-distortion, our results show significant evidence of income-driven adaptation to temperatures (Figures 4.3 and 4.4). As countries become richer, they exhibit a different response function to temperatures compared to poor and middle-income countries (see Figure 4.3 for the estimated non-linear relationship at different income percentiles). This is apparent by the coefficients on temperature-income interactions having the opposite sign to the coefficients on temperatures alone. This suggests that as incomes increase, countries may be better able to cope with the impacts of higher temperatures. Such a dynamic could further exacerbate existing cross-country inequality with continued climate change. However – similar to the base model – controlling for outlying observations dampens the impacts of temperatures onto economic growth across all income levels. Notably, controlling for outliers effectively removes the positive impacts associated with higher temperatures for countries with low annual average temperatures (e.g. countries in the Northern Hemisphere). The coefficients on squared temperatures (with and without income interactions) are not statistically different from zero, casting some doubts on the quadratic relationship between growth and temperatures.

Table 4.1: OLS and IIS Panel Regression Results together with their difference in coefficients and the resulting outlier distortion test statistic. We detect ouliers using a cut-off $c = 2.57$ with an expected false positive rate of 1% for a Normal reference distribution. Coefficients on control variables are omitted.

| | Base | Base IIS | Base Outlier Distortion Test | Adaptation | Adaptation IIS | Adaptation Outlier Distortion Test |
|---|---|---|---|---|---|---|
| Temperature | 0.01734*** | 0.01032*** | 108.95 | -0.06224*** | -0.03384*** | 186.25 |
| | (0.00348) | (0.00233) | [<0.001] | (0.01041) | (0.0072) | [<0.001] |
| Temperature$^2$ | -0.00059*** | -0.00039*** | 99.05 | 0.00070 | 0.00001 | 88.57 |
| | (0.0001) | (0.00007) | [<0.001] | (0.00037) | (0.00026) | [<0.001] |
| Precipitation | 0.00043 | 0.00073 | 1.98 | 0.01018 | 0.01328*** | 7.86 |
| | (0.00111) | (0.00074) | [0.160] | (0.00563) | (0.00383) | [0.005] |
| Precipitation$^2$ | -0.00004 | -0.00004 | 0.46 | -0.00019 | -0.00035** | 17.55 |
| | (0.00003) | (0.00002) | [0.496] | (0.00019) | (0.00013) | [<0.001] |
| Temperature x $GDP_{pc}$ | | | | 0.00811*** | 0.00416*** | 317.59 |
| | | | | (0.00111) | (0.00077) | [<0.001] |
| Temperature$^2$ x $GDP_{pc}$ | | | | -0.00012** | -0.00002 | 148.17 |
| | | | | (0.00004) | (0.00003) | [<0.001] |
| Precipitation x $GDP_{pc}$ | | | | -0.00121 | -0.00155*** | 6.86 |
| | | | | (0.00066) | (0.00045) | [0.009] |
| Precipitation$^2$ x $GDP_{pc}$ | | | | 0.00002 | 0.00004* | 17.52 |
| | | | | (0.00002) | (0.00002) | [<0.001] |
| Num. Outliers | | | 165 | | | 170 |
| Outlier Distortion test statistic for Temp. Variables | | | $\chi^2_2 = 111.27$ [<0.001] | | | $\chi^2_4 = 770.69$ [<0.001] |
| Num.Obs. | 7716 | 7716 | | 7716 | 7716 | |
| BIC | -18483.2 | -23533.1 | | -18801.4 | -23776.0 | |
| Log.Lik. | 11774.742 | 15038.149 | | 11951.758 | 15199.897 | |
| Fixed Effects | Country & Year | Country & Year | | Country & Year | Country & Year | |

* p < 0.05, ** p < 0.01, *** p < 0.001
(Standard Errors) and [p-values]
Please note that the estimated coefficients on Precipitation$^2$ by OLS and IIS are very close but not exactly equal in the base model. Thus, its outlier distortion test statistics is not zero.

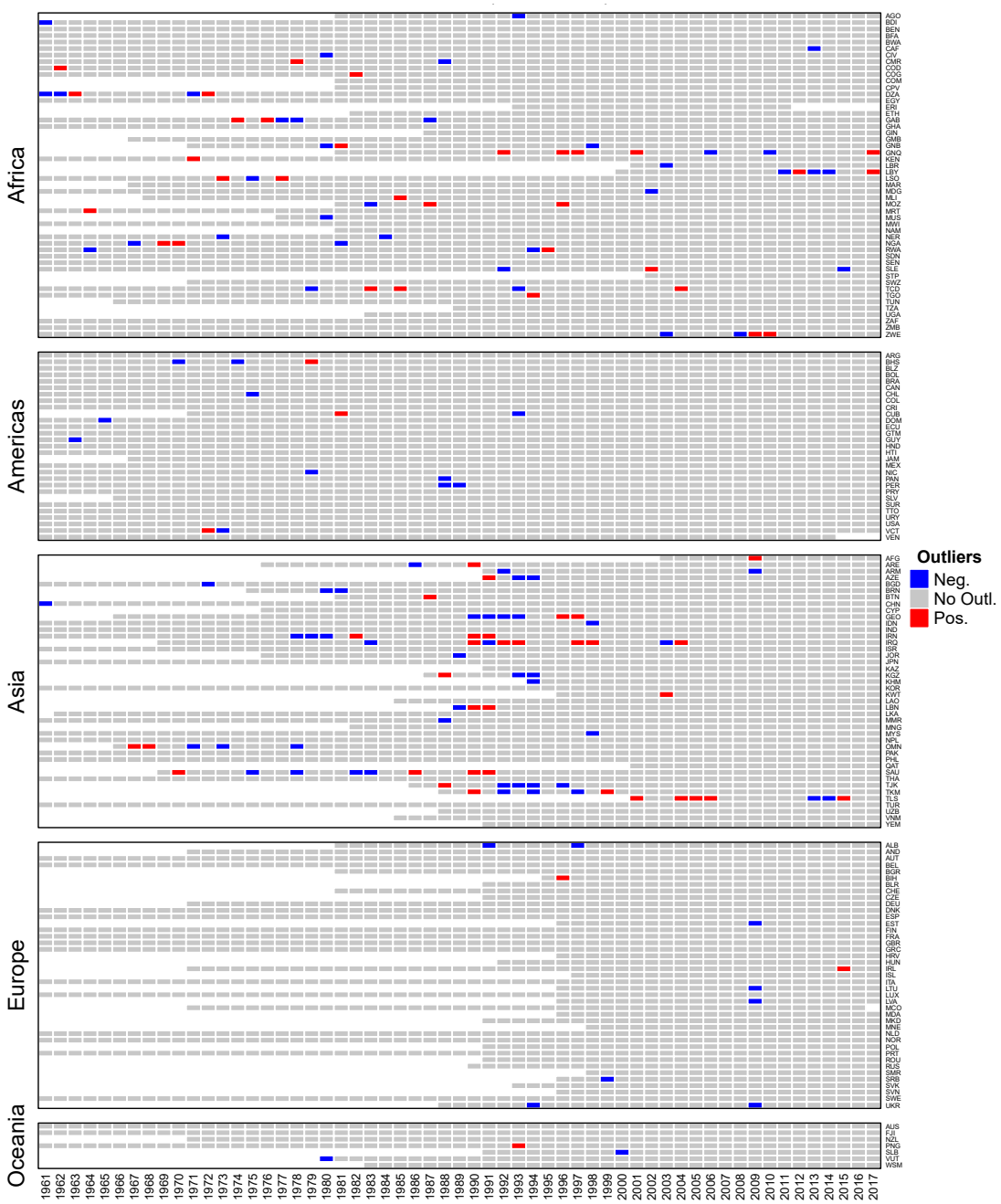


Figure 4.1: Detected outliers using the robust IIS estimator across countries and time in the global cross-country panel from 1961-2017 in the adaptation model (for a cut-off $c = 2.57$ with an expected false positive rate of 1% for a Normal reference distribution). The figure shows country-year observations as gray when not outlying, blue when there is a negative outliers, and red for positive outliers.

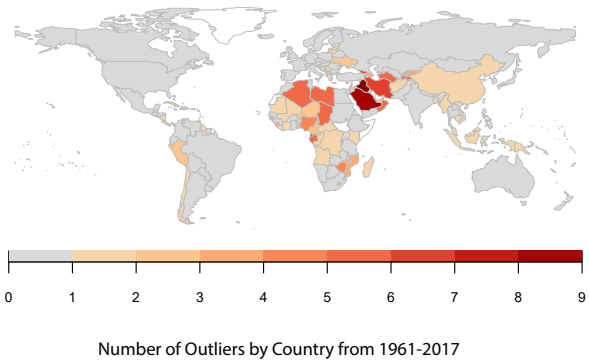


Figure 4.2: Detected outliers aggregated over the full sample of 1961 - 2017 by country in the panel in the adaptation model (for a cut-off $c = 2.57$ with an expected false positive rate of 1% for a Normal reference distribution). Gray denotes no outliers detected.

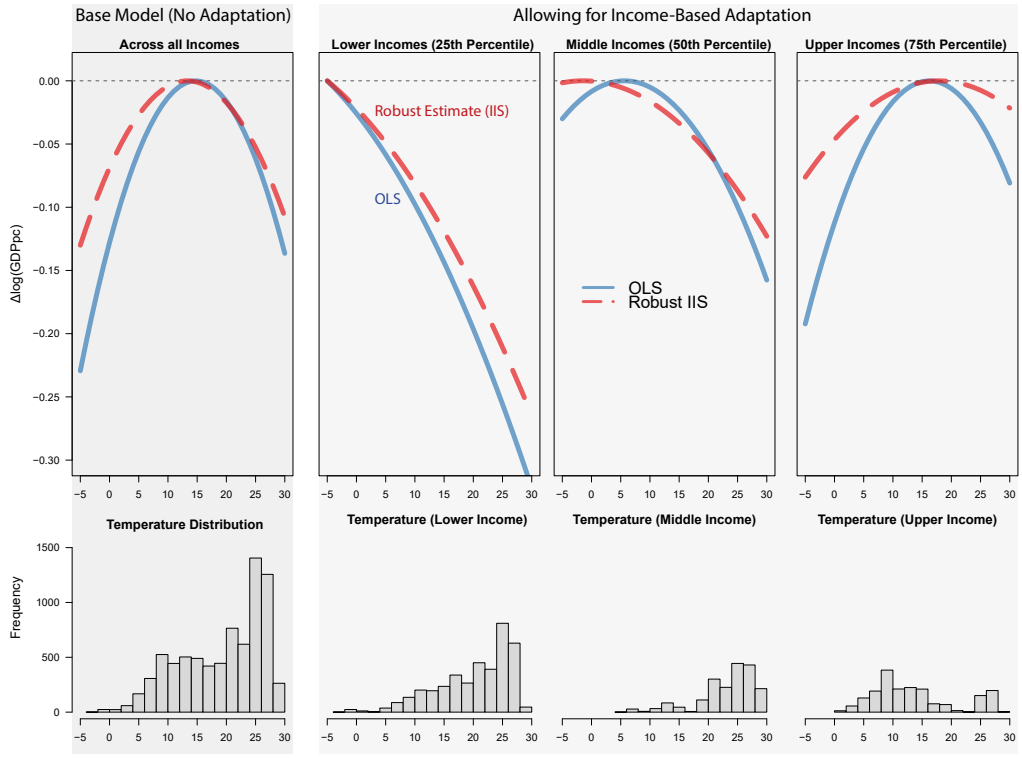


Figure 4.3: Estimated Impact of Temperatures on GDP Per Capita Growth Allowing for Adaptation and Controlling for Outliers. OLS-estimated relationship shown in blue, robust IIS estimated relationship shown in red (for a cut-off $c = 2.57$ with an expected false positive rate of 1% for a Normal reference distribution). Estimated non-linear impact function for the overall base model (top left) and at three different income levels (other top panels). Observed temperatures for the entire sample and across income ranges are shown in the bottom panel.

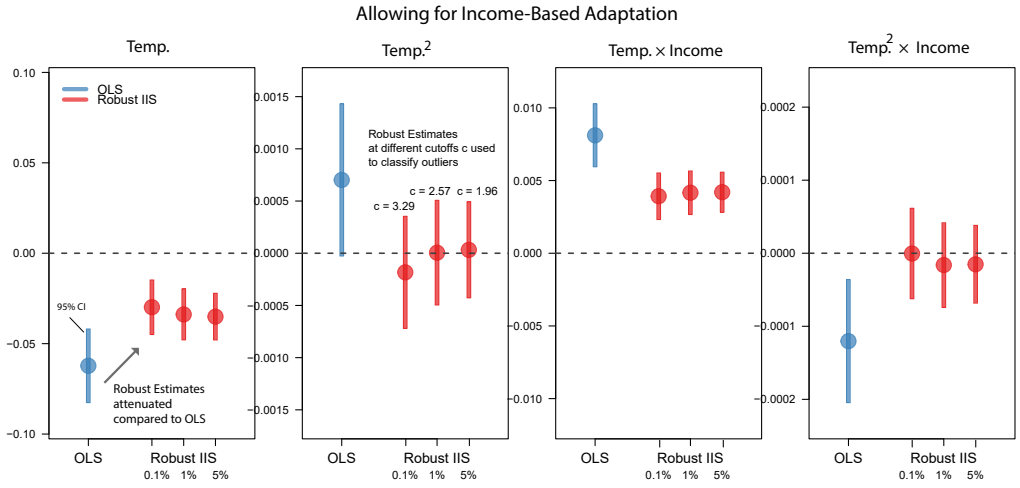


Figure 4.4: Estimated Coefficients of Temperatures on GDP Per Capita Growth Allowing for Adaptation and Controlling for Outliers (at varying cut-off levels used to classify outliers). Coefficients on temperatures and the income interaction terms are shown using OLS (blue) and the robust IIS estimator (red). Cut-offs to classify outliers in IIS are set to yield expected false positive rates ranging from 5% ($c = 1.96$) to 0.1% ($c = 3.29$) for a Normal reference distribution. Note that scale of the y-axis differs across plots.

Overall, our results provide macro-econometric evidence of income-driven adaptation. Temperatures have a significant effect on GDP per capita growth, but this relationship varies by income. Poor countries face large negative impacts of warmer temperatures, while rich countries see significant dampening of temperature effects – see Figure (4.4). Moreover, our estimation results highlight the importance of controlling for outlying observations and testing for subsequent distortion. Idiosyncratic shocks in the model distort coefficient estimates, with robust results showing dampened climate impacts onto GDP per capita growth compared to conventional OLS estimates. Notably, the positive impacts on growth for rich countries are all but removed when controlling for outlying observations.

## 5 Conclusion

We introduce a formal test to assess outlier distortion in regression models by comparing OLS estimates to those obtained using outlier-robust Huber-skip estimators (including Robustified Least Squares – RLS, and Impulse Indicator Saturation – IIS). To develop the distortion test, we fully establish asymptotic theory of RLS and IIS. Our analysis is valid in cross sectional, time series, and panel settings, with stationary or deterministically-trending regressors. Based on our empirical process theory, our results and distortion test can be generalised to derive asymptotic theory for other types of robust estimators, such as winsorized, other Huber, M-type, and L-type estimators (though analysis of S-type estimators is beyond the theoretical framework shown in this paper).

Our proposed outlier distortion test performs well in simulations with size close to nominal levels for large samples ($n > 200$) and high power under a range of alternatives, including vertical outliers and bad leverage points. Our proposed bootstrap implementation of re-sampling the L2 norm on outlier-removed data, or re-sampling the variance using the raw data can improve the performance of the test in finite samples or when the assumed error distribution is different to the underlying error distribution. However, these bootstrap results stem from a small set of simulations – a complete analysis of the bootstrap will form part of our future research.

Our application of the outlier distortion test to the macro-economic impacts of climate change highlights the importance of assessing the influence of outliers in regression models. We find that estimates of climate impacts are sensitive to outlying observations that necessitate robust estimation in all model specifications. When using a robust estimator and formally testing for distortion, the estimated impacts of temperatures onto GDP per capita growth are significantly different from those obtained when using OLS. While our econometric estimates support a significant relationship between temperatures and GDP per capita growth, we also show evidence of income-driven adaptation. The relationship between growth and temperatures varies by income levels with higher incomes mitigating temperature impacts. Nonetheless, all outlier-robust estimates suggest large negative effects of increased temperatures on economic growth even under adaptation, re-emphasizing the need to achieve the temperature goals of the Paris Agreement to avoid large and significant losses, especially in poor countries.

More generally, our proposed test allows for the assessment of outlier-driven coefficient distortion in a wide set of regression models and can be readily implemented using the R-package 'gets' (Pretis, Reade, & Sucarrat 2018).

## References

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American economic review*, *91*(5), 1369–1401.

Acemoglu, D., Johnson, S., & Robinson, J. A. (2012). The colonial origins of comparative development: An empirical investigation: Reply. *American Economic Review*, *102*(6), 3077–3110.

Acemoglu, D., Naidu, S., Restrepo, P., & Robinson, J. A. (2019). Democracy does cause growth. *Journal of Political Economy*, *127*(1), 47–100.

Acevedo, S., Mrkaic, M., Novta, N., Pugacheva, E., & Topalova, P. (2020). The effects of weather shocks on economic activity: what are the channels of impact? *Journal of Macroeconomics*, *65*, 103207.

Aguiar, F. C., Bentz, J., Silva, J. M., Fonseca, A. L., Swart, R., Santos, F. D., & Penha-Lopes, G. (2018). Adaptation to climate change at local level in Europe: An overview. *Environmental Science & Policy*, *86*, 38-63. Retrieved from `https://www.sciencedirect.com/science/article/pii/S146290111731153X` doi: https://doi.org/10.1016/j.envsci.2018.04.010

Albouy, D. Y. (2012). The colonial origins of comparative development: an empirical investigation: comment. *American economic review*, *102*(6), 3059–76.

Anundsen, A. K. (2015). Econometric regime shifts and the US subprime bubble. *Journal of Applied Econometrics*, *30*(1), 145–169.

Auerbach, A. J., Hassett, K. A., & Oliner, S. D. (1994). Reassessing the social returns to equipment investment. *The Quarterly Journal of Economics*, *109*(3), 789–802.

Auffhammer, M. (2022). Climate Adaptive Response Estimation: Short and long run impacts of climate change on residential electricity and natural gas consumption. *Journal of Environmental Economics and Management*, *114*, 102669. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0095069622000432` doi: https://doi.org/10.1016/j.jeem.2022.102669

Barreca, A., Clay, K., Deschenes, O., Greenstone, M., & Shapiro, J. S. (2016). Adapting to climate change: The remarkable decline in the US temperature-mortality relationship over the twentieth century. *Journal of Political Economy*, *124*(1), 105–159.

Berenguer-Rico, V., Johansen, S., & Nielsen, B. (2019). The analysis of marked and weighted empirical processes of estimated residuals. *Working Paper*.

Berenguer-Rico, V., & Nielsen, B. (2018). Marked and weighted empirical processes of residuals with applications to robust regressions. *Working Paper*.

Berenguer-Rico, V., & Wilms, I. (2021). Heteroscedasticity testing after outlier removal. *Econometric Reviews*, *40*(1), 51–85.

Berrang-Ford, L., Siders, A., Lesnikowski, A., Fischer, A. P., Callaghan, M. W., Haddaway, N. R., ... others (2021). A systematic global stocktake of evidence on human adaptation to climate change. *Nature Climate Change*, *11*(11), 989–1000.

Billingsley, P. (1968). *Convergence of probability measures*. John Wiley & Sons.

Burke, M., & Emerick, K. (2016). Adaptation to climate change: Evidence from US agriculture. *American Economic Journal: Economic Policy*, *8*(3), 106–40.

Burke, M., Hsiang, S. M., & Miguel, E. (2015). Global non-linear effect of temperature on economic production. *Nature*, *527*(7577), 235–239.

Bühlmann, P., & Künsch, H. R. (1999). Block length selection in the bootstrap for time series. *Computational Statistics & Data Analysis*, *31*(3), 295-310. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0167947399000146` doi: https://doi.org/10.1016/S0167-9473(99)00014-6

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge University Press.

Cameron, A. C., & Trivedi, P. K. (2010). *Microeconometrics using stata* (Vol. 2). Stata Press College Station, TX.

Carleton, T. A., Jina, A., Delgado, M. T., Greenstone, M., Houser, T., Hsiang, S. M., ... others (2020). *Valuing the global mortality consequences of climate change accounting for adaptation costs and benefits* (Tech. Rep.). National Bureau of Economic Research.

Castle, J. L., & Hendry, D. F. (2009). The long-run determinants of UK wages, 1860–2004. *Journal of Macroeconomics*, *31*(1), 5–28.

Chen, S., & Gong, B. (2021). Response and adaptation of agriculture to climate change: Evidence from China. *Journal of Development Economics*, *148*, 102557. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0304387820301322` doi: https://doi.org/10.1016/j.jdeveco.2020.102557

CIESIN, F. (2016). Gridded population of the world, version 4 (GPWv4): population count grid. *Center for International Earth Science Information Network (CIESIN), Columbia University*.

Davidson, R., & MacKinnon, J. G. (1999). The size distortion of bootstrap tests. *Econometric Theory*, *15*(3), 361–376.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press. doi: 10.1017/CBO9780511802843

Dehon, C., Gassner, M., & Verardi, V. (2012). Extending the Hausman test to check for the presence of outliers. *Advances in Econometrics*.

Dell, M., Jones, B. F., & Olken, B. A. (2012). Temperature shocks and economic growth: Evidence from the last half century. *American Economic Journal: Macroeconomics*, *4*(3), 66–95.

De Long, J. B., & Summers, L. H. (1991). Equipment investment and economic growth. *The Quarterly Journal of Economics*, *106*(2), 445–502.

De Long, J. B., & Summers, L. H. (1994). Equipment investment and economic growth: reply. *The Quarterly Journal of Economics*, *109*(3), 803–807.

Deschênes, O., & Greenstone, M. (2011). Climate change, mortality, and adaptation: Evidence from annual fluctuations in weather in the US. *American Economic Journal: Applied Economics*, *3*(4), 152–85.

Dovonon, P., Gonçalves, S., Hounyo, U., & Meddahi, N. (2019). Bootstrapping high-frequency jump tests. *Journal of the American Statistical Association*, *114*(526), 793–803.

Dreger, C., & Wolters, J. (2014). Money demand and the role of monetary indicators in forecasting euro area inflation. *International Journal of Forecasting*, *30*(2), 303–312.

Durbin, J. (1954). Errors in variables. *Revue de l'institut International de Statistique*, 23–32.

Ericsson, N. R. (2017). How biased are US government forecasts of the federal debt? *International Journal of Forecasting*, *33*(2), 543–559.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: the approach based on influence functions*. New York: Wiley.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the econometric society*, 1251–1271.

Hausman, J. A., & Taylor, W. E. (1981). A generalized specification test. *Economics Letters*, *8*(3), 239–245.

Hendry, D. F., Johansen, S., & Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, *23*(2), 317–335.

Hendry, D. F., & Mizon, G. E. (2011). Econometric modelling of time series with outlying observations. *Journal of Time Series Econometrics*, *3*(1).

Holly, A. (1982). A remark on Hausman's specification test. *Econometrica: Journal of the Econometric Society*, 749–759.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Statistics*, *35*, 73–101.

Jiao, X. (2019). A simple robust procedure in instrumental variables regression. *Working Paper*.

Jiao, X., & Kurle, J. (2021). An asymptotic study of the false outlier detection rate in robust two stage least squares models. *Working Paper*.

Jiao, X., & Nielsen, B. (2015). Asymptotic analysis of iterated 1-step Huber-skip M-estimators with varying cut-offs. In *Workshop on analytical methods in statistics* (pp. 23–52).

Jiao, X., & Pretis, F. (2022). Testing the presence of outliers in regression models. *Oxford Bulletin of Economics and Statistics*.

Johansen, S., & Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. *Castle, and Shephard (2009)*, 1–36.

Johansen, S., & Nielsen, B. (2013). Outlier detection in regression using an iterated one-step approximation to the Huber-skip estimator. *Econometrics*, *1*(1), 53–70.

Johansen, S., & Nielsen, B. (2016a). Analysis of the Forward Search using some new results for martingales and empirical processes. *Bernoulli*, *22*(2), 1131–1183.

Johansen, S., & Nielsen, B. (2016b). Asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics*, *43*(2), 321–348.

Johansen, S., & Nielsen, B. (2019). Boundedness of M-estimators for linear regression in time series. *Econometric Theory*, *35*(3), 653–683.

Kahn, M. E., Mohaddes, K., Ng, R. N., Pesaran, M. H., Raissi, M., & Yang, J.-C. (2021). Long-term macroeconomic effects of climate change: A cross-country analysis. *Energy Economics*, *104*, 105624.

Kaji, T. (2018). Switching to the new norm: from heuristics to formal tests using integrable empirical processes. *Working Paper*.

Kalkuhl, M., & Wenz, L. (2020). The impact of climate conditions on economic production. Evidence from a global panel of regions. *Journal of Environmental Economics and Management*, *103*, 102360.

Martinez, A. B. (2020). Forecast Accuracy Matters for Hurricane Damage. *Econometrics*, *8*(2), 18.

Matsuura, K., & Willmott, C. J. (2018). Terrestrial precipitation: 1900–2017 gridded monthly time series. *Electronic. Department of Geography, University of Delaware, Newark, DE*, *19716*.

Newell, R. G., Prest, B. C., & Sexton, S. E. (2021). The GDP-temperature relationship: implications for climate change damages. *Journal of Environmental Economics and Management*, *108*, 102445.

Nymoen, R., & Sparrman, V. (2015). Equilibrium unemployment dynamics in a panel of OECD countries. *Oxford Bulletin of Economics and Statistics*, *77*(2), 164–190.

Pretis, F., Mann, M. L., & Kaufmann, R. K. (2015). Testing competing models of the temperature hiatus: assessing the effects of conditioning variables and temporal uncertainties through sample-wide break detection. *Climatic Change*, *131*(4), 705–718.

Pretis, F., Reade, J., & Sucarrat, G. (2018). Automated General-to-Specific (GETS) regression modeling and indicator saturation methods for the detection of outliers and structural breaks. *Journal of Statistical Software*, *86*(3).

Pretis, F., Schwarz, M., Tang, K., Haustein, K., & Allen, M. R. (2018). Uncertain impacts on economic growth when stabilizing global temperatures at 1.5 C or 2 C warming. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2119), 20160460.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association*, *79*(388), 871–880.

Ruppert, D., & Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, *75*(372), 828–838.

Salibián-Barrera, M., Van Aelst, S., & Willems, G. (2008). Fast and robust bootstrap. *Statistical Methods and Applications*, *17*(1), 41–71.

Schneider, L., Smerdon, J. E., Pretis, F., Hartl-Meier, C., & Esper, J. (2017). A new archive of large volcanic events over the past millennium derived from reconstructed summer temperatures. *Environmental Research Letters*, *12*(9), 094005.

Schwarz, M., & Pretis, F. (2020). Modelling Historical Adaptation Rates to Inform Future Adaptation Pathways. In *EGU General Assembly Conference Abstracts* (p. 21004).

Singh, K. (1998). Breakdown theory for bootstrap quantiles. *The Annals of Statistics*, *26*(5), 1719–1732.

Stillwagon, J. R. (2016). Non-linear exchange rate relationships: An automated model selection approach with indicator saturation. *The North American Journal of Economics and Finance*, *37*, 84–109.

Tol, R. S. (2017). Population and trends in the global mean temperature. *Atmósfera*, *30*(2), 121–135.

Varga, R. S. (2000). *Matrix Iterative Analysis*. Berlin: Springer.

Welsh, A. H., & Ronchetti, E. (2002). A journey in single steps: robust one-step M-estimation in linear regression. *Journal of Statistical Planning and Inference*, *103*(1-2), 287–310.

World Bank, W. (2019). World development indicators. *Washington, DC. http://wdi. worldbank. org/table/4.2*, *10*, 2001–2002.

Wu, D.-M. (1973). Alternative tests of independence between stochastic regressors and disturbances. *Econometrica: journal of the Econometric Society*, 733–750.

# Paper 2: An Empirical Climate Damage Function accounting for Climate Extremes and Adaptation

# An empirical climate damage function accounting for climate extremes and adaptation

Moritz P. Schwarz[1,2,3] and Felix Pretis[2,3,4]

## Abstract

Quantifying the economic impacts of climate change is crucial to inform mitigation and adaptation policy but faces challenges surrounding impacts of climate extremes and uncertainty around the range of plausible adaptation pathways. We overcome these by using machine learning and econometric model selection to construct an empirically-derived climate damage function allowing for impacts of a range of climate extremes under adaptation independent of any specific emission scenario. We use a novel baseline of forecasts of future economic development until the end of the century and – in absence of adaptation – project a decline in median country-level GDP per capita of up to 66% for warming beyond 4.5°C relative to no climate change. Projected marginal impacts under no adaptation suggest an approximate 12% decline in median country-level GDP per capita for each additional °C warming. We further show that the damage curve is not invariant to adaptation and provide empirical evidence of historical adaptation over time and incomes at a macro-economic level. Instability over time lowered the level of projected median GDP per capita impacts by approximately 20 percentage points relative to no-adaptation. Income-driven adaptation could reduce the marginal impacts of an additional degree of warming by a half to around 6% of GDP per capita per additional °C. Nevertheless, projected damages remain high and unequal even in the presence of adaptation reiterating the urgent case for stringent mitigation policy.

[1] Smith School of Enterprise and the Environment, University of Oxford
[2] Institute for New Economic Thinking at the Oxford Martin School, University of Oxford
[3] Climate Econometrics, Nuffield College
[4] Department of Economics, University of Victoria

# 1 Introduction

There is considerable uncertainty about future climate impacts on economic production. Quantifying climate impacts in terms of economic outcomes is crucial to inform mitigation and adaptation policy decisions (Roughgarden and Schneider, 1999; Stern, 2008). Existing economic impact projections are commonly derived using calibrated Integrated Assessment Models (IAMs) or empirically estimated econometric models. IAMs predominantly rely on constant damage functions with little grounding in empirical evidence (Pindyck, 2013; Stern, 2016), while the majority of empirical macro-economic models do not account for adaptation and commonly use aggregated climate measures (see e.g. Dell *et al.*, 2012; Burke *et al.*, 2015; Pretis *et al.*, 2018b; Kahn *et al.*, 2019), possibly masking the effects of sub-annual extreme weather events such as floods, droughts, and heatwaves (see e.g. Field and Barros, 2014; Moriondo *et al.*, 2011; Hanson *et al.*, 2011). Using machine learning and econometric model selection, we construct an empirically-derived damage function allowing for the potential impact of climate extremes and accounting for plausible adaptation pathways. The resulting damage function can be disaggregated to a country level as a function of global mean surface temperature and is independent of any specific emission scenario. In absence of adaptation, combined with a novel baseline using long-run empirical forecasts of future economic development, our empirical damage function projects median GDP per capita level impacts up to a 66% reduction for very high levels of warming beyond 4.5°C degrees relative to long run forecasts without additional climate change. Despite allowing for non-linearities, projected marginal impacts in absence of adaptation suggest a roughly constant 12% decline in median country-level GDP per capita for each additional degree of warming under no adaptation. Comparing our projected impacts to earlier estimates, we find that projected losses in GDP might be higher by around a factor of two when accounting for the impact of extremes (see Supplementary Material Section 7 for a comparison). This translates to additional climate change essentially eradicating future economic development relative to present conditions for a large number of countries. However, we show that the damage function has not remained constant over time or income levels. We quantify historical climate adaptation at a macro-economic level, where climate impacts are changing over time and attenuated by higher incomes. Over time, we find that observed adaptation to-date has reduced the level of projected climate impacts by around 20 percentage points (to 30-40% for warming beyond 4.5°C). In turn, projecting estimates of income-driven adaptation into the future, the marginal impacts of an additional degree of warming are roughly halved, suggesting a 6% loss per additional °C. While the exact channels of adaptation (and resulting shape of the damage curve) are difficult to estimate precisely, even under optimistic adaptation scenarios projected impacts remain high with extreme tail risk, strengthening the case for stringent mitigation policy.

Sources of uncertainty in climate impacts stem from model uncertainty (i.e. which climate phenomena should be included in an empirical impact model, see Cui *et al.*, 2018), estimation uncertainty about model parameters, uncertainty about future climate itself, uncertainty about adaptation and the stability of the estimated model parameters, and scenario uncertainty about any baseline of economic and population growth against which impacts are evaluated. To date, existing empirical macro-economic impact studies have focused primarily on climate- and estimation-uncertainty, placing relatively less emphasis on model uncertainty, the potential impacts of adaptation, and socio-economic forecasts instead of scenarios as baselines. Here we attempt to consider all five sources of uncertainty. To account for model uncertainty - rather than *a-priori* imposing which climate variables are important - we employ model selection strategies including Bayesian model selection and recent developments in machine learning to identify relevant climate determinants including extremes beyond annual average metrics (see Methods). We account for climate uncertainty using a large ensemble of CMIP5 climate model outcomes at a wide range of target global mean surface temperature (GMST) levels. We quantify estimation uncertainty by re-sampling our model parameters and consider adaptation uncertainty by

estimating historical rates of adaptation over time and income levels. We subsequently project climate impacts on future levels of incomes using a new baseline derived from long-run empirical forecasts (Müller *et al.*, 2020, henceforth MSW) of future per-capita economic growth on a country level.

We construct a country-level dataset of 27 population-weighted extreme climate variables measuring temperature and precipitation extremes, such as maximum night-time temperature or maximum consecutive 5-day precipitation, for 169 countries from 1962 to 2011 (see Supplementary Material Table A2 for an overview). We then estimate empirical impact models in-line with the quasi-experimental approach of modelling country-level GDP per-capita growth as a function of weather and climate observations (see e.g. Dell *et al.*, 2012; Burke *et al.*, 2015; Pretis *et al.*, 2018b). When addressing model uncertainty, we focus on uncertainty in the choice of variables included in panel data models with a range of fixed effects and non-linear controls, rather than uncertainty about the type of model (e.g. stochastic frontier models – see Tol, 2020). By allowing for non-linear functional forms, we also mitigate some concerns about mis-matched time-series properties of climate variables and economic growth (Miller and Park, 2010).[5]

We employ three model selection approaches to identify 'relevant' climate variables out of the set of 58 possible covariates, which include indicators for extreme climate and non-linear transformations of all variables (see Methods and Supplementary Material Table A1 and A2). We carefully control the inclusion of variables in the final model by not targeting goodness of fit and removing outlying observations using outlier-robust initial estimators due to potential outlier-driven distortion (Jiao *et al.*, 2021). First, we use a general-to-specific (gets) consistent model selection approach that tightly controls the false-positive rate of retention of variables by choosing a target level of significance (Pretis *et al.*, 2018a; Hendry and Doornik, 2014; Campos *et al.*, 2003). Second, we use the adaptive LASSO (Zou, 2006), a consistent model selection procedure with oracle properties. Third, following the literature identifying the macroeconomic drivers of economic growth (Fernández *et al.*, 2001; Sala-i-Martin *et al.*, 2004), we also use Bayesian model selection (BMS) to identify the climate determinants of GDP per capita growth. As an additional comparison we also construct the damage function implied by earlier estimates of a quadratic impact of temperatures and precipitation on growth (Burke *et al.*, 2015, here referred to as AATPsq: annual-average temperature and precipitation in both linear and squared form).

Existing impact projections primarily characterise damages under specific emission scenarios (such as RCP 8.5), creating a disconnect to impact statements in the wider policy framework which focus on temperature targets instead (such as 1.5°C as a result of the Paris Agreement, see Pretis *et al.*, 2018b; Burke *et al.*, 2018). Thus, to link our empirically-estimated impacts on economic growth to GMST levels rather than a particular emissions scenario, we use a consistent set of projections of climate extremes derived from CMIP5 climate model outputs independent of their forcing scenario (see Methods). This allows us to construct an empirically-derived damage function for a given level of GMST and the associated extremes projected under any particular temperature level (comparable to Jackson *et al.*, 2018 for sea level). To provide GDP per capita *level* projections of the economic impacts we move beyond using the Shared Socio-Economic Pathways (SSPs, Riahi *et al.*, 2017) and use the novel MSW long-run empirical forecasts of economic development as baselines. These provide an empirically-derived reference of forecasted GDP per capita growth (for comparison we also report results using the SSPs in the supplementary material). Notably, the MSW forecasts of future growth are estimated over the entire 20$^{th}$ century, and therefore can act as a baseline scenario that represents growth under climate conditions to-date.

[5] Non-linear transformations of potentially non-stationary I(1) time series can result in what appears to be stationary I(0) outcomes – see Miller and Park (2010).

We project the resulting damage curve as the relationship between GMST and projected GDP per capita change relative to the baseline scenario, accounting for differential realisations of climate for different countries at various temperature levels. Unlike previous approaches that assume a linear (or logarithmic) progression of country-level temperatures over time until 2100, our estimation method takes climate model stochasticity along the temporal evolution of temperatures into account (Calel *et al.*, 2020). In other words, we consider the climate model variation of climate variables in each year up to 2099 when calculating cumulative damages for a given target level of GMST.

Our approach differs notably from earlier empirical macro-economic studies which have analysed how economic growth varies with average annual climate variables. We estimate a damage function that specifies impacts for a given level of policy-relevant GMST anomaly at the end of the century. This global mean increase translates into country-level impacts through the empirical models (derived using model selection) accounting for the potential impact of climate extremes. Combined with the near linear relationship between cumulative emissions and temperatures (Goodwin *et al.*, 2018) this allows us to also relate climate impacts to any specified carbon budget.

Previous individual-country estimates and local case studies suggest significant adaptation to climate change ranging from the decline of heat-mortality (Carleton *et al.*, 2020; Barreca *et al.*, 2016) to improvements in agricultural practises (Burke and Emerick, 2016; Schlenker and Roberts, 2009). However, adaptation has received less attention in empirical macro-economic cross-country panel analyses of economic growth. To estimate the degree of historical adaptation, we use two approaches.

First, we conduct an unconditional analysis where adaptation is solely a function of time. We estimate the selected empirical panel impact models in 30-year windows at the start and at the end of our sample period and assess the resulting damage functions. While this fixed-window approach allows us to characterise change in the climate-growth relationship observed to date, it does not explain the underlying factors driving adaptation. It is likely that higher incomes mitigate climate impacts (see e.g. Acevedo *et al.*, 2020), partly by increasing resilience, partly by richer countries shifting to less climate-vulnerable industries.[6] We therefore consider a second approach exploring income as a channel of adaptation where we allow the selected climate impacts to vary by per-capita income. Specifically, we estimate the selected impact models interacting retained climate variables with the observed log of GDP per capita, which is subsequently projected to the end of the 21st century using the MSW long-run econometric forecasts. To avoid extrapolation of adaptation out-of-sample we constrain the projected adaptation to the highest in-sample income levels across countries. We further restrict growth impacts under adaptation to not exceed the baseline growth rate where non-adaptation impacts in a particular year would otherwise be negative. To reflect that adaptation is a choice, we restrict countries not to be worse-off under adaptation compared to the no-adaptation projection (see Methods).

# 2 Results

### Specification and Projection of the Baseline Damage Function in Absence of Adaptation

Across all 170 CMIP5 model runs considered, our results define a damage function for the range of 1.35°C to 5.90°C of GMST anomalies relative to pre-industrial temperatures. In absence of adaptation, projected reductions in median country-level GDP per capita range from 23% for 1.5°C to 66% for warming above 4.5°C in our 'gets' projections compared to a no further climate change counterfactual

[6] Dell *et al.* (2012) indirectly explored this by interacting a 0/1 measure for low-income countries with temperatures. Here we go beyond this initial approach by allowing the impact of climate variables to vary continuously with income levels.

given by long-run forecasts. Estimates for global average (rather than median country-level) GDP per capita changes are included in the Supplementary Material. Median projected impacts are robust across model selection approaches as well across different functional forms chosen to link impacts to GMST, with the median of the LASSO and the AATPsq models reaching 52% and 57% reductions for temperatures exceeding 4.5°C. The 90% range of projected impacts indicates considerable uncertainty in damages not ruling out the (albeit small) possibility of positive impacts for some countries. Nonetheless, LASSO and BMS estimates suggest lower likelihoods of positive growth as a function of future climate change.[7] The lower tail of the projection distribution (5th percentile) indicates a risk of potentially catastrophic damages which could reduce GDP per capita by 73-94% (relative to the no-climate change scenario) depending on the selected model under high levels of warming.

When mapping selected extreme indicators to GMST, the median damage function (in absence of adaptation) at the end of the century (using the 'gets' estimate) can be approximated as equation (1)

$$\frac{GDPpc_{Climate} - GDPpc_{Base}}{GDPpc_{Base}} = -0.021 - 0.14\,Temp + 0.0047\,Temp^2, \quad \text{for Temp} \in [1.35°C, 5.9°C] \qquad (1)$$

where $Temp$ refers to the GMST anomaly to pre-industrial temperatures at the end of the century (see Supplementary Materials for different functional forms). We emphasise that this function should not be extrapolated outside the final temperature levels observed in the CMIP5 model record used (1.35°C – 5.9°C). As no CMIP5 model is compatible with a GMST warming lower than 1.35°C by the end of the century, we restrict our estimates to these values contrary to earlier studies where the inclusion of the origin (i.e. 0°C warming resulting in 0% difference to the baseline) biases shape of any estimated damage curve. Despite the flexible non-linear functional form, marginal impacts of an additional warming appear nearly constant over projected temperatures, suggesting around a -12% change in median GDP per capita per additional °C under no adaptation.

This relationship describes projected future median GDP per capita level impacts as a function of GMST of our country-level growth estimates driven by a range of climate indicators, not solely the quadratic relationship of country-level temperatures and growth estimated previously (Burke et al., 2015). The quadratic relationship between growth and temperature remains present in some of the underlying panel model estimates (associated with negative growth impacts primarily in the Tropics – see Figure 2) even when accounting for climate extremes. However, unlike prior macro-empirical studies, our model selection results suggest detectable negative growth effects of some climate extremes as these are retained in the final selected models (see Figure 2 for relative contributions of retained variables in the final selected models). Specifically, warm nights (tr in Figure 2) are associated with lower growth globally, consistent with physiological effects and heat-stress effects of night-time temperatures, (see e.g. Fischer et al., 2012; Murage et al., 2017; Obradovich et al., 2017). In turn, mild winters (tnx, tx10p in Figure 2) drive positive growth in Northern latitudes, while high frequency precipitation days (r1mm), and impacts of precipitation extremes (rx1day) are associated with minor reductions in growth throughout the globe, although overall effects are less certain.

[7] This is likely driven by temperatures only entering the final model in a squared functional form with a negative coefficient. The linear temperature term is not retained post-selection in these models.

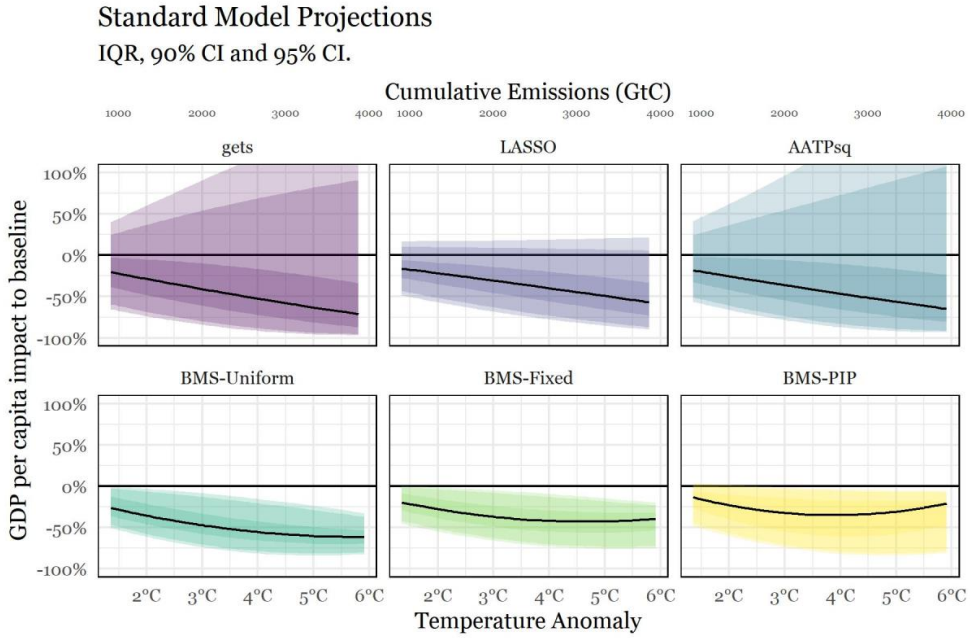


**Figure 1**: Projected damage curve (change in median GDP per capita) based on model-selection panel estimates combined with CMIP5 climate projections relative to MSW baseline GDP per capita growth forecasts. Results indicate robust and large reductions in median country-level GDP per capita with further warming across estimation method and specification. Figure shows estimated damage curve of the median country-level GDP per capita compared to long-run economic forecasts under no-additional warming. Shaded regions denote 25%-75% interquartile range (IQR), 90% and 95% projection confidence interval (CI) across estimation and climate uncertainty estimated using quantile regression over the distribution of impacts. Solid lines show median of projections. The titles above each panel refer to the model selection method used: 'gets' refers to using general-to-specific model selection, 'LASSO' refers to the least absolute shrinkage and selection operator, while 'BMS' refers to Bayesian Model Selection. For each BMS model, a prior must be specified with a fixed (Fixed), uniform (Uniform) or custom prior inclusion probabilities (PIP) being used as specified by the BMS R package function 'bms'. Cumulative emissions are expressed in Gigatonnes of Carbon (GtC).

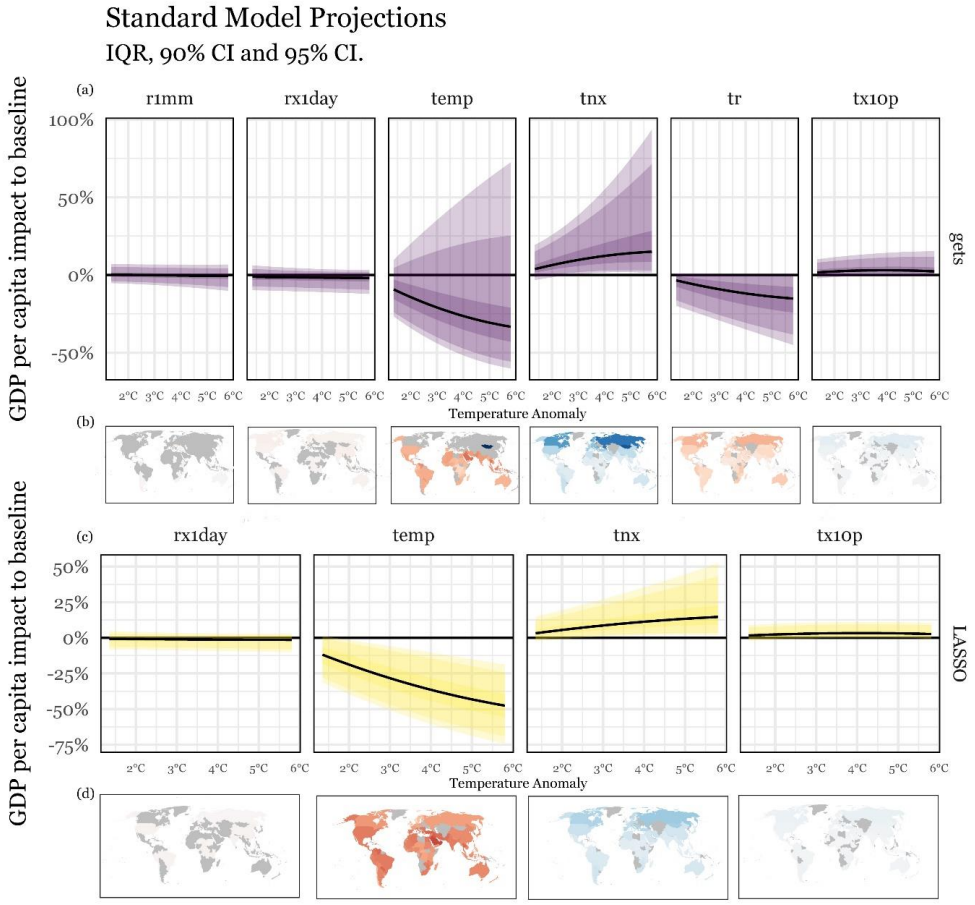


**Figure 2**: Relative Contributions of retained climate variables to the median country-level GDP per capita damage function in 'gets' (general-to-specific model selection) and adaptive LASSO (least absolute shrinkage and selection operator) models over target temperatures in 2090-2099 (panels a & c) and over space for warming beyond 5.5°C (panels b & d). Retained climate variables include population-weighted high frequency precipitation days (r1mm), precipitation extremes (rx1day), annual-average temperatures and its square (temp; only squared coefficient selected in the LASSO), mild winters (tnx and tx10p), and warm nights (tr). See Supplementary Material Table A1 for exact variable definitions.

Our modelling approach allows us to compute country-level damage functions using local observations and projections of extremes retained in the model (Figure 3). Projected impacts are highly unequal across the globe and could result in considerably higher levels of GDP per capita for a very small number of countries for moderate warming in the Northern Hemisphere. These findings are consistent with previous evidence on country-level social costs of carbon (Ricke *et al.*, 2018). For high levels of warming, only few countries see positive projected impacts. Predominantly developing countries could face catastrophic levels of damage with increasing climate change. For more than a dozen countries, median GDP per capita projections exceed 80% reductions under global mean warming of 4.5°C when compared to a scenario in absence of climate impacts – with even higher levels of warming as well as 5$^{th}$ percentile impacts suggesting potentially catastrophic impacts for many countries.

It is worth highlighting that these projections do not suggest an 80% reduction relative to current levels of GDP per capita, but instead a reduction in GDP per capita relative to forecasts in absence of climate impacts. This translates to higher levels of warming eradicating forecasted economic development beyond current conditions for many countries. Bangladesh, India, and Nigeria, for instance, stand to increase their GDP per capita from about US$ 3,600, US$ 6,300, and US$ 5,300 GDP per capita in 2017 (according to the values used in MSW; real values are lower) to US$ 21,600, US$ 31,000, and US$ 29,300 in the MSW baseline. However, with high levels of warming, much of the projected development is wiped out, with GDP per capita only standing at about US$ 4,800, US$ 6,400, and US$ 4,500 for all models with a GSMT anomaly of more than 4°C (Figure 3, and Figure 4 lower panel). An open question remains about the practice of relying on long-run counterfactual forecasts (such as MSW or SSP) as baselines. Even under extreme climate change, future levels of projected incomes are multiples of present observed levels for many countries, which is driven primarily by the underlying baseline growth assumptions rather than by our estimates.

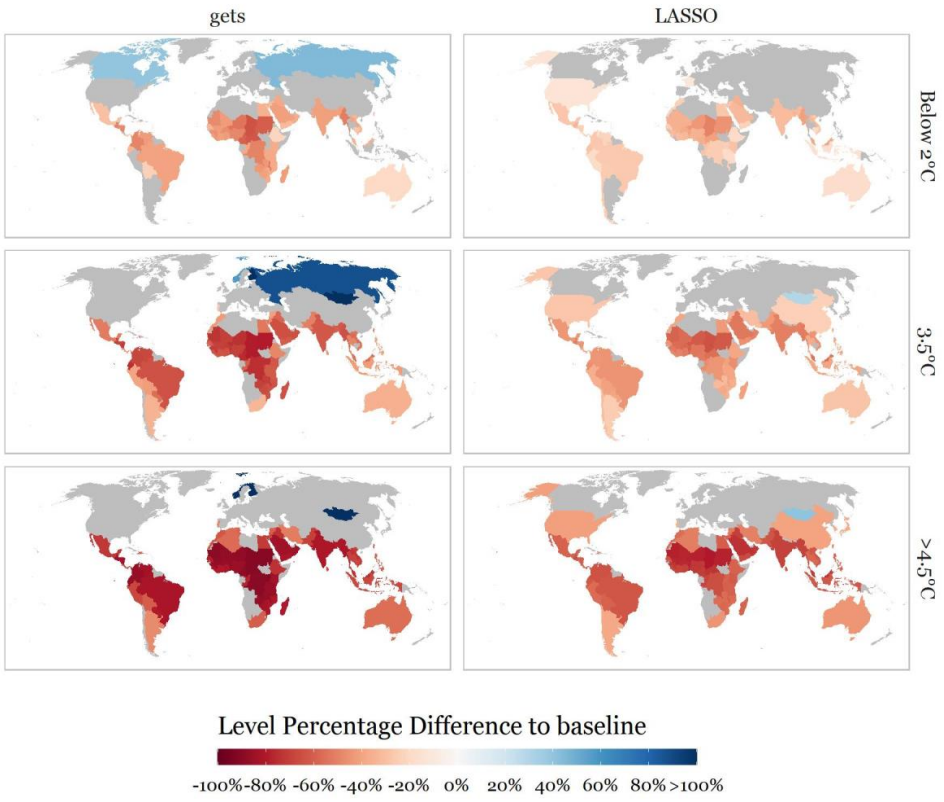


**Figure 3:** Projected difference in GDP per capita under different GMST warming levels in 2090-2100 relative to 'no future climate change' long-run forecasts from Muller, Stock, and Watson (MSW) GDP per capita forecasts for 'gets' (general-to-specific model selection) and LASSO (least absolute shrinkage and selection operator) selected models. Countries are shaded grey if 90% projection interval of impacts includes zero or no value is available. Values were calculated as means across all realizations below 2°C, between 3°C and 4°C and above 4.5°C.

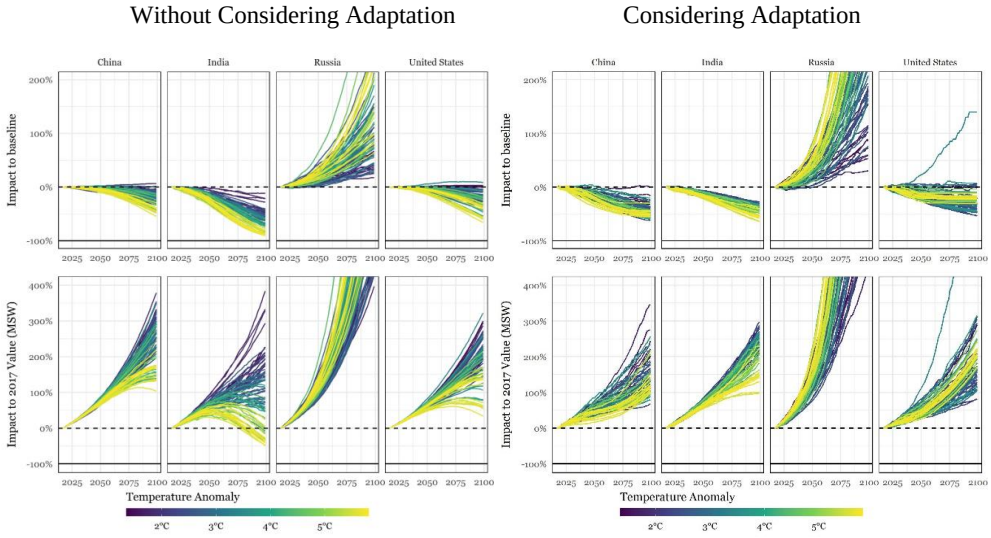


**Figure 4**: Projected country-level climate impacts relative to the 'no future climate change' baseline from MSW (top panels; Muller, Stock, and Watson GDP per capita forecasts) and relative to the baseline in 2017 (bottom panel) for China, India, Russia, and the United States. Colour-gradient denotes level of GMST warming in 2090-2099 relative to pre-industrial temperatures. Left hand set of panels present projections without considering adaptation. Right hand set of figures consider Income Adaptation.

**Damage Curve in Presence of Adaptation**

While the above results show large potential negative climate impacts on economic output, the damage function itself is unlikely to have remained constant over time or income levels. Our moving window estimates of the damage function provide evidence of considerable historical adaptation in-sample over time (Figure 5, middle and right panels). While gradual changes in coefficients may not be detectable for a single country (e.g. as in Kahn *et al.*, 2019), in our cross-country panel we find that the coefficient estimates of most climate variables are attenuated over the sample period. In other words, the impact of climate on economic outcomes in-sample dampens over time, providing macro-economic evidence of a partial decoupling of climatic conditions and the economy. Had the damage function remained at its initial estimate from 1962-1992, projected impacts would have been higher by roughly 20 percentage points relative to the most recent estimate (Figure 5). Part of this historical adaptation trend can be explained by increasing incomes. Re-estimating the empirical damage function using climate-income interactions shows that climate impacts are attenuated by higher incomes (see Supplementary Information Table A7). The coefficients on all income-interaction terms have the opposite sign of the non-interacted terms, moving them closer to zero. Notably, the non-linear relationship between temperatures and economic growth shows flattening. We project the resulting climate-income relationship under adaptation using the MSW long-run GDP per capita growth forecasts as a baseline. The resulting marginal impacts of an additional degree of warming at the median are roughly halved, falling from 12% to 6% per additional °C under adaptation for moderate levels of warming (Figure 5, left panel comparing the slope of the solid and dashed projections). This reduced marginal impact results in level damages for 4.5°C warming being lowered by around a third (Figure 5, left panel and Figure 4). These results are robust to selecting over climate-income interactions jointly (rather than imposing interactions post-selection), as well as interacting climate variables with the lag of incomes rather than the contemporaneous income (see Supplementary Material).

Thus, unconditional estimation of adaptation over time shows overall attenuation of the level of climate impacts. In turn, allowing for income-driven adaptation, we see notable flattening of the damage curve with reduced marginal impacts of each additional degree of warming. Nevertheless, the uncertainty remains large and adaptation itself is unlikely to be costless. Even under income-driven adaptation projected damages suggest a median reduction in GDP levels around 30-40% for warming around 4.5°C relative to a no climate-change baseline.

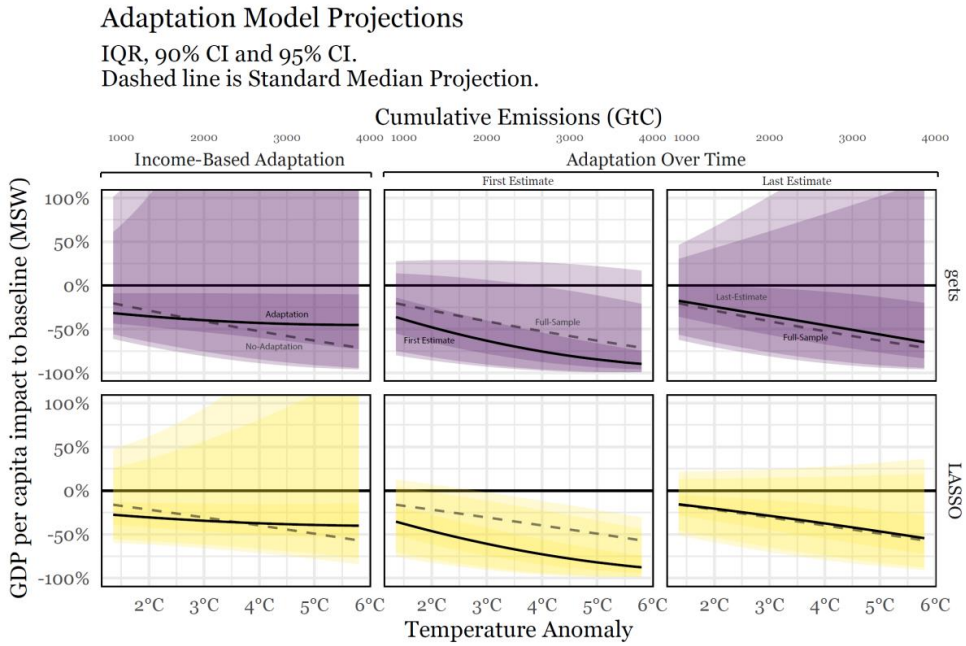


**Figure 5**: Projected damage curve using 'gets' (general-to-specific model selection) and LASSO (least absolute shrinkage and selection operator) accounting for adaptation. Income-based adaptation is shown in left panels, with middle and right panels showing adaptation over time for the first (1962-1992) and last (1981-2011) 30-year estimation sample. Solid line shows median projected impacts under adaptation, while dashed line shows median full sample non-adaptation estimates from Figure 1. Cumulative emissions are expressed in Gigatonnes of Carbon (GtC).

# 3 Discussion

Existing climate damage functions used to guide policy face multiple constraints. They predominantly rely on estimated impacts of mean annual global temperature – ignoring considerable regional variation as well as indicators for extreme climate dynamics. Further, most existing impact assessments assume a constant damage function which likely overstates impacts in-light of ongoing adaptation.

Here, we present an empirically-derived damage function that is specified for target temperatures instead of a particular emission scenario, consider possible effects of local climate extremes, and highlight the dynamic element of historical adaptation, while also considering model, climate, estimation, and scenario uncertainty. The estimated damage function can be disaggregated onto a country level and is readily compatible with existing models for the assessment of the economic impacts of climate change. Our method allows the formulation of a damage function for any period up to 2100 and could be extended beyond that with more long-term economic baselines and climate model runs.

Even though there is some variation across the retained set of climate variables (which is expected when selecting over highly correlated covariates – see Methods), the overall final shape and magnitude of the estimated damage function is not sensitive to the model selection method employed. Resulting projections provide strong evidence of the economic impacts of climate change over the historical estimation period, albeit with considerable uncertainty (Figure 1). Specifically, the large uncertainties in our estimates imply wide tail risks – a feature highlighted in the theoretical climate-economic literature (Weitzman, 2012) but often underappreciated in empirical studies. While climate models devote much attention to the immense challenge of incorporating estimates of potential feedback effects that could result in climatic tipping points (Drijfhout *et al.*, 2015), it is worth noting that the tail risk estimates given by our damage function do not take future socio-economic tipping points into account(see e.g., Dietz *et al.*, 2021), which might exacerbate damages further.

While our projected damage function is robust to the model selection procedures, the function itself changes over time and incomes. Moving window estimates show evidence of historical adaptation, providing cross-country macro-economic support for local case studies on adaptation. Instability of the damage function over time suggests an absolute reduction in the level of damages, while increasing incomes suggest reductions in marginal impacts, with each degree of additional warming resulting in lower damages compared to no adaptation. We emphasise three important take-aways from these adaptation results. First, we have seen significant dampening of climate effects onto economic outcomes already in the observed record. Second, income adaptation is uncertain and sensitive to imposed restrictions and subsequent projection conditional on baseline scenarios. The historical evidence of income-driven adaptation emphasizes the need to consider plausible adaptation pathways when using long run baseline economic and institutional scenarios (Andrijevic *et al.*, 2020) as benchmark for climate impacts, especially for warming levels that represent current mitigation ambition. Projected incomes under climate change in SSPs and long-run economic forecasts (Müller *et al.*, 2020) exceed current incomes considerably even for the world's poorest countries largely due to the underlying assumptions of continued economic growth (see Figure 4). Third, even in the presence of income-driven adaption, projected damages remain high and unequal. Countries that may be successful on their adaptation pathway could see incomes rise, while predominantly poor countries could see comparative stagnation or increased damages, exacerbating global climate-driven inequality further (Diffenbaugh and Burke, 2019).

Even though climate models give us some degree of confidence over the temperature range considered here, our damage curve estimates are unlikely to be robust for warming beyond this range. A key uncertainty also remains the likelihood of non-linear socio-economic tipping points that could occur

once certain impact thresholds are crossed. Finally, we have not modelled the costs of adaptation which are likely to be significant, especially for high levels of warming. Adaptation will not be a panacea to avoid significant impacts under high levels of warming due to hard limits to adaptability, given by wet bulb temperatures which could render much of our economic activity unfeasible (Xu *et al.*, 2020; Sherwood and Huber, 2010). Transitioning to a post-carbon society is a herculean feat – but our damage curve estimates reaffirm that it is beneficial to do so sooner rather than later to minimize economic impacts along the way.

# 4 Methods

### Overview

We model GDP per capita growth $\Delta\log(GDPpc)_{i,t}$ for country *i* in year *t* as a function of historic climate variables in country *i* and year *t* (see following sections for information on data and model) specifying $\Delta\log(GDPpc)_{i,t} = f(Climate_{i,t})$ where we estimate *f()* using model selection and machine learning robust to outlying observations (see Estimation and Selection section below). To create projections under particular temperature targets, we link the CMIP5 projections of retained climate variables to the level of global mean surface temperature (GMST) of the associated climate model exhibits at the end of the century, $Climate_{i,t} = h(GMST_t)$, where *h()* maps GMST to local country-level climate and is approximated using CMIP5 simulations (see 4.4). We link GMST to cumulative emissions, $GMST_t = g(CCe_t)$, using the transient response to cumulative carbon emissions from Goodwin *et al.* (2018), which incorporates forcing from aerosols and non-$CO_2$ greenhouse gases as well. We allow for adaptation over time *t* using in-sample window estimates of *f()* and adaptation over income by interacting $Climate_{i,t}$ with $\ln(GDPpc_{i,t})$(4.7).

### Data

We combine a globally consistent dataset of climate extremes indicators as defined by ETCCDI[8] (Sillmann *et al.*, 2013a) based on climate reanalysis models with monthly average temperatures and precipitation (Matsuura and Willmott, 2018) and country-level log real GDP per capita (World Bank, 2019). Spatial variables were population-weighted (in line with Tol, 2017) using 2015 data from CIESIN (2016) before their aggregation to a country-level. End-of-century projections were based on CMIP5 (Taylor *et al.*, 2012) temperature and precipitation data obtained from ESGF[9], while future climate extremes were obtained from Sillmann *et al.* (2013b). Hundred year GDP per capita forecasts were obtained from Müller *et al.* (2020, MSW), while GDP and population projections for the Shared-Socio-economic pathways were obtained from the IIASA SSP database and interpolated to annual values using a spline function (Riahi *et al.*, 2017).

### Estimation & Selection

We estimate a fixed effects panel model consistent with the existing literature (Dell *et al.*, 2012; Burke *et al.*, 2015; Pretis *et al.*, 2018b) modelling the year-on-year change of the log of GDP per capita as a function of the set of climate indicators, country and time fixed effects, as well as including linear and quadratic country-specific time trends. To address potential measurement error and outlying observations driven by unknown factors other than climate, we use impulse indicator saturation at

[8] Expert Team on Climate Change Detection and Indices
[9] See https://esgf-index1.ceda.ac.uk/projects/cmip5-ceda/

$p<0.001$ to ensure that our estimated results are robust to un-modelled outliers (Hendry *et al.*, 2008). Existing macro-econometric studies on the impact of climate on economic growth focus on highly aggregated and pre-selected measures of average temperature and precipitation. To address model uncertainty about the choice of climate variables we include 27 additional indicators of extreme climate (and non-linear transformations thereof) to assess whether macro-economic climate impacts can be detected beyond their signal in country-level aggregates of temperature and precipitation. Which climate variables are relevant in describing this relationship cannot be judged *a priori* (see e.g. Cui *et al.*, 2018), so we therefore employ machine learning and automated model selection methods (gets, adaptive Lasso, and Bayesian Model Selection - BMS) to inform our choice. All model selection procedures first select over quadratic regressors and only in a second step over retained quadratic regressors as well as all linear regressors.

Selection using 'gets' takes place at the target significance levels of 5%, with a resulting expected false-positive retained number of variables of $58 \times 0.05 = 2.9$. The 'gets' approach provides a consistent model selection procedure (Campos *et al.*, 2003) and is implemented using the R-package 'gets' (Pretis *et al.*, 2018a).

The Adaptive LASSO is implemented using the R-package 'glmnet' (Friedman *et al.*, 2010) and is estimated in three steps. In the first step the penalty weights for the adaptive LASSO are determined using OLS. In the second step, the adaptive LASSO is estimated using the OLS penalty weights und country-level cross-validation to determine the penalty parameter. Finally, the LASSO model is estimated post-selection using OLS which yields acceptable properties in terms of bias (Belloni and Chernozhukov, 2013; Zhao *et al.*, 2021).

Bayesian Model Selection is conducted using an MCMC sampler and uses different priors, including a fixed common prior inclusion probability for each regressor ("Fixed"), consistent with the literature on identifying determinants of economic growth (Sala-i-Martin *et al.*, 2004). Further we show results for a BMS model with a uniform prior model ("Uniform") as well as with pre-defined prior inclusion probabilities using a Normal prior on the model size (number of included variables) centred around the mean number of candidate variables ("PIP"). BMS is implemented using the R-package 'BMS' (Zeugner and Feldkircher, 2015). As projection (and not inference on climate variables) is the primary aim of the present study, we do not focus on post-selection inference for gets and the LASSO, though bias-correction could be applied. Overfitting is not a major concern as none of the selection methods targets goodness of fit. Selection using 'gets' targets the false-positive rate of retention, meaning the spurious detection of variables is easily controllable using low significance levels, while the adaptive LASSO targets prediction and penalises the over-inclusion of variables. BMS allows us to identify which variables are important by assessing the posterior inclusion probabilities from the selection procedure, again where fit is not a target.

### Projection of the Damage Function

We use 170 model ensembles from the full CMIP5 range using RCP2.6, 4.5, 6.0 and 8.5[10] to project future climate for each retained climate variable (see Supplementary Material for an overview). We consider the global mean surface temperature anomaly to pre-industrial temperatures of each model ensemble in 2090-2099 as the target temperature for which impacts are computed. This allows us to abstract the impacts from particular emission scenarios and express them as functions of global mean surface temperature anomalies to pre-industrial temperatures instead (an approach used in Jackson et al., 2018). To adequately consider uncertainty in the estimation, we draw 100 coefficient samples from

[10] See Supplementary Methods for more information.

the estimated multi-variate Normal distribution using the estimated covariance matrix for our projections.

The country-level impacts that determine our final damage function shown in Figure 1 are computed by multiplying the sampled draws of the selected panel model coefficients with corresponding country- and year-specific mean-bias corrected climate variable anomalies relative to their 2000-2010 average as determined by CMIP5 simulations. For each projected model specification, we can therefore analyse several dozen million projection estimates, stemming from taking 100 coefficient draws, using 70 to 160 climate model ensemble runs depending on the selected variables and projecting 110 to 160 countries depending on the baseline for nearly nine decades (projections start in 2012 for SSP and in 2018 for MSW). Subsequently we compare the projected level of GDP per capita relative to its baseline (MSW or SSP) GDP per capita levels and take a country-coefficient specific mean across all annual values for the period 2090-2099. The impacts are then ordered according to the GMST anomaly to pre-industrial temperatures that the associated CMIP5 reaches in the period of 2090-2099. We then estimate a non-linear (quadratic) quantile regression over the distribution of impacts at each target temperature level (Figure 1) to derive our final damage function (see Supplementary Material Tables A13 and A14).

### Linking Damages to Cumulative Emissions

We use the near-linear relationship between cumulative emissions and changes in GMST as estimated by Goodwin *et al.* (2018). Here, we use their best estimate of 1.5 K per (1,000 PgC) to relate our projections to cumulative emissions (shown in the top x-axis in Figure 1).

### Adaptation

To determine the degree of historical adaptation we take two approaches. First, we estimate the selected model in two windows to assess the evolution of the climate-growth relationship over time. Second, we investigate the driver of change over time by assessing whether changes in income can explain parameter changes in the model.

Fixed Window Estimates

We re-estimate the models selected over the full sample from section 4.3 in two windows of 30 years (1962-1992 and 1981-2011).

Estimating Income-Climate Interactions to Assess Adaptation

We re-estimate the selected panel models where each retained climate variable is interacted by the observed log of GDP per capita in each given year for each country in the sample. To project this relationship to 2099 we use the in-sample coefficient estimates and constrain the degree of adaption to the highest in-sample income (across countries) to avoid out-of-sample projections. We note, of course, that this could underestimate eventual adaptation. We restrict growth impacts under adaptation to not exceed the baseline growth rate in absence of climate change for countries where the non-adaptation impact would otherwise be negative. We further restrict countries not to be worse-off under adaptation compared to no-adaptation. We then project future income levels using the MSW GDP per capita estimates under the above restrictions and estimate the resulting damage function using non-linear (quadratic) quantile regression over the distribution of impacts. As additional robustness checks, we also consider interactions with the first lag of GDP per capita (to address concerns about potential endogeneity), and also select over the full range of interactions of all possible climate variables and per-capita income (see Supplementary Material Section 8).

# 5 Supplementary Information

### Data accessibility

All code and data generated by the authors is available in a GitHub repository under the link https://github.com/moritzpschwarz/empirical_damage_curve. All other data is available from the referenced sources. Electronic supplementary material is available online.

### Authors' contributions

M.S. and F.P. conducted the statistical analysis, wrote the paper and produced the figures. M.S. constructed the climate model data and created the level and growth projections.

### Acknowledgments

We gratefully acknowledge funding by the Environmental Change Institute at the University of Oxford, the Clarendon Fund and the Robertson Foundation.

We thank Jesus Crespo Cuaresma, Luke Jackson, David F. Hendry, Simon Dietz, Jennifer Castle, Sam Rowan, Xiyu Jiao, Simona Sulikova, and Richard Tol for their helpful comments throughout the research process.

## Reference List

Acevedo, Sebastian; Mrkaic, Mico; Novta, Natalija; Pugacheva, Evgenia and Topalova, Petia (2020), The Effects of Weather Shocks on Economic Activity. What are the Channels of Impact?, *Journal of Macroeconomics*, p. 103207.

Andrijevic, Marina; Crespo Cuaresma, Jesus; Muttarak, Raya and Schleussner, Carl-Friedrich (2020), Governance in socioeconomic pathways and its role for future adaptive capacity, *Nature Sustainability*, Vol. 3 No. 1, pp. 35–41.

Barreca, Alan; Clay, Karen; Deschenes, Olivier; Greenstone, Michael and Shapiro, Joseph S. (2016), Adapting to Climate Change. The Remarkable Decline in the US Temperature-Mortality Relationship over the Twentieth Century, *Journal of Political Economy*, Vol. 124 No. 1, pp. 105–159.

Belloni, Alexandre and Chernozhukov, Victor (2013), Least squares after model selection in high-dimensional sparse models, *Bernoulli*, Vol. 19 No. 2, pp. 521–547.

Burke, Marshall; Davis, W. M. and Diffenbaugh, Noah S. (2018), Large potential reduction in economic damages under UN mitigation targets, *Nature*, Vol. 557 No. 7706, pp. 549–553.

Burke, Marshall and Emerick, Kyle (2016), Adaptation to Climate Change. Evidence from US Agriculture, *American Economic Journal: Economic Policy*, Vol. 8 No. 3, pp. 106–140.

Burke, Marshall; Hsiang, Solomon M. and Miguel, Edward (2015), Global non-linear effect of temperature on economic production, *Nature*, Vol. 527 No. 7577, pp. 235–239.

Calel, Raphael; Chapman, Sandra C; Stainforth, David A. and Watkins, Nicholas W. (2020), Temperature variability implies greater economic damages from climate change, *Nature communications*, Vol. 11 No. 1, p. 5028.

Campos, Julia; Hendry, David F. and Krolzig, Hans-Martin (2003), Consistent Model Selection by an Automatic Gets Approach*, *Oxford Bulletin of Economics and Statistics*, Vol. 65 s1, pp. 803–819.

Carleton, Tamma; Jina, Amir; Delgado, Michael; Greenstone, Michael; Houser, Trevor; Hsiang, Solomon; Hultgren, Andrew; Kopp, Robert; McCusker, Kelly; Nath, Ishan; Rising, James; Rode, Ashwin; Seo, Hee K; Viaene, Arvid; Yuan, Jiacan and Zhang, Alice T. (2020), Valuing the Global Mortality Consequences of Climate Change Accounting for Adaptation Costs and Benefits, Series: National Bureau of Economic Research, Cambridge, MA.

CIESIN (2016), Gridded Population of the World, Version 4 (GPWv4): Population Count: *http://dx.doi.org/10.7927/H4X63JVC. Accessed DAY MONTH YEAR.,* Center for International Earth Science Information Network - Columbia University - NASA Socioeconomic Data and Applications Center (SEDAC), Series: Palisades, NY.

Cui, Xiaomeng; Ghanem, Dalia and Kuffner, Todd (2018), On Model Selection Criteria for Climate Change Impact Studies.

Dell, Melissa; Jones, Benjamin F. and Olken, Benjamin A. (2012), Temperature Shocks and Economic Growth. Evidence from the Last Half Century, *American Economic Journal: Macroeconomics*, Vol. 4 No. 3, pp. 66–95.

Dietz, Simon; Rising, James; Stoerk, Thomas and Wagner, Gernot (2021), Economic impacts of tipping points in the climate system, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 118 No. 34.

Diffenbaugh, Noah S. and Burke, Marshall (2019), Global warming has increased global economic inequality, *Proceedings of the National Academy of sciences*, Vol. 116 No. 20, pp. 9808–9813.

Drijfhout, Sybren; Bathiany, Sebastian; Beaulieu, Claudie; Brovkin, Victor; Claussen, Martin; Huntingford, Chris; Scheffer, Marten; Sgubin, Giovanni and Swingedouw, Didier (2015), Catalogue of abrupt shifts in Intergovernmental Panel on Climate Change climate models, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 112 No. 43, E5777-86.

Fernández, Carmen; Ley, Eduardo and Steel, Mark F.J. (2001), Model uncertainty in cross-country growth regressions, *Journal of Applied Econometrics*, Vol. 16 No. 5, pp. 563–576.

Field, Christopher B. and Barros, Vicente R. (2014), Climate change 2014: *Impacts, adaptation and vulnerability / edited by Christopher B. Field, Working Group II Co-Chair, Department of Global Ecology, Carnegie Institution for Science, Vicente R. Barros, Working Group II Co-Chair, Centro de Investigaciones del Mar y la Atmósfera, Universidad de Buenos Aires [and 14 others],* Series: Cambridge University Press, New York, NY, Intergovernmental Panel on Climate Change. Working Group II; United Nations Environment Programme; World Meteorological Organization, ISBN: 978-1-107-05807-1.

Fischer, Erich M; Oleson, Keith W. and Lawrence, David M. (2012), Contrasting urban and rural heat stress responses to climate change, *Geophysical research letters*, Vol. 39 No. 3.

Friedman, Jerome; Hastie, Trevor and Tibshirani, Robert (2010), Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, Vol. 33 No. 1, pp. 1–22.

Goodwin, Philip; Katavouta, Anna; Roussenov, Vassil M; Foster, Gavin L; Rohling, Eelco J. and Williams, Richard G. (2018), Pathways to 1.5 °C and 2 °C warming based on observational and geological constraints, *Nature Geoscience*, Vol. 11 No. 2, pp. 102–107.

Hanson, Susan; Nicholls, Robert; Ranger, N; Hallegatte, S; Corfee-Morlot, J; Herweijer, C. and Chateau, J. (2011), A global ranking of port cities with high exposure to climate extremes, *Climatic Change*, Vol. 104 No. 1, pp. 89–111.

Hendry, David F. and Doornik, Jurgen A. (2014), Empirical model discovery and theory evaluation: *Automatic selection methods in econometrics,* Series: MIT Press, ISBN: 0262324423.

Hendry, David F; Johansen, Søren and Santos, Carlos (2008), Automatic selection of indicators in a fully saturated regression, *Computational Statistics*, Vol. 23 No. 2, pp. 337–339.

Jackson, Luke P; Grinsted, Aslak and Jevrejeva, Svetlana (2018), 21st Century Sea-Level Rise in Line with the Paris Accord, *Earth's Future*, Vol. 6 No. 2, pp. 213–229.

Jiao, Xiyu; Pretis, Felix and Schwarz, Moritz (2021), Testing for Coefficient Distortion due to Outliers with an Application to the Economic Impacts of Climate Change, *SSRN Electronic Journal*.

Kahn, Matthew E; Mohaddes, Kamiar; Ng, Ryan N.C; Pesaran, M. H; Raissi, Mehdi and Yang, Jui-Chung (2019), Long-term macroeconomic effects of climate change: *A cross-country analysis,* Series: National Bureau of Economic Research, ISBN: 0898-2937.

Matsuura, Kenji and Willmott, Cort J. (2018), Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1901 - 2017): *Version 5. Data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA,* University of Delaware, Series: available at: https://www.esrl.noaa.gov/psd/ (accessed 26 February 2019).

Miller, J. I. and Park, Joon Y. (2010), Nonlinearity, nonstationarity, and thick tails. How they interact to generate persistence in memory, *Journal of Econometrics*, Vol. 155 No. 1, pp. 83–89.

Moriondo, M; Giannakopoulos, C. and Bindi, M. (2011), Climate change impact assessment. The role of climate extremes in crop yield simulation, *Climatic Change*, Vol. 104 3-4, pp. 679–701.

Müller, Ulrich K; Stock, James H. and Watson, Mark W. (2020), An Econometric Model of International Growth Dynamics for Long-horizon Forecasting.

Murage, Peninah; Hajat, Shakoor and Kovats, R. S. (2017), Effect of night-time temperatures on cause and age-specific mortality in London, *Environmental Epidemiology*, Vol. 1 No. 2, e005.

Obradovich, Nick; Migliorini, Robyn; Mednick, Sara C. and Fowler, James H. (2017), Nighttime temperature and human sleep loss in a changing climate, *Science advances*, Vol. 3 No. 5, e1601555.

Pindyck, Robert S. (2013), Climate Change Policy. What Do the Models Tell Us?, *Journal of Economic Literature*, Vol. 51 No. 3, pp. 860–872.

Pretis, Felix; Reade, J. J. and Sucarrat, Genaro (2018a), Automated General-to-Specific (GETS) Regression Modeling and Indicator Saturation for Outliers and Structural Breaks, *Journal of Statistical Software*, Vol. 86 No. 3.

Pretis, Felix; Schwarz, Moritz; Tang, Kevin; Haustein, Karsten and Allen, Myles R. (2018b), Uncertain impacts on economic growth when stabilizing global temperatures at 1.5°C or 2°C warming, *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, Vol. 376 No. 2119.

Riahi, Keywan; van Vuuren, Detlef P; Kriegler, Elmar; Edmonds, Jae; O'Neill, Brian C; Fujimori, Shinichiro; Bauer, Nico; Calvin, Katherine; Dellink, Rob; Fricko, Oliver; Lutz, Wolfgang; Popp, Alexander; Cuaresma, Jesus C; KC, Samir; Leimbach, Marian; Jiang, Leiwen; Kram, Tom; Rao, Shilpa; Emmerling, Johannes; Ebi, Kristie; Hasegawa, Tomoko; Havlik, Petr; Humpenöder, Florian; Da Silva, Lara A; Smith, Steve; Stehfest, Elke; Bosetti, Valentina; Eom, Jiyong; Gernaat, David; Masui, Toshihiko; Rogelj, Joeri; Strefler, Jessica; Drouet, Laurent; Krey, Volker; Luderer, Gunnar; Harmsen, Mathijs; Takahashi, Kiyoshi; Baumstark, Lavinia; Doelman, Jonathan C; Kainuma, Mikiko; Klimont, Zbigniew; Marangoni, Giacomo; Lotze-Campen, Hermann; Obersteiner, Michael; Tabeau, Andrzej and Tavoni, Massimo (2017), The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications. An overview, *Global Environmental Change*, Vol. 42, pp. 153–168.

Ricke, Katharine; Drouet, Laurent; Caldeira, Ken and Tavoni, Massimo (2018), Country-level social cost of carbon, *Nature Climate Change*, Vol. 8 No. 10, pp. 895–900.

Roughgarden, Tim and Schneider, Stephen H. (1999), Climate change policy. Quantifying uncertainties for damages and optimal carbon taxes, *Energy Policy*, Vol. 27 No. 7, pp. 415–429.

Sala-i-Martin, Xavier; Doppelhofer, Gernot and Miller, Ronald I. (2004), Determinants of Long-Term Growth. A Bayesian Averaging of Classical Estimates (BACE) Approach, *American Economic Review*, Vol. 94 No. 4, pp. 813–835.

Schlenker, Wolfram and Roberts, Michael J. (2009), Nonlinear temperature effects indicate severe damages to US crop yields under climate change, *Proceedings of the National Academy of sciences*, Vol. 106 No. 37, pp. 15594–15598.

Sherwood, Steven C. and Huber, Matthew (2010), An adaptability limit to climate change due to heat stress, *Proceedings of the National Academy of sciences*, Vol. 107 No. 21, pp. 9552–9555.

Sillmann, J; Kharin, V. V; Zhang, X; Zwiers, F. W. and Bronaugh, D. (2013a), Climate extremes indices in the CMIP5 multimodel ensemble. Part 1. Model evaluation in the present climate, *Journal of Geophysical Research: Atmospheres*, Vol. 118 No. 4, pp. 1716–1733.

Sillmann, J; Kharin, V. V; Zwiers, F. W; Zhang, X. and Bronaugh, D. (2013b), Climate extremes indices in the CMIP5 multimodel ensemble. Part 2. Future climate projections, *Journal of Geophysical Research: Atmospheres*, Vol. 118 No. 6, pp. 2473–2493.

Stern, Nicholas (2008), The Economics of Climate Change, *American Economic Review*, Vol. 98 No. 2, pp. 1–37.

Stern, Nicholas (2016), Economics. Current climate models are grossly misleading, *Nature*, Vol. 530 No. 7591, pp. 407–409.

Taylor, Karl E; Stouffer, Ronald J. and Meehl, Gerald A. (2012), An Overview of CMIP5 and the Experiment Design, *Bulletin of the American Meteorological Society*, Vol. 93 No. 4, pp. 485–498.

Tol, Richard S.J. (2017), Population and trends in the global mean temperature, *Atmósfera*, Vol. 30 No. 2, pp. 121–135.

Tol, Richard S.J. (2020), The economic impact of weather and climate, *Preprint submitted to Elsevier*.

Weitzman, Martin L. (2012), What is the "Damages Function" for Global Warming - and what Difference might it make?, *Climate Change Economics*, Vol. 01 No. 01, pp. 57–69.

World Bank (2019), World Development Indicators: *WDI,* Series: available at: http://databank.worldbank.org/data/download/WDI_excel.zip.

Xu, Chi; Kohler, Timothy A; Lenton, Timothy M; Svenning, Jens-Christian and Scheffer, Marten (2020), Future of the human climate niche, *Proceedings of the National Academy of sciences*, Vol. 117 No. 21, pp. 11350–11355.

Zeugner, Stefan and Feldkircher, Martin (2015), Bayesian Model Averaging Employing Fixed and Flexible Priors. The BMS Package for R, *Journal of Statistical Software*, Vol. 68 No. 4, pp. 1–37.

Zhao, Sen; Witten, Daniela and Shojaie, Ali (2021), In defense of the indefensible. A very naive approach to high-dimensional inference, *Statistical Science*, Vol. 36 No. 4, pp. 562–577.

Zou, Hui (2006), The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association*, Vol. 101 No. 476, pp. 1418–1429.

# Paper 3: Finding what nobody records: A Proof of Concept study to identify Weather Impact Events using machine learning

*Paper 3: Finding what nobody records: A Proof of Concept study to identify Weather Impact Events using machine learning*

# Finding what nobody records: A Proof of Concept study to identify Weather Impact Events using machine learning

Moritz Schwarz[1,2*]

[1]Climate Econometrics, Nuffield College, University of Oxford
[2]School of Geography and the Environment, University of Oxford

**Abstract**

Current bottom-up disaster databases are inadequate to provide a full picture of the true number of weather impact events across the world and especially in developing countries. This is largely driven by constrained local capacity to record and track weather impact events, which makes adequate infrastructure planning and adaptation measures more difficult. Here, I use existing disaster databases and intersect them with meteorological, geophysical, and socio-economic data. Using machine learning algorithms, I am then able to recover the recorded classification in the databases for the occurrence of drought, landslide, storm, flood, and extreme temperature events with at least 89% accuracy out of sample and without any on-the-ground knowledge. By comparing my predictions with the data recorded in the EM-DAT/GDIS database, I identify a total of 21,651 grid-cell events that have not been recorded in these databases. To illustrate the significance of these events, I estimate that existing disaster databases have underestimated the true global death toll by up to 1,57 million deaths and could have missed up to 7.6 trillion US$ in total damages over the period of 2010 – 2018. While there remain numerous challenges in this literature and a number of improvements will be made in the near future, this revelation re-emphasizes the gap between reported weather impacts and the true societal impacts which must be taken into account when attempting to solve the climate crisis in an equitable manner.

## 1 Introduction

The exacerbating societal impacts of climate change are becoming more and more relevant for national and international climate policy making. While international climate negotiations have recently agreed to establish a Loss and Damage Fund for developing countries at COP27, consistently and holistically assessing the occurrence of such Loss and Damage events is much more difficult. This political development occurs while the demand for reliable data for weather impact and disasters is growing both for policy planning objectives as well as for industries like insurance (Panwar & Sen, 2020). While numerous approaches have been developed to illustrate the on the ground impact that climate change can have on societal dynamics, the perhaps most frequently used approach focuses on recording disasters and weather impact events in a bottom-up manner

*Corresponding author: moritz.schwarz@ouce.ox.ac.uk

to collect the results in databases like the international Emergency Events Database (EM-DAT), Sigma by the insurance company SwissRe, or NatCat from the insurance company MunichRe, the international Natural Hazards Assessment Network (NATHAN), or the US based Spatial Hazard Events and Losses Database for the United States (SHELDUS), or the National Weather Service's Storm Events Database (Panwar & Sen, 2020; Gall et al., 2009).

However, these bottom-up approaches are likely to greatly underestimate the true incidence of weather impact events and suffer from various inconsistencies (Panwar & Sen, 2020). Based on one of the most prominent disaster databases, EM-DAT (EM-DAT, 2020), heat waves have killed just 71 people in Sub-Saharan Africa from 1900 to 2019. The same database suggests that European heatwaves since 1980 have resulted in more than 140,000 deaths and more than US$12 billion in damages (Harrington & Otto, 2020). Should we therefore conclude that heatwaves in Sub-Saharan Africa are less of a concern? A more likely explanation is that existing databases - which rely heavily on the manual recording of data - fail to register most socio-economic impacts of extreme weather and climate events, which could prove catastrophic in the face of more frequent events caused by future climate change. This lack of data on important societal impacts hinders policy makers in their responses and likely biases estimates of the economic impacts of future climate change. Furthermore, there is a large literature on the lack of comparable socio-economic data that these databases suffer from (see e.g. Edmonds & Noy, 2018; Gall et al., 2009; Moriyama et al., 2018; Guha-Sapir & Below, 2002; Panwar & Sen, 2020).

Felbermayr & Gröschl (2014) show that a country's GDP per capita has an effect on the likelihood of an event being recorded in a database like EM-DAT, which highlights the issue of providing reliable data for developing countries even more. Similar concerns with EM-DAT motivated Felbermayr & Gröschl (2014) to produce an independent database called initially GeoMet and later renamed to ifo GAME, although it was, to my knowledge, never updated beyond 2010. In the context of this literature, the UN Sendai Framework for Disaster Risk Reduction called for the establishment of a "systematic and comparable disaster database to help build resilience against natural disasters" (Panwar & Sen, 2020, p. 296) and established the DesInventar database – a benchmark data collection tool that is not covering the entire globe though.

While the difficulty of identifying weather impacts and disasters has been well understood for at least two decades, the COVID-19 pandemic has, notwithstanding the terrible societal consequences of the pandemic, shined a light on the potential value of using near-real-time data to tailor policy responses for recovery measures (Chetty et al., 2020). In a similar spirit, reliable data is also essential in the context of climate change to guide optimal adaptation investments and appropriate disaster responses, especially in developing countries.

In this paper, I show that recent advances in machine learning and econometric modelling can provide one avenue of progress for illustrating the true societal effects of climate change in an effort to improve my adaptation measures against weather impacts. I do so by combining several existing data sets on disasters and combine them with meteorological and societal data to attempt to predict the occurrence of weather impact events. This paper therefore attempts to use these methods to supplement rather than replace manual efforts to understand the socio-economic impacts of extreme

weather and climate.

The remainder of this paper is structured as follows. First, I introduce the conceptual approach of the paper and provide details on the methodology used in section 2. Subsequently, I provide details on the data used in section 3, before I present and discuss the results in section 4. In the section 5 I conclude. My overall results show the immense potential of machine learning in providing more holistic disaster event data.

## 2 Conceptual Approach and Methods

### 2.1 Conceptual Approach

This paper combines methods from economics and statistics in an effort to improve the global record of disasters and weather impacts. Using various machine learning algorithms, I predict the occurrence of five disaster categories. These are the occurrence of drought, extreme temperature, flood, landslides, and storms.

Classification of socio-economic impacts of weather events will be formulated as a function of observed weather variables, geographical features (such as elevation), and socio-economic characteristics (such as population distribution).

Conceptually, I formulate a model resembling:

$$Event_{c,t,j} = f(Weather_{c,t,j}, Physical\ Features_{c,l}, SocioEconomic\ Features_{c,m}) \quad (1)$$

where *Event* is a 0/1 dummy to indicate whether a weather impact event of type $j$ is taking place at time $t$ in grid cell $c$. It is worth further elaborating on the definition of these disaster and weather events. I define a key characteristic of those events to be temporally and spatially-distinct disruptions of society due to prevailing weather conditions. This, in turn, clarifies that the mere meteorological occurrence of extreme weather does not constitute a weather impact event e.g., if that weather occurs in remote or uninhabited locations or if a community is well prepared for such conditions and no societal disruption occurs or no costs are evident. This is consistent with the collection logic of the input databases considered. $Weather$ represents a number of meteorological features (e.g. several different ERA5 variables represented by $j$) at the relevant time (as well as some preceding time periods) in the relevant grid cell (as well as in neighbouring cells). The different physical and socio-economic features $l$ and $m$ are constant over the time frame considered here.

Crucially, finding the optimal classification function $f()$, and indeed one that achieves high precision predictions, is the goal of this paper. To this end, I utilize historic weather impact event data and use machine-learning classification algorithms in various specifications.

### 2.2 Positive-Unlabelled Problems

Given my input data, I can be reasonably confident that an event has occurred when an event has been recorded – the likelihood of a false-positive in the input data is relatively small (EM-DAT, 2020; Panwar & Sen, 2020; Jones et al., 2022). However, the absence of an event, given all documented issues with conventional disaster databases (Panwar & Sen, 2020), does not necessarily constitute a negative i.e. proof that no weather impact event has occurred. Therefore, postulating that these situations are true negatives might not be appropriate. In other words, I have data that is labelled as "positive" and data that is simply unlabelled (Bekker et al., 2019). Conceptually, my study is therefore formulated as a classification problem of positive and unlabelled (PU) data where my goal is to correctly assign each observation to the right classification. I do not strive to provide estimates of weather impacts for causal inference or attribution – rather I aim to test whether this machine-learning approach could viably predict the occurrence of these events.

Therefore when using the PU approach, it is important to note that all estimated models are trained on a training dataset in which the set of non-events contains both true situations of non-events but also events where an event has actually taken place. For this reason, I expect any of my estimates to be an underestimate of the true effect. This effect is exacerbated by the inherent selection bias that is pertinent in the weather impact databases due to the on-the-ground capacities to record such events. Given this concern, I must assume that there is substantial bias in the labelling of the data – with weather event impacts in socio-economically wealthy areas much more likely to be identified and recorded. While this fact is part of the motivation for this study, below I discuss several methods that attempt to alleviate these concerns.

While this literature is relatively new, there are a few proposed approaches (see e.g. Bekker & Davis, 2020, for an overview on PU classification and learning) and a few applications that can be compared to this problem at hand (McDonald et al., 2021).

### 2.3 Implementation and Algorithms

My approach follows these studies conceptually and is implemented as follows: First, training traditional classification algorithm using positive cases assuming that unlabelled are negative. In my context this can be feasibly regarded as mostly accurate within regions where there is sufficient reporting e.g. if a number of weather impact events within a highly developed country are reported and included in a disaster database, I can perhaps assume that all non-events (at least using a consistent event definition) in this country are truly negatives. Second, biased learning where unlabelled events are coded as negative, albeit with specific noise. Third, taking the first approach estimating models on positive cases, but then adjusting the resulting classification probabilities using baseline estimates of the probability of being positive.

Here, I operationalise this approach and test different machine learning algorithms and techniques. Overall, I define four different modelling algorithms (labelled A to D). I use all four modelling approaches used by McDonald et al. (2021) and follow their approach closely. I use A) a

traditional naïve classifier using a Support Vector Machine (SVM) radial basis function algorithm as well as B) an SVM that is trained using a biased learning approach using data bagging (see below). I further use a Random Forest (RF) approach, both C) in a naïve way as well as D) using the RF algorithm with a similar bias learning bagging approach. For the RF approach, a parameter setting the number of 'decision trees' must be chosen – with each of those decision trees drawing a different estimation sample and resulting in a single classification (see below for a discussion on the appropriate number of trees). For all algorithms, I use the statistical programming package R and in particular the packages tidymodels, kernlab, and ranger (Kuhn & Wickham, 2020; Karatzoglou et al., 2004; Wright & Ziegler, 2017).

The bagging approach is essentially a bootstrapping technique that resamples the overall sample with replacement. An individual bag is therefore a single draw in this resampling process. In all estimated models, all positive cases are contained in the sample while the unlabelled cases are downsampled. The downsampling ratio is a hyperparameter and is set as an integer ranging from 1 to 5 and refers to the relative occurrence of positive and unlabelled cases. In the case of a downsampling ratio of 1, positive and unlabelled cases are equal – in the case of a downsampling ratio of 5, there are 5 times as many unlabelled cases. As is common with boostrapping techniques, results are then averaged across bags i.e. across different samples (McDonald et al., 2021; Mordelet & Vert, 2014). Models B and D are estimated using up to 100 bags.

While Boehmke & Greenwell (2019) suggest to initially estimate models using the rule of thumb of 10 trees per predictor, which in this case would result in 220 trees, I formulate Random Forest models with 100 trees due to computational capacity constraints for Models C and D. However, to alleviate concerns of not using sufficient trees in the random forest estimation, the final models were run using 1000 trees as well without a substantial accuracy loss, see Figure 3. For the same reason, Model A was eventually removed as a viable model, as models were unable to be estimated even on advanced high performance computing systems due to their memory requirements. All other hyperparameters for the machine learning algorithms are used in their default setting.

While much of this approach is inspired by the conceptual estimation set-up of the PNAS study by McDonald et al. (2021), there are a number of differences that are worth stressing. Firstly, my dataset is much larger than and secondly my share of positive cases is much higher than the one of the study authors – with my positive prevalence ranging from 0.5% to 7.2% of the data, see Table 1, compared to a prevalence of 0.1% in the study of McDonald et al. (2021). Given that my dataset is sufficiently large, I am able to reserve 20% of my data as testing data. I follow what Swartz et al. (2021), in their critique of the PNAS study, calls the"golden rule of machine learning" that test data cannot influence the training of the model by dividing my data using a training/testing split of 80%/20%. I are therefore able to evaluate the performance of my selected algorithm out-of-sample. Nonetheless, I also use the 10-fold cross-validation (CV) technique used by the same authors in their study. Using the CV results, I am able to tune the three relevant hyperparameters: number of bags, downsampling ratio (both of these for models B and D) as well as the cut-off value where a case would be considered to be a positive case. I follow McDonald et al. (2021) and Lee & Liu (2003) in maximizing the mean of the modified F1 score while aiming for a small standard deviation and a parsimonious model formulation.

## 3 Data

My training data consists of the Geocoded Disasters (GDIS) dataset, v1 (1960 -– 2018) (Rosvold & Buhaug, 2021), which is made up of 39,953 locations for 9,924 disasters that occurred worldwide for the years 1960 to 2018. I subset this data to the timeframe of January 2010 to December 2018 as the recorded disasters before this time are only recorded on an annual level and merge it with the EM-DAT database using their unique disaster ID to supplement the GDIS database with further information on each event (EM-DAT, 2020).

I combine this data with various weather indicators from the ERA5 reanalysis data (Hersbach et al., 2020), such as temperature (the monthly mean of daily mean temperature, the monthly maximum of daily maximum temperature, and the monthly minimum of daily minimum temperatures), precipitation (total monthly precipitation, the monthly mean of daily total precipitation, the maximum daily total precipitation, and the minimum of daily total precipitation) as well as an indicator for the number of dry days each month (i.e. days where total daily precipitation is equal to 0). Those variables are entered in their contemporaneous form as well as with a time lag of up to three months. I also add geographical features, such as elevation (Hollister et al., 2021) using the 'AWS Terrain Tiles' option, which uses a global dataset of terrain heights assembled by Amazon Web Services (Larrick et al., 2020) and socio-economic indicators such as population (CIESIN, 2016, using version 4 Revision 11). I further add time specific features, such as the year, month, and day of the week where the event occurred alongside a fixed effect for the precise month-year combination.

I then subset all data sets to exclude any non-populated areas (i.e. grid cells with a population of 0), including any non-land grid cells using the outlines of the Natural Earth dataset at 110m resolution without Antarctica as a land-area mask (South, 2017).

Combining the entirety of the information laid out above, I create a geospatial data series on a $2.5° \times 2.5°$ resolution that results in a total of up to 266,868 observations (108 months from January 2010 to December 2018 with 2,471 land-based grid-cells). An example slice for the data considered is contained in Figure 1 and a set of summary statistics is contained in Table 1 and for all predictors in Table 3.

Before estimating the models described above in section 2, I pre-process the relevant data by removing all predictors with a near-zero variance and all predictors with a correlation coefficient exceeding 0.75 (as in McDonald et al., 2021). I then center to a mean of 0 and scale all predictors to a standard deviation of 1 and remove any remaining missing values.

| | Mean | SD | Min | Max | N |
|---|---|---|---|---|---|
| Drought | 0.053 | 0.224 | 0.000 | 1.000 | 266,868 |
| Flood | 0.100 | 0.300 | 0.000 | 1.000 | 266,868 |
| Landslide | 0.007 | 0.085 | 0.000 | 1.000 | 266,868 |
| Storm | 0.058 | 0.234 | 0.000 | 1.000 | 266,868 |
| Extreme Temperature | 0.045 | 0.208 | 0.000 | 1.000 | 266,868 |

Table 1: Summary Statistics for the used data.

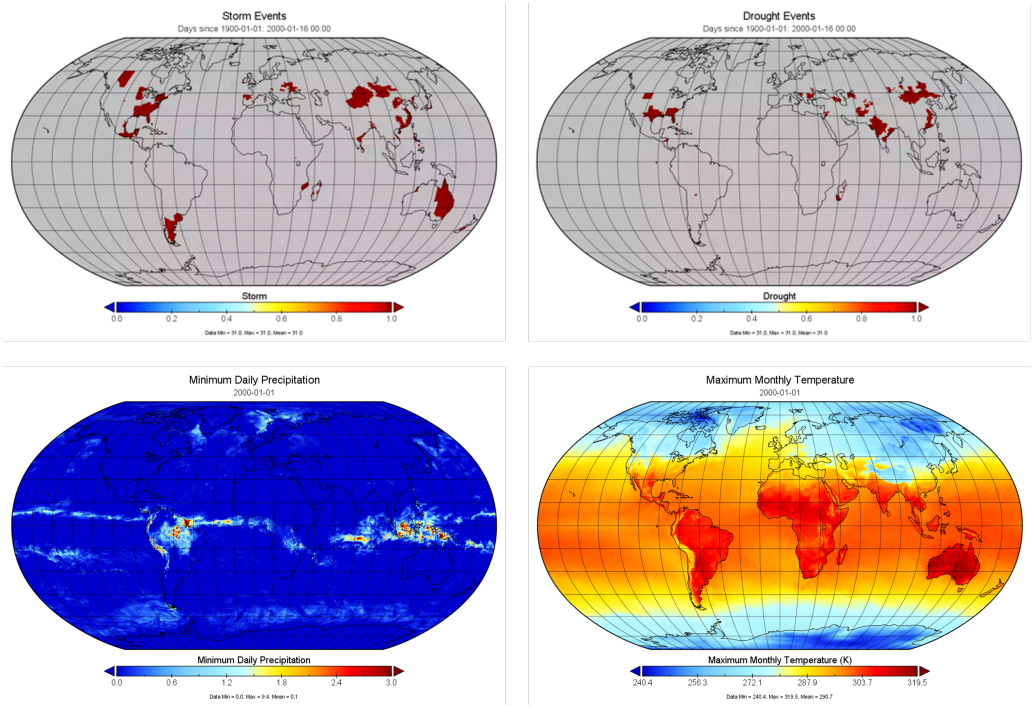


Figure 1: Example data for selected variables for January 2000.

## 4 Results and Discussion

The algorithm was tested in up to 1,002 specifications. The performance of the algorithm is displayed in Appendix Figures 2 for the drought case (the other events are displayed in Appendix section B. These performance comparisons clearly show that Model C, the naïve random forest classifier performs best, as the modified F1 is significantly higher than all other specifications. While the SVM results seem to perform better in recalling true positives, the SVM models seem to suffer from many cases of false-positives, evident in their higher detection prevalence, which in turn reduces their modified F1 statistic. The bagging approach does also not seem to increase model performance – neither across bags nor across an increasing downsampling ratio. Hence, the models used and compared below all follow the approach of Model C.

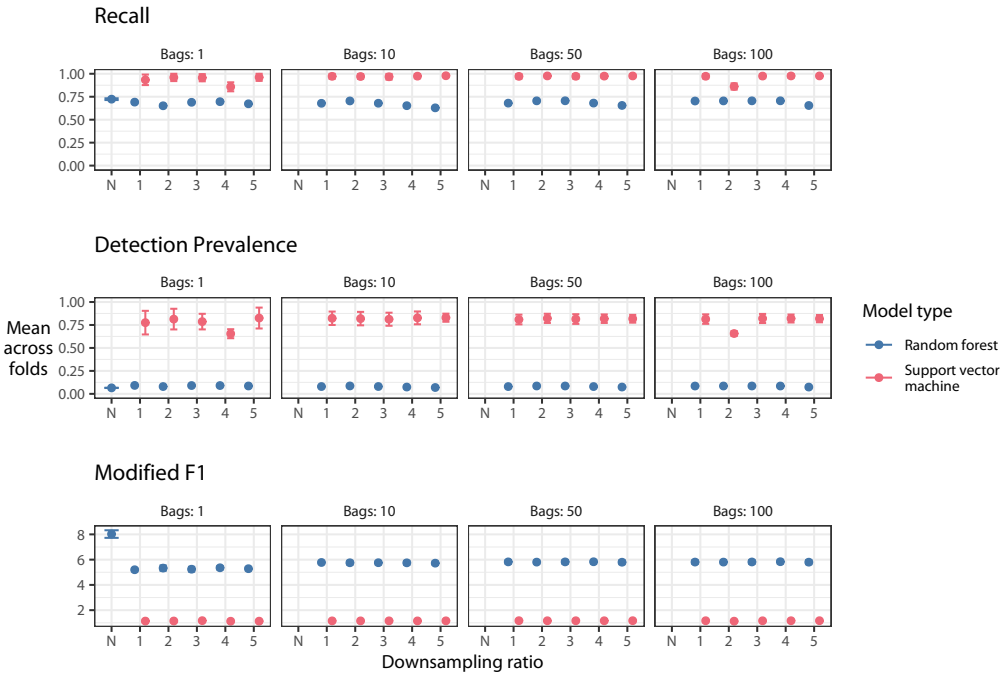


Figure 2: Algorithm performance comparison for Drought Events. N denotes the naïve classifier. Note that the y-axis scale varies for each sub-plot and is bounded from 0 to 1 for the two upper plot rows.

While the demand for high quality disaster data is higher than ever, traditional bottom-up databases are failing to provide high quality, reliable data that can be used by policy makers. My results show the immense potential that these methods can play in improving weather impact estimation, see Figure 3. Across different event types, all event types display a very high prediction accuracy; generally exceeding at least 89%, which has greatly exceeded the authors expectations.

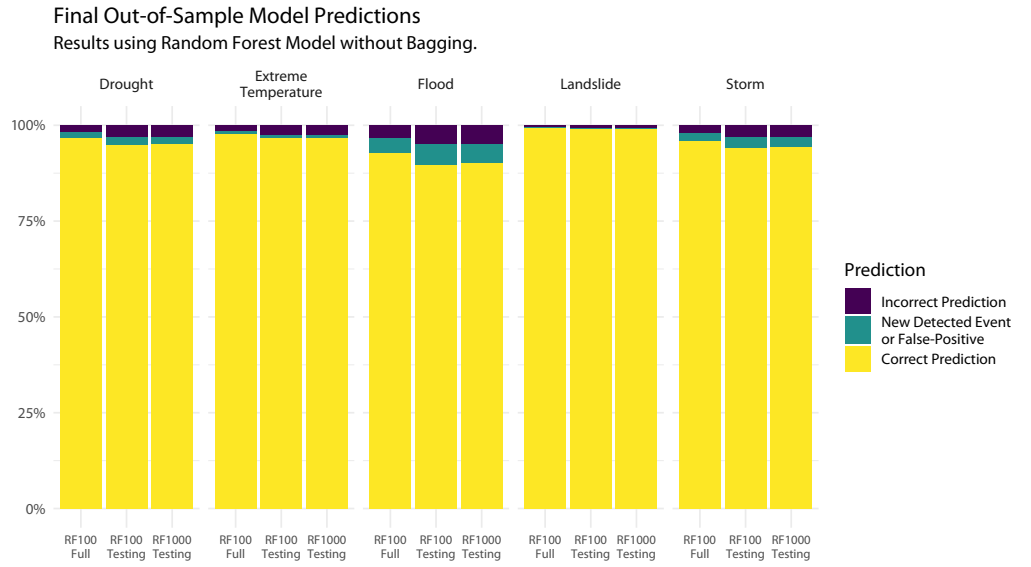


Figure 3: Final prediction results for Model C with different samples and different numbers of Random Forest Trees. RF100 stands for a Random Forest Model of type C run with 100 trees. RF1000 stands for the same model with 1000 trees. The distinction Full and Testing stands for the sample that was used for prediction. Testing therefore refers to an out-of-sample prediction.

The highest out-of-sample prediction accuracy can be achieved looking at landslide events, while flood event predictions are relatively the worst when simply considering the input data as being correct. Using a 20% out of sample prediction, the chosen algorithm correctly predicts the occurrence of an event with 99% and 89.7% accuracy respectively, with the other disaster types falling within that range (see Figure 3 Table 2 for the full results). While determining the importance of individual variables, especially in a causal sense, is neither the objective of this study nor a suitable outcome of many prediction algorithms, I present the variable importance of the final prediction algorithm in Appendix section C.

As the objective of this study was to detect previously undetected weather impact events, given my selected algorithm, I estimate that unlabelled events have taken place in up to 5,5% of cases when looking at flood events, but only in about 0.4% of cases for landslides. Naturally, this figure might, however, also include false-positives.

Table 2: Full Prediction Results.

| | | | Correct Positive | | Correct Unlabeled | | Incorrect | | New Events or False Positive | |
|---|---|---|---|---|---|---|---|---|---|---|
| Event Type | RF Trees | Sample | | | | | | | | |
| Drought | 100 | Testing | 2.91% | 1,442 | 91.98% | 45,562 | 2.87% | 1,422 | 2.23% | 1,107 |
| Drought | 100 | Full | 4.14% | 10,239 | 92.72% | 229,522 | 1.57% | 3,897 | 1.57% | 3,878 |
| Drought | 1000 | Testing | 2.72% | 1,349 | 92.36% | 45,735 | 2.99% | 1,481 | 1.93% | 955 |
| Extreme Temperature | 100 | Testing | 2.56% | 1,267 | 94.07% | 46,585 | 2.38% | 1,179 | 0.99% | 489 |
| Extreme Temperature | 100 | Full | 3.31% | 8,189 | 94.46% | 233,826 | 1.55% | 3,845 | 0.68% | 1,676 |
| Extreme Temperature | 1000 | Testing | 2.56% | 1,267 | 94.17% | 46,631 | 2.38% | 1,179 | 0.89% | 443 |
| Flood | 100 | Testing | 5.91% | 2,926 | 83.82% | 41,508 | 4.81% | 2,384 | 5.46% | 2,702 |
| Flood | 100 | Full | 7.56% | 18,710 | 85.09% | 210,640 | 3.18% | 7,880 | 4.16% | 10,306 |
| Flood | 1000 | Testing | 5.94% | 2,943 | 84.20% | 41,694 | 4.79% | 2,373 | 5.07% | 2,510 |
| Landslide | 100 | Testing | 0.22% | 109 | 98.78% | 48,915 | 0.56% | 277 | 0.44% | 219 |
| Landslide | 100 | Full | 0.47% | 1,153 | 98.83% | 244,646 | 0.31% | 774 | 0.39% | 963 |
| Landslide | 1000 | Testing | 0.22% | 111 | 98.85% | 48,950 | 0.56% | 275 | 0.37% | 184 |
| Storm | 100 | Testing | 3.08% | 1,523 | 91.21% | 45,167 | 3.03% | 1,501 | 2.68% | 1,329 |
| Storm | 100 | Full | 4.26% | 10,549 | 91.80% | 227,249 | 1.98% | 4,910 | 1.95% | 4,828 |
| Storm | 1000 | Testing | 3.03% | 1,502 | 91.47% | 45,297 | 3.08% | 1,525 | 2.42% | 1,196 |


Figure 4: Estimated Impacts for the events identified in this study but missing from EM-DAT/GDIS. The different values represent different comparison measures considered. When excluding all events in EM-DAT/GDIS that do not have a value recorded for the three presented values, a mean/median is calculated for the remaining values. I also present one calculation metric where I form the mean across all events, assuming that each event with a missing value features a true value of 0.

When assuming that all of the identified events are true positives, however, and by using a simplified comparison metric, namely the average death count, average number of people affected, and the total economic cost of a single event in the EM-DAT database (see also Table 4), I estimate that conventional databases under-report the effects of approximately between 67,000 and 1,57 million deaths and under-report the damages associated with extreme events by approximately between 0.93 to 7.62 trillion US$ in weather impact damages and has affected between 452 million and 6.7 billion people across the entire time period considered, see Figure 4. The total number of affected people can include individuals being affected more than once. The calculation was adjusted for the average grid cell size of a unique EM-DAT event while the given range of values depends on the calculation measure used, with the median effect of an EM-DAT event providing the lower bound and the mean effect per EM-DAT event providing an upper bound .[1]

The insights gained using my approach could allow policy makers to make more informed decisions on weather and climate adaptation and mitigation policy, both locally as well as globally. Using and improving upon my methodological framework could be crucial for providing short- and

[1]This average grid cell size was obtained by comparing the number of unique events for each disaster type in the EM-DAT database with the number of unique positive grid cell dummies in the input estimation data.

medium-range forecasts of socio-economic impacts of extreme weather and climate events, given reliable weather forecasts. This could allow authorities to improve disaster response preparation e.g. in the context of hurricane or storm predictions, ultimately improving humanitarian assistance.

While the approaches presented in this paper never claim causality and need to face up to legitimate scrutiny, such as by Swartz et al. (2021), my paper highlights that the advances in machine learning have immense potential to provide crucial disaster data reliably. I have also shown that the issue of few labelled cases and a generally small number of positive cases is not necessarily an issue for advanced machine learning algorithms.

Considering the results presented in this paper, I note several avenues for improving upon the estimation presented here. Firstly, I have only considered four different model specifications across Random Forest and SVM models. Several other machine learning algorithms could be considered in the future. Incidentally, the fact that Model C, the naïve Random Forest classifier, a fairly common classifier, proved superior to all other algorithms to us suggests that further work on algorithm choice could provide further progress in increasing prediction accuracy. Due to computational constraints, my models were estimated at a grid resolution of $2.5° \times 2.5°$ at a monthly frequency, although my data can also be used at a $1° \times 1°$ resolution and at a daily frequency. Lastly, further data, such as satellite data, should be added to the predictor variables to improve results even further (Rolf et al., 2021).

For these results, two concrete follow-up processes are envisaged. Firstly, the EM-DAT input database should be complemented by the DesInventar database, which, similarly misses large areas of the world, but is a good complement. Secondly, the identified additional disaster events will be analysed further using text mining and manual inspection to confirm whether these events have actually occurred, following Walker et al. (2019).

## 5 Conclusion

Current bottom-up disaster databases are inadequate to provide a full picture of the true number of weather impact events across the world and especially in developing countries. This is largely driven by constrained local capacity to record and track weather impact events, which makes adequate infrastructure planning and adaptation measures more difficult.

In this paper, I combine two databases on weather impact events and intersect this data with meteorological, geophysical, and socio-economic data. Using machine learning, I am then able to recover the recorded classification in the database for the occurrence of drought, landslide, storm, flood, and extreme temperature events with at least 89% accuracy out of sample.

By comparing my predictions with the data recorded in the EM-DAT/GDIS database, I identify a total of 21,651 grid-cell events that have not been recorded in these databases. To illustrate the significance of these events, I estimate that existing disaster databases have underestimated the true global death toll by up to 1,57 million deaths and could miss up to 7.6 trillion US$ in total

damages over the period of 2010 – 2018. While there remain numerous challenges in this literature and a number of improvements will be made in this literature in the near future, this revelation re-emphasizes the gap between reported weather impacts and the true societal impacts which must be taken into account when attempting to solve the climate crisis in an equitable manner. While of course not all identified and recorded events in this context constitute Loss and Damage events as they are understood within the UN Framework Convention on Climate Change (UNFCCC), the potential of the presented methods to support identification of such instances could be immense and hence clearly warrants further research into this field.

## Acknowledgements

I gratefully acknowledge funding from the Canadian Social Sciences and Humanities Research Council through their Insight Development Grant.

## Conflicts of Interest

The author declares no conflicts of interest.

## Data and Code Availability

Both the data and the code used in this study are available from github.com/moritzpschwarz/weather-impact-monitor.

## References

Bekker, J., & Davis, J. (2020). Learning from positive and unlabeled data: A survey. *Machine Learning*, *109*(4), 719–760.

Bekker, J., Robberechts, P., & Davis, J. (2019). Beyond the selected completely at random assumption for learning from positive and unlabeled data. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 71–85).

Boehmke, B., & Greenwell, B. (2019). *Hands-on machine learning with r*. Chapman and Hall/CRC.

Chetty, R., Friedman, J. N., Hendren, N., & Stepner, M. e. a. (2020). Real-time economics: A new platform to track the impacts of covid-19 on people, businesses, and communities using private sector data. *NBER Working Paper*, *27431*, 36–46.

CIESIN, F. (2016). Gridded population of the world, version 4 (gpwv4): population count grid. *Center for International Earth Science Information Network (CIESIN), Columbia University*.

Edmonds, C., & Noy, I. (2018). The economics of disaster risks and impacts in the pacific. *Disaster Prevention and Management: An International Journal*.

EM-DAT. (2020). Em-dat: The ofda/cred international disaster database. *Centre for Research on the Epidemiology of Disasters, Universidad Católic a de Lovaina, Bruselas*.

Felbermayr, G., & Gröschl, J. (2014). Naturally negative: The growth effects of natural disasters. *Journal of development economics*, *111*, 92–106.

Gall, M., Borden, K. A., & Cutter, S. L. (2009). When do losses count? six fallacies of natural hazards loss data. *Bulletin of the American Meteorological Society*, *90*(6), 799–810.

Guha-Sapir, D., & Below, R. (2002). The quality and accuracy of disaster data: A comparative analyse of 3 global data sets. *Centre for Research on the Epidemiology of Disasters (CRED) Working Paper, Brussels: CRED*.

Harrington, L. J., & Otto, F. E. (2020). Reconciling theory with the reality of African heatwaves. *Nature Climate Change*, *10*(9), 796–798. (Publisher: Nature Publishing Group)

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... others (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049.

Hollister, J., Shah, T., Robitaille, A. L., Beck, M. W., & Johnson, M. (2021). elevatr: Access elevation data from various apis [Computer software manual]. Retrieved from `https://github.com/jhollist/elevatr/` (R package version 0.4.2) doi: 10.5281/zenodo.5809645

Jones, R. L., Guha-Sapir, D., & Tubeuf, S. (2022). Human and economic impacts of natural disasters: can we trust the global data? *Scientific data*, *9*(1), 572.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, *11*(9), 1–20. doi: 10.18637/jss.v011.i09

Kuhn, M., & Wickham, H. (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. [Computer software manual]. Retrieved from `https://www.tidymodels.org`

Larrick, G., Tian, Y., Rogers, U., Acosta, H., & Shen, F. (2020). Interactive visualization of 3d terrain data stored in the cloud. In *2020 11th ieee annual ubiquitous computing, electronics & mobile communication conference (uemcon)* (pp. 0063–0070).

Lee, W. S., & Liu, B. (2003). Learning with positive and unlabeled examples using weighted logistic regression. In *Icml* (Vol. 3, pp. 448–455).

McDonald, G. G., Costello, C., Bone, J., Cabral, R. B., Farabee, V., Hochberg, T., ... Zahn, O. (2021). Satellites can reveal global extent of forced labor in the world's fishing fleet. *Proceedings of the National Academy of Sciences*, *118*(3).

Mordelet, F., & Vert, J.-P. (2014). A bagging svm to learn from positive and unlabeled examples. *Pattern Recognition Letters*, *37*, 201–209.

Moriyama, K., Sasaki, D., & Ono, Y. (2018). Comparison of global databases for disaster loss and damage data. *Journal of Disaster Research*, *13*(6), 1007–1014.

Panwar, V., & Sen, S. (2020). Disaster damage records of em-dat and desinventar: A systematic comparison. *Economics of Disasters and Climate Change*, *4*(2), 295–317.

Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., ... Hsiang, S. (2021). A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, *12*(1), 1–11.

Rosvold, E. L., & Buhaug, H. (2021). Gdis, a global dataset of geocoded disaster locations. *Scientific data*, *8*(1), 1–7.

South, A. (2017). rnaturalearthdata: World vector map data from natural earth used in 'rnaturalearth' [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=rnaturalearthdata` (R package version 0.1.0)

Swartz, W., Cisneros-Montemayor, A. M., Singh, G. G., Boutet, P., & Ota, Y. (2021). Ais-based profiling of fishing vessels falls short as a "proof of concept" for identifying forced labor at sea. *Proceedings of the National Academy of Sciences*, *118*(19).

Walker, A. J., Pretis, F., Powell-Smith, A., & Goldacre, B. (2019). Variation in responsiveness to warranted behaviour change among nhs clinicians: novel implementation of change detection methods in longitudinal prescribing data. *BMJ*, *367*.

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. doi: 10.18637/jss.v077.i01

# Paper 4: Discovering What Mattered: Answering Reverse Causal Questions by Detecting Unknown Treatment Assignment and Timing as Breaks in Panel Models

*Paper 4: Discovering What Mattered: Answering Reverse Causal Questions by Detecting Unknown Treatment Assignment and Timing as Breaks in Panel Models*

# Discovering What Mattered: Answering Reverse Causal Questions by Detecting Unknown Treatment Assignment and Timing as Breaks in Panel Models

Felix Pretis[1,2] and Moritz Schwarz[2,3*]

[1]Department of Economics, University of Victoria
[2]Climate Econometrics, Nuffield College, University of Oxford
[3]Smith School of Enterprise and the Environment, University of Oxford

**Abstract**

Much of empirical research focuses on forward causal questions ("Does X cause Y?") while reverse causal questions ("What causes Y?") can provide invaluable insights but is difficult to implement in practice. Here we operationalise the modelling of reverse causal questions through the detection of unknown treatment assignment and timing as structural breaks in fixed effects panel models. We show that conventional treatment evaluation of known interventions in a two-way fixed effects panel (often interpreted as difference-in-differences) is equivalent to allowing for heterogeneous structural breaks in the treated units' fixed effects. Using machine learning, we can thus detect previously unknown heterogeneous treatment effects as structural breaks in individual fixed effects corresponding to unit-specific treatment which can be subsequently attributed to potential causes (such as policy interventions). We demonstrate the feasibility of our approach by detecting the impact of ETA terrorism on Spanish regional GDP per capita without prior knowledge of its occurrence. Our proposed method to detect breaks in panel models can be readily implemented using our open-source R-package 'gets' with the 'getspanel' update or using the (adaptive) LASSO.

**JEL Classification:** C21, C23, C52
**Keywords:** Panel Data, Two-Way Fixed Effects, Treatment, Policy Evaluation, Difference-in-Differences; Break Detection, Indicator Saturation, Adaptive LASSO, Machine Learning;

*Author contact: fpretis@uvic.ca and moritz.schwarz@ouce.ox.ac.uk. We thank Anders Bredahl Kock, Sophia Carodenuto, Nicolas Koch, Andrew B Martinez, and Nolan Ritter for helpful comments and suggestions.

*Paper 4: Discovering What Mattered: Answering Reverse Causal Questions by Detecting Unknown Treatment Assignment and Timing as Breaks in Panel Models*

# 1 Introduction

Informed policy decisions require knowledge of which interventions have an impact. To assess the effectiveness of any policy – or the impacts of shocks in general – the wider empirical literature has been primarily concerned with forward causal questions to assess the 'effects of causes' (Gelman 2011, Gelman & Imbens 2013, Mill 1843). For example, does terrorism affect GDP per capita, or is a carbon tax successful in reducing emissions? Such forward causal questions ('Did $X$ have an effect on $Y$?') place specific interventions at the centre of their investigations and attempt to identify the effects of that known specific event on an outcome of choice. Forward causal questions therefore rely on pre-existing knowledge of interventions having taken place. This approach thus risks missing interventions that are *a-priori* unknown.

Policy makers regularly have no prior knowledge of the specific policies, trends, or shocks that have contributed to a certain outcome, prompting them to instead look for answers to *reverse causal questions*: rather than finding the 'effects of causes', they attempt to find the 'causes of effects' ('What has affected $Y$?'). For example, what affected GDP per capita, or what has reduced emissions? While the introduction of a carbon tax may not have had a strong impact on emissions, a local policy intervention incentivising energy efficiency improvements might have. If this particular intervention is not immediately apparent to policy makers, its contribution will be unaccounted for.

Even though such a 'reverse causal' approach is highly relevant to identify potentially unknown impactful policies or interventions, it has not been extensively operationalised in causal empirical modelling. This may be because forward causal questions are comparatively easy to evaluate using the range of available tools for programme evaluation, ranging from matching, difference-in-differences, to synthetic controls. In contrast, it is less obvious how reverse causal questions can be answered in practice. As Gelman & Imbens (2013) put it: "*Reverse causal reasoning is different; it involves asking questions and searching for new variables that might not yet even be in our model*".

Here we introduce a formal approach to answer reverse causal questions. We expand on the idea of "*searching for new variables*", and place the concept of reverse causal questions into the domain of variable and model selection, and more specifically break (or anomaly) detection. We propose tackling reverse causal questions by detecting anomalies in the form of structural breaks in the familiar setting of two-way fixed effects (TWFE) panel estimators allowing for heterogeneous treatment effects which are commonly used to evaluate policy. In such a setting, the conventional approach (often interpreted as difference-in-differences) is to include a dummy variables that denote the interaction of treated units with the post-treatment time period. We illustrate that this is equivalent to allowing for step-shift changes in the treated units' intercepts. If interventions are unknown, treatment assignment and timing can thus be detected as step-shifts conditional on treatment effects being non-zero. In our closely-related paper, Koch et al. (2022) we apply this approach to detect effective climate policies. Here, we provide the theoretical basis formally linking structural breaks to heterogeneous treatment effects.

Specifically, we formulate the detection of structural breaks as a problem of variable selection, where we saturate a TWFE panel model with a full set of step-shift break functions denoting potential treatment of each individual at any point in time. We then apply machine learning selection methods allowing for more candidate variables than observations to identify relevant step-shifts to detect treatment without prior knowledge of its occurrence. Once a break has been identified, it can be interpreted as a treatment dummy for the relevant unit (see Figure 1 for a stylised example). We can then – in a post-estimation analysis – attempt to attribute the treatment dummy to an event that is likely to have affected the treated unit at the detected time.

This operationalises the idea of reverse causal questions – we start with a model commonly used for policy evaluation, but rather than testing a particular policy intervention (in the form of a treatment dummy), we search for structural breaks in the individual fixed effects, which can subsequently be attributed ex-post to events that took place. Thus, rather than assessing effects of causes (as in the forward causal approach), our approach provides a data-driven method to identify effects using break detection which can then be attributed to potential causes. Note that our discussion on reverse causal questions should not be confused with the concept of reverse causality. Reverse causal questions refer to the modelling process of discovering the causes of effects, and do not refer to the direction of causal relations between variables.

### 1.1 Related Literature & Contribution

Break detection to assess the impact of policy has been commonly used in time series analysis. However, most time series applications do not have control groups, making a causal interpretation of any break difficult. A causal interpretation in time series is possible where breaks occur in some conditioning variables under super exogeneity (Bazinas & Nielsen, 2015). This has been shown first in Engle et al. (1983) as causal relations invariant to shocks (referred to as super exogeneity). Under such super exogeneity causal identification is possible, see e.g. Martinez (2020), Mukanjari & Sterner (2018), or Pretis (2021) for relevant examples of this.

Where super exogeneity does not hold or is difficult to establish, however, a causal interpretation of structural breaks is more difficult. Examples in the time series literature range from Perron (1989) detecting breaks in GNP time series attributed to the Great Depression and an oil price shocks, Hendry (2020) identifying policy interventions in UK $CO_2$ per capita emissions, Estrada et al. (2013) quantifying the impact of the Montreal Protocol on CFC emissions and subsequently temperatures, to Apergis & Lau (2015) identifying whether breaks in Australian electricity markets align with policy interventions. Piehl et al. (2003) also use the detection of breaks in time series to assess treatment effectiveness of a youth homicide prevention programme in Boston.

Compared to time series applications, fewer papers tackle structural breaks in a panel setting, and to the best of our knowledge, no paper has formally considered the link between structural breaks and treatment effects in a panel, or the detection of breaks in a TWFE panel to detect treatment. Attributing breaks in panels as treatment was first explored in Pretis (2019) assessing the impact of carbon taxes, but not formally linked to heterogeneous treatment effects. In our related paper, Koch et al. (2022) we apply the break detection approach to EU $CO_2$ road emissions to identify effective policies in a causal framework, albeit with the focus on the application rather than the econometric theory underlying the approach.

Panel methods for the detection of breaks range from estimating break dates using least-squares to detecting breaks by selecting over break dates using model selection. In the least-squares literature, Chan et al. (2008) extend the Andrews (1993) structural change with unknown change point test (Sup-test) for simple structural breaks to panels in a setting that focuses on detecting changes in coefficients on random variables (rather than on changes in the fixed effect as would be necessary in a policy evaluation framework). De Wachter & Tzavalis (2012) test for common breaks in dynamic panels and Baltagi et al. (2016) study heterogeneous panels with structural breaks using a Bai (1997)-type approach. Recent work has focused on structural breaks in panels using common factors, these include Zhu et al. (2020) who study a single break in dynamic panel with a common factor structure, as well as Cheng et al. (2016) who consider breaks in the form of changing latent factor loadings. Bai et al. (2020) develop a least squares approach to detect breaks in factor loadings in panel factor models.

The above papers use the classical approach of estimating break dates by least squares, while now there is a growing literature using variable selection to identify breaks. In this selection-based literature, Qian & Su (2016b) propose to use the LASSO (Least Absolute Shrinkage and Selection Operator) to detect breaks in simple time series models, with Qian & Su (2016a) extending this approach to detect common breaks in panels. Their method is related to our approach presented here, however, we focus on individual breaks in fixed effects and thus treatment rather than on common breaks. In the factor literature (not focusing on policy evaluation), Li et al. (2016) propose to use the LASSO to detect common breaks in interactive fixed effects models. Conley & Taber (2011) compare inference when the number of treatment groups is small to inference in structural break detection, but they do not explore this link further.

Our paper is perhaps closest in-spirit to Okui & Wang (2021) on the selection side, and Wooldridge (2021) on the treatment estimation side. Okui & Wang (2021) detect heterogeneous (group-specific) structural breaks in coefficients on random variables using the adaptive LASSO. Relative to our approach, Okui & Wang (2021) do not focus on treatment effects and they partial out the individual fixed effects rather than studying breaks in them. Their analysis also focuses on grouped rather than individual structural breaks. We instead concentrate on breaks in unit fixed effects to detect treatment and explore alternative selection methods (in addition to the LASSO we also use the general-to-specific – gets – selection method) which can be embedded in an outlier-robust estimation framework. Nevertheless, the group-specific method of Okui & Wang (2021) may be a promising avenue of future research in the case of multiple (unknown) treated units.

In a standard 'known-treatment' setting, Wooldridge (2021) shows that heterogeneous and time-varying treatment effects can be identified and estimated consistently using a TWFE estimator in a common timing and staggered setting using interactions of treatment times and dummies. We show that the starting point of our break detection approach nests Wooldridge's interacted TWFE specification as a special case where we relax the knowledge around treatment assignment and timing, as well as homogeneity of treatment effects over treated individuals. Thus, the interaction-augmented TWFE estimator proposed in Wooldridge (2021) constitutes the target of model selection in our case, and the final retained models identify heterogeneous treatment effects. Specifically, we show that 'known' policy dummy variables in a TWFE panel model are equivalent to step-shifts in the individual fixed effects of the treated units which can be detected using break detection methods. Similarly, we demonstrate that time-varying and heterogeneous treatment effects using interactions are equivalent to allowing for unit-specific impulse dummies which capture single-period structural breaks. In other words, as we show, treatment dummies in a TWFE setting are equivalent to structural breaks taking the form of a step-shift in the individual fixed effects of the panel units. Using this equivalence between step-shifts in the unit-specific intercept (i.e. fixed effect) and known treatment dummies, we therefore propose an alternative estimation approach based on reverse causal questions: rather than simply evaluating a known intervention, we instead estimate a TWFE panel model while searching for potential structural breaks (step-shifts) in the unit-specific intercepts. Notably, this approach identifies unit-heterogeneous fully-time-varying or piece-wise constant treatment effects. The final model (conditional on having identified treatment breaks) corresponds to a heterogeneous treatment effects model where treatment effect heterogeneity is identified using interactions as in Wooldridge (2021) and thus does not suffer from the concerns around heterogeneous treatment effects in staggered interventions (see Goodman-Bacon 2021, Callaway & Sant'Anna 2020, De Chaisemartin & d'Haultfoeuille 2020, 2021, Baker et al. 2021).

Overall, relative to the time series literature using breaks for indicative policy evaluation, we expand the break-detection approach to a panel setting where units without breaks act as a control group against

which a break (i.e. treatment intervention) can be identified. Relative to the existing panel break literature, we focus on machine learning methods to detect breaks in single individual fixed effects and their attribution as treatment interventions. Relative to the vast existing TWFE literature on policy evaluation, our approach implements a reverse-causal estimation strategy detecting previously unknown treatment (or events) while at the same time allowing known interventions to be embedded. Recently the detection of structural breaks has also been applied to detect unknown discontinuities in regression discontinuity design (RDD). For instance, Porter & Yu (2015) use a simple Andrews (1993)-type structural break test to identify a regression discontinuity without prior knowledge of its existence. Similarly to their argument of an unknown discontinuity, we explore the use of break detection to identify previously unknown treatment in TWFEs.

There are a range of nuances to our proposed approach. First, if treatment assignment and timing is known (and happens to have a large effect), then imposing interacted treatment dummies allowing for heterogeneous treatment effects for the known intervention in a TWFE estimator is effectively equivalent to agnostically detecting a break in this fixed effect (if that is the only break retained) and estimating the model post-break detection. The estimated model with an imposed break or a single retained break is *identical*. In other words, if the intervention was known, we could simply run a TWFE estimator, which will be equivalent to having found the one particular intervention and then assessing the estimated model.

Second, our idea is modular with respect to known treatment. If there is a known intervention, we can impose it into the model without selection and estimate its impact, while at the same time searching for additional breaks. This allows us to assess the impact of a known policy while also detecting potentially unknown interventions, effectively implementing the theory-embedding approach described in Hendry & Johansen (2015). It is worth noting that our approach concentrates on causes of effects by first identifying effects. If there are no effects, naturally we cannot find any corresponding cause. Thus, for unknown treatment we cannot distinguish between no treatment or a zero treatment effect. For known treatment, however, this is not a concern as it can easily be embedded as a forced a-priori treatment variable that is introduced into the model independent of the selection. We can also restrict the search for breaks and treatment to a subset of units if we suspect that some units may be treated and are certain that others are not.

Third, our conceptual approach is also modular in terms of the choice of detection method. We can use different machine learning methods of our choice to detect breaks (i.e. treatment), depending on the preferred properties of the selection algorithm. For example, if our main concern is the false-positive detection rate, then we can choose to use methods that control the false discovery rate (such as general-to-specific selection methods, henceforth 'gets'). If instead we care about computational speed, we could use regularised estimators, such as the (adaptive) LASSO.

There are of course some constraints to our methods. First, when detecting breaks in individual fixed effects, each treated unit will be identified with a separate treatment dummy. While this allows for straight-forward heterogeneous treatment effects, it means that we do not gain power if there are multiple units that received the same treatment. Thus, our TWFE break detection approach mirrors the use of interactions to identify known heterogeneous treatment effects (Wooldridge, 2021) and lends itself to panels with longer time series and smaller cross-sectional dimensions with heterogeneous treatment (similar to settings encountered when using synthetic controls).

Second, all break detection methods evaluate the presence of breaks relative to a specified underlying model. If the model is not well-specified, then breaks that we detect may simply reflect model misspecification. This is of course also a problem in conventional TWFE difference-in-differences settings,

however, can be amplified in our setting if we attempt to attribute a 'spurious' break to an event. This effect can be mediated by selecting at tight significance levels to control the false-positive rate (when using gets, see section 3.1.1) or by making use of robust estimators less sensitive to observations falling outside the specified model (such as embedding break detection in a wider outlier-robust estimation framework, e.g. Impulse Indicator Saturation, IIS – see Hendry et al. 2008; Jiao & Pretis 2020, Jiao et al. 2021).

Third, once a treatment effect is detected in the form of a break, it has to be attributed to a potential cause by manually referring to the existing literature and records. While this can be a challenge and requires subject-specific knowledge, it offers an opportunity to learn from the data. A search for a potential cause of an effect is comparable to arguing that a known intervention was exogenous (or as-if randomly assigned) in a conventional programme evaluation application.

We demonstrate the feasibility of our break-detection approach using a well-known dataset on the economic impacts of terrorism in Spain, where the onset of ETA terrorism in the Basque Country depressed regional GDP per capita relative to unaffected regions. The dataset was originally analysed by Abadie & Gardeazabal (2003) in their seminal paper introducing synthetic control methods. We show that we can detect the 'treatment' of Basque terrorism as breaks in individual fixed effects in models of GDP per capita without prior knowledge of its occurrence, in line with the original results by Abadie & Gardeazabal (2003). The treatment intervention can be detected both in a simple TWFE setting with two regions (where one region is unknowingly treated), as well as in a wider panel model of all of Spain's mainland regions. Beyond ETA terrorism, we also detect breaks in other regions which we attribute to an industrial crisis in Madrid, and increased regional autonomy in the post-Franco era. Hence, our approach directly operationalises the idea of reverse causal questions – we start with a simple TWFE estimator, use machine-learning selection to identify significant interventions which can be interpreted as treatment effects and subsequently attribute them to events that took place. Our proposed panel break detection approach can be readily implemented using our accompanying R-package 'gets' (Pretis et al. 2018) and the 'getspanel' update (Schwarz et al. 2021).

The roadmap for the remainder of the paper is as follows. In section 2.2 we first provide a simple illustration of how structural breaks are closely linked to treatment evaluation in two-way fixed effects estimators. We consider the standard case of known treatment assignment and illustrate its equivalence to a step-shift break in the treated units' intercept. In section 2.3 we then consider the case where treatment assignment and timing is unknown, and we establish that unknown treatment assignment and timing can be identified using impulse dummies in a saturated regression (for fully time-varying effects) and step-dummies (for piece-wise constant effects). Using recent results from Wooldridge (2021), we show how time-varying and unit-heterogeneous treatment effects can be nested in dummy-saturated models with more variables than observations. We further show that if we detect multiple treatment breaks (in multiple different units), they can be interpreted as time-varying treatment effects in a staggered treatment intervention setting. We discuss our approach in a balanced panel without explicitly discussing control variables, however, the results should generalise to the inclusion of other covariates. In the following section 3 we then briefly discuss two estimation approaches using general-to-specific (gets) selection and the adaptive LASSO (and provide some simulations in the Supplementary material). Finally, we apply our methods to models of Spanish regional GDP per capita in section 4.

## 2 Conceptual Approach: Break Detection to Detect Unknown Treatment Assignment & Timing

We consider the detection of treatment and subsequent estimation of treatment effects when both treatment assignment and treatment timing are unknown. We show that if treatment assignment and timing are unknown, such treatment can be identified by allowing for potential structural breaks at any point in time for any unit in a model including individual and time fixed effects. Applying machine learning methods allowing for model selection for more variables than observations, we then remove all irrelevant treatment dummies and are left with the resulting model that identifies treatment assignment, timing, and estimates treatment effects conditional on the treatment effects being non-zero.

To aid the reader's understanding, we present a stylised example in Figure 1, which provides guidance for the various aspects of our conceptual approach we touch upon in the following sections.

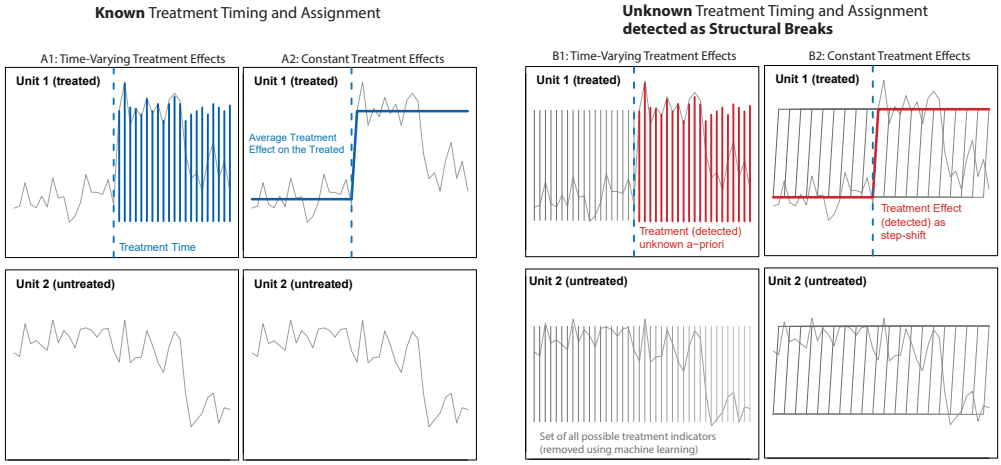


Figure 1: Detecting unknown treatment timing and assignment as structural breaks – a stylised Example using artificial data. Left: 'Known' Treatment baseline for time-varying and constant treatment effects. Right: Detecting treatment as breaks using impulses for time-varying and step-shifts. All possible impulse and step-indicators shown in grey, a subset of which (red) identify the true underlying treatment (blue in left panels).

### 2.1 Setting

To illustrate the overall motivation and the close link between structural breaks and treatment evaluation, consider a panel of $N$ units over $T$ time periods where one group is treated with a single treatment from time $t = q$ onwards. We initially consider the baseline case of known treatment assignment and timing, where the treatment indicator $d = 1$ for the treated group (or individual) and $d = 0$ for the untreated. Using the notation in Wooldridge (2021), we denote by $y_t(0)$ the outcome in the untreated control group, and $y_t(1)$ the outcome in the treated group at time $t$. The treatment effect at time $t$ due to treatment occurring from time $t = q$ onwards is given by the difference $y_t(1) - y_t(0)$. As is convention in the literature, we focus on the average treatment effect on the treated $\tau_t$:

$$\tau_t = E[y_t(1) - y_t(0)|d = 1] \tag{1}$$

To identify the average treatment effect we assume there is no anticipation of treatment, in other words, the potential outcome for a unit prior to treatment is identical to the untreated units:

$$E[y_t(1) - y_t(0)|d = 1] = 0, \text{ for t} < \text{q} \tag{2}$$

Further we rely on the common trend assumption which is standard in much of the treatment effects literature:

$$E[y_t(0) - y_{t=1}(0)|d] = E[y_t(0) - y_{t=1}(0)] = \theta_t, \text{ for } t = 2, ..., T \tag{3}$$

Finally, we also assume there is at least one untreated unit. We then write the observed outcome as:

$$y_t = y_t(0) + d[y_t(1) - y_t(0)] \tag{4}$$

The expected outcome conditional on treatment is:

$$E[y_t|d] = E[y_t(0)|d] + d \times \tau_t \tag{5}$$

We define the change in $y_t$ over time in absence of treatment as:

$$g_t(0) = y_t(0) - y_{t=1}(0) \tag{6}$$

and under the common trend assumption we have that $E[g_t(0)|d] = E[g_t(0)] = \theta_t$. We thus have that:

$$E[y_{t=1}(0)|d] = \lambda + \xi d \tag{7}$$

where $\xi$ denotes the average pre-treatment difference between the treated and untreated groups and $\lambda$ denotes the average level of $y$ for the untreated. Combining all above yields the expected value of $y_t$ conditional on treatment as:

$$E[y_t|d] = \lambda + \xi d + \theta_t + d \times \tau_t \tag{8}$$

For illustration purposes, assume the treatment effect is constant over time, $\tau_t = \tau$. Under the assumption of no anticipation we have that:

$$\begin{aligned} E[y_t|d] &= \lambda + \xi d + \theta_t, \text{ for } t < q & (9)\\ &= \lambda + \xi d + \theta_t + d \times \tau, \text{ for } t \geq q & (10) \end{aligned}$$

This is a standard result in the treatment effects literature and the above model can be consistently estimated using a TWFE estimator (see e.g. Wooldridge 2021):

$$y_{i,t} = c_i + g_t + \tau w_{i,t} + u_{i,t} \tag{11}$$

with $w_{i,t} = d_i \times q_t$ where $d_i$ is an indicator for whether the individual is treated, $q_t$ an indicator for the post treatment period, $c_i$ denote individual fixed effects, and $g_t$ time fixed effects. Note that Wooldridge (2021) groups the untreated mean into a single intercept, however, the treatment effect estimates are unaffected by whether we include a common intercept or allow for unit-specific intercepts (i.e. fixed effects). Notably, the above model shows the close link to structural breaks as we identify the average

treatment effect as a step-shift of magnitude $\tau$ at time $q$ in the treated unit's intercept:

$$\begin{aligned} E[y_{i,t}|d_i = 1] &= c_i + \tau \times 1_{\{t \geq q\}} + g_t \quad (12) \\ &= c_{i,t} + g_t \\ \text{where } c_{i,t} &= \begin{cases} c_i & \text{for } t < q \\ c_i + \tau & \text{for } t \geq q \end{cases} \quad (13) \end{aligned}$$

Figure 1 (column A2, left) shows a stylised example illustrating how a constant treatment effect corresponds to a simple step-shift in the individual-specific intercept.

### 2.2 Known Treatment Assignment with Unknown Timing

Now suppose we know which units are treated, but the timing of treatment is unknown. This may be the case when we suspect some intervention or event took place in some regions/countries, but the actual date of the intervention is uncertain. Let $H$ denote the set of treated individuals and $1_{\{i \in H,\, t \geq q\}}$ an indicator function equal to one when $i$ is part of the treated group and $t$ falls in the post-treatment period. When treatment timing is unknown, we can interpret the identification of treatment effects as a break detection problem where we detect a structural break in the treated unit's specific intercept conditional on there being a non-zero treatment effect:

$$y_{i,t} = c_i + \tau \times 1_{\{i \in H,\, t \geq q\}} + g_t + u_{i,t} \quad (14)$$

When treatment is known, the above model (14) corresponds to a partial structural change model (see e.g. Perron 2006) with $c_{i,t}$ being allowed to break for treated individuals in the sample, and we estimate the break date $q$ as well as treatment effect $\tau$. If there is only a single treated unit and we detect a structural break in its intercept at the time of treatment, then the resulting model with a structural break is *identical* to the treatment effect model (11). There is thus a close link between break detection and the estimation of treatment effects in TWFE estimators.

However, the above model (14) may be overly restrictive as it assumes a single treatment with known-assignment and unknown-timing. In practice there may exist a myriad of possibly unknown interventions and we may face uncertainty around both treatment assignment as well as timing. In other words, we may not know which (if any) units are treated, and at what time such treatment may have occurred. In addition, treatment effects may also be heterogeneous over treated units as well as over time.

We therefore now turn to the setting where we allow for both treatment assignment and timing to be unknown (see section 2.3), and also relax the assumption of time-constant and homogeneous treatment effects over treated units (Section 2.3.1). Subsequently we consider multiple treatments (which could also be interpreted as staggered adoption) in section 2.3.3.

### 2.3 Unknown Treatment Assignment & Timing for a Single Treatment

We now consider detecting treatment when treatment assignment and timing are unknown and treatment effects may be heterogeneous over treated units and time. We begin by relaxing the assumption that treatment effects are constant over time in the known treatment setting. Allowing for time-varying

treatment effects $\tau_t$ we can write the expected outcome conditional on treatment as:

$$\begin{aligned} E[y_t|d] &= \lambda + \xi d + \theta_t, t < q \\ &= \lambda + \xi d + \theta_t + d \times \tau_t, t \geq q \end{aligned} \tag{15}$$

If treatment assignment and timing is known, the above can be consistently estimated using interactions in a TWFE estimator (see Wooldridge 2021 for the 'known treatment' case), where again we here allow for unit-specific intercepts:

$$y_{i,t} = c_i + g_t + \sum_{s=q}^{T} \tau_s (d_i \cdot 1_{\{t=s\}}) + u_{i,t} \tag{16}$$

This is equivalent to a set of step-shifts of duration 1 (with common coefficient over $i$ in $H$) at times $q, q+1, ..., T$, where each time step is denoted as an index $s$, which ranges from $s_1, s_2, ..., S$. We now relax the assumption of having common coefficients over $i$, in other words, we allow for heterogeneity in treatment effects over $i$. Let $H = \{m_1, m_2, ..., m_M\}$ denote the set of $M$ treated units. For example if units $i = 2$ and $i = 3$ are treated, then there are two treated units $M = 2$, and $m_1 = 2$ and $m_2 = 3$. The above model (16) can then be written in a general specification allowing for unit-heterogeneous treatment effects at every time as:

$$y_{i,t} = c_i + g_t + \sum_{j \in H} \sum_{s=q}^{T} \tau_{j,s} 1_{\{i=j,t=s\}} + u_{i,t} \tag{17}$$

where $1_{\{i=j,t=s\}}$ denotes an indicator function equal to one for all treated $i$ in the set of treated units $H$ and $t = s$ in the post-treatment period $s \geq q$, and zero otherwise. This specification relaxes the restriction of homogeneous treatment effects across treated units. Figure 1 (column A1, left) shows a stylised example of individual impulses capturing a treatment effect. Specifically, each treated post-treatment observation is captured by a single time-period dummy. While these cannot be estimated consistently because they capture single observations, such dummy variables can be estimated unbiasedly (see Hendry & Santos 2005) and, as we showed here, identify unit- and time-specific treatment effects.

To relate the known case to the unknown treatment setting, we further refine our notation. We define an index of the timing of non-zero treatment effects for each treated unit $j \in H$ denoted as $R_j = \{q_{j,s=1}, q_{j,s=2}, ..., q_{j,S_j}\}$ where $S_j$ denotes the number of treatment indicators for unit $j$. For example, suppose that in a 3-unit panel with $T = 20$ observations units $i = 2$ and $i = 3$ are treated ($m_1 = 2, m_2 = 3$) with non-zero treatment effects from $t = q, ..., T$. Then $H = \{2, 3\}$ with corresponding treatment effects at $R_2 = \{q_{2,1} = q, q_{2,2} = q+1, ..., q_{2,S_2} = T\}$ and $R_3 = \{q_{3,1} = q; q_{3,2} = q+1, ..., q_{3,S_3} = T\}$. With common treatment timing and effects this implies that $R_2 = R_3$. Thus the known-treatment and known-timing baseline in (17) can be written as:

$$y_{i,t} = c_i + g_t + \sum_{j=m_1}^{m_M} \sum_{s=q_{j,1}}^{q_{j,S_j}} \tau_{j,s} 1_{\{i=j,t=s\}} + u_{i,t} \tag{18}$$

or by simplifying notation as:

$$y_{i,t} = c_i + g_t + \sum_{j \in H} \sum_{s \in R_j} \tau_{j,s} 1_{\{i=j,t=s\}} + u_{i,t} \tag{19}$$

Now what if treatment assignment and timing are unknown? The above model in (19) constitutes the 'known' intervention baseline, i.e. the target of model selection/break detection. In 2.3.1 we now consider the detection of treatment assignment and timing allowing for unit-heterogeneous and time-varying treatment effects as in (19) which we will show is matched by a saturating set of unit-time-specific impulse dummies. We then consider treatment detection allowing for unit-heterogeneous but piece-wise constant treatment effects over time, which we will show is nested by a saturating set of unit-specific step-shift breaks.

### 2.3.1 Detecting Unknown Treatment When Treatment Effects are Fully-Time Varying

If treatment assignment and treatment timing is unknown, the 'known' treatment model (19) can be embedded in a general model allowing for potential treatment of any unit at any point. The most flexible specification that nests the 'known' treatment specification (19) as a special case is a fully-saturated model allowing for a treatment dummy for each individual at every point in time:

$$y_{i,t} = c_i + g_t + \sum_{j=1}^{N}\sum_{s=1}^{T} \tau_{j,s} 1_{\{i=j,t=s\}} + u_{i,t}. \tag{20}$$

The model in (20), which identifies unit-specific treatment effects for each unit for each time period, however, cannot be estimated as such because the number of parameters matches (or exceeds) the number of observations. Effectively there are $NT$ possible indicator variables added to the balanced panel. Figure 1 (column B1, right) shows the full set of these impulse indicators, a subset of which identify the true treatment effect shown in panels on the right.

The aim is thus to reduce the general model (20) to a sparse model, ideally coinciding with the underlying target of the known baseline (19). Thus, we require the additional assumption that treatment effects are sparse, we have at least one untreated unit and some untreated time-periods for treated units – assumptions that are very common in the wider treatment evaluation literature. Starting with this general model, we then apply machine learning/model selection to remove all but 'relevant' dummy variables using selection algorithms capable of handling more variables than observations. We discuss two possible machine learning algorithms in more detail in section 3.1. In fact, the dummy-saturated model in (20) is equivalent to an outlier-robust Huber-Skip estimator, where the retained impulse dummies detecting 'outliers' relative to the model capture the time-varying unit-specific treatment effects (see e.g. Jiao et al. 2021, Hendry et al. 2008, Johansen & Nielsen 2009, Johansen & Nielsen 2016a). We write the sparse final selected model as:

$$y_{i,t} = \hat{c}_i + \hat{g}_t + \sum_{j\in\widehat{H}}\sum_{s\in\widehat{R}_j} \hat{\tau}_{j,s} 1_{\{i=j,t=s\}} \tag{21}$$

where we effectively estimate treatment assignment $\widehat{H} = \{\hat{m}_1, \hat{m}_2, ..., \hat{m}_{\hat{M}}\}$, together with the index of treatment occurrence $\widehat{R}_j = \{\hat{q}_{j,1}, \hat{q}_{j,2}, ..., \hat{q}_{j,\hat{S}_j}\}$ where $\hat{S}_j$ denotes the number of treatment indicators for unit $j$, and the the time-varying and unit-specific treatment effects $\hat{\tau}_{j,s}$ conditional on having non-zero treatment effects. Note that we detect treatment when it has an effect, i.e. we detect treatment effects conditional on them being non-zero. Using the resulting estimated treatment effects $\tau_{j,s}$ it is straightforward to compute the average treatment effects for the treated (ATTs) for specific units or time periods. As impulse indicators are orthogonal, it is trivial to compute the standard error for the resulting ATTs. For example, we can compute the estimated ATT for individual $j$ over the entire period of non-zero

detected treatment effects as:

$$\widehat{ATT}_j = \frac{1}{\hat{S}_j}\sum_{s=1}^{\hat{S}_j}\hat{\tau}_{j,s}, \text{ with standard error } se\left(\widehat{ATT}_j\right) = \sqrt{\frac{1}{\hat{S}_j}\sum_{s=1}^{\hat{S}_j} se\left(\hat{\tau}_{j,s}\right)^2} \tag{22}$$

If we are interested in a subset of treated periods we could simply restrict the ATT to those relevant time periods (or units). A remaining issue, however, is that detecting individual impulses may suffer from low power if treatment effects are small and actually constant over some period of time. If we are interested in ATTs over time for some treated units, allowing for piece-wise constant treatment effects may yield higher power of detection which we discuss in the next section 2.3.2.

### 2.3.2 Detecting Unknown Treatment With Piece-Wise Constant Treatment Effects

While treatment effects may be heterogeneous over individuals $i$, they may be constant for some time periods. Such constancy over time can lead to higher power to detect treatment. Consider treatment effects in (15) that are constant over time following treatment from $t = q$ onwards, but allowed to vary over treated individuals:

$$y_{i,t} = c_i + g_t + \sum_{j\in H}\tau_j 1_{\{i\,=\,j,\,t\,\geq\,q\}} + u_{i,t} \tag{23}$$

Then for each treated unit in $H$ (where $d_i = 1$) with time-invariant treatment effect $\tau_{i,t} = \tau_i,$ for all $t$, the change from pre-treatment to post-treatment is given by a step-shift change in the unit-specific intercept of magnitude $\tau_i$. For example, for treated unit $i$ with time-invariant treatment effect, the expected outcome is given by:

$$\begin{aligned} E[y_{i,t}|d_i = 1] &= c_i + g_t + \tau_i \times 1_{\{t\geq q\}} \\ &= c_{i,t} + g_t \\ \text{where } c_{i,t} &= \begin{cases} c_i & \text{for } t < q \\ c_i + \tau_i & \text{for } t \geq q \end{cases} \end{aligned} \tag{24}$$

which is just a step-shift in the unit-specific intercept (i.e. fixed effect), equal to $c_i$ prior to treatment, and $c_i + \tau_i$ post treatment. Estimates of $\tau_i$ then correspond to the unit-specific average treatment effect over time. If treatment timing and assignment are unknown, we can generalise the impulse-dummy approach to nest known treatment as a special case in a general model now allowing for step-shifts at any point in time as:

$$y_{i,t} = c_i + g_t + \sum_{j=1}^{N}\sum_{s=2}^{T}\tau_{j,s}1_{\{i=j,t\geq s\}} + u_{i,t} \tag{25}$$

where a subset of the step-functions $1_{\{i=j,\,t\,\geq\,s\}}$ correspond to the actual treatment effects model in (23). This allows for any unit to be potentially treated at any point in time – with $s$ starting at 2 rather than 1, so as not to coincide with the fixed effect in $c_i$. We then aim to remove treatment indicators such that we only retain the subset of truly treated units and time periods. Under sparsity of treatment effects i.e., there remain units and time periods without treatment, we write the final selected model as:

$$y_{i,t} = \hat{c}_i + \hat{g}_t + \sum_{j\in\widehat{H}}\sum_{s\in\widehat{R}_j}\hat{\tau}_{j,s}1_{\{i=j,t\geq s\}} \tag{26}$$

where we estimate treatment assignment by detecting those units $i$ that have at least one break indicator retained, and break times are estimated by the starting date of each retained break function. Figure 1 (column B2, right) shows the full set of step-functions, a subset of which identify the true treatment effect.

Note that this setup does not impose that treatment effects have to be strictly constant over time post-treatment, as a linear combination of step-functions can capture time-varying treatment effects.

### 2.3.3 Unknown Treatment Assignment and Timing For Multiple Treatments

If there is a single underlying treatment and break detection identifies a single intervention then the interpretation and attribution of detected effects is straight forward. However, in practice there may be multiple treatments detected as breaks at different times for multiple different units. What do we identify if we detect multiple such treatment occurrences at different times for different units? Irrespective of the selection algorithm employed (see section 3.1), consider the following final retained model with a range of detected treatment impulse dummies:

$$y_{i,t} = \hat{c}_i + \hat{g}_t + \sum_{j \in \widehat{H}} \sum_{s \in \widehat{R}_j} \hat{\tau}_{j,s} 1_{\{i=j, t=s\}} \tag{27}$$

or step-functions:

$$y_{i,t} = \hat{c}_i + \hat{g}_t + \sum_{j \in \widehat{H}} \sum_{s \in \widehat{R}_j} \hat{\tau}_{j,s} 1_{\{i=j, t \geq s\}} \tag{28}$$

What is identified if the detected treatment time varies across units in the panel? For example, what if we find both $j = 1$ and $j = 2$ to be in the treated group, but their treatment timing differs, i.e. $R_{j=1} \neq R_{j=2}$ for both $j = 1$ and $j = 2$? We show that the final retained models with heterogeneous treatment dummy variables (27) and (28) are equivalent to staggered treatment with heterogeneous effects where heterogeneity and staggered adoption are captured through interactions. In other words, the impulse indicator estimator identifies unit and time-specific staggered treatment effects conditional on the treatment effect being non-zero. If treatment effects are constant over time, then a saturating set of step-functions nests the known-treatment assignment and timing model as a special case even when treatment is staggered. To illustrate this equivalence, we follow the discussion in a known-treatment setting by Wooldridge (2021) on how interaction terms identify treatment effects in a staggered treatment setting. Subsequently we show that this is nested by the IIS and SIS break detection estimators in an unknown treatment setting, establishing that detected breaks identify unit- and time-specific treatment effects.

For exposition, consider a staggered treatment DGP where we denote the time of the first intervention by $q$. We define treatment cohort dummies similar to Wooldridge as $d_q, ..., d_Q$ where $Q$ denotes the final time of intervention, which would be equal to $T$ when treatment lasts until the end of the sample. We refer to the time of each intervention as $r \in \{q, q+1, ..., Q\}$. The potential outcome at time $t$ for unit treated at time $r$ is given by $y_t(r)$, with the outcome for the never treated unit referred to as $y_t(\infty)$ i.e. treated at no point in time. The quantities of interest are the treatment effects of each unit first receiving treatment at time $r$ given by the difference in outcomes $y_t(r) - y_t(\infty), r = q, ..., Q$. In a staggered setting we hope to identify the average treatment effects on the treated ATT for each intervention (given by different cohorts which we will relax to different individuals):

$$\tau_{r,t} = E[y_t(r) - y_t(\infty) | d_r = 1], r = q, ..., Q; t = r, ..., T. \tag{29}$$

Under no anticipation and common trends in a standard known-treatment setting, Wooldridge (2021) demonstrates that heterogeneous treatment effects can be consistently estimated in a staggered treatment setting using interactions in a TWFE estimator (we replicate the derivation in Supplementary Section 6.1). Specifically, the expected outcome in a staggered treatment setting can be written as

$$\begin{aligned} E[y_t|\mathbf{d}] &= \eta + \lambda_q d_q + ... + \lambda_Q d_Q + \theta_t \text{ (pre-treatment } t < q) \\ &= \eta + \lambda_q d_q + ... + \lambda_Q d_Q + \theta_t + \tau_{q,t} d_q + ... + \\ &\quad + \tau_{Q,t} d_Q \text{ (post-treatment } t \geq q) \end{aligned} \tag{30}$$

where $\eta$ is the average level of $y$ for the untreated group and $\lambda_q$ refers to the average level of $y$ for the treated cohorts pre-treatment. This can be consistently estimated using a TWFE estimator with time-cohort interactions as:

$$y_{i,t} = c_i + g_t + \sum_{r=q}^{Q} \sum_{s=r}^{T} \tau_{r,s} (d_{i,r} \cdot 1_{\{t=s\}}) + u_{i,t} \tag{31}$$

In the above equation each cohort has a set of time-varying treatment effect estimates. Now consider each treated unit in the panel being allowed its own treatment effects (i.e. each unit in each cohort receives its own treatment effect or each cohort is of size one). As before, consider $H = \{m_1, m_2, ..., m_M\}$ as the set of $i$ that are treated at some time, where treatment timing is not exclusive. In other words, $m_1$ and $m_2$ may be treated at the same time (but may also be treated at different times). Then relaxing the above assumption that each treatment cohort has the same treatment effect, the above model (31) can be written as:[1]

$$y_{i,t} = c_i + g_t + \sum_{j \in H} \sum_{s=r}^{T} \tau_{j,s} 1_{\{i=j, t=s\}} + u_{i,t} \tag{33}$$

This is identical to the interaction of the treatment dummy $d_{i,r}$ and time dummies $1_{\{t=s\}}$ above, except we disaggregate treatment cohorts into individual units. Using simplifying notation, we can write (33) as:

$$y_{i,t} = c_i + g_t + \sum_{j \in H} \sum_{s \in R_j} \tau_{j,s} 1_{\{i=j, t=s\}} + u_{i,t} \tag{34}$$

This matches the impulse-dummy saturated final specification where treatment assignment and timing is estimated in (27). Similarly, if treatment effects are piece-wise constant over time we can write (33) as:

$$y_{i,t} = c_i + g_t + \sum_{j \in H} \sum_{s \in R_j} \tau_{j,s} 1_{\{i=j, t \geq s\}} + u_{i,t} \tag{35}$$

Estimating this heterogeneous staggered-treatment model matches the post-selection step-function model in (28). Thus, detecting multiple treatments through impulses or step-indicators is equivalent to the estimation of treatment effects in a staggered-intervention setting when heterogeneous treatment effects are identified using interactions. We identify average treatment effects (over time) for each treated unit relative to the never treated cases, conditional on the treatment effect being large enough to be detected. If a single unit experiences more than one treatment – then this can be interpreted as time-varying treatment (where the sum of effects is the treatment effect relative to the never treated), or a separate treatment

[1] To recover (31) we could restrict equation (33) as:

$$\tau_{m_l,s} = \tau_{m_k,s} = \tau_{r,s}, \text{ for } k \neq l \text{ and } (m_l, m_k) \in \text{ the same treatment cohort } r \tag{32}$$

event relative to treatment received earlier.

### 2.4 Challenges

Naturally there are challenges to our proposed approach to detect treatment, and properties of the final identified model will depend on the machine learning/model selection algorithms employed. In section 3, we briefly discuss the general properties of using 'gets' (through impulse- and step-indicator saturation, IIS and SIS respectively) and the (adaptive) LASSO (Tibshirani 1996, Zou 2006) and how they relate to the power of identifying treatment correctly, controlling the false-positive rate of retained break variables, and conducting valid inference.

First, we may miss that treatment occurred (a relevant break variable is not retained). However, our approach allows a researcher to embed a known or suspected treatment just like in a difference-in-difference treatment evaluation setting (see section 3.2). Additionally, varying the acceptable false-positive rate can also result in identifying more potential treatments.

Second, we may detect spurious treatment as false-positives by retaining irrelevant break variables. Though, the 'gets' approach described in section 3.1.1 allows an explicit control of the false-positive rate.

Third, we may face challenges with post-selection inference (effects may be biased as large breaks are more likely to be retained than small ones). Some of these concerns can be mitigated through bias-correction and adjustment for post-selection inference.[2]

Fourth, if treatment affected all units under analysis, then the treatment effect will be subsumed into the year fixed effects $g_t$ and not detectable as such – but again such a treatment would also not be identified with comparable treatment evaluation methods.

## 3 Operationalising the Detection of Treatment Assignment and Timing

### 3.1 Detection Methods and Their Approximate Properties

The idea of detecting structural breaks to identify treatment can be operationalised by applying break-detection in a panel setting, starting with the general saturated models (20) or (25) for fully time-varying and piece-wise constant treatment effects respectively. We emphasise that the idea of detecting treatment by detecting breaks is separate from the method of implementation – there are numerous possible machine learning detection/selection methods available and their properties will determine the effectiveness of detecting previously unknown treatment. Here we briefly consider two model selection approaches: general-to-specific selection using impulse-indicators (interpreted as an outlier-robust estimator) or step-indicators (for piece-wise constant treatment), and the (adaptive) LASSO, though we are not limited to these in practice. For example, the group-specific break detection approach in Okui and Wang (2021) could be a promising future avenue of detecting treated groups rather than individuals. Note that in the outlier-robust/general-to-specific setting, the problem of variable selection is generally studied under the null of no treatment – i.e. we focus on the false positive rate of detection which is also the main calibration parameter. In turn, in the shrinkage-based model selection literature (e.g. the LASSO) the focus has been on consistent selection, with less attention being paid to the false positive rate. We also consider the performance of each in a simulation study reported in Supplementary Section 6.2.

[2] See e.g., the coefficient bias correction function in the 'gets' package.

### 3.1.1 Treatment Detection using 'gets' and Impulse- or Step- Indicator Saturation

The impulse-indicator saturated model (20) is equivalent to impulse indicator saturation (IIS, see Hendry, Johansen and Santos, 2008) in a panel and can be interpreted as a Huber-skip outlier-robust estimator (see e.g. Jiao et al. 2021, Johansen & Nielsen 2009, Johansen & Nielsen 2016a). Coefficients on dummies that are used to determine outliers correspond thus to individual- and time-specific treatment effects. IIS has well-established properties under the null of no outliers (here interpreted as no treatment/zero treatment effects) where the false positive rate can be easily controlled by specifying the relevant tuning parameter. IIS corresponds to a robust Huber Skip estimator and targets the false-positive rate of detection by removing impulse indicators up to the chosen level of significance $\gamma_c$. For example, under a normal reference distribution, choosing $c = 1.96$ would correspond to a target level of significance of $\gamma_{1.96} = 0.05$. We denote the observed false positive rate $\hat{\gamma}_c$ as the proportion of spuriously retained indicators at the chosen cut-off $c$ out of all possible break variables considered:

$$\hat{\gamma}_c = \frac{L_c}{L} \tag{36}$$

where $L_c = \sum_{j=1}^{\hat{M}} \hat{S}_j$ is the number of retained indicator variables at cut-off $c$ and $L$ denotes the total number of potential treatment variables selected over, usually equal to the total sample size $L = n = NT$ in a balanced panel allowing for treatment at any point in time for every unit. The asymptotic properties of IIS under the null of no breaks as the total sample size $n \to \infty$ are explored in Hendry et al. (2008) and Johansen and Nielsen (2009; 2016b), who show that when there are no breaks (and accounting for multiple testing), the false positive rate of retained breaks (i.e. the number of retained indicator dummies $L_c$ relative to all possible indicators $L$) converges to the chosen nominal level of significance of selection $\gamma_c$:

$$\hat{\gamma}_c = \frac{L_c}{L} \to \gamma_c, \text{ as } n \to \infty \tag{37}$$

where $n$ denotes the sample size (in a balanced panel $n = NT$). In other words, if there is no treatment effect (i.e. if there are no true underlying breaks), then the proportion of spuriously detected indicators converges to the chosen level of significance, e.g. 1% for $\gamma_c = 0.01$. In the present context of detecting treatment at any point in time for any unit in a balanced panel, selecting at $\gamma_c = 0.01$ yields an expected number of $0.01 \times NT$ spuriously retained indicators. Thus, IIS in a Huber-Skip robust interpretation makes it straight forward to control the false discovery rate of breaks (and thereby treated units) by varying $\gamma_c$.

We can estimate the set of treated units $\widehat{H}$ as those that have at least one treatment indicator (i.e. impulse dummy) retained:

$$i \in \widehat{H} \text{ if } \hat{Q}_i > 0 \tag{38}$$

For practical purposes, this definition could also be made more stringent to differentiate between 'outliers' and actual treatment that can be attributed to potential causes. In other words, we could restrict identification of treatment to some minimum of consecutive impulse dummies. The number of estimated treatment breaks for unit $j$ is given by $\hat{S}_j$ (with $E[\hat{S}_j] = \gamma_c \times T$). The probability of a particular unit in the panel being falsely-classified as ever-treated depends on the number of time series observations for each unit. Consider a panel of $N$ individuals over $T$ time periods. IIS adds $NT$ dummies, with an expected number of retained dummies of $\gamma_c \times NT$. The probability of at least one break per individual will depend on the number of time periods in the sample and the cut-off $\gamma_c$. The probability of a particular unit $i$ being spuriously classified as ever-treated is given the probability of at least one observation

of unit $i$ being falsely-classified as treated:

$$P(i \in \widehat{H}|d_i = 0) = 1 - (1 - \gamma_c)^T \tag{39}$$

which increases with $T$ because for larger samples (and fixed $\gamma_c$) the probability of retaining an indicator spuriously increases as the number of indicators increases with $T$. Under the null of no treatment (when in fact no unit is treated), the expected number of falsely detected treated units is then given by:

$$E[\hat{M}] = P(i \in \widehat{H}|d_i = 0) \times N = (1 - (1 - \gamma_c)^T)N \tag{40}$$

If we are worried about the false-positive *rate* of treated units specifically (rather than the false-positive rate $\gamma_c$ of treatment at any point in time for any unit), it is possible to scale $\gamma_c$ to ensure a stable false-positive rate of classifying the treatment group. Let $p_H$ denote the target false positive rate of a unit being incorrectly-classified as treated. Then for any target false positive rate $p_H$, we can choose $\gamma_c$ as:

$$\gamma_c = 1 - (1 - p_H)^{\frac{1}{T}} \tag{41}$$

This controls the false positive rate of being assigned to the ever treated group to $p_H$ in expectation. For example if $T = 50$, and we aim for a false-positive rate of a single unit incorrectly being classified as treated of 5% i.e. $p_H = 0.05$, then we should set the nominal level of selection to $\gamma_c = 0.001 = 1 - (1 - 0.05)^{\frac{1}{50}}$. Similarly, we could set the target level of significance to maintain a stable *number* of false-positive treated units. If on average we are willing to accept a total of $N_0 = E[\hat{M}]$ false-positive treated units in expectation (where $\hat{M}$ is the estimated number of treated units), the above results imply that:

$$N_0 = (1 - (1 - \gamma_c)^T)N \tag{42}$$

which can be targeted by setting $\gamma_c$ to:

$$\gamma_c = 1 - (1 - \frac{N_0}{N})^{\frac{1}{T}} \tag{43}$$

and which will yield $N_0$ expected treated units in expectation when there are in fact no treated units in the true underlying DGP. For example, if we have a panel of $N = 20,\ T = 50$, and we are willing to accept one unit to be falsely-classified as treated on average ($N_0 = 1$), then we can set $\gamma_c = 0.001 \approx 1 - (1 - \frac{1}{20})^{\frac{1}{50}}$. Thus, if we are concerned about the false positive rate of treatment classification, then treatment detection in a panel perhaps warrants tighter target significance levels $\gamma_c$ than conventionally used in the selection/break detection literature.

If we consider the piece-wise constant treatment effects model matched by step indicators (25), then selecting over treatment variables using the tree search 'gets' is equivalent to applying step-indicator saturation (SIS, Castle et al., 2015) in a fixed effects panel where blocks of steps are included for each individual. SIS uses a near exhaustive tree-search based on a specified level of significance $\gamma_c$ up to which individual step-functions are removed. The properties of SIS are reasonably well-understood (see Castle et al., 2015; Nielsen & Qian, 2018), and transfer to the panel setting when interpreted as a least-squares dummy variable estimator. The asymptotic properties of SIS under the null of no breaks as $n \to \infty$ are explored in Nielsen & Qian (2018), who show that when there are no breaks (and accounting for multiple testing), the false positive rate of retained breaks (i.e. the number of detected break indicators $L_c$ relative to all possible break variables $L$) converges to the chosen nominal level of significance of selection $\gamma_c$:

$$\hat{\gamma}_c = \frac{L_c}{L} \to \gamma_c, \text{ as } n \to \infty \tag{44}$$

Specifically, if we allow for possible treatment of each unit at every point in time, then – in absence of treatment – the expected value of detected breaks is $\gamma_c \times N(T-1)$ in a balanced panel.[3] Which again translates into a probability of being classified as treated as above, with an exponent of (T-1):

$$P(i \in \widehat{H} | d_i = 0) = 1 - (1 - \gamma_c)^{(T-1)} \tag{45}$$

Then for any target false positive rate of being classified as treated $p_H$, we could choose $\gamma_c$ as:

$$\gamma_c = 1 - (1 - p_H)^{(1/(T-1))} \tag{46}$$

This matches the properties of IIS except we are searching over $N(T-1)$ rather than $NT$ possible indicators in an exhaustive search – of course the first indicator would coincide with the fixed effects.

Under the alternative (i.e. in the presence of actual treatment), for simple cases (where the number of variables does not exceed the number of observations), 'gets' has been shown to be a consistent model selection procedure retaining all relevant variables with probability equal to one as $n \to \infty$ (see e.g. Campos et al., 2003). In our setting where the number of variables can exceed the number of observations, we investigate the performance under the alternative (in the presence of structural breaks/treatment) using a range of simulations (see section 6.2). As the selection rule is pre-specified, coefficients on impulse and step-indicators could be bias-corrected to address concerns about post selection inference (see Pretis et al. 2018 for an implementation of bias correction in SIS). The 'gets' selection approach using IIS or SIS can be readily implemented to detect treatment breaks in panels using the R-package 'gets' with the 'getspanel' update.

### 3.1.2 Treatment Detection using the (adaptive) LASSO

As a second possible selection approach we briefly consider one variant of the LASSO to detect unknown treatment in the TWFE panel model. Unlike 'gets', the LASSO does not target the false-positive rate, instead penalising the L1-norm of possible coefficients. The simple LASSO itself is not a consistent model selection method, however, the adaptive LASSO which modifies the weights on coefficients is consistent and exhibits oracle properties (Huang et al., 2008). In particular, Huang et al. (2008) show the oracle properties of adaptive LASSO in high-dimensional problems where the number of regressors increases with the sample size.

To implement the LASSO to identify treatment we require different weights $v$ on the coefficients that will be penalised. We specify the weights on control variables such that these are never removed from the model (e.g. the individual and time fixed effects), while the potential treatment variables will receive penalty weights that allow them to be dropped from the model. Since the base models with impulses (20) or steps (25) may contain more variables than observations we cannot use conventional OLS as an initial estimator to determine the penalty weights for the adaptive LASSO. Instead, we follow the fixed effects panel approach of Kock (2013) and Kock (2016) to use the conventional LASSO as an initial estimator, and subsequently take the inverse of the initial LASSO coefficients as the penalty weights in the second step of the adaptive LASSO. The least squares objective function for the adaptive LASSO implementation in our setting is given by:

$$\underset{c,g,\tau}{\arg\min} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - c_i - g_t - \sum_{j=1}^{N} \sum_{s=2}^{T} \tau_{j,s} 1_{\{i=j, t \geq s\}})^2 + \lambda \sum_{j=1}^{N} \sum_{s=2}^{T} v_{j,s} |\tau_{j,s}| \tag{47}$$

[3] Albeit simulation results show a higher false-positive rate for SIS in small samples, warranting perhaps a more conservative choice of $\gamma_c$.

where the second term denotes the penalty term on the break coefficients $\tau$ with tuning parameter $\lambda$ and penalty weights $v$ corresponding to the inverse of the coefficients in an initial LASSO estimator. The tuning parameter $\lambda$ can be chosen using using cross-validation or information criteria. Closely related to our work, albeit not focused on fixed effects, Qian & Su (2016a) use the adaptive LASSO to estimate common breaks across individuals. Okui & Wang (2021) show that the adaptive LASSO – albeit using a fused structure – can further be used to estimate breaks that are heterogeneous across groups. However, they do not focus on breaks in fixed effects or treatment evaluation. Larger breaks, i.e. larger treatment effects, are more likely to be retained in the final model – akin to the gets approach in section 3.1.1 – potentially complicating inference on the final retained model. Post-selection inference has received a fair amount of attention in the LASSO literature. Simple data-splitting approaches (such as Cox 1975) are not feasible in our setting as the treatment variables only apply to a subset of observations. Lee et al. (2016) propose a post-selection inference correction for the LASSO. Alternatively, Zhao et al. (2021) show that the naive approach of re-estimating an OLS model post-selection can perform surprisingly well in many settings.

### 3.2 Embedding Known Interventions

There are two ways in which break detection to identify treatment can be implemented: either as an agnostic way to detect fully unknown treatment assignment and timing, or as a robustness check embedding known treatment and searching for additional previously-unknown interventions. Above we outlined the case where we detect treatment as a purely agnostic data-driven approach to identify interventions without any prior knowledge of their occurrence. While the approach is agnostic and any unit may be treated at any point in time, a potential downside is a loss in power if treatment assignment and timing is known and there are multiple treated units with a homogeneous treatment effects, since each treated unit would have to be identified individually.

If treatment assignment and timing is known for a particular intervention then break detection can be adapted as a robustness check for additional unknown treatment in conventional TWFE difference-in-differences models. In this case we force the known treatment dummy (or dummies for interactions) to be included in the model, and select over additional treatment indicators. This corresponds to the Hendry & Johansen (2015) theory-embedding approach where fixed regressors are embedded in a wider information set that we select over.

Then selection takes place over the break variables to detect additional treatment (omitting the break variables perfectly coinciding with known treatment dummies), known treatment dummies remains in the model without being selected over. This allows additional unknown treatment to be detected, while the coefficient on the forced (not-selected-over) break variable yields an estimate of the conventional treatment effect in a TWFE panel. It is worth highlighting that we do not necessarily need to allow for a break at every point in time (or individual). If there is a strong reason that the break should be localised in particular time periods (or among particular individuals), then only those could be included in the candidate set of break variables selected-over.

### 3.3 Ex-Post Attribution of Detected Events

Having identified treatment as structural breaks the remaining challenge is then to attribute the detected effects to possible causes. Much of the difference-in-differences TWFE literature is dedicated to justifying that specific known interventions were exogenous (or as-if randomly assigned). Similar to such subject-specific justifications, the ex-post attribution of events to possible causes will require subject-specific knowledge. Ultimately, in absence of a randomised experiment, making the case for a known

intervention to be exogenous is comparable to searching for a potential cause of a detected effect. Particularly, once a potential cause of a detected effect has been identified, we could have simply estimated a conventional difference-in-differences model using the 'known' intervention. Thus, in the proposed reverse causal approach we expect that much of the discussion will be dedicated to arguing that a particular detected break coincides with a particular event that was discovered after the effect was observed. Naturally there may be many such events that took place and it can be difficult to attribute the observed effect to that single event. However, the same challenge applies in 'known' treatment evaluation – treatment has to occur in isolation without other events taking place at the same time affecting the treated units. So while the search for causes is different than arguing that a cause was unique, subject-specific knowledge will be necessary in both settings.

# 4 Illustrative Application: Detecting the Impacts of Terrorism on GDP per Capita

We demonstrate our break detection approach to identify unknown treatment assignment and timing using a well-known dataset on Spanish regional GDP per capita (see Abadie & Gardeazabal, 2003). We purposely choose a well-known example to illustrate our methods. We also provide a policy-focused application with novel data in our closely-related paper in Koch et al. (2022). The dataset for our illustrative application here spans all of mainland Spain's 15 regions (where we exclude the Canary and Balearic Islands) over 31 years from 1965 to 1995 for a total of 465 region-year observations. In their seminal paper, Abadie & Gardeazabal (2003) used a forward causal approach to study the effect of ETA terrorism on regional economic output. The authors find a substantial reduction in regional GDP in response to local terrorism introducing synthetic control methods. Here we ask the reverse causal question: what affected regional GDP per capita in the Basque Country (or wider Spain)? We show that the "treatment" taking the form of ETA terrorism (alongside a number of other previously unidentified treatments) can be detected without prior knowledge of its occurrence using our proposed break detection approach.

To illustrate our methods, we first consider a simple TWFE panel setting with two regions (the Basque Country and Madrid) where we search for breaks to detect treatment in GDP per capita.[4] We then expand this into a multi-region panel of mainland Spain to assess breaks in a wider context. Our results show that we can detect the effect of ETA terrorism without prior knowledge of its occurrence and obtain treatment effect estimates that are near-identical to a known-treatment model. Our break detection approach also provides evidence that the treatment effects of GDP impacts of ETA terrorism were transitory and are no longer detectable post-1990. In addition, in the panel with more than two regions we also detect breaks which we attribute to an industrial crisis and increased autonomy following the Franco era in other regions.

## 4.1 Detecting Treatment in a Panel with Two Regions

We first consider a simple panel with two regions: the Basque country and Madrid ($N = 2, T = 31$, $NT = 62$). For comparison, we initially estimate the forward causal 'infeasible' model of log GDP per capita (controlling for log investments $Inv$ similar to Abadie & Gardeazabal 2003) using a TWFE estimator with a known intervention of Basque terrorism to provide a baseline relative to our break

[4] For completeness we also show that a simple time series model of Basque GDP per capita is unable to identify ETA terrorism impacts due to the lack of control groups – see Supplementary Material 6.3).

detection approach. We then demonstrate that we can directly detect the terrorism 'treatment' without prior knowledge using our reverse causal approach.

As a baseline, consider a TWFE estimator with a 'known' intervention of ETA terrorism. We estimate baseline models first allowing for time-varying treatment effects using interactions in (48), then assuming time-invariant treatment effects in (49) specified as a dummy variable for the Basque region in the 'post-treatment' period, defined here as 1979 onwards, as Abadie & Gardeazabal (2003) found that the impact of terrorism was notable in GDP per capita from the end of the 1970s.

'Known' Treatment (fully time-varying treatment effects):

$$\begin{aligned} log(GDPpc_{i,t}) &= \alpha_i + \phi_t + \sum_{s=1979}^{1995} d_i \tau_s 1_{\{t=s\}} + \beta_1 log(Inv)_{i,t} + u_{i,t} \\ &\text{where } d_i = 1_{\{i=\text{Basque}\}} \end{aligned} \tag{48}$$

'Known' Treatment (time-constant treatment effects):

$$\begin{aligned} log(GDPpc_{i,t}) &= \alpha_i + \phi_t + d_{i,t}\tau + \beta_1 log(Inv)_{i,t} + u_{i,t} \\ &\text{where } d_{i,t} = 1_{\{i=\text{Basque}, t \geq 1979\}} \end{aligned} \tag{49}$$

Estimation results for the 'known' baseline models are shown in Tables 1 and 2, under the columns "Known TWFE". The results of the known baseline show an approximate 5% reduction in GDP per capita in the Basque country relative to Madrid in response to ETA terrorism in this simple two-region model. This result is similar across the time-varying model (see equation 20) (where the estimated ATT is given by the average of the impulse coefficients) as well as the piece-wise constant treatment effects model (see equation 25). Specifically, the ATT across impulses in the known baseline in (48) is -0.0496 (se=0.0197), and the time-constant estimate given by the coefficient in (49) on the known step-function is -0.0495 (se=0.006).

Now suppose the "treatment" of ETA terrorism in the Basque country was unknown, and we approached the data with our reverse causal question of 'what affected GDP per capita? We demonstrate how treatment interventions can be detected without prior knowledge of their occurrence.

#### 4.1.1 Unknown Treatment with Fully Time-Varying Effects

We now estimate a model allowing for the potential treatment of any unit at any point in time first using impulse dummies capturing time-varying treatment and select over them using the 'gets' selection algorithm (we consider the LASSO for the piece-wise constant setting below). The model is saturated with a full set of impulse dummies in (50) which are selected over at a target level of significance $\gamma_c$. We consider three different target significance levels, $\gamma_c = 0.05$ as well as $0.025$ and $0.01$ to illustrate the impact of the calibration choice on treatment detection.

'Unknown' Treatment:

$$log(GDPpc_{i,t}) = c_i + g_t + \sum_{j=1}^{2} \sum_{s=1966}^{1995} \tau_{j,s} 1_{\{i=j,t=s\}} + \beta_1 log(Inv)_{i,t} + u_{i,t} \tag{50}$$

The resulting detected impulses, which we interpret as unit-specific time-varying treatment effects, are shown in Figure 2 (for $\gamma_c = 0.05$) and Table 1 (for all three values of $\gamma_c$). We detect the treatment of

Basque terrorism without prior knowledge of its occurrence as individual impulses in the Basque region from 1980 to 1990. Each coefficient provides an estimate of the unit- and time-specific treatment effect. We can easily compute our estimates of the ATT by taking the mean of the impulses over time. Standard errors for the ATT are also straight-forward to compute as impulses are orthogonal. Computing the ATT over the time period from 1980 to 1990 from the model with $\gamma_c = 0.05$ yields an estimate of the ATT of -0.059 (se=0.016) which is near identical (and not significantly different) to the known-treatment baseline estimate of -0.0496 (se=0.0197). The fact that the ATT using detected impulses is marginally larger than the known baseline ATT can be explained by the fact that the impulses are only retained up to 1990 while the 'known' baseline time-varying treatment considers treatment effects up until the end of the sample in 1995. Indeed, we only detect treatment breaks up until 1990, suggesting that the impacts of ETA terrorism on GDP were transitory and no longer detectable post-1990. This is consistent with the known-treatment baseline which finds predominantly insignificant time-varying treatment effects after 1990.

Varying $\gamma_c$, we successfully detect the intervention at relatively loose levels of significance $\gamma_c = 0.05$ or $\gamma_c = 0.025$. The loss of power for more conservative levels of the target false positive rate becomes apparent when we set $\gamma_c = 0.01$, where we do not detect any treatment as impulse dummies coinciding with ETA terrorism. However, this reduction in power can be tackled by specifying piece-wise constant treatment effects using step functions as we demonstrate in the following section 4.1.2. Note that in this N=2 panel, the treatment effects are relative to the single control region and one could achieve the same detected treatment if Basque country was selected as the 'control', in which case the treatment effects would be detected for Madrid and opposite-signed. We would then interpret them as the effect of the absence of terrorism.

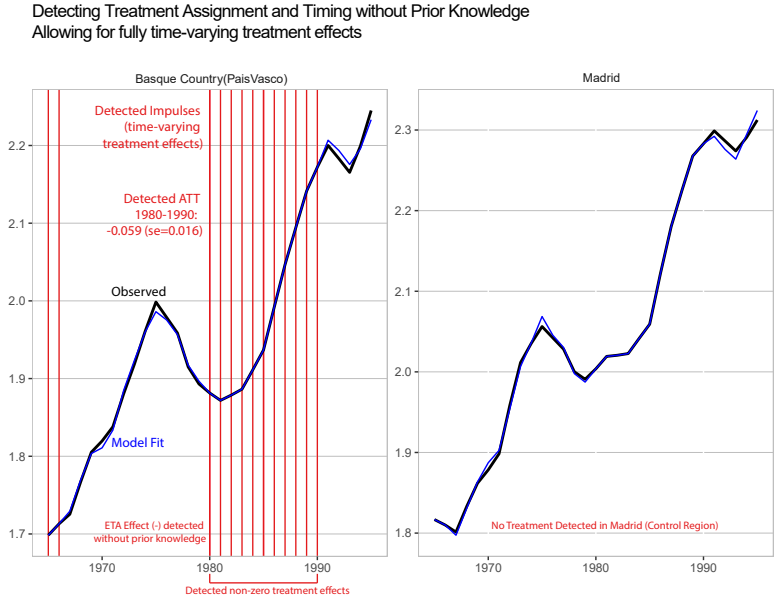


Figure 2: TWFE Panel: GDP per Capita, Basque Country & Madrid – Time-Varying Treatment detected using IIS in 'gets' for a target significance level of $\gamma_c = 0.05$. Red vertical lines denote detected impulses (identifying time-varying treatment effects).

Table 1: Detecting Fully-Time-Varying Treatment: Two-Region Panel Model

| Dependent Variable: | log(GDPpc) (Basque & Madrid) | | | |
|---|---|---|---|---|
| | **Unknown Treatment** | | | **Known Treatment** |
| Model: | gets ($\gamma_c = 0.05$) | gets ($\gamma_c = 0.025$) | gets ($\gamma_c = 0.01$) | 'Known' TWFE |
| *Variables* | | | | |
| log(Invest) | 0.1048*** | 0.0777* | -0.0234 | -0.0638 |
| | (0.0329) | (0.0376) | (0.0534) | (0.0563) |
| $\tau$: $i$=Basq.,$t$=1965 | -0.0657*** | -0.0559*** | -0.0156 | |
| | (0.0163) | (0.0193) | (0.0320) | |
| $\tau$: $i$=Basq.,$t$=1966 | -0.0352** | | | |
| | (0.0151) | | | |
| $\tau$: $i$=Basq.,$t$=1979 | | | | -0.0348 |
| | | | | (0.0213) |
| $\tau$: $i$=Basq.,$t$=1980 | -0.0514*** | -0.0463** | | -0.0474** |
| | (0.0143) | (0.0173) | | (0.0183) |
| $\tau$: $i$=Basq.,$t$=1981 | -0.0859*** | -0.0783*** | | -0.0664*** |
| | (0.0151) | (0.0181) | | (0.0189) |
| $\tau$: $i$=Basq.,$t$=1982 | -0.0661*** | -0.0622*** | | -0.0699*** |
| | (0.0142) | (0.0172) | | (0.0185) |
| $\tau$: $i$=Basq.,$t$=1983 | -0.0679*** | -0.0621*** | | -0.0598*** |
| | (0.0145) | (0.0175) | | (0.0183) |
| $\tau$: $i$=Basq.,$t$=1984 | -0.0743*** | -0.0652*** | | -0.0455** |
| | (0.0158) | (0.0189) | | (0.0199) |
| $\tau$: $i$=Basq.,$t$=1985 | -0.0630*** | -0.0549*** | | -0.0401* |
| | (0.0154) | (0.0184) | | (0.0193) |
| $\tau$: $i$=Basq.,$t$=1986 | -0.0615*** | -0.0557*** | | -0.0530** |
| | (0.0145) | (0.0175) | | (0.0183) |
| $\tau$: $i$=Basq.,$t$=1987 | -0.0514*** | -0.0492** | | -0.0655*** |
| | (0.0142) | (0.0173) | | (0.0194) |
| $\tau$: $i$=Basq.,$t$=1988 | -0.0475*** | -0.0456** | | -0.0632*** |
| | (0.0142) | (0.0173) | | (0.0196) |
| $\tau$: $i$=Basq.,$t$=1989 | -0.0483*** | -0.0451** | | -0.0561** |
| | (0.0142) | (0.0172) | | (0.0188) |
| $\tau$: $i$=Basq.,$t$=1990 | -0.0342** | | | -0.0395* |
| | (0.0142) | | | (0.0186) |
| $\tau$: $i$=Basq.,$t$=1991 | | | | -0.0328 |
| | | | | (0.0200) |
| $\tau$: $i$=Basq.,$t$=1992 | | | | -0.0354* |
| | | | | (0.0194) |
| $\tau$: $i$=Basq.,$t$=1993 | | | | -0.0443* |
| | | | | (0.0207) |
| $\tau$: $i$=Basq.,$t$=1994 | | | | -0.0332 |
| | | | | (0.0237) |
| $\tau$: $i$=Basq.,$t$=1995 | | | | -0.0046 |
| | | | | (0.0214) |
| *Fixed-effects* | | | | |
| Region | Yes | Yes | Yes | Yes |
| Year | Yes | Yes | Yes | Yes |
| *Fit statistics* | | | | |
| Observations | N=2, T=31 | N=2, T=31 | N=2, T=31 | N=2, T=31 |
| Within $R^2$ | 0.87 | 0.79 | 0.023 | 0.84 |

*Standard-errors in parentheses*
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

### 4.1.2 Unknown Treatment with Piece-Wise Constant Effects

Treatment effects may be piece-wise constant and thus detected with greater likelihood (due to the higher power of step-functions). To illustrate this, we estimate a TWFE panel (51) saturated with a full set of step-functions denoting potential treatment in either region (Basque or Madrid) at any point in time:

'Unknown' Treatment:

$$log(GDPpc_{i,t}) = c_i + g_t + \sum_{j=1}^{2}\sum_{s=1966}^{1995} \tau_{j,s} 1_{\{i=j,t\geq s\}} + \beta_1 log(Inv)_{i,t} + \epsilon_{i,t} \tag{51}$$

We select over treatment functions using 'gets' at two target levels of $\gamma_c = 0.001$ and $\gamma_c = 0.01$ as well as using the adaptive LASSO where we penalise the possible treatment coefficients $\tau$, with penalty weights chosen using the simple LASSO as an initial estimator, and the tuning parameter selected using cross-validation.

Table 2 and Figure 3 shows the results of break detection. The adaptive LASSO estimates are reported using the 'naive' approach of re-estimating the selected model using OLS (see e.g. Zhao et al., 2021). Detecting treatment using 'gets' at $\gamma_c = 0.001$ results in a single treatment indicator being retained for the Basque Country from 1979 onwards. The resulting selected model is *identical* (in absence of any bias-correction due to selection) to the TWFE estimator with known treatment intervention imposed, with the estimated coefficient on the retained break variable of -0.0496 (se=0.0197) matching the estimated treatment effect in the TWFE difference-in-differences model. In other words – without knowing that treatment occurred – we are able to detect the treatment intervention and estimate a model effectively *identical* to the known intervention panel. Similarly, the adaptive LASSO is able to identify treatment (detecting a negative intervention in Basque country in 1980), with the estimated 'naive' post-LASSO treatment effect near identical to the 'known' imposed intervention in 1979. The adaptive LASSO further detects additional earlier breaks which is unsurprising as it can be often less conservative than 'gets' with low levels of $\gamma_c$. Relaxing the target level of $\gamma_c$ to a less conservative level of 0.01 results in additional breaks being detected which can be interpreted as time-varying treatment effects: the negative break in 1981 suggests that the initial impact of ETA terrorism became larger in the early 1980s, however, the opposite-signed break in 1990 provides evidence of the transitory nature of the impact. Consistent with our results from the fully-time-varying specification (and known baseline), treatment effects post-1990 are closer to zero (see section 4.1.1).

Overall, both 'gets' and the adaptive LASSO implementation of our proposed break detection approach detect the 'treatment' without prior knowledge of its occurrence. Break detection estimates to detect treatment suggest a roughly 5% reduction in GDP per capita in response to terrorism in the Basque region relative to Madrid as the control region, which is identical to the known intervention TWFE estimator. Further, it is worth noting that the break detection approach suggests a reduction in GDP per capita from around 1979/1980 onwards, which is consistent with Abadie and Gardeazabal's finding that GDP per capita reductions occurred with a lag relative to the onset of terrorism in the mid 1970s.

Thus, not only are we able to detect treatment without prior knowledge on which regions were treated and when treatment occurred, but the estimated break dates also provide insights into the lagged onset of the economic impacts of terrorism.

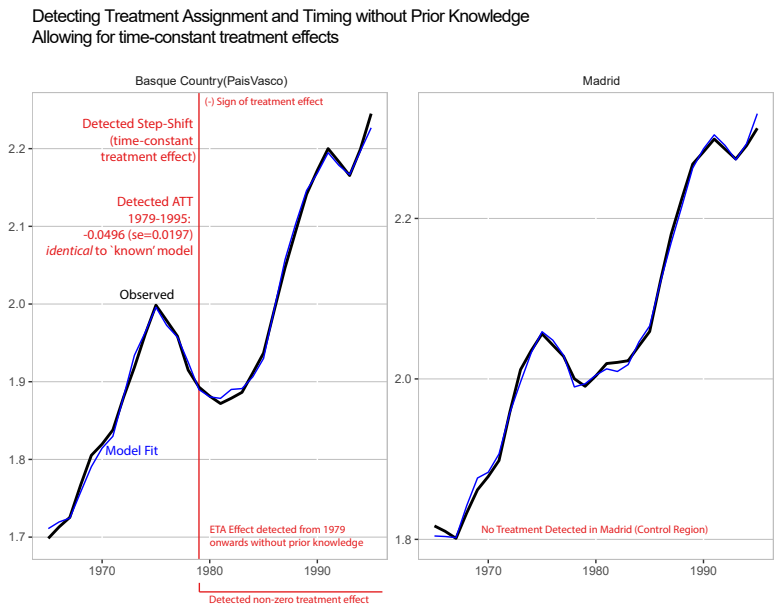


Figure 3: TWFE Panel: GDP per Capita, Basque Country & Madrid – Treatment detected using SIS in 'gets' at $\gamma_c = 0.001$. Red vertical lines denote detected step-shifts (identifying treatment effects).

Table 2: Detecting Piece-Wise Constant Treatment: Two-Region Panel Model

| Dependent Variable: | log(GDPpc) (Basque & Madrid) | | | |
|---|---|---|---|---|
| | **Unknown Treatment** | | | **Known Treatment** |
| Model: | gets ($\gamma_c = 0.001$) | gets ($\gamma_c = 0.01$) | Adapt. LASSO | 'Known' TWFE |
| *Variables* | | | | |
| log(Invest) | -0.1065*** | -0.0540* | -0.0624* | -0.1065*** |
| | (0.0294) | (0.0314) | (0.0320) | (0.0294) |
| $\tau$: Break ($i$=Basq, $t \geq 1966$) | | | 0.0324 | |
| | | | (0.0190) | |
| $\tau$: Known ETA ($i$=Basq, $t \geq 1979$) | | | | **-0.0495***** |
| | | | | (0.0063) |
| $\tau$: Break ($i$=Basq, $t \geq$1979) | **-0.0495***** | **-0.0401***** | | |
| | (0.0063) | (0.0115) | | |
| $\tau$: Break ($i$=Basq, $t \geq$1980) | | | **-0.0471***** | |
| | | | (0.0065) | |
| $\tau$: Break ($i$=Basq, $t \geq$1981) | | -0.0176 | | |
| | | (0.0119) | | |
| $\tau$: Break ($i$=Basq, $t \geq$1990) | | 0.0277*** | | |
| | | (0.0092) | | |
| *Fixed-effects* | | | | |
| Region | Yes | Yes | Yes | Yes |
| Year | Yes | Yes | Yes | Yes |
| *Fit statistics* | | | | |
| Observations | N=2, T=31 | N=2, T=31 | N=2, T=31 | N=2, T=31 |
| Within $R^2$ | 0.69 | 0.77 | 0.671 | 0.69 |

*Standard-errors in parentheses*
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

### 4.2 Detecting Treatment in a Panel with Multiple Regions

We repeat the above analysis for a panel covering all of mainland Spain using 'gets'. We now include all $N = 15$ regions of mainland Spain over $T = 31$ years (for a total sample size of $NT = 465$). Just as before, we compare the detected treatment in this larger panel to the benchmark of a known intervention by imposing the 'treatment' as a dummy variable for the Basque region from 1979 onwards in a TWFE estimator. The 'known treatment' baseline yields an estimated treatment effect of -0.155 (se=0.018) relative to the control regions in wider Spain (see Table 3).[5]

Our break detection results using gets at $\gamma_c = 0.001$ show that even in this more general setting we are able to detect the treatment of ETA terrorism through the impacts on GDP in the Basque Country without prior knowledge of its occurrence (see Figure 4 and Table 3). The ETA treatment is detected in 1978 (close to the imposed intervention in the known TWFE estimator in 1979) with an estimated treatment effect of -0.156 (se=0.012) which is *near identical* to the 'known treatment' benchmark.

In addition to the ETA break in 1978, we also detect a small number of possible treatment effects through breaks in the fixed effects of other regions.[6] It is worth noting though that the break associated with the ETA 'treatment' is the single largest break in magnitude compared to all detected breaks. Given the set of detected effects (captured through breaks) for some of the regions, the next step of our approach (see section 3.3) is to investigate the relevant literature for potential causes.

A brief review of the literature on Spanish economic history suggests that the positive breaks (i.e. positive treatment effects on GDP per capita) in Extremadura, Galicia, and Rioja, may correspond to the increased autonomy of the regions awarded in the post-Franco era. The negative break in Madrid in 1970 coincides with an industrial crisis that hit Madrid disproportionally relative to other regions (Rodríguez-Pose & Hardy 2021, and Tobío 1989).

The fact that ex-post attribution is not always straight forward is highlighted by the fact that we have yet to identify likely causes for the positive break in Castilla-La Mancha in 1972 (though it is worth noting that the film adaptation of the highly popular musical "Man of la Mancha" was released in that year), and the negative breaks in Asturias (in 1986) and Madrid (in 1990).

[5] This estimate in the known benchmark and the detected break setting is larger than the two-region panel because the control group is different. The two-region panel only included Madrid as a control region)

[6] To control for outlying observations we also combine our selection over step functions with selection over impulse dummies, where impulses could capture outliers or can also be interpreted as single-period time-varying treatment indicators. Only a single outlying observation is identified: Madrid, 1965.

Table 3: Detecting Piece-Wise Constant Treatment: 15-Region Panel Model

| Dependent Variable: | log(GDPpc) | |
|---|---|---|
| | **Unknown Treatment** | **Known Treatment** |
| Model: | gets ($\gamma_c = 0.0001$) | 'Known' TWFE |
| *Variables* | | |
| log(Invest) | 0.1171*** | 0.1377*** |
| | (0.0121) | (0.0175) |
| $\tau$: Break ($i$=Basq, $t \geq$1978) | **-0.1560***** | |
| | (0.0120) | |
| $\tau$: Known ETA ($i$=Basq, $t \geq$1979) | | **-0.1553***** |
| | | (0.0182) |
| $\tau$: Break: ($i$=Castilla-La Mancha, $t \geq$ 1972) | 0.1169*** | |
| | (0.0143) | |
| $\tau$: Break: ($i$=Extremadura, $t \geq$ 1987) | 0.1350*** | |
| | (0.0127) | |
| $\tau$: Break: ($i$=Galicia, $t \geq$ 1976) | 0.0980*** | |
| | (0.0121) | |
| $\tau$: Break: ($i$=Madrid, $t \geq$ 1970) | -0.1256*** | |
| | (0.0176) | |
| $\tau$: Break: ($i$=Madrid, $t \geq$ 1990) | -0.0903*** | |
| | (0.0150) | |
| $\tau$: Break: ($i$=Princip. De Asturias, $t \geq$ 1986) | -0.1220*** | |
| | (0.0123) | |
| $\tau$: Break: ($i$=La Rioja, $t \geq$ 1981) | 0.0796*** | |
| | (0.0117) | |
| $\tau$: Impulse: ($i$=Madrid, $t = 1965$) | 0.0914** | |
| | (0.0356) | |
| *Fixed-effects* | | |
| Region | Yes | Yes |
| Year | Yes | Yes |
| Observations | N=15, T=31, NT=465 | N=15, T=31, NT=465 |
| Within $R^2$ | 0.71 | 0.29 |

*Standard-errors in parentheses*
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

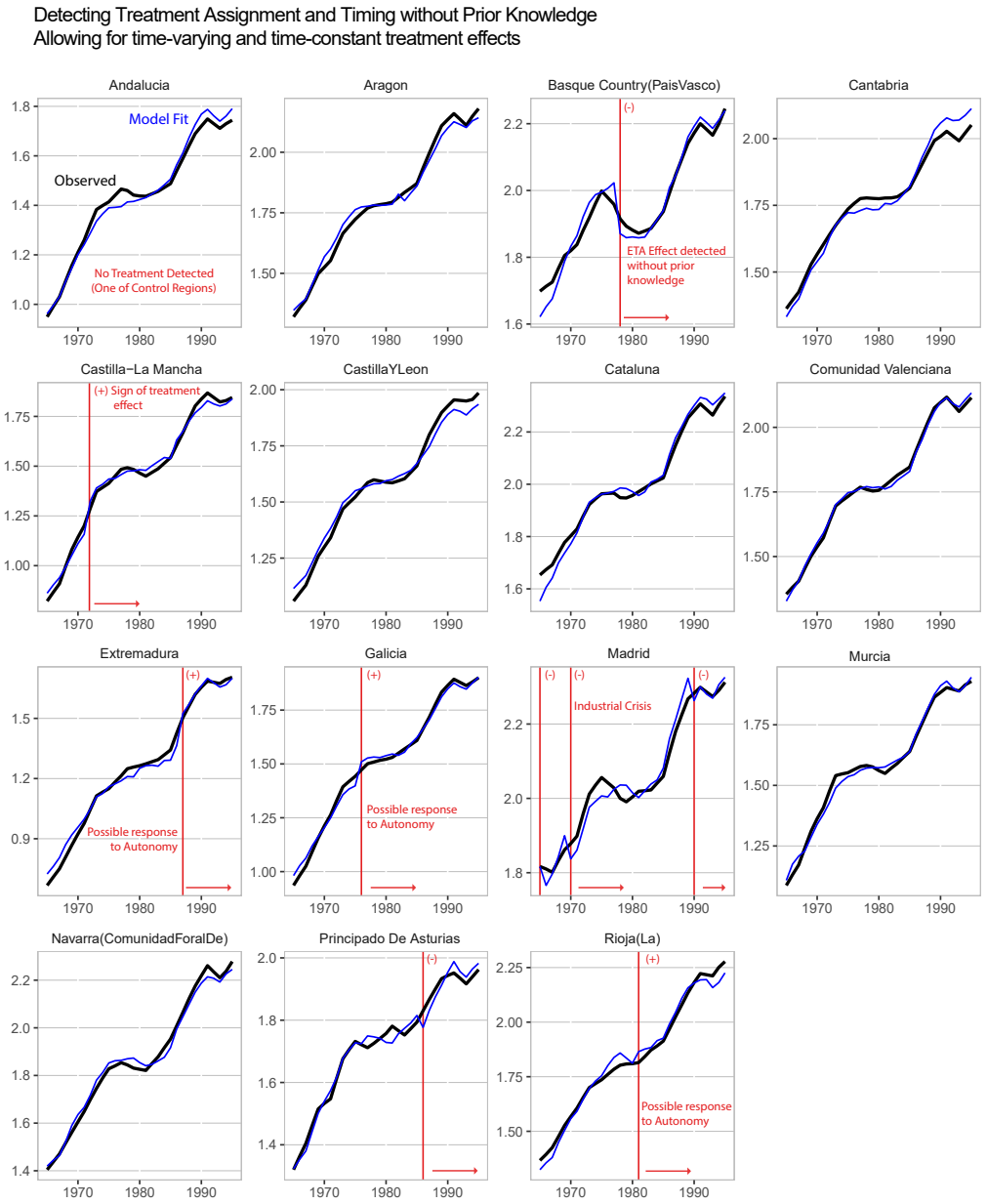


Figure 4: TWFE Panel: GDP per Capita in 15 Regions of Mainland Spain – Treatment detected using 'gets' and $\gamma_c = 0.001$. Red vertical lines denote detected step-shifts (identifying treatment effects).

## 5 Conclusion

We operationalise the modelling of reverse causal questions by searching for structural breaks in fixed effects panel models identifying previously unknown treatment effects which can subsequently be attributed to potential causes. We show that the two-way fixed effects estimator, which identifies heterogeneous treatment effects through interactions, can be nested as a special case of impulse- or step-dummy saturated models – a subset of which identifies underlying treatment effects.

We demonstrate the feasibility of detecting previously unknown treatment assignment and timing by using two machine learning methods suitable for selection over more candidate variables than observations (gets and the adaptive LASSO).

Our application to the economic impacts of terrorism in Spain demonstrates that we can detect the effects of 'treatment' (taking the form of terrorist activity) on GDP per capita without prior knowledge of its occurrence. The estimated treatment effects, when the assignment of treatment and its timing is unknown are near identical to imposing the same treatment as a known intervention *a-priori*. More broadly, our proposed approach is modular and allows for the detection of structural breaks in fixed effects panels with flexible choices for the machine learning algorithms employed. Crucially, using machine learning this allows for the detection of effective policies without prior knowledge of their occurrence or effectiveness. When using gets or the adaptive LASSO, the approach can be readily applied using our freely-available open-source R-packages 'gets' and 'getspanel'.

## References

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American economic review*, *93*(1), 113–132.

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, 821–856. (Publisher: JSTOR)

Apergis, N., & Lau, M. C. K. (2015). Structural breaks and electricity prices: Further evidence on the role of climate policy uncertainties in the Australian electricity market. *Energy Economics*, *52*, 176–182. (Publisher: Elsevier)

Bai, J. (1997). Estimating multiple breaks one at a time. *Econometric theory*, 315–352. (Publisher: JSTOR)

Bai, J., Han, X., & Shi, Y. (2020). Estimation and inference of change points in high-dimensional factor models. *Journal of Econometrics*, *219*(1), 66–100. (Publisher: Elsevier)

Baker, A., Larcker, D. F., & Wang, C. C. (2021). How much should we trust staggered difference-in-differences estimates? *Available at SSRN 3794018*.

Baltagi, B. H., Feng, Q., & Kao, C. (2016). Estimation of heterogeneous panels with structural breaks. *Journal of Econometrics*, *191*(1), 176–195. (Publisher: Elsevier)

Bazinas, V., & Nielsen, B. (2015). *Causal transmission in reduced-form models*. Citeseer.

Callaway, B., & Sant'Anna, P. H. (2020). Difference-in-differences with multiple time periods. *Journal of Econometrics*. (Publisher: Elsevier)

Campos, J., Hendry, D. F., & Krolzig, H.-M. (2003). Consistent model selection by an automatic Gets approach. *Oxford Bulletin of Economics and Statistics*, *65*, 803–819. (Publisher: Wiley Online Library)

Castle, J., Doornik, J., Hendry, D., & Pretis, F. (2015). Detecting location shifts during model selection by step-indicator saturation. *Econometrics*, *3*(2), 240–264. doi: 10.3390/econometrics3020240

Chan, F., Mancini-griffoli, T., Pauwels, L. L., & others. (2008). Testing structural stability in heterogeneous panel data. Christchurch, New Zealand. (Publisher: Citeseer)

Cheng, X., Liao, Z., & Schorfheide, F. (2016). Shrinkage estimation of high-dimensional factor models with structural instabilities. *The Review of Economic Studies*, *83*(4), 1511–1543. (Publisher: Oxford University Press)

Conley, T. G., & Taber, C. R. (2011). Inference with "difference in differences" with a small number of policy changes. *The Review of Economics and Statistics*, *93*(1), 113–125. (Publisher: MIT Press)

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, *62*(2), 441–444. (Publisher: Oxford University Press)

De Chaisemartin, C., & d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964–96.

De Chaisemartin, C., & d'Haultfoeuille, X. (2021). Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey. *SSRN*.

De Wachter, S., & Tzavalis, E. (2012). Detection of structural breaks in linear dynamic panel data models. *Computational Statistics & Data Analysis*, *56*(11), 3020–3034. (Publisher: Elsevier)

Engle, R. F., Hendry, D. F., & Richard, J.-F. (1983). Exogeneity. *Econometrica*, *51*(2), 277–304. Retrieved 2023-02-19, from http://www.jstor.org/stable/1911990

Estrada, F., Perron, P., & Martínez-López, B. (2013). Statistically derived contributions of diverse human influences to twentieth-century temperature changes. *Nature Geoscience*, *6*(12), 1050–1055. (Publisher: Nature Publishing Group)

Gelman, A. (2011). *Causality and statistical learning.* University of Chicago Press Chicago, IL.

Gelman, A., & Imbens, G. (2013, November). *Why ask Why? Forward Causal Inference and Reverse Causal Questions* (Working Paper No. 19614). National Bureau of Economic Research. Retrieved 2021-10-25, from https://www.nber.org/papers/w19614 (Series: Working Paper Series) doi: 10.3386/w19614

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*. (Publisher: Elsevier)

Hendry, D. F. (2020). *First in, first out: Econometric modelling of UK annual CO2 emissions, 1860–2017* (Tech. Rep.). Oxford: Economics Group, Nuffield College, University of Oxford. Retrieved from https://ideas.repec.org/p/nuf/econwp/2002.html

Hendry, D. F., & Johansen, S. (2015). Model discovery and trygve haavelmo's legacy. *Econometric Theory*, *31*(1), 93–114. doi: 10.1017/S0266466614000218

Hendry, D. F., Johansen, S., & Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, *23*(2), 317–335. (Publisher: Springer)

Hendry, D. F., & Santos, C. (2005). Regression models with data-based indicator variables. *Oxford Bulletin of Economics and statistics*, *67*(5), 571–595. (Publisher: Wiley Online Library)

Huang, J., Ma, S., & Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 1603–1618. (Publisher: JSTOR)

Jiao, X., & Pretis, F. (2020). Testing the presence of outliers in regression models. *Working Paper*.

Jiao, X., Pretis, F., & Schwarz, M. (2021, August). *Testing for Coefficient Distortion due to Outliers with an Application to the Economic Impacts of Climate Change* (SSRN Scholarly Paper No.

ID 3915040). Rochester, NY: Social Science Research Network. Retrieved 2021-11-03, from https://papers.ssrn.com/abstract=3915040 doi: 10.2139/ssrn.3915040

Johansen, S., & Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. *Castle, and Shephard (2009)*, *1*, 1–36.

Johansen, S., & Nielsen, B. (2016a). Analysis of the Forward Search using some new results for martingales and empirical processes. *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, *22*(2), 1131–1183. (Publisher: Bernoulli Society for Mathematical Statistics and Probability)

Johansen, S., & Nielsen, B. (2016b). Asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics*, *43*(2), 321–348. doi: 10.1111/sjos.12174

Koch, N., Naumann, L., Pretis, F., Ritter, N., & Schwarz, M. (2022). What reduces road $CO_2$ emissions? Policy attribution using break detection. *Working Paper*.

Kock, A. B. (2013). Oracle efficient variable selection in random and fixed effects panel data models. *Econometric Theory*, 115–152. (Publisher: JSTOR)

Kock, A. B. (2016). Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions. *Econometric Theory*, *32*(1), 243. (Publisher: Cambridge University Press)

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., & others. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, *44*(3), 907–927. (Publisher: Institute of Mathematical Statistics)

Li, D., Qian, J., & Su, L. (2016). Panel data models with interactive fixed effects and multiple structural breaks. *Journal of the American Statistical Association*, *111*(516), 1804–1819. (Publisher: Taylor & Francis)

Martinez, A. B. (2020). Forecast accuracy matters for hurricane damage. *Econometrics*, *8*(2), 18.

Mill, J. S. (1843). *A system of logic* (Vol. 1). London: Parker.

Mukanjari, S., & Sterner, T. (2018). Do markets trump politics? evidence from fossil market reactions to the paris agreement and the us election.

Nielsen, B., & Qian, M. (2018). *Asymptotic properties of the gauge of step-indicator saturation.*

Okui, R., & Wang, W. (2021). Heterogeneous structural breaks in panel data models. *Journal of Econometrics*, *220*(2), 447–473. (Publisher: Elsevier)

Perron, P. (1989). The great crash, the oil price shock, and the unit root hypothesis. *Econometrica: journal of the Econometric Society*, 1361–1401. (Publisher: JSTOR)

Perron, P. (2006). Dealing with structural breaks. *Palgrave handbook of econometrics*, *1*(2), 278–352.

Piehl, A. M., Cooper, S. J., Braga, A. A., & Kennedy, D. M. (2003). Testing for structural breaks in the evaluation of programs. *Review of Economics and Statistics*, *85*(3), 550–558. (Publisher: MIT Press)

Porter, J., & Yu, P. (2015). Regression discontinuity designs with unknown discontinuity points: Testing and estimation. *Journal of Econometrics*, *189*(1), 132–147. (Publisher: Elsevier)

Pretis, F. (2019). Does a carbon tax reduce CO2 emissions? Evidence from british columbia. *Working Paper*.

Pretis, F. (2021). Exogeneity in climate econometrics. *Energy Economics*, *96*, 105122. Retrieved from https://www.sciencedirect.com/science/article/pii/S014098832100027X doi: https://doi.org/10.1016/j.eneco.2021.105122

Pretis, F., Reade, J., & Sucarrat, G. (2018). Automated General-to-Specific (GETS) regression mod-

eling and indicator saturation methods for the detection of outliers and structural breaks. *Journal of Statistical Software*, *86*(3). (Publisher: Foundation for Open Access Statistics)

Qian, J., & Su, L. (2016a). Shrinkage estimation of common breaks in panel data models via adaptive group fused lasso. *Journal of Econometrics*, *191*(1), 86–109. (Publisher: Elsevier)

Qian, J., & Su, L. (2016b). Shrinkage estimation of regression models with multiple structural changes. *Econometric Theory*, *32*(6), 1376. (Publisher: Cambridge University Press (CUP): HSS Journals)

Rodríguez-Pose, A., & Hardy, D. (2021). Reversal of economic fortunes: Institutions and the changing ascendancy of Barcelona and Madrid as economic hubs. *Growth and Change*, *52*(1), 48–70. (Publisher: Wiley Online Library)

Schwarz, Moritz, & Pretis, F. (2021). *getspanel.* GitHub. Retrieved from `https://github.com/moritzpschwarz/getspanel`

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. (Publisher: Wiley Online Library)

Tobío, C. (1989). Economie and social restructuring in the Metropolitan Area of Madrid (1970–85). *International journal of urban and regional research*, *13*(2), 324–338. (Publisher: Wiley Online Library)

Wooldridge, J. (2021). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Available at SSRN 3906345*.

Zhao, S., Witten, D., & Shojaie, A. (2021). In defense of the indefensible: A very naive approach to high-dimensional inference. *Statistical Science*, *36*(4), 562–577. (Publisher: Institute of Mathematical Statistics)

Zhu, H., Sarafidis, V., & Silvapulle, M. J. (2020). A new structural break test for panels with common factors. *The Econometrics Journal*, *23*(1), 137–155. (Publisher: Oxford University Press)

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

# Paper 5: Attributing agnostically detected large reductions in road $CO_2$ emissions to policy mixes

*Paper 5: Attributing agnostically detected large reductions in road $CO_2$ emissions to policy mixes*

nature energy

ARTICLES

https://doi.org/10.1038/s41560-022-01095-6

Check for updates

# Attributing agnostically detected large reductions in road $CO_2$ emissions to policy mixes

Nicolas Koch[1,2,3] ✉, Lennard Naumann[1,4], Felix Pretis[5,6], Nolan Ritter[1,2] and Moritz Schwarz[6,7]

**Policymakers combine many different policy tools to achieve emission reductions. However, there remains substantial uncertainty around which mixes of policies are effective. This uncertainty stems from the predominant focus of ex post policy evaluation on isolating effects of single, known policies. Here we introduce an approach to identify effective policy interventions in the EU road transport sector by detecting treatment effects as structural breaks in $CO_2$ emissions that can potentially occur in any country at any point in time from any number of a priori unknown policies. This search for 'causes of effects' within a statistical framework allows us to draw systematic inference on the effectiveness of policy mixes. We detect ten successful policy interventions that reduced emissions between 8% and 26%. The most successful policy mixes combine carbon or fuel taxes with green vehicle incentives and highlight that emissions reductions on a magnitude that matches the EU zero emission targets are possible.**

Although ever more countries commit to net-zero greenhouse gas emissions, it remains unclear how to achieve them. Canada, the European Union, Japan, New Zealand and the United Kingdom have passed net-zero emissions targets into law while others such as China and the United States made similar pledges in official policy documents[1]. As a major emitter, the transportation sector is indispensable for achieving net-zero emissions. Globally, it emitted nearly 8.5 Gt $CO_2$ in 2019 or one-quarter of all GHG emissions, while the International Energy Agency suggests that annual global sector emissions need to fall below 1 Gt $CO_2$ by 2050 to reach net-zero overall emissions[1]. So far, the sector has proven resilient to emissions reduction efforts, especially in road transport[2].

The question how to best achieve net-zero emission targets triggers fierce policy debates about the appropriate means as exemplified in the context of the European Green Deal. To become the first climate-neutral continent by 2050, the European Union is currently revising its climate, energy and transport-related legislation under the so-called 'Fit-for-55 package'. One building block is the ambitious increase of EU member states' emissions reduction targets for 2030 from 30% to 40% compared with 2005 under the EU Effort Sharing Regulation (ESR). Under the ESR, each EU member state must meet binding annual emissions reduction targets for the agriculture, buildings and transport sectors by implementing national policies. Yet, according to the latest national projections available, most EU members will miss their targets pursuing current policy instruments. For transport specifically, emissions under the existing policies are projected to be at nearly the same level in 2030 as they were in 2020[3,4]. Thus, the ESR exerts considerable pressure on EU Member States to strengthen climate policies in transport. Policymakers have to choose from myriads of promising policies to achieve emissions reductions. There is a controversial policy discussion about whether the ambitious climate targets are best achieved by using a policy mix that emphasizes tax policies, such as carbon taxes, or green spending, such as electric vehicle subsidies, or command-and-control measures, such as speed limits and efficiency labels[5,6].

There remains substantial empirical uncertainty around which policy mixes are effective in actually achieving the objective they were designed for. Part of this uncertainty remains because empirical policy evaluation in the existing literature predominantly focuses on evaluating single, known interventions in isolation by posing the forward causal question of what happens as a consequence of a particular policy[7–10]. This 'effects of causes' approach[11] runs the risk of missing a priori unknown or underappreciated interventions. It also requires a context that allows for isolating a single policy's effects from simultaneously implemented and potentially confounding ones. Such contexts are rare because policymakers routinely legislate mixes of many interventions simultaneously[2,12]. When having to choose from many interacting available policy interventions, it can, however, be more intuitive to ask a reverse causal question looking for 'causes of effects'[11] to find what caused reductions in emissions (rather than whether a single policy is effective). Such a question is highly relevant to identify either unknown but effective policies or, more importantly, effective mixes of interacting policy interventions. However, in terms of technical implementation, it is less obvious how this kind of question may be tackled.

Here we introduce an approach to implement and answer the reverse causal question of 'What reduced $CO_2$ emissions?' in the EU road transport sector between 1995 and 2018 by first detecting substantial changes in emissions relative to a control group using machine learning and subsequently attributing them to likely causes such as single or interacting policy interventions. Because detection is separate from policy attribution, our approach neither requires any a priori knowledge of reductions in emissions, nor does it require a priori knowledge of the number of policies that caused these. Therefore, we are able to identify previously unknown policies or policy *mixes* that effectively reduce $CO_2$ emissions. While the EU transport sector is a policy-relevant test bed, our approach is readily applicable in many other contexts.

[1]Mercator Research Institute on Global Commons and Climate Change (MCC), Berlin, Germany. [2]Potsdam Institute for Climate Impact Research (PIK), Potsdam, Germany. [3]IZA Institute of Labor Economics, Bonn, Germany. [4]WZB Berlin Social Science Center, Berlin, Germany. [5]Department of Economics, University of Victoria, Victoria, British Columbia, Canada. [6]Nuffield College, University of Oxford, Oxford, United Kingdom. [7]Smith School of Enterprise and the Environment, University of Oxford, Oxford, United Kingdom. ✉e-mail: koch@mcc-berlin.net

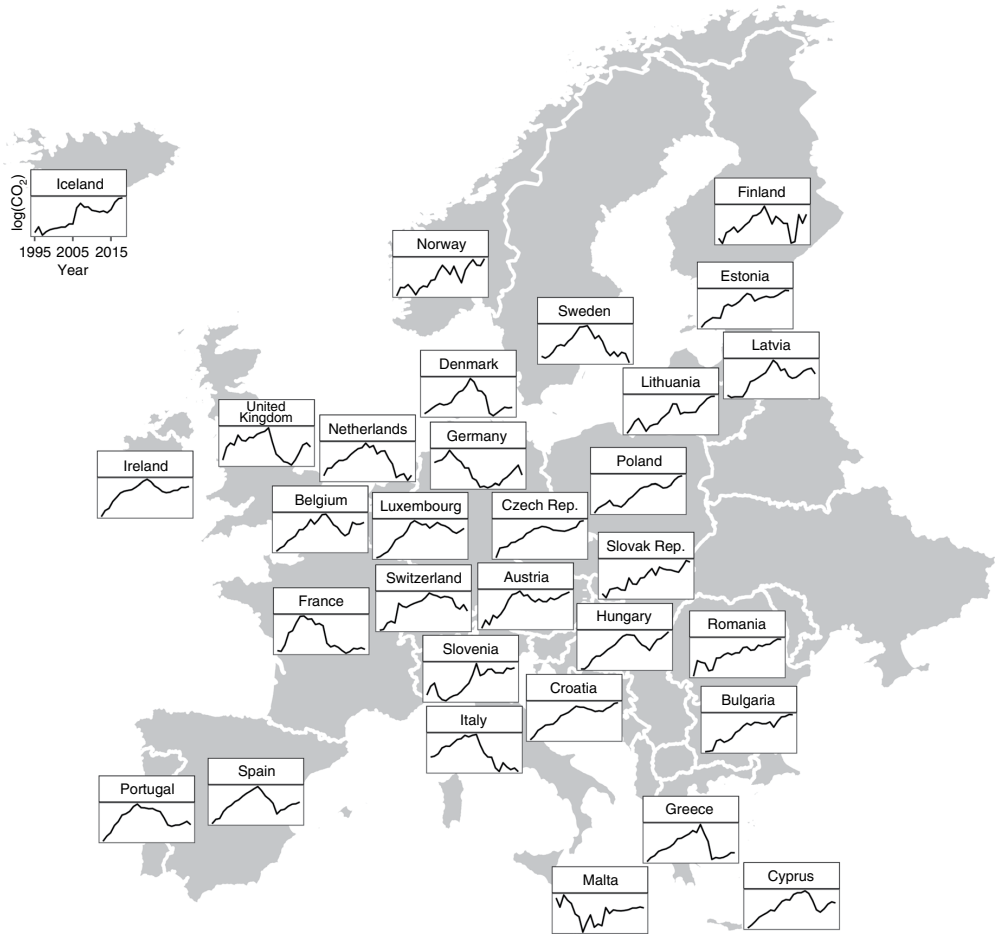


**Fig. 1 | Emissions in road transport in Europe.** The natural logarithm of $CO_2$ emissions (log($CO_2$); relative) between 1995 and 2018 by country. Please refer to Iceland for year indicators on the horizontal axis. The *y* axis indicates log($CO_2$). Rep., Republic. Background map from http://www.efrainmaps.es.

We produce three key findings. First, we detect ten successful policy interventions that reduced emissions between 8% and 26%. We link all these reductions to at least one tax that increases driving costs; we link seven breaks to carbon taxes, four breaks to fuel taxes and three to road tolls. Second, we link eight of the ten breaks to policy mixes that combine the aforementioned taxes with either $CO_2$-based vehicle taxes or subsidies for low-emissions vehicles. Third, we link the breaks with the highest level of confidence and the greatest effect sizes of up to 26% to increases of existing but moderate carbon or fuel taxes. Altogether, the ten policy interventions we identified between 1995 and 2018 reduced emissions in the EU-15 (Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, United Kingdom) by up to 35.9 Mt $CO_2$. In comparison, the current ESR requires a reduction of 480 Mt $CO_2$ for the same region between 2021 and 2030. Even if we conservatively assume that agriculture, buildings and transport contribute in equal measure to these reductions, the transport policies implemented to date seem rather inadequate—even more so when we account for the imminent tightening of the ESR targets proposed in the EU's Fit-For-55 package. At the same time, the relative reductions of up to 26% for certain breaks indicate considerable potential for future reductions. The most successful intervention implemented in Finland in 2000 (−17%), Sweden in 2001 (−11%), Ireland in 2011 (−13%) and Luxembourg in 2015 (−26%) combine increasing carbon or fuel taxes to curb mileage with complementary financial incentives to support the transition to greener vehicles.

## Using break detection to assess policies

Existing ex post policy evaluations predominantly focus on the forward causal question of 'What happens as a consequence of a particular known policy?' It is reasonably straightforward to evaluate these with time-tested, quasi-experimental tools from programme evaluation, ranging from difference-in-differences[13,14] and matching[15] to synthetic control methods[9,10,16]. However, drawing systematic inference is difficult because the available evidence is scattered across countries and policies and because the study of 'effects of causes' runs the risk of missing effective but unknown interventions or those falsely deemed ineffective. Moreover, the forward causal approach needs to ensure that treatments are independent and unconfounded, which is challenging because policymakers routinely implement mixes of many simultaneous interventions with common goals.

It is less obvious how to answer reverse causal questions such as 'What has reduced emissions?'. As Gelman and Imbens[11] put it: 'Reverse causal reasoning is different; it involves asking questions

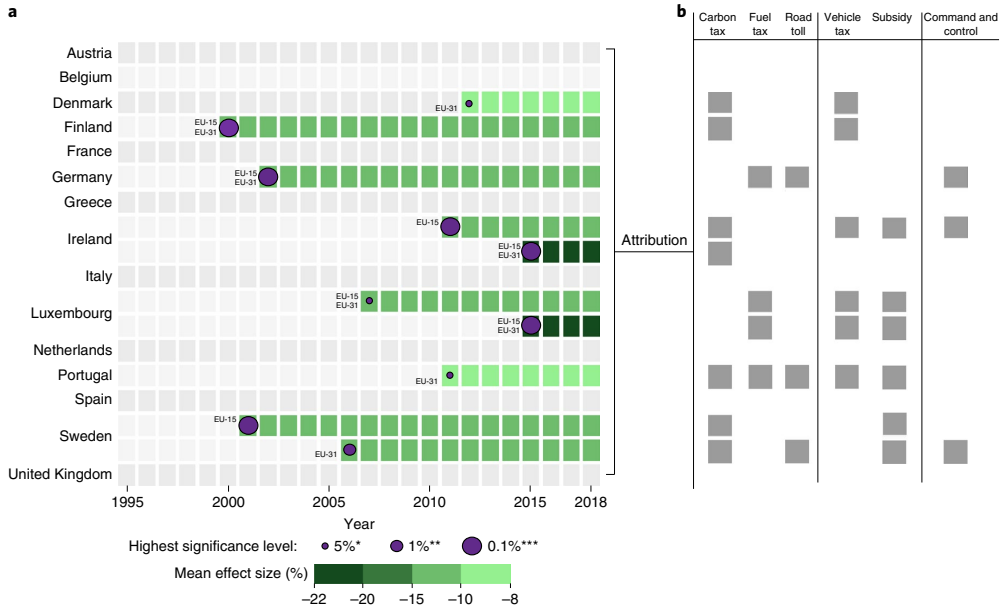


**Fig. 2 | Detected breaks in road $CO_2$ emissions and their attribution. a**, Significant breaks in $CO_2$ emissions. The markers' sizes indicate the highest significance level at which a break was detected. ***, **, and * indicates statistical significance at the 0.1%, 1%, and 5% level, respectively. The colour indicates effect sizes in percent. We find, at most, two detected breaks in any given country (this is an estimation result and not imposed by our Methods, which do not impose an upper limit on the number of detected breaks). Those countries occupy two lines while all others occupy a single line. **b**, The types of policy that caused the reductions.

and searching for new variables that might not yet even be in our model.' We formalize this approach by expanding on the idea of '*searching for new variables*' and placing reverse causal analysis into the domain of variable selection, and more specifically break detection. We tackle the question of 'What reduced emissions?' by identifying notable reductions in $CO_2$ emissions, which we identify as structural breaks. We use familiar two-way fixed effects (TWFE) panel estimators to detect these breaks and estimate a separate treatment effect for each identified break in each country (Methods and refs. [17,18] for the relevant R package 'getspanel'). Our approach identifies country- and time-specific treatment effects on the treated by detecting breaks for each policy and treated country. It reveals when an emissions break occurs within an approximate margin of error. Any policy implemented within this margin potentially caused the break. We place no restriction on the number of potential treatments nor do we impose a minimum break (that is, treatment) length. A causal interpretation rests on the assumption that there were no other influencing factors than the attributed policies themselves.

The idea of scrutinizing data for structural breaks is firmly established in the time series literature on policy evaluation (for example, ref. [19] on the Montreal Protocol, ref. [20] on UK $CO_2$ emissions or ref. [21] on homicides). However, time series methods lack control groups, making causal interpretations difficult. The combination of conservative significance levels (to control the false positive rate of detection) and the use of control groups in the panel setting give the reverse causal approach credibility and reduces the risk of spuriously identifying false positive results. We propose the approach to complement the traditional forward causal analysis. While the latter excels at recovering causal effects of known individual policies, the proposed reverse causal approach simplifies the identification of efficient mixes of policies with large effects that may not have been known a priori.

Specifically, we model the log of $CO_2$ emissions (Fig. 1) as a function of log gross domestic product (GDP) and log population and allow for potential breaks in emissions in any country at any point in time that are captured by 'indicators': interactions of country and year fixed effects. Altogether, with an EU-15 sample and 23 time periods, the maximum number of 345 potential treatments exceeds the number of observations. However, countries are treated sparsely so that most indicators are statistically insignificant. We rely on machine learning to remove all but the significant ones (Methods). Those remaining show treatments that significantly reduced country-specific $CO_2$ emissions relative to a control group conditional on log GDP and log population. It is important to emphasize that these breaks are detected relative to the specified model conditional on the control variables. For example, unconditional visual inspection of Fig. 1 might suggest a break in Greece's $CO_2$ emissions around 2009. However, the visual 'break' in Greek emissions could be explained by the drop in economic activity due to Greece's sovereign debt crisis. Once we condition on GDP (by including it as a control variable), there is no unexpected change in emissions (and thus no break detected), as emissions were falling in line with GDP.

Having identified a series of breaks, we subsequently attribute the significant indicators to policies and disregard those that show increases for this paper. We construct approximate confidence intervals around an indicator's timing to accommodate for uncertainty. These may be as short as a single year or may span several. Then, we search for policies implemented in these confidence intervals (Methods). In the rare event that an interval incorporates only a single policy intervention, attribution is made with respect to a single policy. Otherwise, attribution is made for a policy mix. Attribution using this approach is no different from arguing that a known intervention is exogenous or as-if random when addressing forward causal questions (discussion in Supplementary Note 1).

**Table 1 | Detected breaks, break dates and magnitudes**

| Country | | Model | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| | | EU-15 | EU-15 | EU-15 | EU-31 | EU-31 | EU-31 |
| Significance level in break detection | | 5% | 1% | 0.1% | 5% | 1% | 0.1% |
| Denmark | Effect | | | | −0.080 | | |
| | SE | | | | (0.020) | | |
| | Year | | | | 2012 | | |
| | 95% CI | | | | ±6 | | |
| Finland | Effect | −0.103 | −0.123 | −0.128 | −0.156 | −0.171 | |
| | SE | (0.020) | (0.022) | (0.024) | (0.024) | (0.028) | |
| | Year | 2000 | 2000 | 2000 | 2000 | 2000 | |
| | 95% CI | ±2 | ±2 | ±2 | ±1 | ±2 | |
| Germany | Effect | −0.105 | −0.131 | −0.108 | −0.112 | −0.112 | |
| | SE | (0.018) | (0.020) | (0.022) | (0.021) | (0.025) | |
| | Year | 2002 | 2002 | 2002 | 2003 | 2003 | |
| | 95% CI | ±2 | ±1 | ±3 | ±3 | ±4 | |
| Ireland | Effect | −0.087 | | −0.127 | | | |
| (1st break) | SE | (0.020) | | (0.023) | | | |
| | Year | 2011 | | 2011 | | | |
| | 95% CI | ±3 | | ±2 | | | |
| Ireland | Effect | −0.148 | −0.192 | | −0.247 | −0.244 | −0.229 |
| (2nd break) | SE | (0.028) | (0.028) | | (0.030) | (0.034) | (0.037) |
| | Year | 2015 | 2015 | | 2015 | 2015 | 2015 |
| | 95% CI | ±1 | ±1 | | ±0 | ±1 | ±1 |
| Luxembourg | Effect | −0.136 | | | −0.108 | | |
| (1st break) | SE | (0.024) | | | (0.031) | | |
| | Year | 2007 | | | 2007 | | |
| | 95% CI | ±1 | | | ±3 | | |
| Luxembourg | Effect | | | −0.214 | −0.193 | −0.227 | −0.262 |
| (2nd break) | SE | | | (0.031) | (0.030) | (0.035) | (0.038) |
| | Year | | | 2015 | 2015 | 2015 | 2015 |
| | 95% CI | | | ±1 | ±1 | ±1 | ±1 |
| Portugal | Effect | | | | −0.094 | | |
| | SE | | | | (0.021) | | |
| | Year | | | | 2011 | | |
| | 95% CI | | | | ±4 | | |
| Sweden | Effect | −0.095 | −0.103 | −0.110 | | | |
| (1st break) | SE | (0.017) | (0.019) | (0.022) | | | |
| | Year | 2001 | 2001 | 2001 | | | |
| | 95% CI | ±2 | ±2 | ±3 | | | |
| Sweden | Effect | | | | −0.108 | −0.115 | |
| (2nd break) | SE | | | | (0.019) | (0.022) | |
| | Year | | | | 2006 | 2006 | |
| | 95% CI | | | | ±3 | ±4 | |

This table shows treatment effects on $CO_2$ emissions, standard errors (SE), the year of the break and its half interval. CI, confidence interval. All treatment effects are statistically significant at the 0.1% level.

The combination of European Union-wide, technological standards but largely diverse national policies across EU member states provides an ideal test bed to learn about effective policy mixes. EU $CO_2$ efficiency standards for new vehicles have been in place since 1998 and they became mandatory for each EU member state in 2009. In contrast, tax policies, which are the main instruments to achieve national ESR targets, and command-and-control measures vary considerably across member states and time. Prime examples are frequent changes in fuel taxes, the introduction of carbon taxes and road tolls. Moreover, several changes of $CO_2$-based vehicle taxation

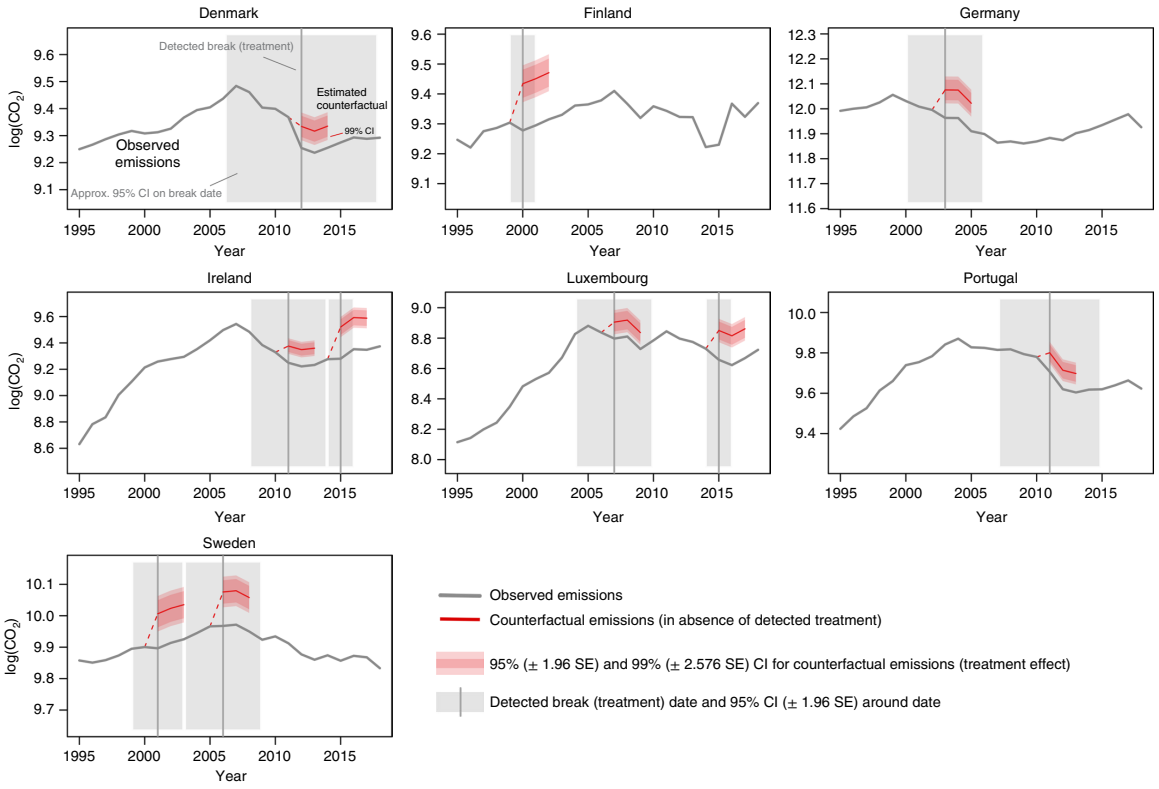


**Fig. 3 | Actual and counterfactual road $CO_2$ emissions.** This Figure contrasts the actual emissions (grey line) with their counterfactuals (red) that would have occurred in the absence of treatment (detected as breaks). Counterfactuals are constructed as log($CO_2$) in absence of detected treatment breaks and plotted for three years following the break date. The 95% confidence intervals around the breaks are ±1.96 standard errors (SE) around their mean estimates; the 99% intervals are ±2.576 SE wide. The vertical lines indicate detected break dates determined using break detection (Methods). The grey shaded areas indicate an approximate 95% or ±1.96 SE confidence interval around the timing of the breaks. CI, confidence interval. Approx., approximate.

intended to nudge consumers to buy more fuel efficient vehicles. In addition, the introduction of tax credits and purchase subsidies for electric vehicles and alternative fuels have been popular policies to increase demand for new technologies. Speed limits, biofuel obligations and efficiency labels exemplify varying command-and-control measures across countries and time.

## Policy mixes that effectively reduced emissions

We estimate reductions in emissions for the EU-15 members using either a sample of (i) EU-15 members or (ii) EU-31 members, which includes Norway, Iceland, Switzerland and the United Kingdom because they were part of the European Single Market and subject to harmonized regulations. The intuition for a second sample is that it increases the credibility of breaks that are consistently detected across both. We investigate the choice of three target levels of significance for detecting treatments. The target levels of 5%, 1% or 0.1% specify the expected false positive rate of the break detection (Methods). The combination of two samples and three target levels leads to a total of six different models.

Figure 2 shows that only ten of the 345 potential treatments across all countries, years and models are significant. The greater the diameter of a circle indicating a break point, the smaller the level of significance at which it is found in any of the models. The darker its colour, the higher the magnitude of its effect. Table 1 presents detailed results. Altogether, the ten reductions in $CO_2$ emissions are from seven different countries. Figure 3 shows these relative to the estimated counterfactual (plotted in red for three years following each break date) given by log($CO_2$) in absence of the estimated breaks (estimated as coefficients on the detected break variables). We show the estimated counterfactuals at the estimated break dates (shown as grey vertical lines); however, there is uncertainty around the break dates (shown as grey shading), thus the counterfactuals may visually appear steeper and more sudden than the actual true underlying (albeit unknown) policy effects.

Altogether, we identify six of the ten reductions at a target significance level (and thus expected false positive rate) of 0.1% (Fig. 2). We regard this rate as the most important criterion in break detection to control the level of confidence we require for a policy to be identified. With a rate of 0.1%, we set a very high bar to dispel any lingering doubts that our findings may be driven by spurious false positives. Second, the more models that detect an intervention, the more confident we are of identifying a break. For instance, we find that five of our six models indicate breaks for Finland in 2000, for Germany in 2002/2003 and for Ireland in 2015 (Table 1). We detect that six out of our ten breaks occur in both the EU-15 and the EU-31 sample. Finally, we consider the stability of our estimated effects across models as a third criterion of robustness (Table 1). For instance, for Luxembourg, the coefficients vary between −0.193 and −0.262.

**Table 2 | Attribution of detected breaks to policies**

| Country | Year | Policy |
|---|---|---|
| Denmark | 2012 | 2008: Carbon tax increase from 13€ $t^{-1}$ $CO_2$e to 23€ $t^{-1}$ $CO_2$e |
| | | 2010: 'Green ownership tax' replaces weight-based taxes for light commercial vehicles |
| | | 2010: Vehicle tax increase for cars without particle filters |
| Finland | 2000 | 1996–1999: Carbon tax increases from 2.3€ $t^{-1}$ $CO_2$e in 1996 to 18€ in 1999 |
| | | 2001: Car ownership tax base changed from total mass to $CO_2$ emissions |
| Germany | 2002/2003 | 1999–2003: 'Ecological Tax Reform' increases motor fuel tax annually by 0.0307€ $l^{-1}$ over five years |
| | | 2001: Harmonization of commuter tax deduction between transport modes |
| | | 2004: Mandatory fuel efficiency labelling for passenger vehicles |
| | | 2005: Road tolls for trucks (originally planned for 2003) |
| Ireland | 2011 | 2008: Vehicle registration tax base and annual motor tax base shifts from engine size to $CO_2$ emissions |
| | | 2009: Tax incentives for purchase of bicycles for commuting of up to 1,000€ |
| | | 2009: Electric vehicle subsidy scheme and vehicle registration tax relief |
| | | 2010: Introduction of a 15€ $t^{-1}$ $CO_2$e carbon tax |
| | | 2010: Biofuel obligations require blending 4% (6%) biofuels in 2010 (2013) |
| Ireland | 2015 | 2014: Carbon tax increase to 20€ |
| Luxembourg | 2007 | 2007: Vehicle tax reform based on $CO_2$ emissions |
| | | 2007: Subsidy for purchase of energy efficient vehicles of 750€ |
| | | 2007–2008: 'Kyoto Cents' law raises fuel tax by 0.02€ $l^{-1}$ for gasoline and 0.025€ $l^{-1}$ for diesel |
| Luxembourg | 2015 | 2013–2014: Subsidies for electric vehicles and vehicles with < 60g $km^{-1}$ $CO_2$ |
| | | 2015: VAT raise from 15% to 17% increases tax burden of fuelling and buying vehicles |
| Portugal | 2011 | 2007: Vehicle ownership tax reform based on $CO_2$ emissions |
| | | 2008: Increase of fuel tax by about 0.025€ $l^{-1}$ |
| | | 2010: Financial incentives to purchase electric vehicles |
| | | 2012: Introduction of nationwide road tolls on motorways and trunk roads |
| | | 2015: Introduction of a 5€ $t^{-1}$ $CO_2$e carbon tax |
| Sweden | 2001 | 2001–2006: 'Green Tax Shift' |
| | | (i) Carbon tax increase from 40€ in 2000 to 57€ in 2001 to 100€ in 2006 |
| | | (ii) Exemptions for biofuels from energy and carbon taxation since 2002 |
| | | (iii) Tax benefits for green company cars since 2002 |
| Sweden | 2006 | 2001–2006: 'Green Tax Shift' |
| | | (i) Carbon tax increase from 57€ $t^{-1}$ $CO_2$e in 2001 to 100€ in 2006 |
| | | (ii) Exemptions for biofuels from energy and carbon taxation since 2002 |
| | | (iii) Tax benefits for green company cars since 2002 |
| | | 2005: Pump Act mandates fuel stations to supply biofuel |
| | | 2006: Introduction of congestion charges in Stockholm |
| | | 2007–2009: Subsidy of up to 1,000€ for eco-friendly vehicles |
| | | 2008–2009: Carbon tax increase from 100€ $t^{-1}$ $CO_2$e in 2006 to 110€ in 2008 to 114€ in 2009 |

Taking the coefficients with the highest magnitude of relative emissions reductions (in %) and the emissions level (in Mt $CO_2$) in a given country at the time of the identified break indicates that the breaks we identified between 1995 and 2018 accounted for total emissions reductions of up to 35.9 Mt $CO_2$. In comparison, the current ESR requires a 30% reduction until 2030 compared with 2005 levels in the sectors that are not subject to EU emissions trading, that is, agriculture, buildings and transport. This target translates into absolute emissions reductions of about 480 Mt $CO_2$ in the EU-15 member states between 2021 and 2030. If we conservatively assume that each sector contributes in equal measure, the magnitude achieved by past transport policies seem inadequate—even more so when we account for the imminent tightening of the ESR targets to a reduction of 40% relative to 2005 emissions under the proposed revision of the ESR under the 'Fit-For-55' policy package. At the same time, the magnitude of three of the ten detected breaks exceeds 17%, which indicates considerable potential for future reductions.

Post-estimation, we can now attribute effects to their likely causes (Table 2) by matching policies with the break points' confidence intervals. Attribution reveals that many interventions are applied simultaneously often by one legislative package.

For example, we attribute the break point in Luxembourg's emissions around 2007 (99% CI: ± 1 year) to three potential policies: a $CO_2$-based vehicle tax reform in 2006, a subsidy scheme for fuel efficient cars in 2007 and a 0.02€ per liter fuel tax increase in 2007 ('Kyoto Cents'). The two breakpoints in Sweden occur in 2001 (99% CI: ± 2–3years) and 2006 (99% CI: ± 3–4years). These correspond

to carbon tax increases implemented in 2001 and tightened annually through 2006. In particular, in 2001, Sweden raised $CO_2$ taxes from 40€ to 57€ per ton and also introduced subsidies for biofuels and green company cars. Annually increasing $CO_2$ taxes reached 100€ per ton in 2006. The break in 2006 (±3–4 years) also coincides with the introduction of biofuel mandates, the implementation of road tolls in Stockholm in 2006, the introduction of subsidies for low-emission vehicles in 2007 and further carbon tax increases to 110€ in 2008 and 114€ in 2009, underlining the importance of considering policy mixes rather than individual policies, specifically as multiple detected breaks in a single country could also capture time-varying treatment effects through tightening emissions targets.

We continue by classifying the identified policies. We differentiate between taxes on carbon, fuel or vehicles, road tolls, subsidies and command-and-control measures in Fig. 2. This helps to assess whether certain policies or mixes are superior. We produce four key findings.

First, we link all detected emissions reductions to at least one tax policy that increases the cost of driving; we link seven cases to carbon, four cases to fuel taxes and three cases to road tolls. Second, we link eight of the ten emissions breaks to policies that combine taxes that increase the cost of driving with reforms that emphasize $CO_2$-based vehicle taxes (six cases) or subsidy schemes for low-emissions vehicles (six cases). For instance, we attribute the ten to 15% emissions reduction in Finland in 2000 to a combination of a carbon tax increase and switching to $CO_2$-based vehicle taxes. Vehicle taxes and subsidies provide incentives to switch to more fuel efficient or zero emissions vehicles, in particular, if consumers either systematically underestimate or discount future savings from increased efficiency. Third, we link the breaks with the highest level of confidence and the greatest magnitude of effect (Finland 2000, Germany 2002/2003, Luxembourg 2015, Ireland 2015) to increases in existing but moderate carbon or fuel taxes.

Fourth, our finding that command-and-control measures relate only to three emissions breaks (mandatory efficiency labels in Germany and biofuel obligation schemes in Ireland and Sweden) potentially indicates that they either play a minor role in reducing $CO_2$ emissions at the national level or that governments did not use them extensively. However, we caution against over-interpreting this finding because key command-and-control measures, such as efficiency standards for new vehicles, are implemented at the EU level. We can detect measures only at the national level, which might be of limited impact. Moreover, our search for potential policy measures relies on databases that hardly include any public transport policies.

Finally, we note our approach's limitations. One concern is that agnostic break detection runs the risk of not detecting real but less effective treatments. To address this concern, we use higher target levels of significance (that is, expected false positive rates) that allow identification of smaller and therefore more potential treatments (Fig. 2). As a further robustness check, we also searched our policy databases for carbon, fuel and road tax interventions that we do not detect (Table 3). Figure 4 compares all actually implemented carbon tax changes to the ones we detected. Overall, we detect all but two. The lack of evidence for any emission break in France despite its 2014 carbon tax introduction may be best explained by the fact that the initial tax of 7€ was offset by an equivalent reduction in the existing energy consumption tax[22]. Similarly, the lack of finding any effects for the 2011 carbon tax increase in Finland may be because of simultaneous reductions in the tax on engine power for cars and trucks that might have weakened its effect[23]. We do not find any major undetected toll increases except for one in Austria in 2004 and the introduction of a vignette system in the United Kingdom in 2014. However, Table 3 shows that we do find a number of undetected (sometimes transitory) changes in fuel taxes that exhibit a wide range of magnitudes. Potentially relevant but undetected fuel tax increases occur in Austria, Belgium, Italy, the Netherlands, Spain and the United Kingdom. The emissions effect of these tax changes is likely too small to be identified by our approach. We draw two conclusions from these robustness checks. First, we are more likely to detect large breaks and cannot detect effective policies that yield small reductions. Thus, our estimates for the effect sizes of our detected policy mixes provide a lower-bound estimate if countries in the control group also experienced smaller emissions reductions that we do not detect. Given the magnitude and urgency of the climate crisis and the ambitious EU climate targets, we believe a focus on interventions with large-scale effects is justified. Second, we caution against generalizing our results and using them as benchmark estimates for particular policy instruments or policy mixes. The provision of such benchmark estimates is an important policy question for future research that can be tackled by combining our proposed reverse causal approach with the standard forward causal approach (Supplementary Note 1 for a more detailed discussion).

Given that we detect breaks relative to a specified model conditional on selected control variables, a related concern is that our model may be mis-specified and might lead to the detection of spurious breaks. However, Supplementary Tables 8–11 in Supplementary Note 3 show that our findings are generally robust to various alternative baseline model specifications (including new controls such as the share of urban population, nonlinear functional forms and linear country-specific time trends), especially based on our preferred EU-15 control group. There are two notable exceptions: (i) the break in Portugal, which already had weak support in our main Table 1 and, as expected with country-specific time trends, (ii) including time trends absorbs the two breaks in Sweden that indicate time-varying treatment effects from the Swedish climate policy package. In addition, a specification test suggested by Oster[24] shows that our results are robust with respect to omitted variable bias (Supplementary Table 12 in Supplementary Note 3).

**Table 3 | Undetected carbon, fuel or road pricing policies**

| Country | Year | Undetected Policies |
|---|---|---|
| Austria | 2004 | Introduction of electronic network-wide road toll system for trucks (which increased costs compared with the previous vignette system) |
| Austria | 2008 | Increase of fuel tax by about 0.03€ $l^{-1}$ |
| Austria | 2012 | Increase of fuel tax by about 0.04€ $l^{-1}$ |
| Belgium | 2006 | Increase of fuel tax by about 0.08€ $l^{-1}$ |
| Belgium | 2010 | Increase of fuel tax by about 0.02€ $l^{-1}$ |
| Finland | 2010–2012 | Increase of carbon tax from about 20€ to about 60€ $t^{-1}$ $CO_2$e |
| France | 2014 | Introduction of a 7€ $t^{-1}$ $CO_2$e carbon tax |
| Greece | 2008–2012 | Gradual increase of fuel tax from about 0.33€ $l^{-1}$ in 2008 to 0.67€ $l^{-1}$ in 2012 |
| Italy | 2006 | Increase of fuel tax by about 0.02€ $l^{-1}$ |
| Italy | 2012 | Increase of fuel tax by about 0.14€ $l^{-1}$ |
| Netherlands | 2005 | Increase of fuel tax by about 0.04€ $l^{-1}$ (and subsequently annual increases by about 0.01–0.02€ $l^{-1}$) |
| Spain | 2010 | Increase of fuel tax by about 0.065€ $l^{-1}$ |
| United Kingdom | 2012 | Increase of fuel tax by about 0.06€ $l^{-1}$ (back to tax level before 2010) |
| United Kingdom | 2014 | Introduction of road toll vignette system for trucks |

Data from ACEA Tax Guide[43], CESifo DICE Report[44], World Bank's Carbon Pricing Dashboard[45] and country-specific sources specified therein.

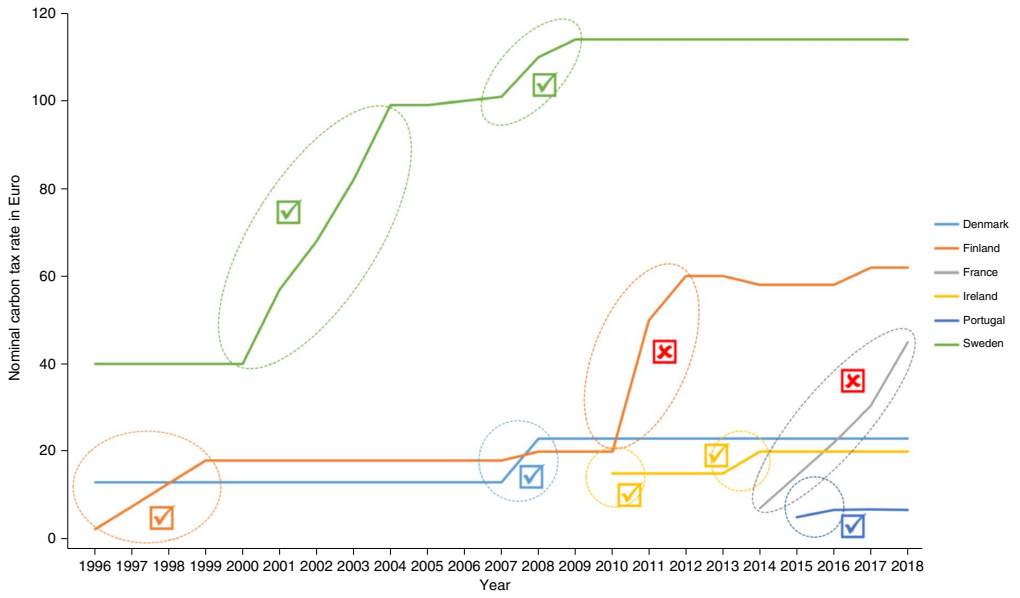


**Fig. 4 | Overview of implemented and detected carbon tax changes.** Nominal carbon tax rates in € $t^{-1}$ $CO_2$ between 1995 and 2018 for all EU-15 countries with a carbon pricing scheme. We encircled relevant changes in carbon tax rates. A tick indicates that we detect the tax changes, while a cross indicates that we do not. Data from World Bank's Carbon Pricing Dashboard[45] and country-specific sources specified therein.

Another concern is that the estimated effects may not be generalizable because (i) we estimate country-specific effects for each intervention and (ii) we are more likely to detect large breaks. To partially address the latter concern, we report bias-adjusted coefficients that do not change the interpretation of our results in Supplementary Table 6 in Supplementary Note 3. The former concern may be addressed by averaging over all identified and similar treatments to approximate the average treatment effect (in the spirit of ref. [25]). But we do not seek to provide benchmark estimates for particular policy instruments here.

A final concern is that national policies may also affect neighbouring countries. In particular, a fuel or carbon tax increase may cause fuel tourism in private consumers in the border region or cause firms to reroute their trucks for refuelling. To evaluate the potential bias from such spillovers, we exclude the two major transit countries with a low fuel tax regime, Austria and Luxembourg, from our sample to estimate our final model. Supplementary Table 7 in Supplementary Note 3 shows that we obtain very similar results in this restricted sample. Moreover, these concerns also apply to forward causal analyses.

## Conclusion

In this study, we propose a complementary approach to ex post policy evaluation. Instead of estimating the effect of a single, known cause on emissions, we seek to identify multiple, unknown causes of an emissions effect. As policymakers implement ever more climate policy packages to meet their obligations under the Paris Agreement or their own net-zero emissions targets, we believe our approach is policy relevant because it enables drawing systematic inference on the effectiveness of such policy *mixes*. We demonstrate this for the EU transport sector, which is a key bottleneck that impedes the European Union's progress to achieve climate neutrality by 2050.

Our results show that relatively few policy interventions effectively curbed $CO_2$ emissions in road transport. We identify ten successful interventions with emissions reductions between 8% and 26% or 35.9 Mt $CO_2$ between 1995 and 2018 and attribute all detected emissions reductions to policy mixes that comprise at least one tax policy intervention that increases the cost of driving. The fact that we detect nearly all carbon price interventions indicates that carbon pricing may be a critical element of effective policy packages. In addition, we attribute the vast majority of emissions reductions to policy mixes that combine carbon, fuel or road-use taxes with additional vehicle taxes or subsidies. The most successful examples of such combinations are policies implemented in Finland in 2000 (−17%) Sweden in 2001 (−11%), Ireland in 2011 (−13%) and Luxembourg in 2015 (−26%). Carbon, fuel or road-use taxes provide incentives to reduce mileage, yet they may not ensure that consumers invest in energy efficient vehicles if consumers are myopic. This effect is known as the energy efficiency gap[26]. Vehicle taxes and subsidies can address myopic consumers and provide incentives to adopt more fuel efficient vehicles. However, they suffer from the rebound effect that describes the unintended side effect that more efficient vehicles cost less to drive and, therefore, encourage additional mileage[27]. Our findings thus provide suggestive evidence that the combinations of policies that simultaneously address the energy efficiency gap and rebound effects are particularly effective. To check the robustness of this evidence based on country-specific effect estimates, future research may combine our proposed reverse causal approach with the standard forward causal approach to provide more systematic benchmark estimates for the effectiveness of particular policy mixes. Our findings are broadly in line with studies that use more structural modelling to evaluate policy mixes[28,29] and the literature suggesting that tax policies can address rebound effects[30,31]. Finally, we also show that the greatest emissions reductions occur when policymakers increase existing but moderate carbon or fuel taxes. This suggests that commitment to staggered, anticipated and permanent tax increases over time may be a strong determinant of emissions reductions.

Altogether, the ambitious country-specific emissions reduction targets under the EU effort sharing regulation require timely action.

We identified policy mixes with emissions reductions on a magnitude that matches the reduction requirements under the net-zero emissions target for seven EU countries. If policymakers in these countries focused on the policy mixes that have been effective in the past, we should expect stronger reductions in road transport emissions. Although policies are context-specific, we yet believe that policymakers in other EU countries may also learn from these successful interventions.

## Methods

**Data.** The data for road transport $CO_2$ emissions is from section 1A3b of the Emissions Database for Global Atmospheric Research (EDGAR) v5.0 (ref. [32]). We retrieved GDP and data on population sizes from World Bank[33,34]. The dependent variable is the natural logarithm of $CO_2$ emissions log($CO_2$). log(GDP), log(GDP)$^2$ and log(population) enter the model as control variables. Supplementary Table 4 in Supplementary Note 3 shows that our findings are robust, if we restrict $CO_2$ emissions to passenger vehicles only (that is, section 1A3bi).

We analysed emissions break in the EU-15 member states (Austria, Belgium, Germany, Denmark, Spain, Finland, France, United Kingdom, Ireland, Italy, Luxembourg, Netherlands, Greece, Portugal and Sweden) because they are subject to largely identical EU regulations (with some minor differences in implementation), in general, and specifically with respect to the European Single Market over our sample period from 1996 to 2018. We disregard years before 1996 due to major historic dissimilarities between these countries. For the control group, we also consider a broader sample of all 27 EU member states and the three European Free Trade Association states and the United Kingdom (EU-15, Croatia, Bulgaria, Cyprus, Czech Republic, Estonia, Hungary, Lithuania, Latvia, Malta, Poland, Romania, Slovakia, Slovenia, Switzerland, Iceland and Norway). Supplementary Tables 1–3 in Supplementary Note 2 provide summary statistics.

**Empirical approach.** We identify effective policy interventions by detecting structural breaks in TWFE panel models of $CO_2$ emissions. Detected breaks identify heterogeneous treatment effects without prior knowledge on treatment assignment or timing. A standard approach to analyse policy effectiveness, often interpreted as difference-in-differences when treatment effects are homogeneous—for example, ref. [13]—is to model emissions using a TWFE estimator as a function of control variables and a binary variable that denotes the interaction of 'treated' countries that are subject to particular policies and the post-treatment period. Such 'known' binary policy variables in a TWFE panel are equivalent to step shifts in the individual fixed effects of the treated countries (more detailed discussion in Supplementary Note 1).

Using the equivalence between step shifts in the unit-specific intercept (that is, fixed effect) and known treatments, we use an alternative approach to evaluate reverse causal questions regarding policy interventions. Rather than exclusively evaluating known interventions while disregarding unknown but effective policies, we estimate a TWFE panel in search of potential structural breaks (step shifts) in the unit-specific intercepts. Once a break has been identified, it can be interpreted as a treatment for the relevant country. We then attempt to attribute the break to a policy that affected the treated country around the detected time. Thus, rather than assessing effects of causes, our approach provides a data-driven method to first identify breaks which can, in a second step, be attributed to policy interventions. Pretis and Schwarz[17] provide a detailed discussion of this modelling approach that was first introduced by Pretis[35].

We formulate the detection of structural breaks as a problem of variable selection similar to ref. [36] but extended the approach to the panel setting, where we saturate a TWFE panel model with a full set of step shifts denoting potential treatment of every country at every point in time. We then apply variable selection methods from machine learning that allow for more candidate variables than observations to identify breaks without prior knowledge of their existence. We saturate a TWFE regression with a full set of break variables (step shifts) denoting potential treatment of each unit at every time period, nesting any specific treatment as a special case. In a balanced panel of $N$ countries and $T$ time periods this adds $N(T-1)$ potential break variables to be selected over. Therefore, we start with a full set of step functions with coefficients $\tau_{j,s}$:

$$\log(\mathrm{CO}_2)_{i,t} = \alpha_i + \phi_t + \sum_{j=1}^{N}\sum_{s=2}^{T}\tau_{j,s}1_{\{i=j,t\geq s\}} + x'\beta + \epsilon_{i,t} \quad (1)$$

where $\alpha_i$ and $\phi_t$ denote individual and time fixed effects, $x_{i,t}$ is a vector of control variables that includes log(GDP), log(GDP)$^2$ and log(population size). The population treatment coefficients $\tau_{j,s}$ are sparse with coefficients of zero for all but the treated countries. This operationalizes the notion of ref. [11] that reverse causal questions require variables '*that might not yet even be in our model.*' The target of model selection is then to remove all but the relevant break variables so that in a final sparse model, the selected breaks correspond to the true underlying, and potentially unknown treatments. Let $\widehat{\mathrm{Tr}}$ denote the set of detected treated countries, with associated detected treatment times $\hat{T}_j$ for each treated country $j \in \widehat{\mathrm{Tr}}$. Then the resulting sparse model is:

$$\widehat{\log(\mathrm{CO}_2)}_{i,t} = \hat{\alpha}_i + \hat{\phi}_t + \sum_{j\in\widehat{\mathrm{Tr}}}\sum_{s\in\hat{T}_j}\hat{\tau}_{j,s}1_{\{i=j,t\geq s\}} + x'\hat{\beta} \quad (2)$$

Coefficients $\hat{\tau}_{j,s}$ correspond to estimates of heterogeneous treatment effects for the detected treated countries. For example, we may detect a break for Sweden in 2006 (Swe$\in$Tr, $T_{\mathrm{Swe}}=2006$), where the associated estimated coefficient $\hat{\tau}_{\mathrm{Swe},2006}$ on the break variable captures the country-specific treatment effect. The estimated treatment effects in the final retained model (2) can be interpreted as heterogeneous treatment effects estimated using interactions of the unit-fixed effects with treatment time for each treated unit as in ref. [25], thus also addressing recent concerns about imposing homogeneous treatment effects in panels with staggered adoption (for example, ref. [37]). The main difference relative to the specification in ref. [25] is that in our application, each treated cohort consists of a single country.

We resort to machine learning to move from the general model (1) that embeds all possible treatment dates for all countries to the sparse model (2). A large set of potential selection algorithms are available. To carefully control the false positive rate of detected breaks, we apply the block search algorithm 'gets'[38] using the 'getspanel' update in ref. [18], which forms part of the general-to-specific family of model selection. Alternatives include shrinkage-based methods such as the LASSO and variants thereof, though these do not target the false positive rate (refs. [39–41]). Supplementary Note 1 provides a more detailed discussion.

The main calibration parameter of 'gets' is the target level of significance $\gamma_\epsilon$ which controls the expected false positive rate of retained breaks and is defined as the number of non-zero treatment coefficients relative to all possible treatment coefficients. Their asymptotic properties are explored in ref. [42], who show that in the absence of breaks and accounting for multiple testing, the false positive rate converges to the chosen nominal level of significance of selection $\gamma_\epsilon$. If there are no true treatment breaks present, then the proportion of spuriously detected breaks converges to the chosen level of significance. For instance, with $\gamma_\epsilon=0.01$, the expected false positive rate is 1% and we expect $0.01\times N(T-1)$ spuriously retained breaks. We consider $\gamma_\epsilon$ equal to 0.05, 0.01 and 0.001 in our models of $CO_2$ emissions to assess the robustness of our results. Supplementary Table 5 in Supplementary Note 3 shows that our findings are robust to using cluster-robust standard errors.

**Attribution.** Our attribution strategy to match policy interventions to the year intervals for which we detect break points involved two primary databases and various supplementary data sources.

First, we searched for interventions in two main databases: (i) the IEA's Policies and Measures Database that provides information on past, existing or planned climate and energy policies. Data is collected from governments, international organizations and IEA analyses, and governments can review the provided information periodically. (ii) The National Communications to the United Nations Framework Convention on Climate Change secretariat that our sample countries are required to submit regularly.

Second, to corroborate the information gained from the IEA and United Nations Framework Convention on Climate Change documents and to double check for any policies these two sources omit, we collected additional information from the European Automobile Manufacturers' Association's Annual Tax Guide that provides detailed information on fuel, vehicle and road tax schedules and subsidy programmes, the World Bank's Carbon Pricing Dashboard that provides detailed information on carbon prices and the Climate Change Laws of the World database of the Grantham Research Institute. In a few cases, we also conducted specific searches on Google.

## Data availability

All publicly available data analysed in this study are available from the corresponding author upon request and are also available from online repository Zenodo (https://doi.org/10.5281/zenodo.6768563).

## Code availability

The code required to replicate our study is available from the corresponding author upon request and is also available from online repository Zenodo (https://doi.org/10.5281/zenodo.6768563).

Received: 28 February 2022; Accepted: 12 July 2022;
Published online: 22 August 2022

## References

1. *Net Zero by 2050—A Roadmap for the Global Energy Sector* (IEA, 2021).
2. Axsen, J., Plötz, P. & Wolinetz, M. Crafting strong, integrated policy mixes for deep $CO_2$ mitigation in road transport. *Nat. Clim.Change* **10**, 809–818 (2020).

3. *Planning for Net Zero: Assessing the Draft National Energy and Climate Plans* (Ecologic & Climact, 2019).
4. *Trends and Projections in Europe 2021* (European Environment Agency, 2021).
5. Graf, A., Graichen, J., Matthes, F. C., Gores, S. & Fallasch, F. *How to Raise Europe's Climate Ambitions for 2030* (Agora Energiewende and Öko-Institut e.V., 2019).
6. Transport and Energy *Why Increasing Ambition under the ESR is Unavoidable* (2021).
7. Grant, D., Bergstrand, K. & Running, K. Effectiveness of US state policies in reducing $CO_2$ emissions from power plants. *Nat. Clim. Change* **4**, 977–982 (2014).
8. Martin, G. & Saikawa, E. Effectiveness of state climate and energy policies in reducing power-sector $CO_2$ emissions. *Nat. Clim. Change* **7**, 912–919 (2017).
9. Andersson, J. J. Carbon taxes and $CO_2$ emissions: Sweden as a case study. *Am. Econ. J.* **11**, 1–30 (2019).
10. Bayer, P. & Aklin, M. The European Union emissions trading system reduced $CO_2$ emissions despite low prices. *Proc. Natl Acad. Sci. USA* **117**, 8804–8812 (2020).
11. Gelman, A. & Imbens, G. *Why Ask Why? Forward Causal Inference and Reverse Causal Questions* (National Bureau of Economic Research, 2013).
12. Eskander, S. M. & Fankhauser, S. Reduction in greenhouse gas emissions from national climate legislation. *Nat. Clim. Change* **10**, 750–756 (2020).
13. Lin, B. & Li, X. The effect of carbon tax on per capita $CO_2$ emissions. *Energy Policy* **39**, 5137–5146 (2011).
14. Klemetsen, M., Rosendahl, K. E. & Jakobsen, A. L. The impacts of the EU ETS on Norwegian plant's environmental and economic performance. *Clim. Change Econ* **11**, 2050006 (2020).
15. Colmer, J., Martin, R., Muûls, M. & Wagner, U. J. *Does Pricing Carbon Mitigate Climate Change? Firm-Level Evidence from the European Union Emissions Trading Scheme*, discussion paper 1728 (Center for Economic Performance, 2020).
16. Rafaty, R., Dolphin, G. & Pretis, F. *Carbon Pricing and the Elasticity of $CO_2$ Emissions*, working Paper No. 140 (Institute for New Economic Thinking, 2020).
17. Pretis, F. & Schwarz, M. Discovering what mattered: Answering reverse causal questions by detecting unknown treatment assignment and timing as breaks in panel models. *Preprint at SSRN* https://doi.org/10.2139/ssrn.4022745 (2022).
18. Schwarz, M. & Pretis, F. getspanel. GitHub repository (2021), https://github.com/moritzpschwarz/getspanel/
19. Estrada, F., Perron, P. & Martínez-López, B. Statistically derived contributions of diverse human influences to twentieth-century temperature changes. *Nat. Geosci.* **6**, 1050–1055 (2013).
20. Hendry, D. F. et al. *First In, First Out: Econometric Modelling of UK Annual $CO_2$ Emissions, 1860–2017* (Economics Group, Nuffield College, Univ. of Oxford, 2020).
21. Piehl, A. M., Cooper, S. J., Braga, A. A. & Kennedy, D. M. Testing for structural breaks in the evaluation of programs. *Rev. Econ. Stat.* **85**, 550–558 (2003).
22. Schubert, K. Carbon taxation: The French experience, 2014–2019. *Coalition of Finance Ministers for Climate Action Workshop on Carbon Taxation*, 2019.
23. *Report of the Working Group on Energy Taxation Reform: A Proposal for Implementing the Intentions and Goals of the Government Programme and for Further Development of Energy Taxation* (Finnish Ministry of Finance, 2021).
24. Oster, E. Unobservable selection and coefficient stability: theory and evidence. *J. Bus. Econ. Stat.* **37**, 187–204 (2019).
25. Wooldridge, J. Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators. Preprint at *SSRN* https://doi.org/10.2139/ssrn.3906345 (2021).
26. Gillingham, K. T., Houde, S. & van Benthem, A. A. Consumer myopia in vehicle purchases: evidence from a natural experiment. *Am. Econ. J.* **13**, 207–38 (2021).
27. Gillingham, K., Kotchen, M. J., Rapson, D. S. & Wagner, G. The rebound effect is overplayed. *Nature* **493**, 475–476 (2013).
28. Ravigné, E., Ghersi, F. & Nadaud, F. Is a fair energy transition possible? Evidence from the French low-carbon strategy. *Ecol. Econ.* **196**, 107397 (2022).
29. Landis, F., Rausch, S., Kosch, M. & Böhringer, C. Efficient and equitable policy design: taxing energy use or promoting energy savings?. *Energy J.* **40**, 73–104 (2019).
30. Vivanco, D. F., Kemp, R. & van der Voet, E. How to deal with the rebound effect? A policy-oriented approach. *Energy Policy* **94**, 114–125 (2016).
31. Freire-González, J. & Ho, M. S. Policy strategies to tackle rebound effects: a comparative analysis. *Ecol. Econ.* **193**, 107332 (2022).
32. Crippa, M. et al. *Population, Total Fossil $CO_2$ and GHG Emissions of All World Countries—2019 Report* (Publications Office of the European Union, 2019).
33. GDP (constant 2010 US$). *World Bank Open Data* (2020), https://data.worldbank.org/indicator/NY.GDP.MKTP.KD?view=chart
34. Population, total. *World Bank Open Data*, https://data.worldbank.org/indicator/SP.POP.TOTL?view=chart (2020).
35. Pretis, F. Does a carbon tax reduce $CO_2$ emissions? Evidence from British Columbia. *Environmental and Resource Economics* (2022), https://doi.org/10.1007/s10640-022-00679-w
36. Castle, J. L., Doornik, J. A., Hendry, D. F. & Pretis, F. Detecting location shifts during model selection by step-indicator saturation. *Econometrics* **3**, 240–264 (2015).
37. Goodman-Bacon, A. Difference-in-differences with variation in treatment timing. *J. Econometrics* **225**, 254–277 (2021).
38. Pretis, F., Reade, J. & Sucarrat, G. Automated general-to-specific (gets) regression modeling and indicator saturation methods for the detection of outliers and structural breaks. *J. Stat. Softw.* **86**(3), 1–44 (2018).
39. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
40. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**, 1418–1429 (2006).
41. Okui, R. & Wang, W. Heterogeneous structural breaks in panel data models. *J. Econometrics* **220**, 447–473 (2021).
42. Nielsen, B. & Qian, M. Asymptotic properties of the gauge of step-indicator saturation. Working Paper, University of Oxford (2018).
43. *ACEA Tax Guide* (ACEA) (2022).
44. Rumscheidt, S. Road user charging in the European Union. *CESifo DICE Rep.* **12**, 54–57 (2014).
45. *Carbon Pricing Dashboard* (World Bank, 2022); https://carbonpricingdashboard.worldbank.org/

## Acknowledgements
We thank O. Edenhofer, A.B. Martinez, R. Tol and the participants at the EC2 Conference 2021, the Federal Reserve Virtual Seminar on Climate Economics and the Climate Econometrics Seminar for valuable feedback and suggestions. F.P. and M.S. gratefully acknowledge funding from the Clarendon Fund and the Robertson Foundation. F.P. is also grateful to funding from Social Sciences and Humanities Research Council of Canada (SSHRC). The views expressed here are those of the authors and not necessarily those of the Ministry of Finance or the Austrian government.

## Author contributions
F.P., L.N., M.S. and N.K. designed the analysis. F.P. and M.S. wrote the core programme code. L.N. collected the data. L.N. and N.K. conducted most of the analyses. All authors interpreted results and designed figures. N.R. and N.K. wrote the manuscript with contributions from all authors.

## Competing interests
The authors declare no competing interests.

## Additional information
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41560-022-01095-6.

**Correspondence and requests for materials** should be addressed to Nicolas Koch.

**Peer review information** *Nature Energy* thanks Patrick Bayer, Edgar Hertwich and Md. Saniul Alam for their contribution to the peer review of this work.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

# Part III. Conclusion

To combat climate change and to deal with the climate crisis in a sustainable, equitable, and efficient manner, we must question the current focus of research critically. While much of the efforts in the macro-economic climate impact community have for decades focused on trying to identify the 'efficient' or 'optimal' level of warming given economic preferences, comparatively little attention of economists has been diverted towards identifying the most effective climate policy instruments, where climate and weather impacts have actually occurred, what their effect has been, and how such impacts could challenge societies in the next few decades.

The relevance of getting the estimation of climate impacts right is immense. Currently, a lack of appropriate data, an inadequate research focus, and misspecified estimation methods have led to a situation where crucial questions in the climate change literature remain unsatisfyingly unresolved. This manifests itself in the fact that we cannot rely on comprehensive assessments of extreme weather event occurrence in developing countries, that we face large uncertainty when considering climate policy instrument choice, and that we have limited understanding of the magnitude of the role of adaptation in reducing future climate damages. Challenges like multi-causal relationships, long forecasting periods, high internal variability systems, and high levels of heterogeneity in impacts and responses combine to make any kind of impact estimation in the context of climate change inherently difficult.

Nevertheless, the literature on both estimating physical climate impacts as well as on policy effects has improved significantly over the past few decades. The use of econometric methods has allowed the climate impact literature to progress beyond theoretical process-based estimates of the relationship between economic output and climate, while advances in machine-learning and statistical methods

have provided us with an opportunity to make use of the increasingly data-rich research context that we operate in.

In this thesis, I made a substantial contribution to this impact estimation literature and have advanced this field in a number of key ways. I have explored how the use of advanced and novel econometric methods can alleviate some of the challenges and gaps that the literature still faces.

In the work presented in the five papers of this thesis, I have provided methodological advances for already existing approaches in Papers 1 and 2 by improving macro-econometric climate impact estimates. Furthermore, I have also highlighted two emerging areas of research where future progress could be immense, by presenting a novel method to detect extreme weather events especially in developing countries in Paper 3 and by presenting new methods to allow climate policy evaluation to ask the questions that could actually help us understand which methods are most effective at reducing carbon emissions in Paper 5, which uses a method developed in Paper 4. This thesis has therefore demonstrated that existing estimation approaches can be improved by using novel misspecification tests, the use of more specialised data, as well as more advanced machine learning type model selection algorithms. This thesis has also demonstrated that asking new questions in the field of policy evaluation and weather impact detection can offer promising avenues for further research.

Overall, the individual pieces of work in this thesis collectively illustrate that there is a significant potential to improve various aspects of climate impact estimation. Advanced and novel econometric techniques can deal with omnipresent issues that have plagued the climate debate for decades; these techniques can quantify and alleviate misspecification concerns, can allow for a more specific and insightful debate surrounding adaptation options and likely future damage while providing a more holistic and objective data environment for climate policy and weather impact events.

Undoubtedly, numerous questions and challenges for this literature remain, as discussed extensively in Part I. And while a number of these questions have been addressed with the methods presented here, it also seems clear that econometric techniques are a key tool in our collective toolbox to resolve them in the long-term.

To achieve the crucial climate goals set by the Paris Agreement, such as deep decarbonisation, a greening of financial flows, and a resilient, well-adapted society, I am convinced that empirical methods will need to play a more prominent role. Overall, I argue that when econometric methods are specified correctly, are applied to the most pressing research questions and make use of the appropriate data, then using these methods can allows us to direct adaptation funding more efficiently, track Loss and Damage events around the world, and allow policy-makers to focus on those policy packages that have the largest chance of making a difference. The methods used and developed in the five papers presented in this doctoral thesis illustrate distinct opportunities to advance our understanding in this field – and it is this understanding that will be crucial in our collective endeavour to create a low-carbon equitable economic system for the future.

# Part IV: Appendices

# Contribution Statement

Here, I outline the specific contributions that I have made to each paper to enable the examiners of this thesis to gain a more thorough understanding of my work.

In Paper 1, my main contribution focused on implementing the climate impact application as well as the computational implementation of the bootstrap tests. The asymptotic theory was developed by Xiyu Jiao, while the overall approach and motivation of the study was developed by Felix Pretis. The resulting paper was written jointly (author order is alphabetical).

In Paper 2, I led the design and approach of the overall study, was responsible for the data management, the econometric estimation, and the projections. My co-author Felix Pretis provided guidance on the econometric methods (especially with regard to the model selection algorithms), and in developing the adaptation components.

Paper 3 is a single author paper, which means I was responsible for all aspects of the article.

In Papers 4 and 5, I was responsible for developing, maintaining, and implementing the break detection algorithm used in both chapters based on earlier work by Felix Pretis. I have also maintained and published this algorithm in the R package `getspanel` on the open-source distribution platform CRAN. Paper 4 was conceptually designed and led jointly with Felix Pretis (author order is alphabetical), while Paper 5 was initiated and led by Nico Koch (author order is alphabetical). A precise contribution statement detailing the individual contributions of all authors is contained at the very end of Paper 5.

# Appendix for Paper 1

# ONLINE APPENDIX

## A Proofs of The Main Theorems

The proofs of the main results proceed as follows: initially the one-step stochastic expansion and tightness results are given for the iterated estimators of $(\beta, \sigma^2)$ computed from Algorithm 2.1. Next, we show the expansion of the iterated estimators in any step in terms of the initial estimators and establish the fixed point of the iterated estimators upon through infinite iterations. Given the two different type of initial estimators by the Robustified Least Squares and Impulse Indicator Saturation algorithms, we then build up the stochastic expansions of their algorithmic estimators in any iterated step. All these arguments hold uniformly in the cut-off $c \in [c_+, \infty)$, so weak convergence theory can be established for the estimators of $(\beta, \sigma^2)$ seen as stochastic processes in terms of the drifting cut-off $c$. The proof combines the tightness and finite dimensional convergence, showing that the weak limit varies with the stochastic properties of regressors. Finally, we propose our outlier distortion tests for checking outlier robustness in coefficients.

To conduct all our asymptotic analysis, we require the empirical process theory recently developed by Berenguer-Rico et al. (2019). Thus, we first present Theorem 4.4 from Berenguer-Rico et al. (2019), which summarizes the main result of their paper and provides the first order asymptotic expansions for a class of one-sided weighted and marked empirical processes.

**Lemma A.1.** *(Berenguer-Rico et al. (2019), Theorem 4.4) Suppose Assumption 2.1$(i, iib)$ holds. We have an expansion*

$$n^{-1/2}\sum_{i=1}^{n} w_{in}\varepsilon_i^p 1_{(\varepsilon_i \le \sigma c + n^{-1/2}ac + x_{in}'b)} = n^{-1/2}\sum_{i=1}^{n} w_{in}\varepsilon_i^p 1_{(\varepsilon_i \le \sigma c)} + \mathcal{B}_n(a,b,c) + R(a,b,c),$$

*where the bias term is expressed as*

$$\mathcal{B}_n(a,b,c) = \sigma^{p-1}c^p \mathsf{f}(c) n^{-1/2}\sum_{i=1}^{n} w_{in}(n^{-1/2}ac + x_{in}'b).$$

*Notice $w_{in}$ can be chosen as $1, n^{1/2}x_{in}, nx_{in}x_{in}'$ and $p$ as either of $0, 1, 2$. For any $B > 0$ and as $n \to \infty$, the remainder term satisfies*

$$\sup_{-\infty < c < \infty}\sup_{|a|,|b| \le n^{1/4-\eta}B} |R(a,b,c)| = \mathsf{o}_{\mathsf{P}}(1).$$

*In addition, the normalized process $n^{-1/2}\sum_{i=1}^{n} w_{in}(\varepsilon_i^p 1_{(\varepsilon_i \le \sigma c)} - \mathsf{E}_{i-1}\varepsilon_i^p 1_{(\varepsilon_i \le \sigma c)})$ is tight in $c \in \mathbb{R}$, where $\mathsf{E}_{i-1}(\cdot) = \mathsf{E}(\cdot|\mathcal{F}_{i-1})$ and $\mathsf{E}_{i-1}\varepsilon_i^p 1_{(\varepsilon_i \le \sigma c)} = \mathsf{E}\varepsilon_i^p 1_{(\varepsilon_i \le \sigma c)}$.*

From (2.6) and (2.7), it is clear that the updated estimators of $(\beta, \sigma^2)$ involve the above weighted and marked empirical processes but with the two-sided indicator

$$v_{i,c}^{(m)} = 1_{(|y_i - x_i'\widehat{\beta}_c^{(m)}| \le \widehat{\sigma}_c^{(m)}c)} = 1_{(|\varepsilon_i - x_{in}'\widehat{b}_c^{(m)}| \le \sigma c + n^{-1/2}\widehat{a}_c^{(m)}c)},$$

where $\widehat{b}_c^{(m)} = N^{-1}(\widehat{\beta}_c^{(m)} - \beta)$ and $\widehat{a}_c^{(m)} = n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ are estimation errors for $m$-step estimator of $(\beta, \sigma^2)$. Thus, the next lemma is to extend the stochastic expansions of Lemma A.1 to the empirical processes with the two-sided indicators.

**Lemma A.2.** *Suppose Assumption 2.1(i, iib) holds. We have an expansion*

$$n^{-1/2}\sum_{i=1}^{n} w_{in}\varepsilon_i^p 1_{(|\varepsilon_i - x_{in}'b| \le \sigma c + n^{-1/2}ac)} = n^{-1/2}\sum_{i=1}^{n} w_{in}\varepsilon_i^p 1_{(|\varepsilon_i|\le \sigma c)} + \mathcal{B}_n'(a,b,c) + R'(a,b,c),$$

*where the bias term is expressed as*

$$\mathcal{B}_n'(a,b,c) = 2\sigma^{p-1}c^p \mathsf{f}(c) n^{-1/2}\sum_{i=1}^{n} w_{in}(1_{(p\ even)}n^{-1/2}ac + 1_{(p\ odd)}x_{in}'b).$$

*Notice $w_{in}$ can be chosen as $1, n^{1/2}x_{in}, nx_{in}x_{in}'$ and $p$ as either of $0,1,2$. For any $B>0$ and as $n\to\infty$, the remainder term satisfies*

$$\sup_{0<c<\infty}\sup_{|a|,|b|\le n^{1/4-\eta}B} |R'(a,b,c)| = o_{\mathsf{P}}(1).$$

*In addition, the normalized process $n^{-1/2}\sum_{i=1}^n w_{in}(\varepsilon_i^p 1_{(|\varepsilon_i|\le\sigma c)} - \mathsf{E}_{i-1}\varepsilon_i^p 1_{(|\varepsilon_i|\le\sigma c)})$ is tight in $c\in\mathbb{R}_+$, where $\mathsf{E}_{i-1}\varepsilon_i^p 1_{(|\varepsilon_i|\le\sigma c)} = \mathsf{E}\varepsilon_i^p 1_{(|\varepsilon_i|\le\sigma c)} = \sigma^p \tau_p^c$.*

**Proof of Lemma A.2**. For any $c>0$, first apply the equality for indicators below

$$1_{(|\varepsilon_i - x_{in}'b|\le\sigma c + n^{-1/2}ac)} = 1_{(\varepsilon_i \le \sigma c + n^{-1/2}ac + x_{in}'b)} - 1_{(\varepsilon_i \le -\sigma c - n^{-1/2}ac + x_{in}'b)}$$

to the two-sided empirical processes $n^{-1/2}\sum_{i=1}^n w_{in}\varepsilon_i^p 1_{(|\varepsilon_i - x_{in}'b|\le\sigma c + n^{-1/2}ac)}$. Then, insert the expansions shown in Lemma A.1 with $c$ and $-c$, and use the symmetric property of $\mathsf{f}(c)$ to obtain the bias term

$$\sigma^{p-1}c^p\mathsf{f}(c)n^{-1/2}\sum_{i=1}^{n} w_{in}[\{1+(-1)^p\}n^{-1/2}ac + \{1-(-1)^p\}x_{in}'b].$$

Tightness of the process $n^{-1/2}\sum_{i=1}^n w_{in}(\varepsilon_i^p 1_{(|\varepsilon_i|\le\sigma c)} - \mathsf{E}_{i-1}\varepsilon_i^p 1_{(|\varepsilon_i|\le\sigma c)})$ simply follows from that of the one-sided process $n^{-1/2}\sum_{i=1}^n w_{in}(\varepsilon_i^p 1_{(\varepsilon_i\le\sigma c)} - \mathsf{E}_{i-1}\varepsilon_i^p 1_{(\varepsilon_i\le\sigma c)})$. ■

Before showing the one-step stochastic expansion of the updated estimators, we first present the following lemma, which is a variation of the delta method required for attaining the expansion of $n^{1/2}(\widehat{\sigma}_c^{(m+1)} - \sigma)$ from $n^{1/2}\{(\widehat{\sigma}_c^{(m+1)})^2 - \sigma^2\}$.

**Lemma A.3.** *Let $\{X_n\}$ be a sequence of random variables and $\theta$ be a deterministic parameter. Assume that a univariate function $g$ has the first and second derivatives $\dot{g}, \ddot{g}$, then we have*

$$n^{1/2}\{g(X_n) - g(\theta)\} = \dot{g}(\theta)n^{1/2}(X_n - \theta) + n^{-1/2}\ddot{g}(\bar{\theta})\{n^{1/2}(X_n-\theta)\}^2,$$

*where $|\bar{\theta} - \theta| \le |X_n - \theta|$.*

**Proof of Lemma A.3**. Approximate $g$ around the point $\theta$ by the linear function using the Taylor expansion and particularly check the approximation at the point $X_n$, then

$$g(X_n) = g(\theta) + \dot{g}(\theta)(X_n - \theta) + \ddot{g}(\bar{\theta})(X_n-\theta)^2,$$

where $|\bar{\theta} - \theta| \le |X_n - \theta|$. Rearranging the above immediately gives the expansion shown in the lemma. ■

Equipped with the empirical processes theory in Lemma A.2 and the delta method in Lemma A.3, we can now study the updated estimator (2.6) and (2.7) and build up its stochastic expansion in terms of the original estimator, a kernel and a small remainder term. Denote $c_+ > 0$ as a small positive number.

**Lemma A.4.** *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 2.1. Suppose Assumption 2.1$(i, ii)$ holds, and that $N^{-1}(\widehat{\beta}_c^{(m)} - \beta)$, $n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ are $\mathrm{O_P}(1)$. Then, uniformly in $c \in [c_+, \infty)$ and as $n \to \infty$*

$$N^{-1}(\widehat{\beta}_c^{(m+1)} - \beta) = \frac{2c\mathsf{f}(c)}{\psi_c} N^{-1}(\widehat{\beta}_c^{(m)} - \beta) + (\psi_c \Sigma_n)^{-1} \sum_{i=1}^n x_{in}\varepsilon_i 1_{(|\varepsilon_i| \le \sigma c)} + \mathrm{o_P}(1),$$

$$n^{1/2}(\widehat{\sigma}_c^{(m+1)} - \sigma) = \frac{c(c^2 - \varsigma_c^2)\mathsf{f}(c)}{\tau_2^c} n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma) + \frac{\sigma}{2\tau_2^c} n^{-1/2} \sum_{i=1}^n (\frac{\varepsilon_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|\varepsilon_i| \le \sigma c)} + \mathrm{o_P}(1).$$

**Proof of Lemma A.4**. The $m+1$ step estimators (2.6) and (2.7) of $\beta$, $\sigma^2$ are least squares estimators for the non-outlying observations and satisfy

$$N^{-1}(\widehat{\beta}_c^{(m+1)} - \beta) = (\sum_{i=1}^n x_{in} x_{in}' v_{i,c}^{(m)})^{-1} (\sum_{i=1}^n x_{in}\varepsilon_i v_{i,c}^{(m)}), \tag{A.1}$$

$$\begin{aligned} n^{1/2}\{(\widehat{\sigma}_c^{(m+1)})^2 - \sigma^2\} = \varsigma_c^{-2} (n^{-1} \sum_{i=1}^n v_{i,c}^{(m)})^{-1} n^{-1/2} \{\sigma^2 \sum_{i=1}^n (\frac{\varepsilon_i^2}{\sigma^2} - \varsigma_c^2) v_{i,c}^{(m)} \\ - (\sum_{i=1}^n \varepsilon_i x_{in}' v_{i,c}^{(m)}) (\sum_{i=1}^n x_{in} x_{in}' v_{i,c}^{(m)})^{-1} (\sum_{i=1}^n x_{in}\varepsilon_i v_{i,c}^{(m)})\}. \end{aligned} \tag{A.2}$$

We express the weight $v_{i,c}^{(m)}$ in (2.5) as

$$v_{i,c}^{(m)} = 1_{(|y_i - x_i'\widehat{\beta}_c^{(m)}| \le \widehat{\sigma}_c^{(m)} c)} = 1_{(|\varepsilon_i - x_{in}'\widehat{b}_c^{(m)}| \le \sigma c + n^{-1/2}\widehat{a}_c^{(m)} c)}, \tag{A.3}$$

where $\widehat{b}_c^{(m)} = N^{-1}(\widehat{\beta}_c^{(m)} - \beta)$ and $\widehat{a}_c^{(m)} = n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ are estimation errors for $\beta$ and $\sigma$ in the $m$ step of the algorithm.

By Assumption 2.1$(i, iib)$ and $|\widehat{b}_c^{(m)}| + |\widehat{a}_c^{(m)}| = \mathrm{O_P}(1)$, we can apply the expansions of the empirical processes in Lemma A.2 to (A.1) and (A.2), so for $\widehat{\beta}_c^{(m+1)}$ we have

$$\widehat{b}_c^{(m+1)} = \frac{2c\mathsf{f}(c)}{\psi_c} \widehat{b}_c^{(m)} + (\psi_c \Sigma_n)^{-1} \sum_{i=1}^n x_{in}\varepsilon_i 1_{(|\varepsilon_i| \le \sigma c)} + R_\beta(\widehat{a}_c^{(m)}, \widehat{b}_c^{(m)}, c),$$

where the remainder $R_\beta(a, b, c)$ vanishes uniformly in $c_+ \le c < \infty$ and $|a|, |b| \le B$. A key to this is that $c$ is bounded away from zero and that $\Sigma_n \xrightarrow{\mathrm{D}} \Sigma$ is almost surely positive definite by Assumption 2.1$(iia)$ so that the denominator $\psi_c$, $\psi_c \Sigma_n$ is bounded away from zero.

For $\widehat{\sigma}_c^{(m+1)}$, first write $n^{1/2}(\widehat{\sigma}_c^{(m+1)} - \sigma) = n^{1/2}[\{(\widehat{\sigma}_c^{(m+1)})^2\}^{1/2} - (\sigma^2)^{1/2}]$ and let $g(x) = x^{1/2}$, $X_n = (\widehat{\sigma}_c^{(m+1)})^2$, $\theta = \sigma^2$, then apply Lemma A.3 to obtain

$$n^{1/2}(\widehat{\sigma}_c^{(m+1)} - \sigma) = \frac{1}{2\sigma} n^{1/2}\{(\widehat{\sigma}_c^{(m+1)})^2 - \sigma^2\} + n^{-1/2}\mathrm{O}[n\{(\widehat{\sigma}_c^{(m+1)})^2 - \sigma^2\}^2].$$

Notice $\dot{g}(x) = x^{-1/2}/2$ and $\ddot{g}(x) = -x^{-3/2}/4$. Next, apply the similar arguments as for $\widehat{\beta}_c^{(m+1)}$ to get

$$\widehat{a}_c^{(m+1)} = \frac{c(c^2-\varsigma_c^2)\mathsf{f}(c)}{\tau_2^c}\widehat{a}_c^{(m)} + \frac{\sigma}{2\tau_2^c}n^{-1/2}\sum_{i=1}^{n}(\frac{\varepsilon_i^2}{\sigma^2}-\varsigma_c^2)1_{(|\varepsilon_i|\leq\sigma c)} + R_\sigma(\widehat{a}_c^{(m)},\widehat{b}_c^{(m)},c),$$

where the remainder $R_\sigma(a,b,c)$ also vanishes uniformly. ■

We then prove tightness of iterated estimators. Then, we show the contraction mapping for the one-step expansion of the updated estimator in terms of the original one. With $\eta = 1/4$ corresponding to a bounded initial estimators, this will be sufficient for tightness proof. Note that $|\cdot|$ refers to the usual Euclidean vector norm, while $\|M\| = \max\{\text{eigen}(M'M)\}^{1/2}$ is the spectral norm for any matrix $M$. The norms are compatible so that $|Mx| \leq \|M\||x|$ for any vector $x$.

**Proof of Theorem 2.1**. To make the proof concise, write the one-step expansion more compactly

$$\widehat{u}_c^{(m+1)} = \Gamma_c\widehat{u}_c^{(m)} + K_c + R_u(\widehat{u}_c^{(m)},c), \tag{A.4}$$

where the remainder term satisfies $\sup_{c_+\leq c<\infty}\sup_{|u|\leq B}|R_u(u,c)| = o_{\mathsf{P}}(1)$ and

$$\widehat{u}_c^{(m)} = \begin{pmatrix}\widehat{b}_c^{(m)}\\ \widehat{a}_c^{(m)}\end{pmatrix}, \qquad \Gamma_c = \begin{Bmatrix}\frac{2c\mathsf{f}(c)}{\psi_c}I_{d_x} & 0_{d_x}\\ 0'_{d_x} & \frac{c(c^2-\varsigma_c^2)\mathsf{f}(c)}{\tau_2^c}\end{Bmatrix}, \tag{A.5}$$

$$K_c = \begin{Bmatrix}(\psi_c\Sigma_n)^{-1} & 0_{d_x}\\ 0'_{d_x} & \frac{\sigma}{2\tau_2^c}\end{Bmatrix}\sum_{i=1}^{n}\begin{Bmatrix}x_{in}\varepsilon_i\\ n^{-1/2}(\frac{\varepsilon_i^2}{\sigma^2}-\varsigma_c^2)\end{Bmatrix}1_{(|\varepsilon_i|\leq\sigma c)}. \tag{A.6}$$

Apply the autoregressive equation (A.4) recursively to obtain the representation

$$\widehat{u}_c^{(m+1)} = \Gamma_c^{m+1}\widehat{u}_c^{(0)} + \sum_{l=0}^{m}\Gamma_c^l\{K_c + R_u(\widehat{u}_c^{(m-l)},c)\}. \tag{A.7}$$

Use the triangle inequality and $|Mx| \leq \|M\||x|$ to get

$$|\widehat{u}_c^{(m+1)}| \leq \|\Gamma_c^{m+1}\||\widehat{u}_c^{(0)}| + \{|K_c| + \max_{0\leq l\leq m}|R_u(\widehat{u}_c^{(l)},c)|\}\sum_{l=0}^{m}\|\Gamma_c^l\|.$$

Assumption 2.1($i$) shows $\sup_{c_+\leq c<\infty}\max\{|2c\mathsf{f}(c)/\psi_c|, |c(c^2-\varsigma_c^2)\mathsf{f}(c)/\tau_2^c|\} < 1$; see Theorem 3.5 in Johansen & Nielsen (2013), so $\sup_{c_+\leq c<\infty}\|\Gamma_c\| < 1$. Thus, by Gelfand's formula, see Theorem 3.4 in Varga (2000), $\lim_{m\to\infty}\|M^m\|^{1/m} = \max|\text{eigen}(M)|$, for some $\omega$ such that $\sup_{c_+\leq c<\infty}\|\Gamma_c\| < \omega < 1$ there exists $m_0 > 0$ so for all $m > m_0$ then

$$\sup_{c_+\leq c<\infty}\|\Gamma_c^m\| < \omega^m < 1. \tag{A.8}$$

Also note $(I_{d_x+1} - \Gamma_c)^{-1} = \sum_{l=0}^{\infty}\Gamma_c^l$. This in turn implies for some $1 < B_0 < \infty$

$$\sup_{0\leq m<\infty}\sup_{c_+\leq c<\infty}\|\Gamma_c^m\| < B_0, \qquad \sup_{c_+\leq c<\infty}\|(I_{d_x+1}-\Gamma_c)^{-1}\| \leq \sum_{l=0}^{\infty}\sup_{c_+\leq c<\infty}\|\Gamma_c^l\| < B_0. \tag{A.9}$$

Therefore, we have for all $m \in [0, \infty)$

$$|\widehat{u}_c^{(m+1)}| < B_0\{|\widehat{u}_c^{(0)}| + |K_c| + \max_{0 \le l \le m} |R_u(\widehat{u}_c^{(l)}, c)|\}. \tag{A.10}$$

For any $c \in [c_+, \infty)$, Assumption 2.1$(iii)$ with $\eta = 1/4$ guarantees tightness of $\widehat{u}_c^{(0)}$. Lemma A.2 shows that the kernel $K_c$ process is tight using Assumption 2.1$(i, iib)$. Thus, for all $\epsilon, \delta > 0$ there exist $n_0, U_0 > 0$ so that for all $n > n_0$ the set

$$\mathcal{A}_n = \{B_0 \sup_{c_+ \le c < \infty} (|\widehat{u}_c^{(0)}| + |K_c|) \le U_0/3, B_0 \sup_{c_+ \le c < \infty} \sup_{|u| \le U_0} |R_u(u, c)| < \delta/2\} \tag{A.11}$$

has probability larger than $1 - \epsilon$.

Mathematical induction over $m$ is used to show $\sup_{0 \le m < \infty} \sup_{c_+ \le c < \infty} |\widehat{u}_c^{(m)}| \le U_0$ on the set $\mathcal{A}_n$. For $m = 0$ as induction starts, $\sup_{c_+ \le c < \infty} |\widehat{u}_c^{(0)}| \le B_0^{-1} U_0/3 < U_0$ holds since $B_0 > 1$. The induction assumption is $\sup_{0 \le l \le m} \sup_{c_+ \le c < \infty} |\widehat{u}_c^{(l)}| \le U_0$ implying $B_0 \max_{0 \le l \le m} |R_u(\widehat{u}_c^{(l)}, c)| < \delta/2$. Then the bound in (A.10) becomes $\sup_{c_+ \le c < \infty} |\widehat{u}_c^{(m+1)}| < 2U_0/3 + \delta/2 < U_0$ so that it follows $\sup_{0 \le l \le m+1} \sup_{c_+ \le c < \infty} |\widehat{u}_c^{(l)}| \le U_0$. ■

Next, we show the expansion of the iterated estimator from Algorithm 2.1 at any step in terms of its starting point.

**Proof of Theorem 2.2**. Directly apply the recursive representation (A.7) in the tightness proof, so uniformly in $c \in [c_+, \infty)$ and for any $m \in [0, \infty)$

$$\widehat{u}_c^{(m+1)} = \Gamma_c^{m+1} \widehat{u}_c^{(0)} + \sum_{l=0}^{m} \Gamma_c^l K_c + \sum_{l=0}^{m} \Gamma_c^l R_u(\widehat{u}_c^{(m-l)}, c),$$

where $\sup_{c_+ \le c < \infty} \sup_{|u| \le B} |R_u(u, c)| = \mathsf{o_P}(1)$. Since the spectral radius of $\Gamma_c$ is bounded by one, see (A.8), then (A.9) shows for $1 < B_0 < \infty$

$$\sup_{c_+ \le c < \infty} \|\sum_{l=0}^{m} \Gamma_c^l\| \le \sup_{c_+ \le c < \infty} \sum_{l=0}^{m} \|\Gamma_c^l\| \le \sum_{l=0}^{\infty} \sup_{c_+ \le c < \infty} \|\Gamma_c^l\| < B_0.$$

Further with tightness $\sup_{0 \le m < \infty} \sup_{c_+ \le c < \infty} |\widehat{u}_c^{(m)}| = \mathsf{O_P}(1)$ shown in Theorem 2.1 due to Assumption 2.1 with $\eta = 1/4$, the third term vanishes in the above recursive representation. Applying the equality

$$\sum_{l=0}^{m} \Gamma_c^l = (I_{d_x+1} - \Gamma_c)^{-1}(I_{d_x+1} - \Gamma_c^{m+1}) = (I_{d_x+1} - \Gamma_c^{m+1})(I_{d_x+1} - \Gamma_c)^{-1}, \tag{A.12}$$

we then rearrange the recursive representation to attain

$$\widehat{u}_c^{(m+1)} = \Gamma_c^{m+1} \widehat{u}_c^{(0)} + (I_{d_x+1} - \Gamma_c^{m+1})(I_{d_x+1} - \Gamma_c)^{-1} K_c + \mathsf{o_P}(1), \tag{A.13}$$

uniformly in $c, m$. Recall the definition of $\Gamma_c$ in (A.5), then

$$\Gamma_c^{m+1} = \begin{bmatrix} \{\frac{2c\mathsf{f}(c)}{\psi_c}\}^{m+1} I_{d_x} & 0_{d_x} \\ 0'_{d_x} & \{\frac{c(c^2-\varsigma_c^2)\mathsf{f}(c)}{\tau_2^c}\}^{m+1} \end{bmatrix},$$

$$(I_{d_x+1} - \Gamma_c^{m+1})(I_{d_x+1} - \Gamma_c)^{-1} = \begin{bmatrix} \frac{\psi_c^{m+1} - \{2c\mathsf{f}(c)\}^{m+1}}{\psi_c^m\{\psi_c - 2c\mathsf{f}(c)\}} I_{d_x} & 0_{d_x} \\ 0'_{d_x} & \frac{(\tau_2^c)^{m+1} - \{c(c^2-\varsigma_c^2)\mathsf{f}(c)\}^{m+1}}{(\tau_2^c)^m\{\tau_2^c - c(c^2-\varsigma_c^2)\mathsf{f}(c)\}} \end{bmatrix}.$$

Finally, substitute these and $\widehat{u}_c^{(0)}$, $K_c$ into (A.13) to establish the expansion of $\widehat{u}_c^{(m+1)}$; see $\widehat{u}_c^{(0)}$ in (A.5) and $K_c$ in (A.6). ■

The next corollary re-expresses Lemma A.4 as a stochastic expansion of the fist step estimators in terms of the initial ones, which is a special case of Theorem 2.2 where $m = 0$.

**Proof of Corollary 2.3**. The proof follows by setting $m = 0$ in Theorem 2.2 such that it holds for $\varrho_{\beta,c}^{(1)} = 2c\mathsf{f}(c)/\psi_c$, $\varrho_{x\varepsilon,c}^{(1)} = \psi_c^{-1}$, $\varrho_{\sigma,c}^{(1)} = c(c^2 - \varsigma_c^2)\mathsf{f}(c)/\tau_2^c$, and $\varrho_{\varepsilon\varepsilon,c}^{(1)} = \sigma/(2\tau_2^c)$. ■

We then establish the fixed point of the iterated one-step Huber-skip M-estimators defined in Algorithm 2.1.

**Proof of Theorem 2.4**. Since the spectral radius of $\Gamma_c$ is strictly smaller than one, see (A.8), we have $\Gamma_c^{m+1} \to 0_{(d_x+1)\times(d_x+1)}$ uniformly in $c \in [c_+, \infty)$ as $m \to \infty$. Further with the boundedness of $\widehat{u}_c^{(0)}$ in probability as $n \to \infty$ by Assumption 2.1$(iii)$ with $\eta = 1/4$, the first term in (A.13) vanishes. Notice that to attain the recursive representation (A.13) in the proof of Theorem 2.2, we also require Assumption 2.1$(i, ii)$. Thus, let $n, m \to \infty$ in (A.13), then we obtain the fixed point

$$\widehat{u}_c^{(*)} = \widehat{u}_c^{(\infty)} = (I_{d_x+1} - \Gamma_c)^{-1} K_c, \tag{A.14}$$

uniformly in $c$. Recall the definition of $\Gamma_c$ in (A.5), we then have

$$(I_{d_x+1} - \Gamma_c)^{-1} = \begin{Bmatrix} \frac{\psi_c}{\psi_c - 2c\mathsf{f}(c)} I_{d_x} & 0_{d_x} \\ 0'_{d_x} & \frac{\tau_2^c}{\tau_2^c - c(c^2-\varsigma_c^2)\mathsf{f}(c)} \end{Bmatrix}.$$

Substitute this and $K_c$ into (A.14) to attain the expression of the fixed point $\widehat{u}_c^{(*)}$; see $K_c$ in (A.6).

The next step is to formally prove that $\widehat{u}_c^{(*)}$ is indeed the fixed point. Replace (A.7) and (A.14) into the deviation $\widehat{\Delta}_c^{(m+1)} = \widehat{u}_c^{(m+1)} - \widehat{u}_c^{(*)}$ and apply (A.12) to attain

$$\widehat{\Delta}_c^{(m+1)} = \Gamma_c^{m+1}\{\widehat{u}_c^{(0)} - (I_{d_x+1} - \Gamma_c)^{-1} K_c\} + \sum_{l=0}^{m} \Gamma_c^l R_u(\widehat{u}_c^{(m-l)}, c).$$

To bound $\widehat{\Delta}_c^{(m+1)}$, use the triangle inequality and $|Mx| \leq \|M\||x|$ to get

$$|\widehat{\Delta}_c^{(m+1)}| \leq \|\Gamma_c^{m+1}\|\{|\widehat{u}_c^{(0)}| + \|(I_{d_x+1} - \Gamma_c)^{-1}\||K_c|\} + \max_{0\leq l\leq m} |R_u(\widehat{u}_c^{(l)}, c)| \sum_{l=0}^{m} \|\Gamma_c^l\|.$$

Further bound above using the inequalities (A.8) and (A.9), so for $m > m_0$

$$|\widehat{\Delta}_c^{(m+1)}| < \omega^{m+1}(|\widehat{u}_c^{(0)}| + B_0|K_c|) + B_0 \max_{0\leq l\leq m} |R_u(\widehat{u}_c^{(l)}, c)|.$$

On the set $\mathcal{A}_n$ as in (A.11), since $\sup_{0\le m<\infty}\sup_{c_+\le c<\infty}|\widehat{u}_c^{(m)}|\le U_0$ by Theorem 2.1, then we have $\sup_{c_+\le c<\infty}|\widehat{\Delta}_c^{(m+1)}|<\omega^{m+1}(B_0^{-1}U_0/3+U_0/3)+\delta/2<\omega^{m+1}U_0+\delta/2$. As $0<\omega<1$, $\omega^{m+1}$ declines exponentially so $m_0$ can be chosen sufficiently large that for all $m>m_0$ then $\omega^{m+1}U_0<\delta/2$. Thus, $\mathsf{P}(\sup_{c_+\le c<\infty}|\widehat{\Delta}_c^{(m+1)}|<\delta)>1-\epsilon$ for $m>m_0$, $n>n_0$. ■

Next considering RLS and IIS, we first build up their stochastic expansions.

**Proof of Theorem 2.5**. We first establish the expansion for the Robustified Least Squares. Then, we show that the Impulse Indicator Saturation has the identical expansion of the initial step updated estimator as the Robustified Least Squares so that they have the same general $m+1$ step expansion for any $m\in[0,\infty)$.

RLS starts with the full sample least squares $(\widetilde{\beta},\widetilde{\sigma})$ which are tight and satisfy

$$N^{-1}(\widetilde{\beta}-\beta)=(\sum_{i=1}^{n}x_{in}x_{in}')^{-1}(\sum_{i=1}^{n}x_{in}\varepsilon_i),\quad n^{1/2}(\widetilde{\sigma}-\sigma)=\frac{\sigma}{2}n^{-1/2}\sum_{i=1}^{n}(\frac{\varepsilon_i^2}{\sigma^2}-1)+\mathrm{O}_{\mathsf{P}}(n^{-1/2}).$$

Substitute the above expansion of the initial estimators $(\widehat{\beta}_c^{(0)},\widehat{\sigma}_c^{(0)})=(\widetilde{\beta},\widetilde{\sigma})$ into Theorem 2.2 using Assumption 2.1$(i,ii)$, then it holds for any $m\in[0,\infty)$,

$$\begin{aligned}N^{-1}(\widehat{\beta}_c^{(m+1)}-\beta)&=\varrho_{\beta,c}^{(m+1)}\Sigma_n^{-1}\sum_{i=1}^{n}x_{in}\varepsilon_i+\varrho_{x\varepsilon,c}^{(m+1)}\Sigma_n^{-1}\sum_{i=1}^{n}x_{in}\varepsilon_i1_{(|\varepsilon_i|\le\sigma c)}+\mathrm{o}_{\mathsf{P}}(1),\\ n^{1/2}(\widehat{\sigma}_c^{(m+1)}-\sigma)&=\varrho_{\sigma,c}^{(m+1)}\frac{\sigma}{2}n^{-1/2}\sum_{i=1}^{n}(\frac{\varepsilon_i^2}{\sigma^2}-1)+\varrho_{\varepsilon\varepsilon,c}^{(m+1)}n^{-1/2}\sum_{i=1}^{n}(\frac{\varepsilon_i^2}{\sigma^2}-\varsigma_c^2)1_{(|\varepsilon_i|\le\sigma c)}+\mathrm{o}_{\mathsf{P}}(1),\end{aligned}$$

uniformly in $c\in[c_+,\infty)$.

To demonstrate that two algorithms RLS and IIS have the same expansion of $N^{-1}(\widehat{\beta}_c^{(m+1)}-\beta)$ for $m\in[0,\infty)$ even they start with different initial estimators when running Algorithm 2.1, it suffices to show that the expansion of $N^{-1}(\widehat{\beta}_c^{(1)}-\beta)$ for IIS is the same as RLS in the above. The updated estimator for $\beta$ from an initial step in IIS is expressed as

$$\begin{aligned}N^{-1}(\widehat{\beta}_c^{(1)}-\beta)=&\{\sum_{j=1,2}(N_{3-j}^{-1}N)'\sum_{i\in\mathcal{I}_{3-j}}x_{in_{3-j}}x_{in_{3-j}}'1_{(|y_i-x_i'\widehat{\beta}_j|\le\widehat{\sigma}_jc)}N_{3-j}^{-1}N\}^{-1}\\&\times\{\sum_{j=1,2}(N_{3-j}^{-1}N)'\sum_{i\in\mathcal{I}_{3-j}}x_{in_{3-j}}\varepsilon_i1_{(|y_i-x_i'\widehat{\beta}_j|\le\widehat{\sigma}_jc)}\}.\end{aligned}$$

Notice $x_{in_j}=N_j'x_i$ denotes the normalized regressors for each subsample $i\in\mathcal{I}_j$, $j=1,2$, where $N_j$ corresponds to the normalization matrix based on the subsample size $n_j$. Argue along the lines of Lemma A.4 using the expansions in Lemma A.2, then it follows

$$\begin{aligned}N^{-1}(\widehat{\beta}_c^{(1)}-\beta)=&(\psi_c\Sigma_n)^{-1}\{2c\mathsf{f}(c)\sum_{j=1,2}(N_{3-j}^{-1}N)'\Sigma_{n_{3-j}}^{\mathcal{I}_{3-j}}(N_{3-j}^{-1}N_j)N_j^{-1}(\widehat{\beta}_j-\beta)\\&+\sum_{i=1}^{n}x_{in}\varepsilon_i1_{(|\varepsilon_i|\le\sigma c)}\}+\mathrm{o}_{\mathsf{P}}(1),\end{aligned}$$

uniformly in $c\in[c_+,\infty)$ and where we denote $\Sigma_{n_j}^{\mathcal{I}_j}=\sum_{i\in\mathcal{I}_j}x_{in_j}x_{in_j}'$ for $j=1,2$. To apply the empirical processes argument in the above, we require Assumption 2.1$(i,iib)$ holding for each

subsample $\mathcal{I}_j$ and the fact that the initial estimators $N_j^{-1}(\widehat{\beta}_j-\beta)$, $n_j^{1/2}(\widehat{\sigma}_j-\sigma)$ are tight for $j=1,2$. Substitute the expansion of least squares on each subsample $\mathcal{I}_j$, $N_j^{-1}(\widehat{\beta}_j-\beta)=(\Sigma_{n_j}^{\mathcal{I}_j})^{-1}\sum_{i\in\mathcal{I}_j}x_{in_j}\varepsilon_i$ for $j=1,2$, then we have

$$N^{-1}(\widehat{\beta}_c^{(1)}-\beta)=\frac{2c\mathsf{f}(c)}{\psi_c}\Sigma_n^{-1}\sum_{j=1,2}(N_{3-j}^{-1}N)'\Sigma_{n_{3-j}}^{\mathcal{I}_{3-j}}(N_{3-j}^{-1}N_j)(\Sigma_{n_j}^{\mathcal{I}_j})^{-1}\sum_{i\in\mathcal{I}_j}x_{in_j}\varepsilon_i$$
$$+(\psi_c\Sigma_n)^{-1}\sum_{i=1}^{n}x_{in}\varepsilon_i 1_{(|\varepsilon_i|\leq\sigma c)}+\mathsf{o_P}(1).$$

Notice that we assume $n_1=\mathrm{int}[n/2]$ and $n_2=n-n_1$ so that $N_{3-j}^{-1}N_j\to I_{d_x}$ and $N_jN_{3-j}^{-1}\to I_{d_x}$ as $n\to\infty$ and $n_j\to\infty$ for $j=1,2$. Combine this fact and Assumption 2.1($iia$) that $\Sigma_{n_j}^{\mathcal{I}_j}\xrightarrow{\mathsf{P}}\Sigma$ for $j=1,2$ to attain

$$N^{-1}(\widehat{\beta}_c^{(1)}-\beta)=\frac{2c\mathsf{f}(c)}{\psi_c}\Sigma_n^{-1}\sum_{i=1}^{n}x_{in}\varepsilon_i+(\psi_c\Sigma_n)^{-1}\sum_{i=1}^{n}x_{in}\varepsilon_i 1_{(|\varepsilon_i|\leq\sigma c)}+\mathsf{o_P}(1).$$

Then, we find that RLS and IIS have the identical expansion of $N^{-1}(\widehat{\beta}_c^{(1)}-\beta)$ by noting that $\varrho_{\beta,c}^{(1)}=2c\mathsf{f}(c)/\psi_c$ and $\varrho_{x\varepsilon,c}^{(1)}=\psi_c^{-1}$, so as for the general $m+1$ step beta estimators for $m\in[0,\infty)$. Using the similar reasoning as on beta, it follows that two algorithms also have the same expansion on sigma. ■

Drifting the cut-off value $c\in[c_+,\infty)$, we derive the weak convergence theory of the processes $\mathbb{G}_n^{(m+1)}(c)=N^{-1}(\widehat{\beta}_c^{(m+1)}-\beta)$ computed by RLS and IIS for any $m\in[0,\infty)$ in the stationary case, where $N=n^{-1/2}I_{d_x}$ such that $x_{in}=n^{-1/2}x_i$ and $\Sigma=\mathsf{E}x_ix_i'$.

**Proof of Theorem 2.6**. By Assumption 2.1($i,ii$), Theorem 2.5 shows that for any $m\in[0,\infty)$ and uniformly in $c\in[c_+,\infty)$

$$\mathbb{G}_n^{(m+1)}(c)=\begin{pmatrix}\varrho_{\beta,c}^{(m+1)}\Sigma_n^{-1}\\ \varrho_{x\varepsilon,c}^{(m+1)}\Sigma_n^{-1}\end{pmatrix}'\sum_{i=1}^{n}\begin{pmatrix}x_{in}\varepsilon_i\\ x_{in}\varepsilon_i 1_{(|\varepsilon_i|\leq\sigma c)}\end{pmatrix}+\mathsf{o_P}(1).$$

Then, by $\Sigma_n\xrightarrow{\mathsf{P}}\Sigma>0$, the limiting distribution of the kernel vector, and Slutsky's theorem, we have for any $c\in[c_+,\infty)$

$$\mathbb{G}_n^{(m+1)}(c)\xrightarrow{\mathsf{D}}\begin{pmatrix}\varrho_{\beta,c}^{(m+1)}\Sigma^{-1}\\ \varrho_{x\varepsilon,c}^{(m+1)}\Sigma^{-1}\end{pmatrix}'\mathsf{N}\left\{\begin{pmatrix}0_{d_x}\\ 0_{d_x}\end{pmatrix},\sigma^2\tau_2^c\begin{pmatrix}\frac{1}{\tau_2^c}\Sigma & \Sigma\\ \Sigma & \Sigma\end{pmatrix}\right\}.$$

Since a transformation of multivariate normal is still normal, it follows

$$\mathbb{G}_n^{(m+1)}(c)\xrightarrow{\mathsf{D}}\mathsf{N}[0_{d_x},\{(\varrho_{\beta,c}^{(m+1)})^2+2\tau_2^c\varrho_{\beta,c}^{(m+1)}\varrho_{x\varepsilon,c}^{(m+1)}+\tau_2^c(\varrho_{x\varepsilon,c}^{(m+1)})^2\}\sigma^2\Sigma^{-1}].$$

Theorem 2.1 demonstrates the tightness of the processes $\mathbb{G}_n^{(m+1)}$ for any $m\in[0,\infty)$, so

$$\mathbb{G}_n^{(m+1)}\rightsquigarrow\mathbb{G}^{(m+1)},$$

where the weak limit $\mathbb{G}^{(m+1)}$ is a zero mean Gaussian process with the variance

$$\mathsf{Var}\{\mathbb{G}^{(m+1)}(c)\} = \{(\varrho_{\beta,c}^{(m+1)})^2 + 2\tau_2^c \varrho_{\beta,c}^{(m+1)} \varrho_{x\varepsilon,c}^{(m+1)} + \tau_2^c (\varrho_{x\varepsilon,c}^{(m+1)})^2\}\sigma^2\Sigma^{-1}. \quad \blacksquare$$

Next, we give the weak convergence of the first step and fixed point estimator of RLS and IIS.

**Proof of Corollary 2.7**. These are two special cases of Theorem 2.6 where $m = 0$ such that $\varrho_{\beta,c}^{(1)} = 2c\mathsf{f}(c)/\psi_c$, $\varrho_{x\varepsilon,c}^{(1)} = \psi_c^{-1}$ and where $m \to \infty$ such that $\varrho_{\beta,c}^{(\infty)} = 0$, $\varrho_{x\varepsilon,c}^{(\infty)} = 1/\{\psi_c - 2c\mathsf{f}(c)\}$. $\blacksquare$

We now turn to proving the results for the outlier distortion test and first show the stochastic expansion and weak limit of the difference processes between the RLS/IIS and full sample OLS. We mainly concentrates on stationary regressions here, and therefore choose $N = n^{-1/2} I_{d_x}$ such that the RLS/IIS and OLS need to be normalised by $N^{-1} = n^{1/2} I_{d_x}$, $x_{in} = n^{-1/2} x_i$, and $\Sigma = \mathsf{E} x_i x_i'$.

**Proof of Theorem 2.8**. Rearrange $\mathbb{H}_n^{(m+1)}(c)$ as

$$\mathbb{H}_n^{(m+1)}(c) = n^{1/2}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) = n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta) - n^{1/2}(\widetilde{\beta} - \beta).$$

Incorporate the expansions of RLS/IIS $n^{1/2}(\widehat{\beta}_c^{(m+1)} - \beta)$ and OLS $n^{1/2}(\widetilde{\beta} - \beta)$ from Theorem 2.5 and its proof into the above term, then the expansion is immediately attained for $\mathbb{H}_n^{(m+1)}(c)$ for any $m \in [0, \infty)$ and $c \in [c_+, \infty)$. We can next obtain the weak Gaussian limit $\mathbb{H}^{(m+1)}$ of a sequence of processes $\mathbb{H}_n^{(m+1)}$ by arguing along the lines of Proof of Theorem 2.6 but replacing $\varrho_{\beta,c}^{(m+1)}$ by $\varrho_{\beta,c}^{(m+1)} - 1$. $\blacksquare$

Next, we establish the outlier distortion test and prove Corollary 2.9.

**Proof of Corollary 2.9**. Given any cut-off values, the Gaussian weak limit from Theorem 2.8 immediately implies that the difference between RLS/IIS and OLS converges pointwisely to a Normal distribution. Thus, the proposed Hausman type test statistics has a limiting chi-squared distribution. $\blacksquare$

Using the relative efficiency argument similar to Hausman (1978), we show in the below lemma that the asymptotic variance of the difference between RLS/IIS and OLS can be given by the difference of their respective asymptotic variances under the null of no outliers.

**Lemma A.5.** *Consider RLS or IIS. Suppose Assumption 2.1*$(i, ii)$ *holds. For any* $m \in [0, \infty)$, $c \in [c_+, \infty)$ *and as* $n \to \infty$, *we have*

$$\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) = \mathsf{avar}(\widehat{\beta}_c^{(m+1)}) - \mathsf{avar}(\widetilde{\beta}).$$

**Proof of Lemma A.5**. Under the null of no outliers, for any $m \in [0, \infty)$, $c \in [c_+, \infty)$ RLS/IIS $\widehat{\beta}_c^{(m+1)}$ and OLS $\widetilde{\beta}$ are both consistent, although $\widehat{\beta}_c^{(m+1)}$ is less efficient than $\widetilde{\beta}$ in terms of having the higher asymptotic variance, see Theorem 2.6. We take a weighted average between RLS/IIS and OLS to construct a new estimator

$$\widehat{\theta}_c^{(m+1)}(\lambda) = \lambda \widehat{\beta}_c^{(m+1)} + (1 - \lambda)\widetilde{\beta},$$

where $\lambda \in [0, 1]$. The class of estimators $\widehat{\theta}_c^{(m+1)}(\lambda)$ are consistent, and the choice of $\lambda$ determines the trade-off between efficiency and robustness. The closer $\lambda$ is to zero, the more efficient $\widehat{\theta}_c^{(m+1)}(\lambda)$

is. When $\lambda = 0$ the constructed estimator becomes OLS, so $\widehat{\theta}_c^{(m+1)}(0) = \widetilde{\beta}$ is the most efficient estimator in the class, that is to say $\mathsf{avar}\{\widehat{\theta}_c^{(m+1)}(\lambda)\}$ attains the minimum at zero. Notice

$$\mathsf{avar}\{\widehat{\theta}_c^{(m+1)}(\lambda)\} = \lambda^2\mathsf{avar}(\widehat{\beta}_c^{(m+1)}) + (1-\lambda)^2\mathsf{avar}(\widetilde{\beta}) + 2\lambda(1-\lambda)\mathsf{acov}(\widehat{\beta}_c^{(m+1)}, \widetilde{\beta}).$$

Then, we check its first and second derivatives

$$\begin{aligned}
&\frac{d}{d\lambda}\mathsf{avar}\{\widehat{\theta}_c^{(m+1)}(\lambda)\} = 2\{\lambda\mathsf{avar}(\widehat{\beta}_c^{(m+1)}) - (1-\lambda)\mathsf{avar}(\widetilde{\beta}) + (1-2\lambda)\mathsf{acov}(\widehat{\beta}_c^{(m+1)}, \widetilde{\beta})\},\\
&\frac{d^2}{d\lambda^2}\mathsf{avar}\{\widehat{\theta}_c^{(m+1)}(\lambda)\} = 2\{\mathsf{avar}(\widehat{\beta}_c^{(m+1)}) + \mathsf{avar}(\widetilde{\beta}) - 2\mathsf{acov}(\widehat{\beta}_c^{(m+1)}, \widetilde{\beta})\} = 2\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) \geq 0.
\end{aligned}$$

Thus, the function $\mathsf{avar}\{\widehat{\theta}_c^{(m+1)}(\lambda)\}$ is convex and minimised at $\lambda = 0$, then it follows that $\frac{d}{d\lambda}\mathsf{avar}\{\widehat{\theta}_c^{(m+1)}(\lambda)\}|_{\lambda=0} = 0$ subsequently implying $\mathsf{acov}(\widehat{\beta}_c^{(m+1)}, \widetilde{\beta}) = \mathsf{avar}(\widetilde{\beta})$. Finally,

$$\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) = \mathsf{avar}(\widehat{\beta}_c^{(m+1)}) + \mathsf{avar}(\widetilde{\beta}) - 2\mathsf{acov}(\widehat{\beta}_c^{(m+1)}, \widetilde{\beta}) = \mathsf{avar}(\widehat{\beta}_c^{(m+1)}) - \mathsf{avar}(\widetilde{\beta}). \quad \blacksquare$$

We provide a different but more direct proof in Remark A.1 to demonstrate the equality $\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) = \mathsf{avar}(\widehat{\beta}_c^{(m+1)}) - \mathsf{avar}(\widetilde{\beta})$ under the null of no outliers. The proof relies on the asymptotics derived for RLS/IIS $\widehat{\beta}_c^{(m+1)}$ shown in Theorem 2.6. Furthermore, the remark indirectly indicates that the normal distribution for errors satisfies the regularity conditions required by Lemma A.5.

**Remark A.1.** *Rearrange the expression of* $\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ *from Corollary 2.9 to attain*

$$\{(\varrho_{\beta,c}^{(m+1)})^2 - 2\varrho_{\beta,c}^{(m+1)} + 1 + 2\tau_2^c\varrho_{\beta,c}^{(m+1)}\varrho_{x\varepsilon,c}^{(m+1)} - 2\tau_2^c\varrho_{x\varepsilon,c}^{(m+1)} + \tau_2^c(\varrho_{x\varepsilon,c}^{(m+1)})^2\}\sigma^2\Sigma^{-1}.$$

*Recall* $\tau_2^c = \psi_c - 2c\mathsf{f}(c)$ *if* $\mathsf{f} \overset{\mathsf{D}}{=} \mathsf{N}(0,1)$ *and from Theorem 2.2 that*

$$\varrho_{\beta,c}^{(m+1)} = \{\frac{2c\mathsf{f}(c)}{\psi_c}\}^{m+1}, \qquad \varrho_{x\varepsilon,c}^{(m+1)} = \frac{\psi_c^{m+1} - \{2c\mathsf{f}(c)\}^{m+1}}{\psi_c^{m+1}\{\psi_c - 2c\mathsf{f}(c)\}}.$$

*Apply these terms to have* $-2\varrho_{\beta,c}^{(m+1)} + 1 - 2\tau_2^c\varrho_{x\varepsilon,c}^{(m+1)} = -1$*, and further notice from Theorem 2.6 that* $\mathsf{avar}(\widehat{\beta}_c^{(m+1)}) = \{(\varrho_{\beta,c}^{(m+1)})^2 + 2\tau_2^c\varrho_{\beta,c}^{(m+1)}\varrho_{x\varepsilon,c}^{(m+1)} + \tau_2^c(\varrho_{x\varepsilon,c}^{(m+1)})^2\}\sigma^2\Sigma^{-1}$ *and* $\mathsf{avar}(\widetilde{\beta}) = \sigma^2\Sigma^{-1}$*. Thus, we finally show*

$$\begin{aligned}
\mathsf{avar}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta}) &= \{(\varrho_{\beta,c}^{(m+1)} - 1)^2 + 2\tau_2^c(\varrho_{\beta,c}^{(m+1)} - 1)\varrho_{x\varepsilon,c}^{(m+1)} + \tau_2^c(\varrho_{x\varepsilon,c}^{(m+1)})^2\}\sigma^2\Sigma^{-1}\\
&= \{(\varrho_{\beta,c}^{(m+1)})^2 + 2\tau_2^c\varrho_{\beta,c}^{(m+1)}\varrho_{x\varepsilon,c}^{(m+1)} + \tau_2^c(\varrho_{x\varepsilon,c}^{(m+1)})^2\}\sigma^2\Sigma^{-1} - \sigma^2\Sigma^{-1}\\
&= \mathsf{avar}(\widehat{\beta}_c^{(m+1)}) - \mathsf{avar}(\widetilde{\beta}).
\end{aligned}$$

Finally, we prove Corollary 2.10, which provides two special cases of the outlier distortion tests comparing $\widetilde{\beta}$ with $\widehat{\beta}_c^{(1)}$ when $m = 0$ and with $\widehat{\beta}_c^{(*)}$ when $m \to \infty$.

**Proof of Corollary 2.10**. Set $m = 0$ and $m \to \infty$ such that $\varrho_{\beta,c}^{(1)} = 2c\mathsf{f}(c)/\psi_c$, $\varrho_{x\varepsilon,c}^{(1)} = \psi_c^{-1}$ and $\varrho_{\beta,c}^{(\infty)} = 0$, $\varrho_{x\varepsilon,c}^{(\infty)} = \{\psi_c - 2c\mathsf{f}(c)\}^{-1}$ for $\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(m+1)} - \widetilde{\beta})$ in (2.12), then we obtain our tests. $\blacksquare$

The argument of the paper still works even when the error distribution becomes asymmetric and the cut-off values $\underline{c}$ and $\overline{c}$ are chosen accordingly, but we need an empirical process theory which

can be adapted to this situation. Thus, our last lemma is to present Theorem 3.1 from Johansen & Nielsen (2009) which demonstrates LLN and CLT for the corresponding class of empirical processes. Let $(\widetilde{\beta},\widetilde{\sigma})$ be the initial estimator of $(\beta,\sigma)$ and denote their estimation errors as $\widetilde{b}=N^{-1}(\widetilde{\beta}-\beta)$ and $\widetilde{a}=n^{1/2}(\widetilde{\sigma}-\sigma)$, then the indicator variables for selecting non-outlying observations can be re-expressed as $v_{i,c}=1_{(\widetilde{\sigma}\underline{c}\leq y_i-x_i'\widetilde{\beta}\leq\widetilde{\sigma}\overline{c})}=1_{(\sigma\underline{c}+n^{-1/2}\widetilde{a}\underline{c}\leq\varepsilon_i-x_{in}'\widetilde{b}\leq\sigma\overline{c}+n^{-1/2}\widetilde{a}\overline{c})}$. Equipped with all these notations, it is now ready to show the lemma.

**Lemma A.6.** *(Johansen & Nielsen (2009), Theorem 3.1) Suppose Assumption 2.1 holds with $\eta=1/4$. Further assume $\mu_n=n^{-1/2}\sum_{i=1}^n x_{in}\xrightarrow{\mathsf{D}}\mu$ and that $\underline{c}$ and $\overline{c}$ are chosen such that $\tau_1^c=0$. Then, it holds*

$$\begin{aligned}
n^{-1}\sum_{i=1}^n v_{i,c}&\xrightarrow{\mathsf{P}}\tau_0^c=1-\gamma_c=\psi_c,\\
n^{-1/2}\sum_{i=1}^n x_{in}v_{i,c}&\xrightarrow{\mathsf{D}}\psi_c\mu,\\
\sum_{i=1}^n x_{in}x_{in}'v_{i,c}&\xrightarrow{\mathsf{D}}\psi_c\Sigma.
\end{aligned}$$

*In addition, denote $\xi_k^c=\overline{c}^k\mathsf{f}(\overline{c})-\underline{c}^k\mathsf{f}(\underline{c})$ then we have expansions*

$$\begin{aligned}
n^{-1/2}\sum_{i=1}^n v_{i,c}&=n^{-1/2}\sum_{i=1}^n 1_{(\sigma\underline{c}\leq\varepsilon_i\leq\sigma\overline{c})}+\frac{\xi_0^c}{\sigma}\mu_n'\widetilde{b}+\frac{\xi_1^c}{\sigma}\widetilde{a}+\mathrm{o}_{\mathsf{P}}(1),\\
\sum_{i=1}^n x_{in}\varepsilon_i v_{i,c}&=\sum_{i=1}^n x_{in}\varepsilon_i 1_{(\sigma\underline{c}\leq\varepsilon_i\leq\sigma\overline{c})}+\xi_1^c\Sigma_n\widetilde{b}+\xi_2^c\mu_n\widetilde{a}+\mathrm{o}_{\mathsf{P}}(1),\\
n^{-1/2}\sum_{i=1}^n \varepsilon_i^2 v_{i,c}&=n^{-1/2}\sum_{i=1}^n \varepsilon_i^2 1_{(\sigma\underline{c}\leq\varepsilon_i\leq\sigma\overline{c})}+\sigma\xi_2^c\mu_n'\widetilde{b}+\sigma\xi_3^c\widetilde{a}+\mathrm{o}_{\mathsf{P}}(1).
\end{aligned}$$

## B Bootstrap Algorithms

Here we provide additional detail on the proposed bootstrap implementations of our tests. As outlined in the main text section 2.4.1, we consider the bootstrap to either approximate the distribution of the L2 norm of the difference between OLS and robust coefficient estimates, or estimate the variance of the coefficient difference. For each approach, we consider different resampling schemes: non-parametric case resampling of the raw data, as well as non-parametric case resampling or a parametric residual bootstrap based on the cleaned, outlier-removed data.

The test statistic we are interested in is

$$\widehat{H}_{n,c}^{(1)} = n(\widehat{\beta}_c^{(1)} - \widetilde{\beta})' \widehat{\mathsf{avar}}(\widehat{\beta}_c^{(1)} - \widetilde{\beta})^{-1}(\widehat{\beta}_c^{(1)} - \widetilde{\beta}) \overset{a}{\sim} \chi^2_{d_x},$$

where

$$\widehat{\mathsf{avar}}(\widehat{\beta}_c^{(1)} - \widetilde{\beta}) = \frac{\{2c\mathsf{f}(c) - \psi_c\}^2 + 2\tau_2^c\{2c\mathsf{f}(c) - \psi_c\} + \tau_2^c}{\psi_c^2}(\widehat{\sigma}_c^{(1)})^2(\widehat{\Sigma}_c^{(1)})^{-1}.$$

The algorithm to construct the L2 bootstrap is given as follows, where re-sampling scheme A, B, or C has to be chosen.

**Algorithm B.1.** ***L2 Bootstrap***. *Choose a cut-off c and bootstrap replication number B.*
*1. Compute the OLS* $\widetilde{\beta}$ *and robust estimate* $\widehat{\beta}_c^{(1)}$ *using Algorithm 2.1 based on the original sample* $\{(y_i, x_i)\}_{i=1}^n$. *Then, construct the L2 norm of the coefficient difference* $T = \|\widehat{\beta}_c^{(1)} - \widetilde{\beta}\|_2$.
*2. Choose one of the following three re-sampling schemes* A, B, C.
**A.** ***Non-parametric case re-sampling from raw data***. *Draw a bootstrap sample* $\{(y_i^*, x_i^*)\}_{i=1}^n$ *by re-sampling pairs of observations* $(y_i, x_i)$ *with replacement from the original sample* $\{(y_i, x_i)\}_{i=1}^n$.
**B.** ***Non-parametric case re-sampling from outlier-removed data***. *Draw a bootstrap sample* $\{(y_i^*, x_i^*)\}_{i=1}^n$ *by re-sampling pairs of observations* $(y_i, x_i)$ *with replacement only from the outlier-removed sample* $\{(y_i, x_i)|v_{i,c}^{(0)} = 1\}$.
**C.** ***Parametric residual bootstrap from outlier-removed data***. *Compute residuals using the robust estimate from step 1 such that* $\widehat{\varepsilon}_{i,c}^{(1)} = y_i - x_i'\widehat{\beta}_c^{(1)}$, $i = 1, 2, \ldots, n$. *Re-sample with replacement from* $\{\widehat{\varepsilon}_{i,c}^{(1)}\}_{i=1}^n$ *to obtain bootstrap residuals* $\{\widehat{\varepsilon}_i^*\}_{i=1}^n$ *and construct a bootstrap sample* $\{(y_i^*, x_i)\}_{i=1}^n$ *through the parametric regression* $y_i^* = x_i'\widehat{\beta}_c^{(1)} + \widehat{\varepsilon}_i^*$, $i = 1, 2, \ldots, n$.
*3. Repeat step 2 B times. For each boostrap replication* $b = 1, 2, \ldots, B$, *compute* $\widehat{\beta}_{c,b}^{(1)*}$ *and* $\widetilde{\beta}_b^*$ *based on the bootstrap sample and construct* $T_b^* = \|\widehat{\beta}_{c,b}^{(1)*} - \widetilde{\beta}_b^*\|_2$.
*4. Compare the test statistic T from step 1 to the bootstrap distribution of* $T^*$ *from step 3, and thus reject the null if T exceeds the chosen percentile of* $\{T_b^*\}_{b=1}^B$.

The idea of another type of bootstrap algorithms is to still follow the Hausman type test statistics with the limiting chi-squared distribution, but to replace the estimated asymptotic variance which depends on $\mathsf{f}$ by an estimate using bootstrap. It shares the same spirit as the bootstrap approach proposed for the regular Hausman test by Cameron & Trivedi (2005) and (2010). The algorithm to test the null hypothesis of no distortion using the boostrap-estimate of the variance is given as follows.

**Algorithm B.2.** ***Variance Bootstrap***. *Choose a cut-off c and bootstrap replication number B.*
*1. Compute the OLS* $\widetilde{\beta}$ *and robust estimate* $\widehat{\beta}_c^{(1)}$ *using Algorithm 2.1 based on the original sample* $\{(y_i, x_i)\}_{i=1}^n$.
*2. Choose and proceed with one of the three re-sampling schemes* A, B, C *shown in the step 2 of Algorithm B.1 to draw the bootstrap sample* $\{(y_i^*, x_i^*)\}_{i=1}^n$.

*3. Repeat step 2 B times. For each boostrap replication* $b = 1, 2, \ldots, B$*, compute* $\widehat{\beta}_{c,b}^{(1)*}$ *and* $\widetilde{\beta}_b^*$ *based on the bootstrap sample. Estimate* $\mathsf{var}(\widehat{\beta}_c^{(1)} - \widetilde{\beta})$ *with*

$$\widehat{\mathsf{var}}(\widehat{\beta}_c^{(1)} - \widetilde{\beta})^* = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\beta}_{c,b}^{(1)*} - \widetilde{\beta}_b^* - \bar{\beta}_{c,\text{diff}}^{(1)*})(\widehat{\beta}_{c,b}^{(1)*} - \widetilde{\beta}_b^* - \bar{\beta}_{c,\text{diff}}^{(1)*})',$$

*where* $\bar{\beta}_{c,\text{diff}}^{(1)*} = (1/B)\sum_{b=1}^{B}(\widehat{\beta}_{c,b}^{(1)*} - \widetilde{\beta}_b^*)$.
*4. Construct the outlier distortion test statistic using the difference of two estimates* $\widehat{\beta}_c^{(1)} - \widetilde{\beta}$ *from step 1 and bootstrap variance estimate* $\widehat{\mathsf{var}}(\widehat{\beta}_c^{(1)} - \widetilde{\beta})^*$ *from step 3 such that*

$$(\widehat{\beta}_c^{(1)} - \widetilde{\beta})'\widehat{\mathsf{var}}(\widehat{\beta}_c^{(1)} - \widetilde{\beta})^{*-1}(\widehat{\beta}_c^{(1)} - \widetilde{\beta}) \overset{a}{\sim} \chi_{d_x}^2.$$

*Thus, reject the null if the above test statistic exceeds the chosen critical value of the* $\chi_{d_x}^2$ *distribution.*

For time series, we perform both parametric residual bootstraps as well as a non-parametric time series block bootstrap. For the parametric residual bootstrap, we use the scheme C shown in step 2 of Algorithm B.1 and generate the bootstrap sample $\{y_t^*\}_{t=0}^T$ iteratively using the estimated autogressive coefficient in the robust regression and the sampled bootstrap residuals $\{\widehat{\varepsilon}_t^*\}_{t=1}^T$ through $y_t^* = \widehat{\beta}_c^{(1)} y_{t-1}^* + \widehat{\varepsilon}_t^*$, $t = 1, 2, \ldots, T$, where we set $y_0^* = y_0$. For the non-parametric time series block bootstraps we follow Bühlmann & Künsch (1999) and use block resampling with a fixed block size $l$ where $l = n^{1/3}$, which is rounded up to the nearest integer and where $n$ is the sample size. Subsequently we use scheme A or B in step 2 of Algorithm B.1 drawing blocks (instead of individual observations) of length $l$ from either the raw or outlier-removed data.

Performance of the different bootstrap schemes are evaluated in a range of simulations, with results reported in section 3 as well as in the Appendix section C below.

# C Additional Simulation Results

We report several figures for additional simulation results in Appendix C.1. Then, Appendix C.2 shows all simulation results presented in tables.

## C.1 Additional Simulation Figures

Here we report additional simulation results of the asymptotic test in the presence of outliers. Figure C.1 shows the simulation results under a range of alternatives for 15% outlier contamination. Figure C.2 shows the simulation results when testing for distortion of a single coefficient under a range of alternatives.

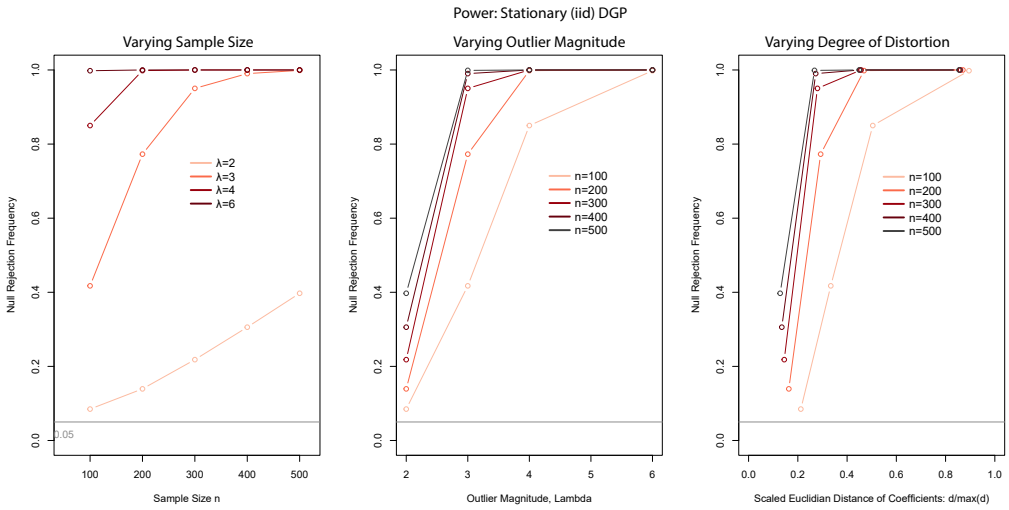


Figure C.1: Simulation performance of the asymptotic test under the alternatives contaminated with vertical outliers for varying sample sizes and outlier magnitudes when the DGP is iid ($\rho = 0$) including five regressors and 15% of the sample is outlier-contaminated.

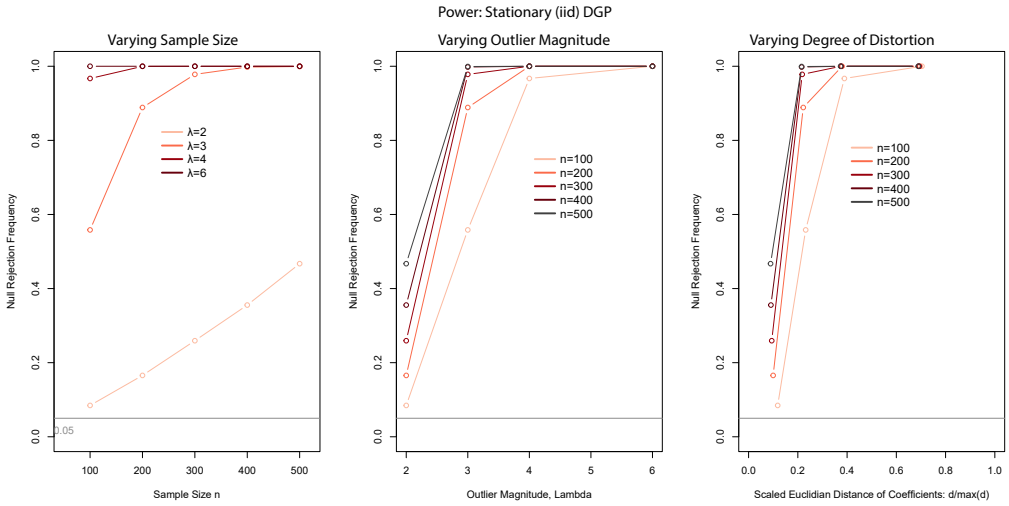


Figure C.2: Simulation performance of the asymptotic test under the alternatives contaminated with vertical outliers for varying sample sizes and outlier magnitudes when the DGP is iid ($\rho = 0$) including a single regressor and 10% of the sample is outlier-contaminated.

Figure C.3 shows the parametric bootstrap simulation results under the null of no outliers. The parametric residual bootstrap does not appear to provide any improvement relative to the asymptotic test.

Figure C.4 shows the non-parametric bootstrap simulation results under the alternative when the DGP is contaminated with bad leverage points. The results in the main text show that the L2 nonparametric bootstrap using the cleaned data or the variance bootstrap using the raw data performs well under the null of no outliers. In the presence of bad leverage points, particularly the non-parametric variance bootstrap performs well with power increasing with sample size. The L2 bootstrap does not exhibit high power when the DGP is contaminated with bad leverage points, further supporting the preference for the variance bootstrap.

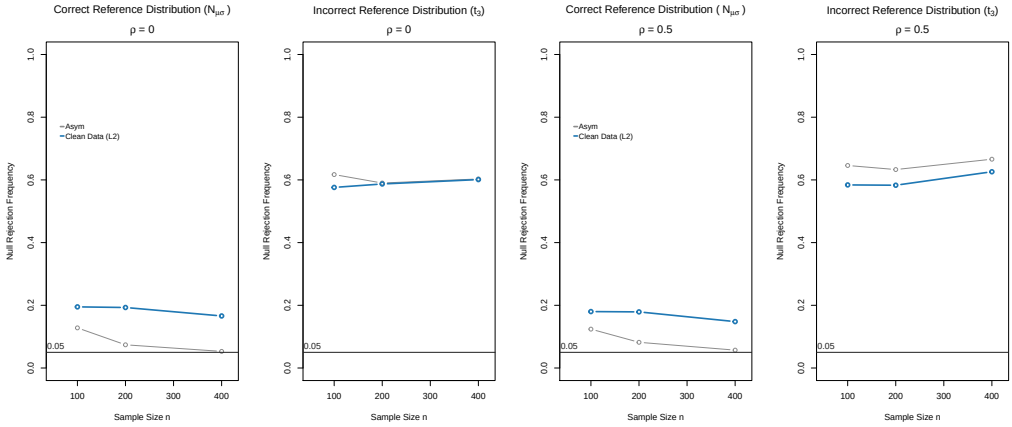


Figure C.3: Simulation performance of the parametric bootstrap test under the null of no outliers.

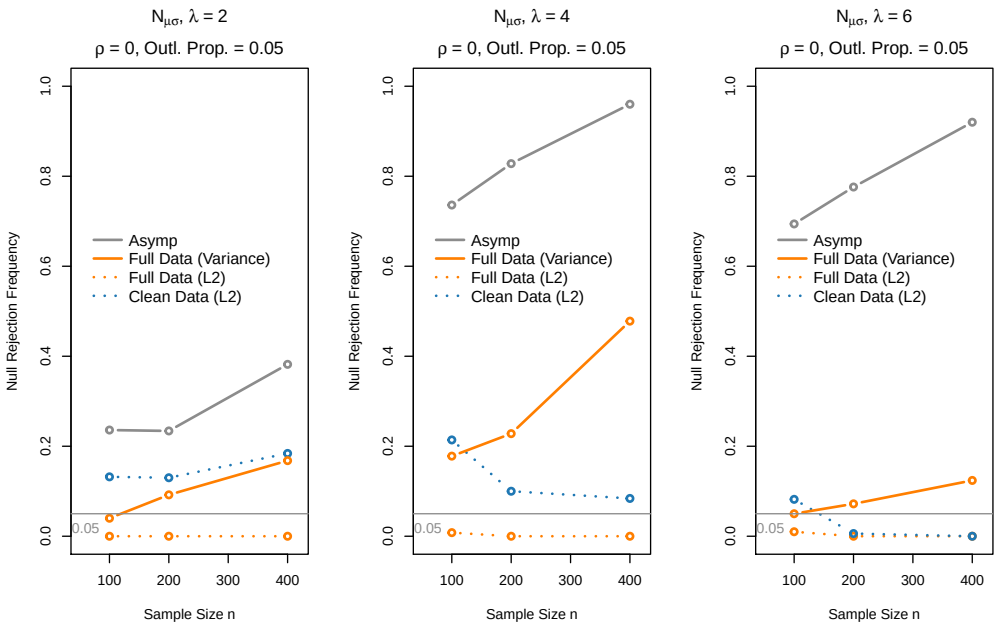


Figure C.4: Simulation performance of the non-parametric bootstrap tests under alternatives contaminated with bad leverage points for varying sample sizes.

## C.2 Simulation Results Presented in Tables

Here we present the full set of simulation results in Tables (Figures in the main text and Appendix visualise the same information). Notably, we also report the results for a bootstrap simulation when the assumed error distribution does not match the asymmetric (log-normal) error distribution in the DGP in Table C.14. Due to the computational complexity of the bootstrap we limit this case to a single specification (with large sample size) – the bootstrap does not perform well in the presence of asymmetric errors in this single setup, thus expanding the simulation specifications was not necessary for this case.

Table C.1: Simulation performance of the asymptotic test under the null of no distortion for varying sample sizes, cut-offs, and test levels when the DGP is iid ($\rho = 0$) with a Normal reference distribution and Monte Carlo replications $m = 10000$.

| | | $\gamma_c$ | |
|---|---|---|---|
| n | # Regressors | 0.01 | 0.05 |
| **Level 0.01** | | | |
| 100 | 5 | 0.107 | 0.029 |
| 200 | 5 | 0.056 | 0.018 |
| 300 | 5 | 0.044 | 0.017 |
| 400 | 5 | 0.034 | 0.014 |
| 500 | 5 | 0.031 | 0.015 |
| **Level 0.05** | | | |
| 100 | 5 | 0.194 | 0.088 |
| 200 | 5 | 0.123 | 0.069 |
| 300 | 5 | 0.103 | 0.063 |
| 400 | 5 | 0.092 | 0.060 |
| 500 | 5 | 0.084 | 0.060 |

Table C.2: Simulation performance of the asymptotic test under the null of no distortion for varying sample sizes and number of regressors when the DGP is iid ($\rho = 0$) with a Normal reference distribution and Monte Carlo replications $m = 10000$.

| | | | # Regressors | | |
|---|---|---|---|---|---|
| n | Level | $\gamma_c$ | 1 | 5 | 10 |
| 100 | 0.05 | 0.05 | 0.060 | 0.088 | 0.107 |
| 200 | 0.05 | 0.05 | 0.057 | 0.069 | 0.081 |
| 300 | 0.05 | 0.05 | 0.051 | 0.063 | 0.069 |
| 400 | 0.05 | 0.05 | 0.052 | 0.060 | 0.064 |
| 500 | 0.05 | 0.05 | 0.049 | 0.060 | 0.061 |

Table C.3: Simulation performance of the asymptotic test under the null of no distortion for varying sample sizes and degrees of persistence ($\rho = 0, 0.5$) in the DGPs with a Normal reference distribution and Monte Carlo replications $m = 10000$.

| | | | | $\rho$ | |
|---|---|---|---|---|---|
| n | Level | $\gamma_c$ | # Regressors | 0 | 0.5 |
| 100 | 0.05 | 0.05 | 5 | 0.088 | 0.090 |
| 200 | 0.05 | 0.05 | 5 | 0.069 | 0.066 |
| 300 | 0.05 | 0.05 | 5 | 0.063 | 0.065 |
| 400 | 0.05 | 0.05 | 5 | 0.060 | 0.058 |
| 500 | 0.05 | 0.05 | 5 | 0.060 | 0.057 |

Table C.4: Simulation performance of the asymptotic test under the alternatives contaminated with vertical outliers for varying sample sizes and outlier magnitudes when the DGP is iid ($\rho = 0$) including a single regressor and 10% of the sample is outlier-contaminated with a Normal reference distribution and Monte Carlo replications $m = 10000$.

| | | | | | | Rejection Frequency for $\lambda$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| n | $\rho$ | Outl. Prop. | Level | $\gamma_c$ | # Regressors | 2 | 3 | 4 | 6 |
| 100 | 0 | 0.1 | 0.01 | 0.05 | 1 | 0.085 | 0.558 | 0.967 | 1 |
| 200 | 0 | 0.1 | 0.01 | 0.05 | 1 | 0.166 | 0.889 | 1.000 | 1 |
| 300 | 0 | 0.1 | 0.01 | 0.05 | 1 | 0.259 | 0.978 | 1.000 | 1 |
| 400 | 0 | 0.1 | 0.01 | 0.05 | 1 | 0.356 | 0.998 | 1.000 | 1 |
| 500 | 0 | 0.1 | 0.01 | 0.05 | 1 | 0.467 | 0.999 | 1.000 | 1 |

Table C.5: Simulation performance of the asymptotic test under the alternatives contaminated with vertical outliers for varying sample sizes and outlier magnitudes when the DGP is iid ($\rho = 0$) including a single regressor and 15% of the sample is outlier-contaminated with a Normal reference distribution and Monte Carlo replications $m = 10000$.

| | | | | | | Rejection Frequency for $\lambda$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| n | $\rho$ | Outl. Prop. | Level | $\gamma_c$ | # Regressors | 2 | 3 | 4 | 6 |
| 100 | 0 | 0.15 | 0.01 | 0.05 | 1 | 0.100 | 0.585 | 0.963 | 1 |
| 200 | 0 | 0.15 | 0.01 | 0.05 | 1 | 0.215 | 0.915 | 1.000 | 1 |
| 300 | 0 | 0.15 | 0.01 | 0.05 | 1 | 0.338 | 0.988 | 1.000 | 1 |
| 400 | 0 | 0.15 | 0.01 | 0.05 | 1 | 0.471 | 0.999 | 1.000 | 1 |
| 500 | 0 | 0.15 | 0.01 | 0.05 | 1 | 0.583 | 1.000 | 1.000 | 1 |

Table C.6: Simulation performance of the asymptotic test under the alternatives contaminated with vertical outliers for varying sample sizes and outlier magnitudes when the DGP is iid ($\rho = 0$) including five regressors and 10% of the sample is outlier-contaminated with a Normal reference distribution and Monte Carlo replications $m = 10000$.

| | | | | | | Rejection Frequency for $\lambda$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| n | $\rho$ | Outl. Prop. | Level | $\gamma_c$ | # Regressors | 2 | 3 | 4 | 6 |
| 100 | 0 | 0.1 | 0.01 | 0.05 | 5 | 0.084 | 0.449 | 0.907 | 1 |
| 200 | 0 | 0.1 | 0.01 | 0.05 | 5 | 0.119 | 0.767 | 0.999 | 1 |
| 300 | 0 | 0.1 | 0.01 | 0.05 | 5 | 0.164 | 0.928 | 1.000 | 1 |
| 400 | 0 | 0.1 | 0.01 | 0.05 | 5 | 0.243 | 0.986 | 1.000 | 1 |
| 500 | 0 | 0.1 | 0.01 | 0.05 | 5 | 0.307 | 0.998 | 1.000 | 1 |

Table C.7: Simulation performance of the asymptotic test under the alternatives contaminated with vertical outliers for varying sample sizes and outlier magnitudes when the DGP is iid ($\rho = 0$) including five regressors and 15% of the sample is outlier-contaminated with a Normal reference distribution and Monte Carlo replications $m = 10000$.

| | | | | | | Rejection Frequency for $\lambda$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| n | $\rho$ | Outl. Prop. | Level | $\gamma_c$ | # Regressors | 2 | 3 | 4 | 6 |
| 100 | 0 | 0.15 | 0.01 | 0.05 | 5 | 0.085 | 0.417 | 0.850 | 0.998 |
| 200 | 0 | 0.15 | 0.01 | 0.05 | 5 | 0.139 | 0.773 | 0.998 | 1.000 |
| 300 | 0 | 0.15 | 0.01 | 0.05 | 5 | 0.218 | 0.950 | 1.000 | 1.000 |
| 400 | 0 | 0.15 | 0.01 | 0.05 | 5 | 0.306 | 0.990 | 1.000 | 1.000 |
| 500 | 0 | 0.15 | 0.01 | 0.05 | 5 | 0.397 | 0.999 | 1.000 | 1.000 |

Table C.8: Simulation performance of the asymptotic test under the alternatives contaminated with vertical outliers for varying sample sizes and outlier magnitudes when the DGP contains a stationary autoregressive process ($\rho = 0.5$) and five regressors and 10% of the sample is outlier-contaminated with a Normal reference distribution and Monte Carlo replications $m = 10000$.

| | | | | | | Rejection Frequency for $\lambda$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| n | $\rho$ | Outl. Prop. | Level | $\gamma_c$ | # Regressors | 2 | 3 | 4 | 6 |
| 100 | 0.5 | 0.1 | 0.01 | 0.05 | 5 | 0.079 | 0.423 | 0.898 | 1 |
| 200 | 0.5 | 0.1 | 0.01 | 0.05 | 5 | 0.117 | 0.751 | 0.998 | 1 |
| 300 | 0.5 | 0.1 | 0.01 | 0.05 | 5 | 0.160 | 0.923 | 1.000 | 1 |
| 400 | 0.5 | 0.1 | 0.01 | 0.05 | 5 | 0.225 | 0.981 | 1.000 | 1 |
| 500 | 0.5 | 0.1 | 0.01 | 0.05 | 5 | 0.303 | 0.997 | 1.000 | 1 |

Table C.9: Simulation performance of the asymptotic test under the alternatives contaminated with vertical outliers for varying sample sizes and outlier magnitudes when the DGP contains a stationary autoregressive process ($\rho = 0.5$) and five regressors and 15% of the sample is outlier-contaminated with a Normal reference distribution and Monte Carlo replications $m = 10000$.

| | | | | | | Rejection Frequency for $\lambda$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| n | $\rho$ | Outl. Prop. | Level | $\gamma_c$ | # Regressors | 2 | 3 | 4 | 6 |
| 100 | 0.5 | 0.15 | 0.01 | 0.05 | 5 | 0.089 | 0.392 | 0.828 | 0.995 |
| 200 | 0.5 | 0.15 | 0.01 | 0.05 | 5 | 0.129 | 0.753 | 0.995 | 1.000 |
| 300 | 0.5 | 0.15 | 0.01 | 0.05 | 5 | 0.207 | 0.932 | 1.000 | 1.000 |
| 400 | 0.5 | 0.15 | 0.01 | 0.05 | 5 | 0.283 | 0.985 | 1.000 | 1.000 |
| 500 | 0.5 | 0.15 | 0.01 | 0.05 | 5 | 0.377 | 0.999 | 1.000 | 1.000 |

Table C.10: Simulation performance of the asymptotic test under the alternatives contaminated with vertical outliers for varying sample sizes and outlier magnitudes when the DGP contains a stochastically trensding process ($\rho = 1$) and five regressors and 10% of the sample is outlier-contaminated with a Normal reference distribution and Monte Carlo replications $m = 10000$.

| | | | | | | Rejection Frequency for $\lambda$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| n | $\rho$ | Outl. Prop. | Level | $\gamma_c$ | # Regressors | 2 | 3 | 4 | 6 |
| 100 | 1 | 0.1 | 0.01 | 0.05 | 5 | 0.084 | 0.420 | 0.887 | 1 |
| 200 | 1 | 0.1 | 0.01 | 0.05 | 5 | 0.117 | 0.742 | 0.998 | 1 |
| 300 | 1 | 0.1 | 0.01 | 0.05 | 5 | 0.162 | 0.920 | 1.000 | 1 |
| 400 | 1 | 0.1 | 0.01 | 0.05 | 5 | 0.227 | 0.981 | 1.000 | 1 |
| 500 | 1 | 0.1 | 0.01 | 0.05 | 5 | 0.289 | 0.996 | 1.000 | 1 |

Table C.11: Simulation performance of the asymptotic test under the alternatives contaminated with vertical outliers for varying sample sizes and outlier magnitudes when the DGP contains a stochastically trensding process ($\rho = 1$) and five regressors and 15% of the sample is outlier-contaminated with a Normal reference distribution and Monte Carlo replications $m = 10000$.

| | | | | | | Rejection Frequency for $\lambda$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| n | $\rho$ | Outl. Prop. | Level | $\gamma_c$ | # Regressors | 2 | 3 | 4 | 6 |
| 100 | 1 | 0.15 | 0.01 | 0.05 | 5 | 0.090 | 0.394 | 0.817 | 0.995 |
| 200 | 1 | 0.15 | 0.01 | 0.05 | 5 | 0.135 | 0.754 | 0.996 | 1.000 |
| 300 | 1 | 0.15 | 0.01 | 0.05 | 5 | 0.202 | 0.931 | 1.000 | 1.000 |
| 400 | 1 | 0.15 | 0.01 | 0.05 | 5 | 0.281 | 0.989 | 1.000 | 1.000 |
| 500 | 1 | 0.15 | 0.01 | 0.05 | 5 | 0.382 | 0.999 | 1.000 | 1.000 |

Table C.12: Simulation performance of the asymptotic test under the alternative contaminated with bad leverage points for varying sample sizes and outlier magnitudes for a Normal Distribution when the dependent variable is a cross-sectional process ($\rho = 0$) with a single regressor and 5% of the sample is outlier-contaminated. Monte Carlo replications is $m = 10000$.

| | | | | | | Rejection Frequency for $\lambda$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| n | $\rho$ | Outl. Prop. | Level | $\gamma_c$ | # Regressors | 2 | 3 | 4 | 6 |
| 100 | 0 | 0.05 | 0.01 | 0.05 | 1 | 0.118 | 0.392 | 0.578 | 0.604 |
| 200 | 0 | 0.05 | 0.01 | 0.05 | 1 | 0.138 | 0.516 | 0.706 | 0.667 |
| 300 | 0 | 0.05 | 0.01 | 0.05 | 1 | 0.181 | 0.646 | 0.836 | 0.793 |
| 400 | 0 | 0.05 | 0.01 | 0.05 | 1 | 0.226 | 0.750 | 0.907 | 0.867 |
| 500 | 0 | 0.05 | 0.01 | 0.05 | 1 | 0.279 | 0.833 | 0.952 | 0.924 |

Table C.13: Simulation performance of the asymptotic test under the alternative contaminated with bad leverage points for varying sample sizes and outlier magnitudes for a Normal Distribution when the dependent variable is a stationary autoregressive process ($\rho = 0.5$) with a single regressor and 5% of the sample is outlier-contaminated. Monte Carlo replications is $m = 10000$.

| | | | | | | Rejection Frequency for $\lambda$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| n | $\rho$ | Outl. Prop. | Level | $\gamma_c$ | # Regressors | 2 | 3 | 4 | 6 |
| 100 | 0.5 | 0.05 | 0.01 | 0.05 | 1 | 0.115 | 0.377 | 0.566 | 0.586 |
| 200 | 0.5 | 0.05 | 0.01 | 0.05 | 1 | 0.124 | 0.482 | 0.677 | 0.641 |
| 300 | 0.5 | 0.05 | 0.01 | 0.05 | 1 | 0.162 | 0.615 | 0.814 | 0.756 |
| 400 | 0.5 | 0.05 | 0.01 | 0.05 | 1 | 0.202 | 0.713 | 0.894 | 0.850 |
| 500 | 0.5 | 0.05 | 0.01 | 0.05 | 1 | 0.245 | 0.797 | 0.941 | 0.908 |

Table C.14: Simulation performance of the bootstrap tests under the null of no distortion when the reference distribution does and does not match the error distribution in the DGP using non-parametric bootstrapping, including the asymmetric log-normal distribution.

| | | | | | | Rejection Frequency | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Asymp. | Raw | | | Clean | | |
| n | # Regressors | $\rho$ | Level | $\gamma_c$ | Parametric | | $L_1$ | $L_2$ | Variance | $L_1$ | $L_2$ | Variance |
| **Normal Distribution** | | | | | | | | | | | | |
| 50 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.433 | 0.001 | 0.002 | 0.002 | 0.029 | 0.026 | 0.175 |
| 100 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.121 | 0.001 | 0.001 | 0.009 | 0.030 | 0.030 | 0.409 |
| 200 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.053 | 0.000 | 0.000 | 0.012 | 0.026 | 0.025 | 0.413 |
| 400 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.063 | 0.000 | 0.000 | 0.028 | 0.033 | 0.036 | 0.334 |
| 50 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.459 | 0.002 | 0.002 | 0.004 | 0.004 | 0.003 | 0.009 |
| 100 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.142 | 0.001 | 0.001 | 0.004 | 0.000 | 0.000 | 0.012 |
| 200 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.071 | 0.000 | 0.000 | 0.023 | 0.000 | 0.000 | 0.029 |
| 400 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.054 | 0.000 | 0.000 | 0.033 | 0.000 | 0.000 | 0.041 |
| $t_3$ **Distribution** | | | | | | | | | | | | |
| 50 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.725 | 0.000 | 0.000 | 0.001 | 0.139 | 0.143 | 0.451 |
| 100 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.614 | 0.000 | 0.000 | 0.011 | 0.268 | 0.279 | 0.721 |
| 200 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.608 | 0.000 | 0.000 | 0.055 | 0.271 | 0.281 | 0.731 |
| 400 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.650 | 0.000 | 0.000 | 0.070 | 0.293 | 0.318 | 0.731 |
| 50 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.758 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.014 |
| 100 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.605 | 0.000 | 0.000 | 0.023 | 0.002 | 0.003 | 0.056 |
| 200 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.614 | 0.000 | 0.000 | 0.084 | 0.001 | 0.001 | 0.108 |
| 400 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.667 | 0.000 | 0.000 | 0.091 | 0.001 | 0.001 | 0.104 |
| **Log-normal Distribution** | | | | | | | | | | | | |
| 400 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 0.992 | NA | NA | NA |

Table C.15: Simulation performance of the bootstrap tests under the null of no distortion when the reference distribution does and does not match the error distribution in the DGP using parametric bootstrapping.

| | | | | | | Rejection Frequency | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Asymp. | Clean | | |
| n | # Regressors | $\rho$ | Level | $\gamma_c$ | Parametric | | $L_1$ | $L_2$ | Variance |
| **Normal Distribution** | | | | | | | | | |
| 50 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 0.397 | 0.178 | 0.184 | 0.344 |
| 100 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 0.128 | 0.186 | 0.195 | 0.315 |
| 200 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 0.074 | 0.179 | 0.193 | 0.273 |
| 400 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 0.053 | 0.147 | 0.166 | 0.234 |
| 50 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 0.441 | 0.207 | 0.222 | 0.395 |
| 100 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 0.124 | 0.182 | 0.180 | 0.317 |
| 200 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 0.082 | 0.174 | 0.179 | 0.284 |
| 400 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 0.057 | 0.149 | 0.148 | 0.223 |
| $t_3$ **Distribution** | | | | | | | | | |
| 50 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 0.722 | 0.447 | 0.452 | 0.624 |
| 100 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 0.617 | 0.556 | 0.576 | 0.702 |
| 200 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 0.590 | 0.563 | 0.587 | 0.688 |
| 400 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 0.603 | 0.577 | 0.601 | 0.685 |
| 50 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 0.730 | 0.467 | 0.488 | 0.632 |
| 100 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 0.646 | 0.551 | 0.584 | 0.704 |
| 200 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 0.633 | 0.565 | 0.583 | 0.691 |
| 400 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 0.666 | 0.589 | 0.626 | 0.716 |

Table C.16: Simulation performance under the alternatives for varying sample sizes when using the bootstrap implementations of our test for vertical outliers (outlier magnitude of 2 SD of the error term and 10% proprtion of outliers) using Non-Parametric Bootstraps.

| | | | | | | | Rejection Frequency | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Asymp. | Raw | | | Clean | | |
| $\lambda$ | n | # Regressors | $\rho$ | Level | $\gamma_c$ | Parametric | | $L_1$ | $L_2$ | Variance | $L_1$ | $L_2$ | Variance |
| **Normal Distribution** | | | | | | | | | | | | | |
| 2 | 50 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.320 | 0.000 | 0.000 | 0.000 | 0.006 | 0.008 | 0.124 |
| 2 | 100 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.102 | 0.000 | 0.000 | 0.002 | 0.022 | 0.022 | 0.368 |
| 2 | 200 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.114 | 0.000 | 0.000 | 0.012 | 0.024 | 0.026 | 0.452 |
| 2 | 400 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.234 | 0.000 | 0.000 | 0.098 | 0.032 | 0.056 | 0.560 |
| 2 | 50 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.488 | 0.000 | 0.000 | 0.002 | 0.024 | 0.028 | 0.194 |
| 2 | 100 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.214 | 0.000 | 0.000 | 0.018 | 0.064 | 0.068 | 0.494 |
| 2 | 200 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.266 | 0.000 | 0.000 | 0.086 | 0.084 | 0.110 | 0.614 |
| 2 | 400 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.462 | 0.000 | 0.000 | 0.230 | 0.146 | 0.208 | 0.732 |
| 2 | 50 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.436 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 |
| 2 | 100 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.124 | 0.000 | 0.000 | 0.002 | 0.000 | 0.000 | 0.014 |
| 2 | 200 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.132 | 0.000 | 0.000 | 0.024 | 0.000 | 0.000 | 0.048 |
| 2 | 400 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.238 | 0.000 | 0.000 | 0.122 | 0.000 | 0.000 | 0.108 |
| 2 | 50 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.564 | 0.000 | 0.000 | 0.002 | 0.002 | 0.002 | 0.008 |
| 2 | 100 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.262 | 0.000 | 0.000 | 0.004 | 0.000 | 0.000 | 0.034 |
| 2 | 200 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.284 | 0.000 | 0.000 | 0.078 | 0.000 | 0.000 | 0.140 |
| 2 | 400 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.430 | 0.000 | 0.000 | 0.276 | 0.000 | 0.000 | 0.318 |
| $t_3$ **Distribution** | | | | | | | | | | | | | |
| 2 | 50 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.554 | 0.000 | 0.000 | 0.000 | | | |
| 2 | 100 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.390 | 0.000 | 0.000 | 0.002 | | | |
| 2 | 200 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.352 | 0.000 | 0.000 | 0.012 | | | |
| 2 | 400 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.330 | 0.000 | 0.000 | 0.010 | | | |
| 2 | 50 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.678 | 0.000 | 0.000 | 0.002 | | | |
| 2 | 100 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.546 | 0.000 | 0.000 | 0.016 | | | |
| 2 | 200 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.528 | 0.000 | 0.000 | 0.048 | | | |
| 2 | 400 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.516 | 0.000 | 0.000 | 0.046 | | | |
| 2 | 50 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.562 | 0.000 | 0.000 | 0.000 | | | |
| 2 | 100 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.418 | 0.000 | 0.000 | 0.000 | | | |
| 2 | 200 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.368 | 0.000 | 0.000 | 0.010 | | | |
| 2 | 400 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.434 | 0.000 | 0.000 | 0.016 | | | |
| 2 | 50 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.694 | 0.004 | 0.000 | 0.002 | | | |
| 2 | 100 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.584 | 0.000 | 0.000 | 0.012 | | | |
| 2 | 200 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.540 | 0.000 | 0.000 | 0.066 | | | |
| 2 | 400 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.618 | 0.000 | 0.000 | 0.068 | | | |

Table C.17: Simulation performance under the alternatives for varying sample sizes when using the bootstrap implementations of our test for vertical outliers (outlier magnitude of 4 SD of the error term and 10% proportion of outliers) using Non-Parametric Bootstraps.

| | | | | | | | Rejection Frequency | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Asymp. | Raw | | | Clean | | |
| $\lambda$ | n | # Regressors | $\rho$ | Level | $\gamma_c$ | Parametric | | $L_1$ | $L_2$ | Variance | $L_1$ | $L_2$ | Variance |
| **Normal Distribution** | | | | | | | | | | | | | |
| 4 | 50 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.840 | 0.000 | 0.000 | 0.002 | 0.040 | 0.036 | 0.524 |
| 4 | 100 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.924 | 0.000 | 0.000 | 0.178 | 0.332 | 0.468 | 0.948 |
| 4 | 200 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 0.938 | 0.478 | 0.738 | 1.000 |
| 4 | 400 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 1.000 | 0.782 | 0.992 | 1.000 |
| 4 | 50 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.912 | 0.002 | 0.002 | 0.002 | 0.194 | 0.196 | 0.686 |
| 4 | 100 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.972 | 0.000 | 0.000 | 0.452 | 0.574 | 0.690 | 0.984 |
| 4 | 200 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 0.992 | 0.728 | 0.904 | 1.000 |
| 4 | 400 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 1.000 | 0.932 | 0.998 | 1.000 |
| 4 | 50 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.868 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 |
| 4 | 100 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.930 | 0.000 | 0.000 | 0.114 | 0.000 | 0.000 | 0.204 |
| 4 | 200 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.998 | 0.000 | 0.000 | 0.902 | 0.000 | 0.000 | 0.900 |
| 4 | 400 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 1.000 |
| 4 | 50 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.926 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.010 |
| 4 | 100 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.972 | 0.000 | 0.000 | 0.372 | 0.000 | 0.000 | 0.434 |
| 4 | 200 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 0.974 | 0.000 | 0.000 | 0.990 |
| 4 | 400 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 1.000 |
| $t_3$ **Distribution** | | | | | | | | | | | | | |
| 4 | 50 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.722 | 0.000 | 0.000 | 0.000 | | | |
| 4 | 100 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.564 | 0.000 | 0.000 | 0.026 | | | |
| 4 | 200 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.636 | 0.000 | 0.000 | 0.140 | | | |
| 4 | 400 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.802 | 0.000 | 0.000 | 0.368 | | | |
| 4 | 50 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.812 | 0.004 | 0.002 | 0.006 | | | |
| 4 | 100 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.692 | 0.000 | 0.000 | 0.090 | | | |
| 4 | 200 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.794 | 0.000 | 0.000 | 0.314 | | | |
| 4 | 400 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.890 | 0.000 | 0.000 | 0.558 | | | |
| 4 | 50 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.708 | 0.000 | 0.000 | 0.000 | | | |
| 4 | 100 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.648 | 0.000 | 0.000 | 0.020 | | | |
| 4 | 200 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.674 | 0.000 | 0.000 | 0.182 | | | |
| 4 | 400 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.808 | 0.000 | 0.000 | 0.380 | | | |
| 4 | 50 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.802 | 0.000 | 0.000 | 0.000 | | | |
| 4 | 100 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.804 | 0.000 | 0.000 | 0.100 | | | |
| 4 | 200 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.798 | 0.000 | 0.000 | 0.344 | | | |
| 4 | 400 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.904 | 0.000 | 0.000 | 0.606 | | | |

Table C.18: Simulation performance under the alternatives for varying sample sizes when using the bootstrap implementations of our test for vertical outliers (outlier magnitude of 6 SD of the error term and 10% porpotion of outliers) using Non-Parametric Bootstraps.

| | | | | | | | Rejection Frequency | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Asymp. | Raw | | | Clean | | |
| $\lambda$ | n | # Regressors | $\rho$ | Level | $\gamma_c$ | Parametric | | $L_1$ | $L_2$ | Variance | $L_1$ | $L_2$ | Variance |
| **Normal Distribution** | | | | | | | | | | | | | |
| 6 | 50 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.998 | 0.000 | 0.000 | 0.000 | 0.372 | 0.392 | 0.938 |
| 6 | 100 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 0.506 | 0.790 | 0.880 | 1.000 |
| 6 | 200 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 1.000 | 0.964 | 0.998 | 1.000 |
| 6 | 400 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 6 | 50 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.998 | 0.000 | 0.000 | 0.004 | 0.694 | 0.766 | 0.976 |
| 6 | 100 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 0.886 | 0.926 | 0.962 | 1.000 |
| 6 | 200 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 1.000 | 0.988 | 1.000 | 1.000 |
| 6 | 400 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 6 | 50 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.990 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.010 |
| 6 | 100 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 0.504 | 0.000 | 0.000 | 0.562 |
| 6 | 200 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 1.000 |
| 6 | 400 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 1.000 |
| 6 | 50 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.994 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.036 |
| 6 | 100 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 0.844 | 0.000 | 0.000 | 0.844 |
| 6 | 200 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 1.000 |
| 6 | 400 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 1.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 1.000 |
| $t_3$ **Distribution** | | | | | | | | | | | | | |
| 6 | 50 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.930 | 0.000 | 0.000 | 0.000 | | | |
| 6 | 100 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.926 | 0.000 | 0.000 | 0.198 | | | |
| 6 | 200 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.978 | 0.000 | 0.000 | 0.756 | | | |
| 6 | 400 | 5 | 0.0 | 0.01 | 0.05 | Non-Parametric | 0.994 | 0.000 | 0.000 | 0.934 | | | |
| 6 | 50 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.964 | 0.000 | 0.000 | 0.006 | | | |
| 6 | 100 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.970 | 0.000 | 0.000 | 0.480 | | | |
| 6 | 200 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.988 | 0.000 | 0.000 | 0.886 | | | |
| 6 | 400 | 5 | 0.0 | 0.05 | 0.05 | Non-Parametric | 0.998 | 0.000 | 0.000 | 0.954 | | | |
| 6 | 50 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.916 | 0.000 | 0.000 | 0.000 | | | |
| 6 | 100 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.956 | 0.000 | 0.000 | 0.170 | | | |
| 6 | 200 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.992 | 0.000 | 0.000 | 0.758 | | | |
| 6 | 400 | 5 | 0.5 | 0.01 | 0.05 | Non-Parametric | 0.996 | 0.000 | 0.000 | 0.934 | | | |
| 6 | 50 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.960 | 0.000 | 0.000 | 0.000 | | | |
| 6 | 100 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.984 | 0.000 | 0.000 | 0.464 | | | |
| 6 | 200 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.994 | 0.000 | 0.000 | 0.866 | | | |
| 6 | 400 | 5 | 0.5 | 0.05 | 0.05 | Non-Parametric | 0.998 | 0.000 | 0.000 | 0.968 | | | |

Table C.19: Simulation performance under the alternatives for varying sample sizes when using the bootstrap implementations of our test for a Bad Leverage Point (outlier magnitude of 2 SD of the error term and 5% proportion of outliers) using Non-Parametric Bootstraps.

| | | | | | | | Rejection Frequency | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Asymp. | Raw | | | Clean | | |
| $\lambda$ | n | # Regressors | $\rho$ | Level | $\gamma_c$ | Parametric | | $L_1$ | $L_2$ | Variance | $L_1$ | $L_2$ | Variance |
| **Normal Distribution** | | | | | | | | | | | | | |
| 2 | 50 | 1 | 0 | 0.01 | 0.05 | Non-Parametric | 0.212 | 0.000 | 0.000 | 0.004 | 0.038 | 0.040 | 0.190 |
| 2 | 100 | 1 | 0 | 0.01 | 0.05 | Non-Parametric | 0.106 | 0.000 | 0.000 | 0.008 | 0.040 | 0.046 | 0.190 |
| 2 | 200 | 1 | 0 | 0.01 | 0.05 | Non-Parametric | 0.148 | 0.000 | 0.000 | 0.014 | 0.062 | 0.056 | 0.176 |
| 2 | 400 | 1 | 0 | 0.01 | 0.05 | Non-Parametric | 0.202 | 0.000 | 0.000 | 0.048 | 0.068 | 0.074 | 0.206 |
| 2 | 50 | 1 | 0 | 0.05 | 0.05 | Non-Parametric | 0.328 | 0.000 | 0.000 | 0.028 | 0.104 | 0.116 | 0.280 |
| 2 | 100 | 1 | 0 | 0.05 | 0.05 | Non-Parametric | 0.234 | 0.000 | 0.000 | 0.040 | 0.140 | 0.132 | 0.304 |
| 2 | 200 | 1 | 0 | 0.05 | 0.05 | Non-Parametric | 0.258 | 0.000 | 0.000 | 0.092 | 0.128 | 0.130 | 0.320 |
| 2 | 400 | 1 | 0 | 0.05 | 0.05 | Non-Parametric | 0.364 | 0.000 | 0.000 | 0.168 | 0.184 | 0.184 | 0.358 |

Table C.20: Simulation performance under the alternatives for varying sample sizes when using the bootstrap implementations of our test for a Bad Leverage Point (outlier magnitude of 4 SD of the error term and 5% proportion of outliers) using Non-Parametric Bootstraps.

| | | | | | | | Rejection Frequency | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Asymp. | Raw | | | Clean | | |
| $\lambda$ | n | # Regressors | $\rho$ | Level | $\gamma_c$ | Parametric | | $L_1$ | $L_2$ | Variance | $L_1$ | $L_2$ | Variance |
| **Normal Distribution** | | | | | | | | | | | | | |
| 4 | 50 | 1 | 0 | 0.01 | 0.05 | Non-Parametric | 0.652 | 0.000 | 0.000 | 0.006 | 0.350 | 0.408 | 0.576 |
| 4 | 100 | 1 | 0 | 0.01 | 0.05 | Non-Parametric | 0.584 | 0.000 | 0.000 | 0.052 | 0.134 | 0.128 | 0.242 |
| 4 | 200 | 1 | 0 | 0.01 | 0.05 | Non-Parametric | 0.698 | 0.000 | 0.000 | 0.066 | 0.062 | 0.052 | 0.238 |
| 4 | 400 | 1 | 0 | 0.01 | 0.05 | Non-Parametric | 0.904 | 0.000 | 0.000 | 0.196 | 0.038 | 0.026 | 0.396 |
| 4 | 50 | 1 | 0 | 0.05 | 0.05 | Non-Parametric | 0.760 | 0.002 | 0.002 | 0.092 | 0.496 | 0.528 | 0.630 |
| 4 | 100 | 1 | 0 | 0.05 | 0.05 | Non-Parametric | 0.668 | 0.008 | 0.008 | 0.178 | 0.212 | 0.214 | 0.376 |
| 4 | 200 | 1 | 0 | 0.05 | 0.05 | Non-Parametric | 0.790 | 0.000 | 0.000 | 0.228 | 0.102 | 0.100 | 0.404 |
| 4 | 400 | 1 | 0 | 0.05 | 0.05 | Non-Parametric | 0.964 | 0.000 | 0.000 | 0.478 | 0.102 | 0.084 | 0.590 |

Table C.21: Simulation performance under the alternatives for varying sample sizes when using the bootstrap implementations of our test for a Bad Leverage Point (outlier magnitude of 6 SD of the error term and 5% proportion of outliers) using Non-Parametric Bootstraps.

| | | | | | | | Rejection Frequency | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Asymp. | Raw | | | Clean | | |
| $\lambda$ | n | # Regressors | $\rho$ | Level | $\gamma_c$ | Parametric | | $L_1$ | $L_2$ | Variance | $L_1$ | $L_2$ | Variance |
| **Normal Distribution** | | | | | | | | | | | | | |
| 6 | 50 | 1 | 0 | 0.01 | 0.05 | Non-Parametric | 0.764 | 0.000 | 0.000 | 0.000 | 0.564 | 0.612 | 0.660 |
| 6 | 100 | 1 | 0 | 0.01 | 0.05 | Non-Parametric | 0.614 | 0.000 | 0.000 | 0.014 | 0.070 | 0.072 | 0.128 |
| 6 | 200 | 1 | 0 | 0.01 | 0.05 | Non-Parametric | 0.672 | 0.000 | 0.000 | 0.022 | 0.002 | 0.002 | 0.070 |
| 6 | 400 | 1 | 0 | 0.01 | 0.05 | Non-Parametric | 0.874 | 0.000 | 0.000 | 0.042 | 0.000 | 0.000 | 0.080 |
| 6 | 50 | 1 | 0 | 0.05 | 0.05 | Non-Parametric | 0.840 | 0.002 | 0.000 | 0.092 | 0.628 | 0.648 | 0.680 |
| 6 | 100 | 1 | 0 | 0.05 | 0.05 | Non-Parametric | 0.732 | 0.010 | 0.010 | 0.050 | 0.092 | 0.082 | 0.194 |
| 6 | 200 | 1 | 0 | 0.05 | 0.05 | Non-Parametric | 0.768 | 0.000 | 0.000 | 0.072 | 0.006 | 0.006 | 0.194 |
| 6 | 400 | 1 | 0 | 0.05 | 0.05 | Non-Parametric | 0.928 | 0.000 | 0.000 | 0.124 | 0.000 | 0.000 | 0.174 |

Table C.22: Simulation performance under the alternatives for varying sample sizes when using the bootstrap implementations of our test for vertical outliers (outlier magnitude of 2 SD of the error term and 5% (or 10%) proportion of outliers) using Parametric Bootstraps.

| | | | | | | | Rejection Frequency | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Asymp. | Clean | | |
| $\lambda$ | n | # Regressors | $\rho$ | Level | $\gamma_c$ | Parametric | | $L_1$ | $L_2$ | Variance |
| **Normal Distribution** | | | | | | | | | | |
| 2 | 50 | 5 | 0.0 | 0.01 | 0.05 | Parametric | 0.360 | 0.112 | 0.128 | 0.314 |
| 2 | 100 | 5 | 0.0 | 0.01 | 0.05 | Parametric | 0.114 | 0.104 | 0.118 | 0.286 |
| 2 | 200 | 5 | 0.0 | 0.01 | 0.05 | Parametric | 0.112 | 0.156 | 0.186 | 0.362 |
| 2 | 400 | 5 | 0.0 | 0.01 | 0.05 | Parametric | 0.232 | 0.216 | 0.304 | 0.532 |
| 2 | 50 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 0.484 | 0.258 | 0.254 | 0.418 |
| 2 | 100 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 0.230 | 0.260 | 0.268 | 0.446 |
| 2 | 200 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 0.270 | 0.350 | 0.406 | 0.560 |
| 2 | 400 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 0.470 | 0.460 | 0.538 | 0.718 |
| 2 | 50 | 5 | 0.5 | 0.01 | 0.05 | Parametric | 0.396 | 0.110 | 0.116 | 0.330 |
| 2 | 100 | 5 | 0.5 | 0.01 | 0.05 | Parametric | 0.160 | 0.128 | 0.130 | 0.316 |
| 2 | 200 | 5 | 0.5 | 0.01 | 0.05 | Parametric | 0.112 | 0.138 | 0.142 | 0.314 |
| 2 | 400 | 5 | 0.5 | 0.01 | 0.05 | Parametric | 0.220 | 0.198 | 0.260 | 0.484 |
| 2 | 50 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 0.526 | 0.272 | 0.270 | 0.460 |
| 2 | 100 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 0.280 | 0.318 | 0.298 | 0.448 |
| 2 | 200 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 0.256 | 0.310 | 0.322 | 0.500 |
| 2 | 400 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 0.410 | 0.410 | 0.470 | 0.664 |

Table C.23: Simulation performance under the alternatives for varying sample sizes when using the bootstrap implementations of our test for vertical outliers (outlier magnitude of 4 SD of the error term and 5% (or 10%) proportion of outliers) using Parametric Bootstraps.

| | | | | | | | Rejection Frequency | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Asymp. | Clean | | |
| $\lambda$ | n | # Regressors | $\rho$ | Level | $\gamma_c$ | Parametric | | $L_1$ | $L_2$ | Variance |
| **Normal Distribution** | | | | | | | | | | |
| 4 | 50 | 5 | 0.0 | 0.01 | 0.05 | Parametric | 0.864 | 0.402 | 0.454 | 0.794 |
| 4 | 100 | 5 | 0.0 | 0.01 | 0.05 | Parametric | 0.954 | 0.710 | 0.832 | 0.970 |
| 4 | 200 | 5 | 0.0 | 0.01 | 0.05 | Parametric | 0.996 | 0.908 | 0.980 | 1.000 |
| 4 | 400 | 5 | 0.0 | 0.01 | 0.05 | Parametric | 1.000 | 0.980 | 0.998 | 1.000 |
| 4 | 50 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 0.920 | 0.672 | 0.706 | 0.868 |
| 4 | 100 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 0.978 | 0.892 | 0.942 | 0.990 |
| 4 | 200 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 1.000 | 0.974 | 0.996 | 1.000 |
| 4 | 400 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 1.000 | 1.000 | 1.000 | 1.000 |
| 4 | 50 | 5 | 0.5 | 0.01 | 0.05 | Parametric | 0.844 | 0.390 | 0.434 | 0.762 |
| 4 | 100 | 5 | 0.5 | 0.01 | 0.05 | Parametric | 0.944 | 0.700 | 0.810 | 0.972 |
| 4 | 200 | 5 | 0.5 | 0.01 | 0.05 | Parametric | 0.996 | 0.878 | 0.972 | 0.998 |
| 4 | 400 | 5 | 0.5 | 0.01 | 0.05 | Parametric | 1.000 | 0.990 | 1.000 | 1.000 |
| 4 | 50 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 0.908 | 0.670 | 0.722 | 0.846 |
| 4 | 100 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 0.984 | 0.892 | 0.952 | 0.990 |
| 4 | 200 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 1.000 | 0.960 | 0.996 | 1.000 |
| 4 | 400 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 1.000 | 0.998 | 1.000 | 1.000 |

Table C.24: Simulation performance under the alternatives for varying sample sizes when using the bootstrap implementations of our test for vertical outliers (outlier magnitude of 6 SD of the error term and 5% (or 10%) proportion of outliers) using Parametric Bootstraps.

| | | | | | | | Rejection Frequency | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Asymp. | Clean | | |
| $\lambda$ | n | # Regressors | $\rho$ | Level | $\gamma_c$ | Parametric | | $L_1$ | $L_2$ | Variance |
| **Normal Distribution** | | | | | | | | | | |
| 6 | 50 | 5 | 0.0 | 0.01 | 0.05 | Parametric | 0.992 | 0.852 | 0.912 | 0.986 |
| 6 | 100 | 5 | 0.0 | 0.01 | 0.05 | Parametric | 1.000 | 0.988 | 0.998 | 1.000 |
| 6 | 200 | 5 | 0.0 | 0.01 | 0.05 | Parametric | 1.000 | 0.998 | 1.000 | 1.000 |
| 6 | 400 | 5 | 0.0 | 0.01 | 0.05 | Parametric | 1.000 | 1.000 | 1.000 | 1.000 |
| 6 | 50 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 0.998 | 0.962 | 0.980 | 0.990 |
| 6 | 100 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 1.000 | 1.000 | 1.000 | 1.000 |
| 6 | 200 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 1.000 | 1.000 | 1.000 | 1.000 |
| 6 | 400 | 5 | 0.0 | 0.05 | 0.05 | Parametric | 1.000 | 1.000 | 1.000 | 1.000 |
| 6 | 50 | 5 | 0.5 | 0.01 | 0.05 | Parametric | 0.994 | 0.846 | 0.890 | 0.976 |
| 6 | 100 | 5 | 0.5 | 0.01 | 0.05 | Parametric | 1.000 | 0.982 | 0.994 | 0.998 |
| 6 | 200 | 5 | 0.5 | 0.01 | 0.05 | Parametric | 1.000 | 1.000 | 1.000 | 1.000 |
| 6 | 400 | 5 | 0.5 | 0.01 | 0.05 | Parametric | 1.000 | 1.000 | 1.000 | 1.000 |
| 6 | 50 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 0.998 | 0.944 | 0.956 | 0.990 |
| 6 | 100 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 1.000 | 0.998 | 0.998 | 1.000 |
| 6 | 200 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 1.000 | 1.000 | 1.000 | 1.000 |
| 6 | 400 | 5 | 0.5 | 0.05 | 0.05 | Parametric | 1.000 | 1.000 | 1.000 | 1.000 |

# D Additional Application Results

## D.1 Additional Results Using the Base and Adaptation Models

Detected outliers when using the base model are plotted in Figures D.1 and D.2, and the number of positive/negative outliers aggregated over countries in Figure D.3 for the base model and in Figure D.4 for the adaptation model.

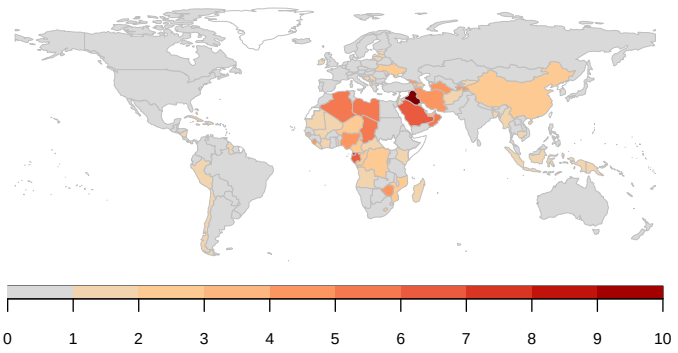


Figure D.1: Detected outliers aggregated over the full sample of 1961 - 2017 by country in the panel in the base model. Gray denotes no outliers detected.

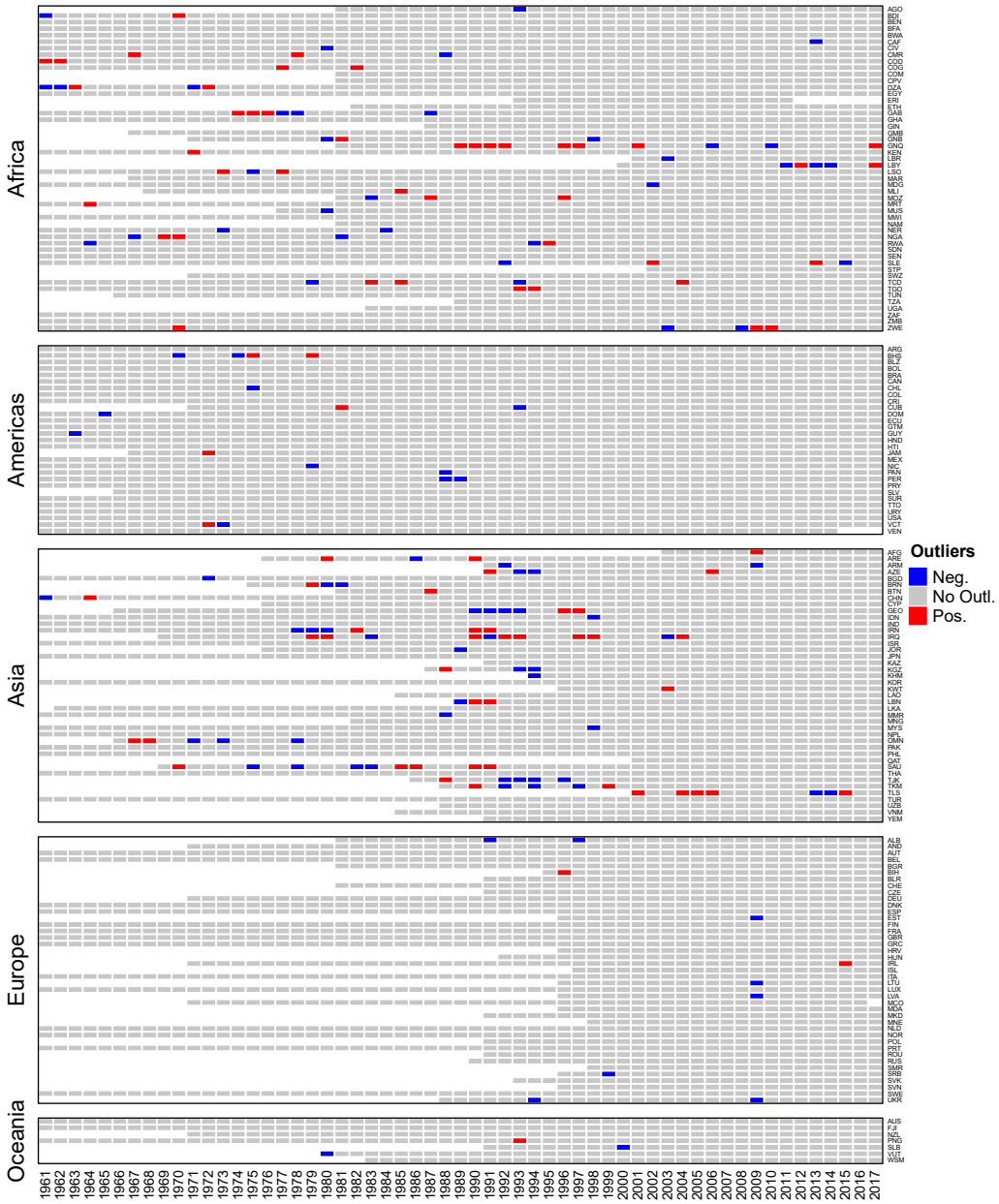


Figure D.2: Detected outliers using the robust IIS estimator across countries and time in the global cross-country panel from 1961-2017 in the base model. The figure shows country-year observations as gray when not outlying, blue when there is a negative outliers, and red for positive outliers.

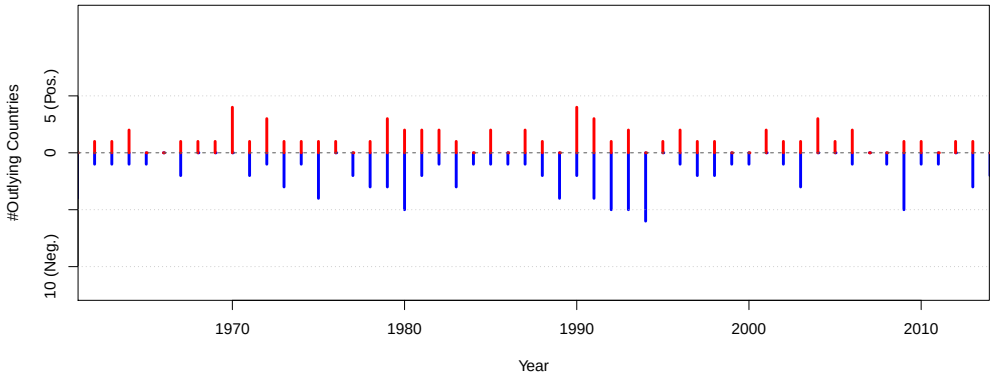


Figure D.3: Detected outliers using the robust IIS estimator across countries and time in the global cross-country panel from 1961-2017 in the base model. The figure shows the number of positive and negative outliers summed over countries for each year.

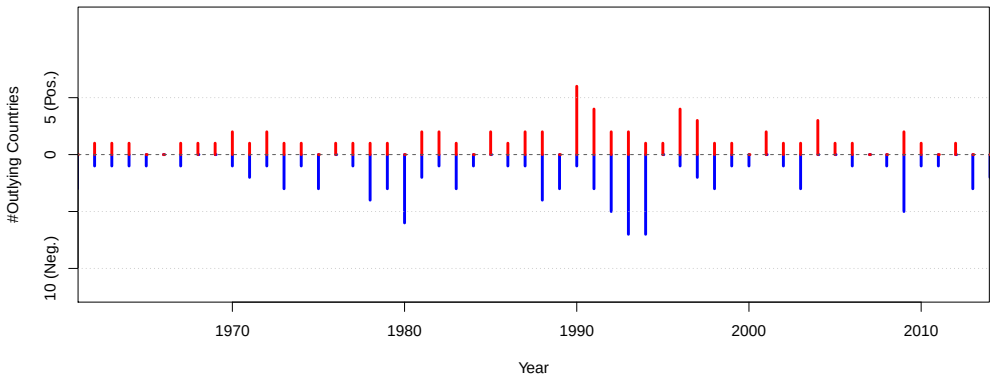


Figure D.4: Detected outliers using the robust IIS estimator across countries and time in the global cross-country panel from 1961-2017 in the adaptation model. The figure shows the number of positive and negative outliers summed over countries for each year.

## D.2 Sensitivity of the Results to the Chosen Cut-off Level

Here we show additional estimation results when varying the cut-off level $c$ used to classify observations as outlying. Tables below provide the results for $c = 1.96$ and $c = 3.29$ with an expected false positive rate of 5% and 0.1% for a Normal reference distribution respectively.

Table D.1: OLS and IIS Panel Regression Results together with their difference in coefficients and the resulting outlier distortion test statistic. Coefficients on control variables are omitted. IIS selection was carried out at $\gamma_c = 0.05$ ($c = 1.96$).

| | Base | Base IIS | Base Outlier Distortion Test | Adaptation | Adaptation IIS | Adaptation Outlier Distortion Test |
|---|---|---|---|---|---|---|
| Temperature | 0.01734*** | 0.00934*** | 50.91 | -0.06224*** | -0.03506*** | 59.99 |
| | (0.00348) | (0.00211) | [<0.001] | (0.01041) | (0.00661) | [<0.001] |
| Temperature$^2$ | -0.00059*** | -0.00034*** | 52.34 | 0.00070 | 0.00003 | 28.78 |
| | (0.0001) | (0.00006) | [<0.001] | (0.00037) | (0.00024) | [<0.001] |
| Precipitation | 0.00043 | 0.00098 | 2.39 | 0.01018 | 0.01554*** | 8.47 |
| | (0.00111) | (0.00067) | [0.122] | (0.00563) | (0.00347) | [0.004] |
| Precipitation$^2$ | -0.00004 | -0.00004* | 0.17 | -0.00019 | -0.00038** | 8.92 |
| | (0.00003) | (0.00002) | [0.680] | (0.00019) | (0.00012) | [0.003] |
| Temperature x $GDP_{pc}$ | | | | 0.00811*** | 0.00420*** | 109.42 |
| | | | | (0.00111) | (0.0007) | [<0.001] |
| Temperature$^2$ x $GDP_{pc}$ | | | | -0.00012** | -0.00002 | 52.95 |
| | | | | (0.00004) | (0.00003) | [<0.001] |
| Precipitation x $GDP_{pc}$ | | | | -0.00121 | -0.00179*** | 6.99 |
| | | | | (0.00066) | (0.00041) | [0.008] |
| Precipitation$^2$ x $GDP_{pc}$ | | | | 0.00002 | 0.00004** | 8.3 |
| | | | | (0.00002) | (0.00001) | [0.004] |
| Num. Outliers | | | 299 | | | 306 |
| Outlier Distortion test statistic for Temp. Variables | | | $\chi^2_2 = 54.8$ [<0.001] | | | $\chi^2_4 = 272.56$ [<0.001] |
| Num.Obs. | 7716 | 7716 | | 7716 | 7716 | |
| BIC | -18483.2 | -24137.8 | | -18801.4 | -24346.8 | |
| Log.Lik. | 11774.742 | 15940.230 | | 11951.758 | 16093.982 | |
| Fixed Effects | Country & Year | Country & Year | | Country & Year | Country & Year | |

Please note that the estimated coefficients on Precipitation$^2$ by OLS and IIS are very close but not exactly equal in the base model. Thus, its outlier distortion test statistics is not zero.

(Standard Errors) and [p-values]

* p < 0.05, ** p < 0.01, *** p < 0.001

Table D.2: OLS and IIS Panel Regression Results together with their difference in coefficients and the resulting outlier distortion test statistic. Coefficients on control variables are omitted. IIS selection was carried out at $\gamma_c = 0.001$ ($c = 3.29$).

| | Base | Base IIS | Base Outlier Distortion Test | Adaptation | Adaptation IIS | Adaptation Outlier Distortion Test |
|---|---|---|---|---|---|---|
| Temperature | 0.01734*** | 0.01135*** | 462.62 | -0.06224*** | -0.02983*** | 1409.44 |
| | (0.00348) | (0.00249) | [<0.001] | (0.01041) | (0.0077) | [<0.001] |
| Temperature$^2$ | -0.00059*** | -0.00041*** | 461.08 | 0.00070 | -0.00018 | 830.53 |
| | (0.0001) | (0.00007) | [<0.001] | (0.00037) | (0.00027) | [<0.001] |
| Precipitation | 0.00043 | 0.00051 | 0.75 | 0.01018 | 0.01238** | 22.93 |
| | (0.00111) | (0.00079) | [0.388] | (0.00563) | (0.00411) | [<0.001] |
| Precipitation$^2$ | -0.00004 | -0.00004 | 0.76 | -0.00019 | -0.00028* | 31.23 |
| | (0.00003) | (0.00002) | [0.384] | (0.00019) | (0.00014) | [<0.001] |
| Temperature x $GDP_{pc}$ | | | | 0.00811*** | 0.00392*** | 2080.06 |
| | | | | (0.00111) | (0.00082) | [<0.001] |
| Temperature$^2$ x $GDP_{pc}$ | | | | -0.00012** | 0.00000 | 1143.24 |
| | | | | (0.00004) | (0.00003) | [<0.001] |
| Precipitation x $GDP_{pc}$ | | | | -0.00121 | -0.00141** | 12.82 |
| | | | | (0.00066) | (0.00049) | [<0.001] |
| Precipitation$^2$ x $GDP_{pc}$ | | | | 0.00002 | 0.00003 | 25.11 |
| | | | | (0.00002) | (0.00002) | [<0.001] |
| Num. Outliers | | | 101 | | | 92 |
| Outlier Distortion test statistic for Temp. Variables | | | $\chi^2_2 = 489.82$ [<0.001] | | | $\chi^2_4 = 3768.21$ [<0.001] |
| Num.Obs. | 7716 | 7716 | | 7716 | 7716 | |
| BIC | -18483.2 | -22967.3 | | -18801.4 | -23103.9 | |
| Log.Lik. | 11774.742 | 14468.820 | | 11951.758 | 14514.735 | |
| Fixed Effects | Country & Year | Country & Year | | Country & Year | Country & Year | |

Please note that the estimated coefficients on Precipitation$^2$ by OLS and IIS are very close but not exactly equal in the base model. Thus, its outlier distortion test statistics is not zero.
(Standard Errors) and [p-values]
* p < 0.05, ** p < 0.01, *** p < 0.001

## D.3 Adaptation Using Lagged Income Levels

Here we provide additional results when estimating the macro-economic impacts of temperatures. To alleviate concerns around feedback between $\Delta log(y)_{i,t}$ and temperature interacted with $log(y)_{i,t}$, we also present a model using $log(y)_{i,t-1}$ to capture adaptation (see equation (D.1) below):

$$\Delta log(y)_{i,t} = \alpha_i + \lambda_t + \beta_1 T_{i,t} + \beta_2 T^2_{i,t} + \beta_3 \left[T_{i,t} \times log(y)_{i,t-1}\right] + \beta_4 \left[T^2_{i,t} \times log(y)_{i,t-1}\right] + x'_{i,t}\gamma + u_{i,t}. \quad \text{(D.1)}$$

Estimation results using (D.1) are shown below in Figure D.5, with the resulting estimated non-linear relationship shown in D.6. Detected outliers are shown in Figures D.7 and D.8. Tables below provide the estimation results using lagged income interactions for varying cut-offs $c = 2.57$ (for a 1% false positive rate for a Normal reference distribution) to $c = 3.29$ (for 0.1% expected false positive rate) and $c = 1.96$ (for a 5% false positive rate).

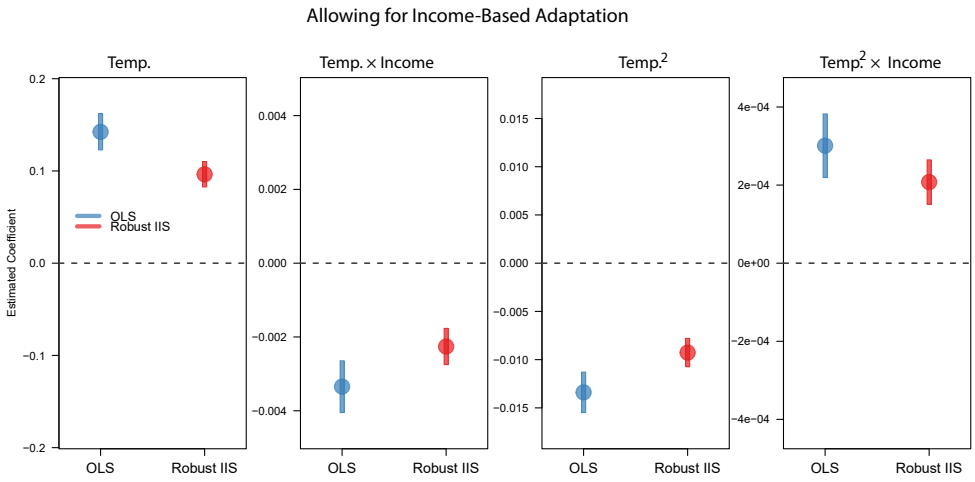


Figure D.5: Model using Lagged GDP per capita. Estimated Impact of Temperatures on GDP Per Capita Growth Allowing for Adaptation and Controlling for Outliers. Coefficients on temperatures and the income interaction terms are shown using OLS and the robust IIS estimator. Note that scale of the y-axis differs across plots.

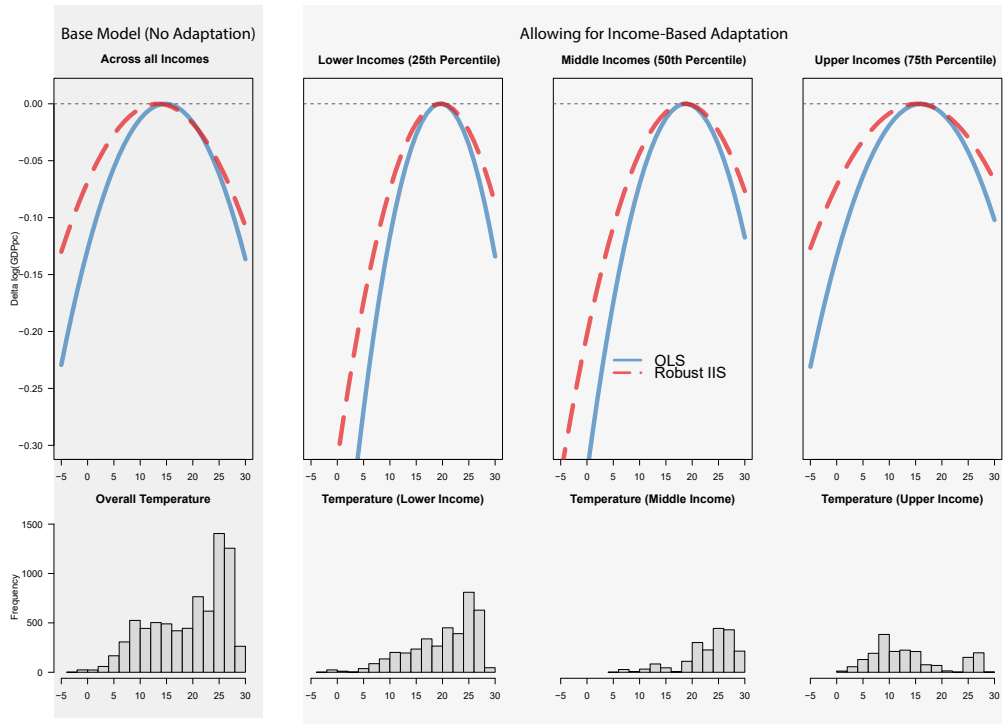


Figure D.6: Projection with Lagged GDP per capita. Estimated Impact of Temperatures on GDP Per Capita Growth Allowing for Adaptation and Controlling for Outliers. OLS-estimated relationship shown in blue, robust IIS estimated relationship shown in red. Estimated non-linear impact function for the overall base model (top left) and at three different income levels (other top panels). Observed temperatures for the entire sample and across income ranges are shown in the bottom panel.

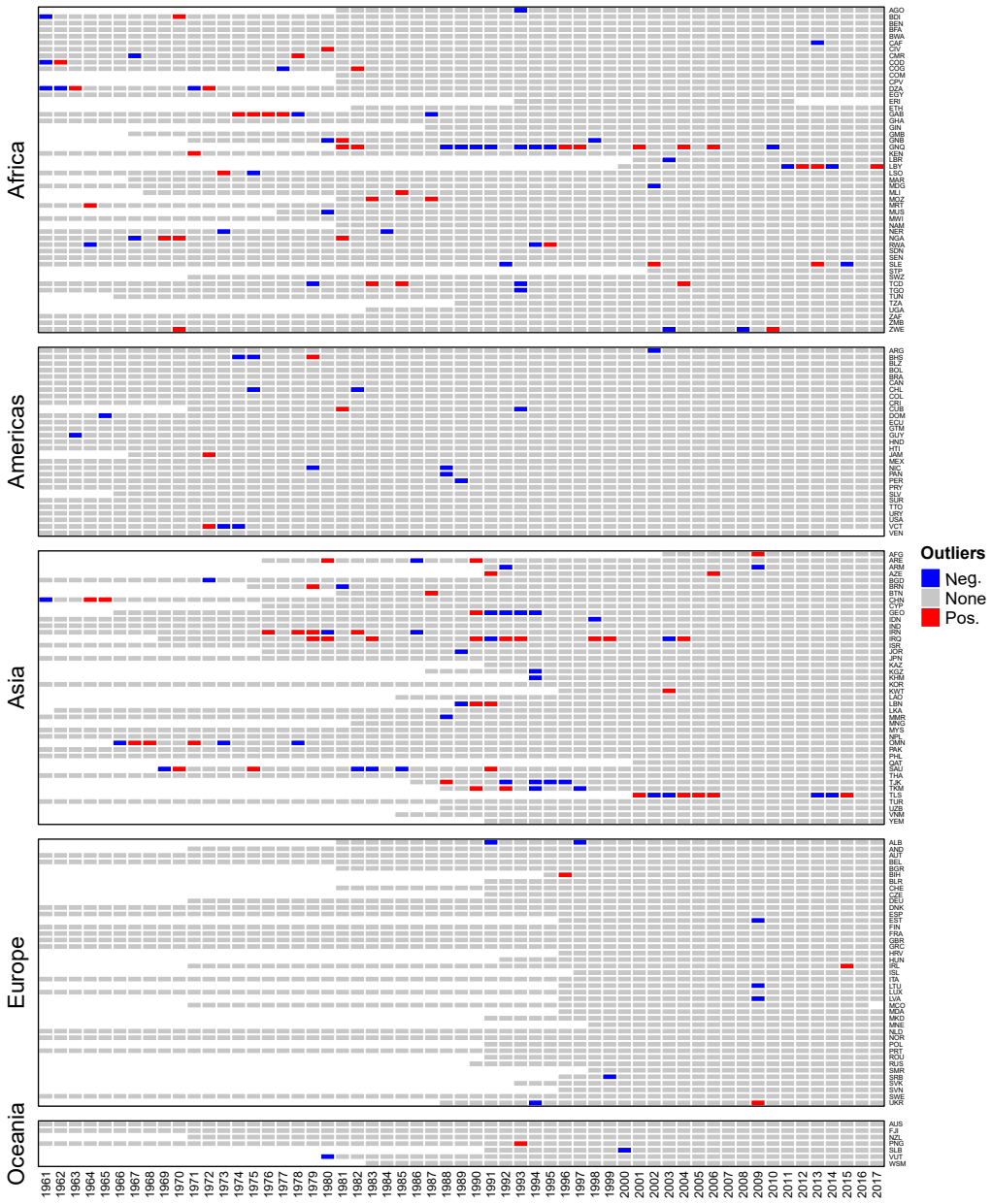


Figure D.7: Detected outliers using the robust IIS estimator across countries and time in the global cross-country panel from 1961-2017 in the adaptation model with lagged GDP per capita. The figure shows country-year observations as gray when not outlying, blue when there is a negative outliers, and red for positive outliers.

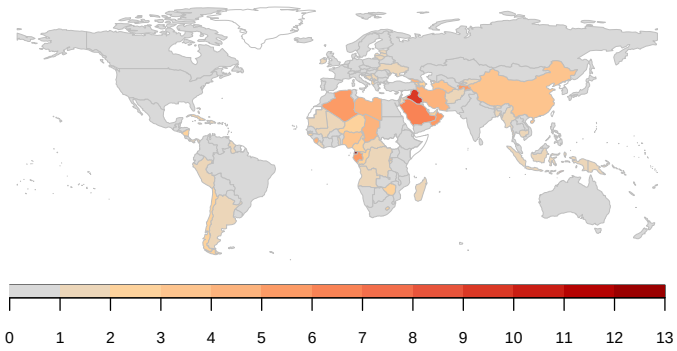


Figure D.8: Detected outliers aggregated over the full sample of 1961 - 2017 by country in the panel in the adaptation model with lagged GDP per capita. Gray denotes no outliers detected.

Table D.3: OLS and IIS Panel Regression Results together with their difference in coefficients and the resulting outlier distortion test statistic. Coefficients on control variables are omitted. IIS selection was carried out at $\gamma_c = 0.01$ ($c = 2.57$).

| | Lagged Adaptation | Lagged Adaptation IIS | Lagged Adaptation Outlier Distortion Test |
|---|---|---|---|
| Temperature | 0.14234*** | 0.09637*** | 186.25 |
| | (0.01003) | (0.00703) | [<0.001] |
| Temperature$^2$ | -0.00335*** | -0.00226*** | 88.57 |
| | (0.00036) | (0.00025) | [<0.001] |
| Precipitation | -0.00212 | 0.01125** | 7.86 |
| | (0.00551) | (0.00382) | [0.005] |
| Precipitation$^2$ | -0.00004 | -0.00032* | 17.55 |
| | (0.00019) | (0.00013) | [<0.001] |
| Temperature x Lag($GDP_{pc}$) | -0.01339*** | -0.00927*** | 317.59 |
| | (0.00107) | (0.00075) | [<0.001] |
| Temperature$^2$ x Lag($GDP_{pc}$) | 0.00030*** | 0.00021*** | 148.17 |
| | (0.00004) | (0.00003) | [<0.001] |
| Precipitation x Lag($GDP_{pc}$) | 0.00035 | -0.00124** | 6.86 |
| | (0.00065) | (0.00045) | [0.009] |
| Precipitation$^2$ x Lag($GDP_{pc}$) | 0.00000 | 0.00003* | 17.52 |
| | (0.00002) | (0.00002) | [<0.001] |
| Num. Outliers | | | 158 |
| Outlier Distortion test statistic for Temp. Variables | | | $\chi^2_4 = 1022.68$ [<0.001] |
| Num.Obs. | 7716 | 7716 | |
| BIC | -19066.8 | -23825.7 | |
| Log.Lik. | 12084.427 | 15171.040 | |
| Two Way Fixed Effects | Yes | Yes | |

(Standard Errors) and [p-values]
* p < 0.05, ** p < 0.01, *** p < 0.001

Table D.4: OLS and IIS Panel Regression Results together with their difference in coefficients and the resulting outlier distortion test statistic. Coefficients on control variables are omitted. IIS selection was carried out at $\gamma_c = 0.001$ ($c = 3.29$).

| | Lagged Adaptation | Lagged Adaptation IIS | Lagged Adaptation Outlier Distortion Test |
|---|---|---|---|
| Temperature | 0.14234*** | 0.09848*** | 1409.44 |
| | (0.01003) | (0.0074) | [<0.001] |
| Temperature$^2$ | -0.00335*** | -0.00222*** | 830.53 |
| | (0.00036) | (0.00026) | [<0.001] |
| Precipitation | -0.00212 | 0.01137** | 22.93 |
| | (0.00551) | (0.00404) | [<0.001] |
| Precipitation$^2$ | -0.00004 | -0.00035* | 31.23 |
| | (0.00019) | (0.00014) | [<0.001] |
| Temperature x Lag($\text{GDP}_{\text{pc}}$) | -0.01339*** | -0.00928*** | 2080.06 |
| | (0.00107) | (0.00079) | [<0.001] |
| Temperature$^2$ x Lag($\text{GDP}_{\text{pc}}$) | 0.00030*** | 0.00020*** | 1143.24 |
| | (0.00004) | (0.00003) | [<0.001] |
| Precipitation x Lag($\text{GDP}_{\text{pc}}$) | 0.00035 | -0.00125** | 12.82 |
| | (0.00065) | (0.00048) | [<0.001] |
| Precipitation$^2$ x Lag($\text{GDP}_{\text{pc}}$) | 0.00000 | 0.00004* | 25.11 |
| | (0.00002) | (0.00002) | [<0.001] |
| Num. Outliers | | | 103 |
| Outlier Distortion test statistic for Temp. Variables | | | $\chi^2_4 = 4778.56$ [<0.001] |
| Num.Obs. | 7716 | 7716 | |
| BIC | -19066.8 | -23267.9 | |
| Log.Lik. | 12084.427 | 14645.982 | |
| Two Way Fixed Effects | Yes | Yes | |

(Standard Errors) and [p-values]
* p < 0.05, ** p < 0.01, *** p < 0.001

Table D.5: OLS and IIS Panel Regression Results together with their difference in coefficients and the resulting outlier distortion test statistic. Coefficients on control variables are omitted. IIS selection was carried out at $\gamma_c = 0.05$ ($c = 1.96$).

| | Lagged Adaptation | Lagged Adaptation IIS | Lagged Adaptation Outlier Distortion Test |
|---|---|---|---|
| Temperature | 0.14234*** | 0.09224*** | 59.99 |
| | (0.01003) | (0.00653) | [<0.001] |
| Temperature$^2$ | -0.00335*** | -0.00224*** | 28.78 |
| | (0.00036) | (0.00023) | [<0.001] |
| Precipitation | -0.00212 | 0.01346*** | 8.47 |
| | (0.00551) | (0.00344) | [0.004] |
| Precipitation$^2$ | -0.00004 | -0.00036** | 8.92 |
| | (0.00019) | (0.00012) | [0.003] |
| Temperature x Lag($GDP_{pc}$) | -0.01339*** | -0.00864*** | 109.42 |
| | (0.00107) | (0.00069) | [<0.001] |
| Temperature$^2$ x Lag($GDP_{pc}$) | 0.00030*** | 0.00020*** | 52.95 |
| | (0.00004) | (0.00003) | [<0.001] |
| Precipitation x Lag($GDP_{pc}$) | 0.00035 | -0.00148*** | 6.99 |
| | (0.00065) | (0.00041) | [0.008] |
| Precipitation$^2$ x Lag($GDP_{pc}$) | 0.00000 | 0.00004** | 8.3 |
| | (0.00002) | (0.00001) | [0.004] |
| Num. Outliers | | | 308 |
| Outlier Distortion test statistic for Temp. Variables | | | $\chi^2_4 = 472.44$ [<0.001] |
| Num.Obs. | 7716 | 7716 | |
| BIC | -19066.8 | -24457.0 | |
| Log.Lik. | 12084.427 | 16158.022 | |
| Two Way Fixed Effects | Yes | Yes | |

(Standard Errors) and [p-values]
* p < 0.05, ** p < 0.01, *** p < 0.001

# Appendix for Paper 2

# An empirical climate damage function accounting for climate extremes and adaptation

Moritz P. Schwarz and Felix Pretis

# Supplementary Material

# 1 Additional Figures

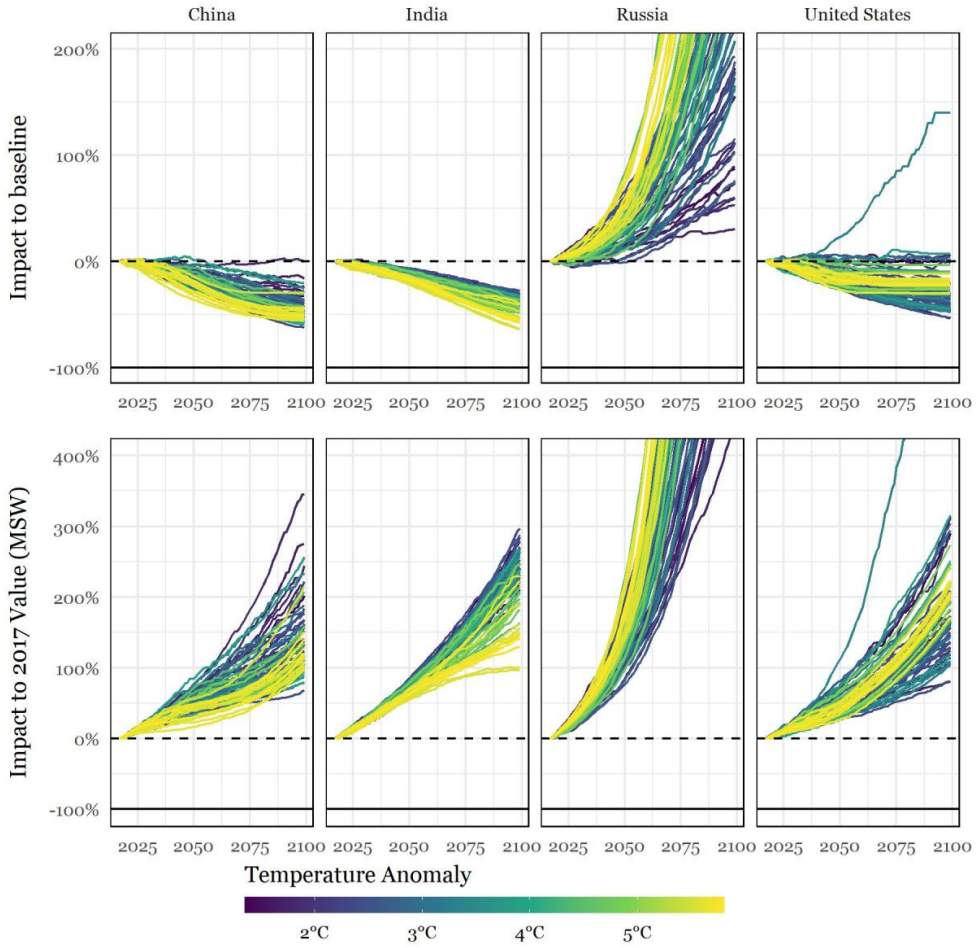


**Figure A1**: Projected country-level climate impacts accounting for Income Adaptation relative to the 'no future climate change' baseline (top panel) and relative to the baseline in 2017 in the Muller, Stock, and Watson (MSW) data (bottom panel) for China, India, Russia, and the United States. Colour-gradient denotes level of Global Mean Surface Temperature (GMST) warming in 2090-2099 relative to pre-industrial temperatures.

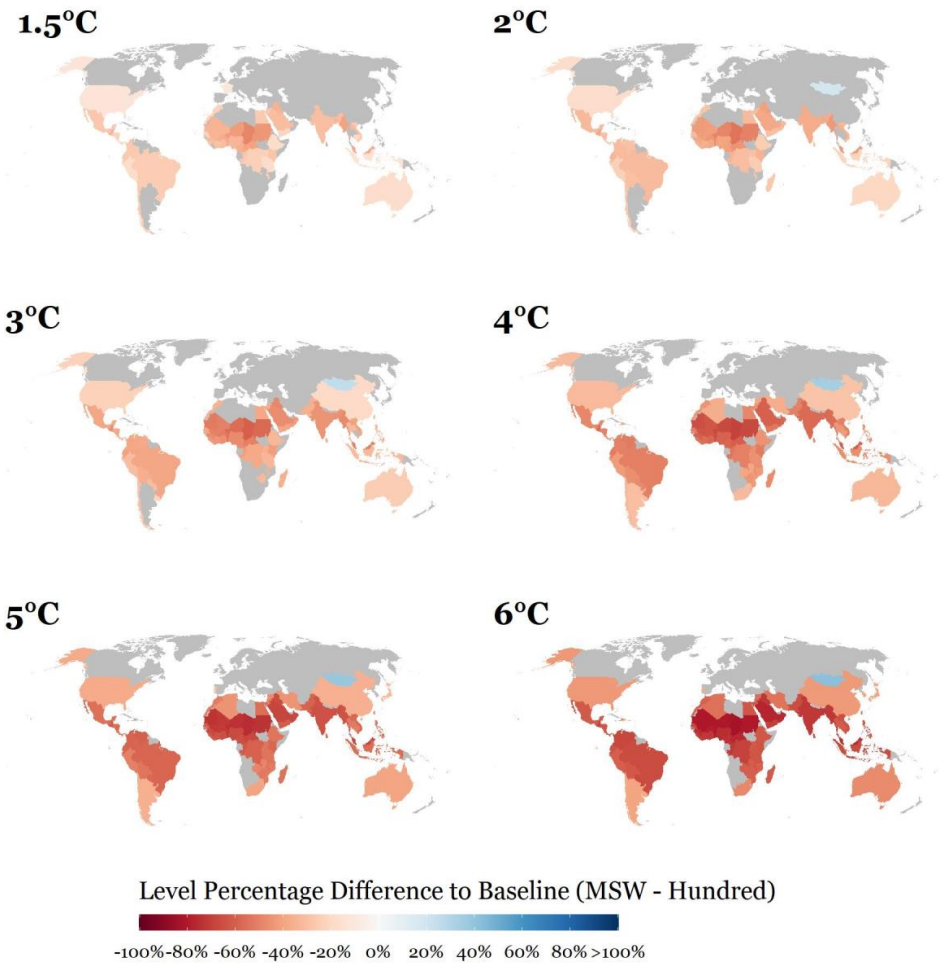


**Figure A2:** Projected difference in GDP per capita under different GMST warming levels in 2090-2100 relative to 'no future climate change' long-run forecasts for LASSO selected models. Countries are shaded grey if 90% projection interval of impacts includes zero or no value is available. Values were calculated as means across all realisations. Values were aggregated using the GMST anomaly of each climate model run, which as rounded to the nearest integer (apart from the 1.5°C map, which includes all model runs below 1.75°C where the 2°C map includes all values above 1.75°C and below 2.5°C). The maximum GMST anomaly of any CMIP5 model considered is 5.9°C, so the map under the label 6°C includes all models above 5.5°C up to the maximum of 5.9°C. Baseline uses Muller, Stock, and Watson (MSW) GDP per capita forecasts using the 100 year forecast value (Hundred).

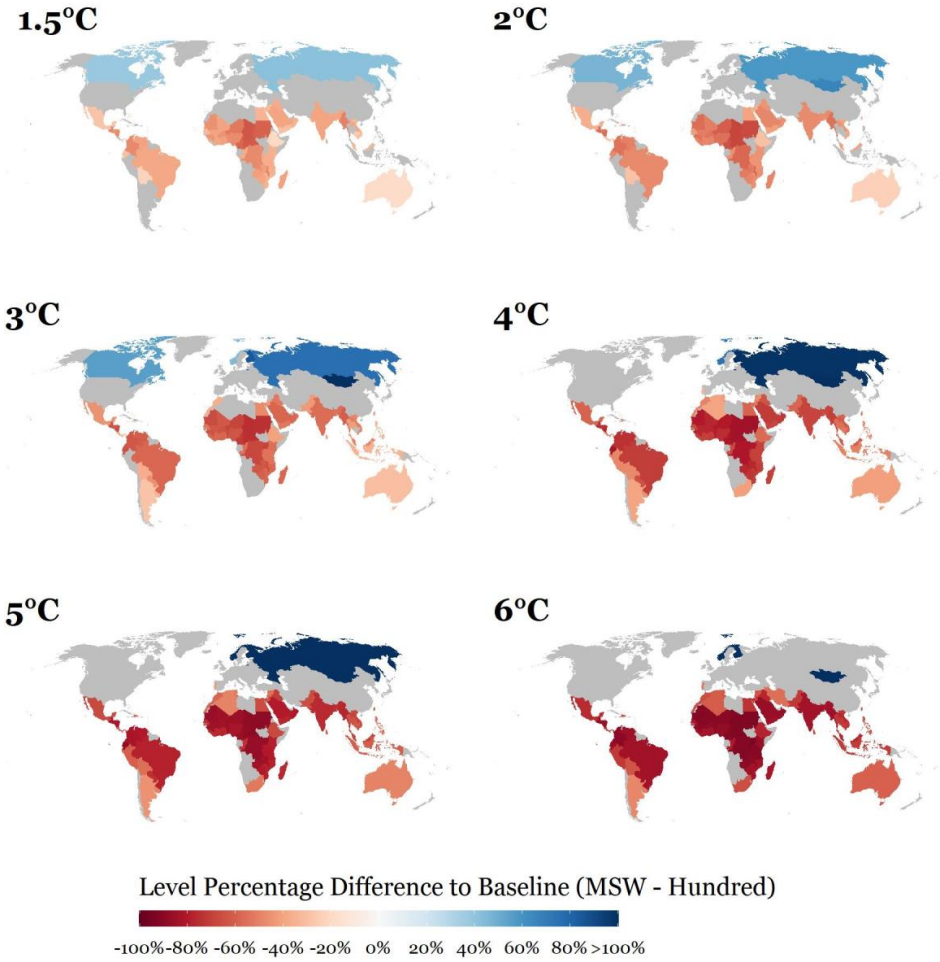


**Figure A3:** Projected difference in GDP per capita under different GMST warming levels in 2090-2100 relative to 'no future climate change' long-run forecasts for gets selected models. Countries are shaded grey if 90% projection interval of impacts includes zero or no value is available. Values were calculated as means across all realisations. Values were aggregated using the GMST anomaly of each climate model run, which as rounded to the nearest integer (apart from the 1.5°C map, which includes all model runs below 1.75°C where the 2°C map includes all values above 1.75°C and below 2.5°C). The maximum GMST anomaly of any CMIP5 model considered is 5.9°C, so the map under the label 6°C includes all models above 5.5°C up to the maximum of 5.9°C. Baseline uses Muller, Stock, and Watson (MSW) GDP per capita forecasts using the 100 year forecast value (Hundred).

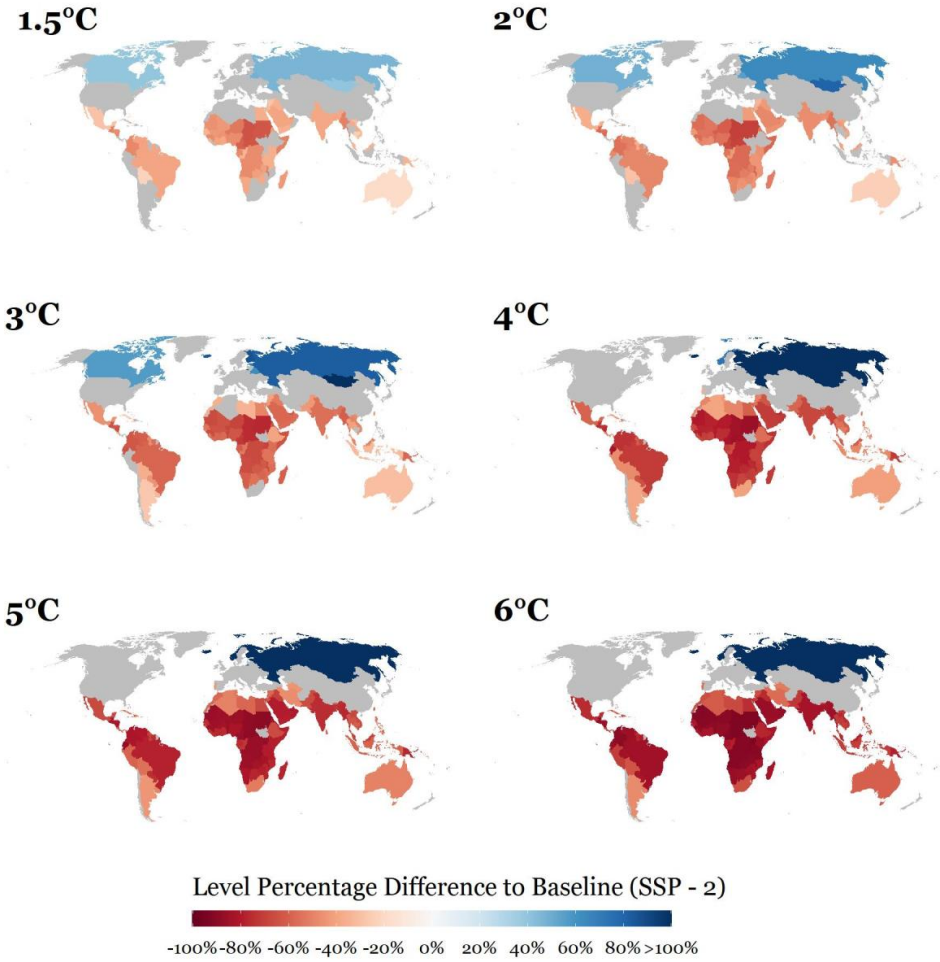


**Figure A4:** Projected difference in GDP per capita under different GMST warming levels in 2090-2100 relative to the SSP2 baseline for gets selected models. Countries are shaded grey if 90% projection interval of impacts includes zero or no value is available. Values were calculated as means across all realisations. Values were aggregated using the GMST anomaly of each climate model run, which as rounded to the nearest integer (apart from the 1.5°C map, which includes all model runs below 1.75°C where the 2°C map includes all values above 1.75°C and below 2.5°C). The maximum GMST anomaly of any CMIP5 model considered is 5.9°C, so the map under the label 6°C includes all models above 5.5°C up to the maximum of 5.9°C. Baseline uses the Shared Socio-Economic Pathways Scenario 2 (SSP2) GDP per capita projections.

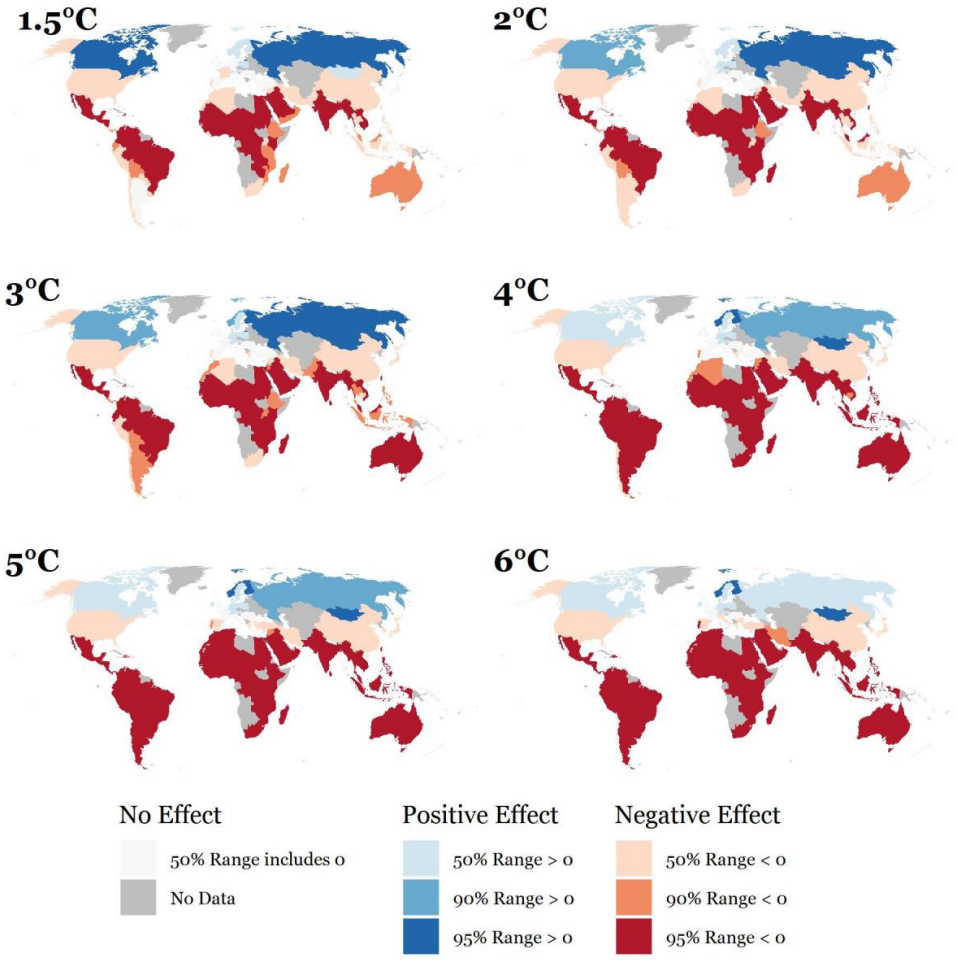


**Figure A5:** Uncertainty in projected difference in GDP per capita under different GMST warming levels in 2090-2100 relative to the MSW baseline for gets selected models. Countries are shaded grey if no value is available. Values represent the consistency of effects across the IQR, 90% Confidence Interval and 95% Confidence Interval calculated as means across all realisations. Values were aggregated using the GMST anomaly of each climate model run, which as rounded to the nearest integer (apart from the 1.5°C map, which includes all model runs below 1.75°C where the 2°C map includes all values above 1.75°C and below 2.5°C). The maximum GMST anomaly of any CMIP5 model considered is 5.9°C, so the map under the label 6°C includes all models above 5.5°C up to the maximum of 5.9°C.

# 2 Climate Variable Overview

A detailed description of all Climate Extreme Indicators can be found at https://www.climdex.org/learn/indices/.

Table A1. Extreme Climate Variables as defined by the Expert Team on Climate Change Detection and Indices (definitions taken from Donat *et al.*, 2013).

| Indicator | Indicator Name | Indicator Definition | Units |
|---|---|---|---|
| CDD | Consecutive dry days | Maximum number of consecutive days when precipitation < 1 mm | days |
| CSDI | Cold spell duration index | Annual count when at least six consecutive days of min temperature < 10th percentile | days |
| CWD | Consecutive wet days | Maximum number of consecutive days when precipitation > 1 mm | days |
| DTR | Diurnal temperature range | Monthly mean difference between daily max and min temperature | °C |
| FD | Frost days | Annual count when daily minimum temperature < 0°C | days |
| GSL | Growing season length | Annual (1st Jan to 31st Dec in NH, 1st July to 30th June in SH) count between first span of at least 6 days with TG > 5°C and first span after July 1 (January 1 in SH) of 6 days with TG < 5°C (where TG is daily mean temperature) | days |
| ID | Ice days | Annual count when daily maximum temperature < 0°C | days |
| PRCPTOT | Annual total wet day precipitation | Annual total precipitation from days > 1 mm | mm |
| R10mm | Number of heavy precipitation days | Annual count when precipitation ≥ 10 mm | days |
| R1mm | Number of precipitation days | Annual count when precipitation ≥ 1 mm | days |
| R20mm | Number of very heavy precipitation days | Annual count when precipitation ≥ 20 mm | days |
| R95p | Very wet days | Annual total precipitation from days > 95th percentile | mm |
| R99p | Extremely wet days | Annual total precipitation from days > 99th percentile | mm |
| Rx1day | Max 1 day precipitation amount | Monthly maximum 1 day precipitation | mm |
| Rx5day | Max 5 day precipitation amount | Monthly maximum consecutive 5 day precipitation | mm |
| SDII | Simple daily intensity index | The ratio of annual total precipitation to the number of wet days (≥ 1mm) | mm/ day |
| SU | Summer days | Annual count when daily max temperature > 25°C | days |
| TN10p | Cool nights | Percentage of time when daily min temperature < 10th percentile | % |
| TN90p | Warm nights | Percentage of time when daily min temperature > 90th percentile | % |
| TNn | Coldest night | Monthly minimum value of daily min temperature | °C |
| TNx | Warmest nigh | Monthly maximum value of daily min temperature | °C |
| TR | Tropical nights | Annual count when daily min temperature > 20°C | days |
| TX10p | Cool days | Percentage of time when daily max temperature < 10th percentile | % |
| TX90p | Warm days | Percentage of time when daily max temperature > 90th percentile | % |
| TXn | Coldest day | Monthly minimum value of daily max temperature | °C |
| TXx | Hottest day | Monthly maximum value of daily max temperature | °C |
| WSDI | Warm spell duration index | Annual count when at least six consecutive days of max temperature > 90th percentile | days |

## 2.1 Summary Statistics

Table A2. Summary Statistics of the Selection Variables used in all Models

**Summary Statistics**

| Statistic | N | Mean | St. Dev. | Min | Pctl(25) | Pctl(75) | Max |
|---|---|---|---|---|---|---|---|
| ΔLog(GDP per capita) | 6,536 | 0.019 | 0.062 | -1.050 | -0.003 | 0.046 | 0.878 |
| Annual Average Temperature | 6,536 | 19.402 | 7.013 | -1.967 | 13.774 | 25.500 | 29.853 |
| Annual Average Precipitation | 6,536 | 9.731 | 6.079 | 0.067 | 5.366 | 12.798 | 37.104 |
| CDD | 6,536 | 66.670 | 71.035 | 2.924 | 21.595 | 84.728 | 767.832 |
| CSDI | 6,536 | 3.438 | 4.815 | 0.000 | 0.586 | 4.578 | 107.293 |
| CWD | 6,536 | 31.873 | 29.424 | 0.642 | 9.809 | 48.624 | 254.683 |
| DTR | 6,536 | 7.813 | 3.526 | 0.875 | 5.404 | 10.178 | 16.696 |
| FD | 6,536 | 29.857 | 48.070 | 0.000 | 0.000 | 44.998 | 237.323 |
| GSL | 6,536 | 332.225 | 56.569 | 45.107 | 315.173 | 365.000 | 366.000 |
| ID | 6,536 | 9.250 | 22.068 | 0.000 | 0.000 | 4.564 | 160.502 |
| PRCPTOT | 6,536 | 1,213.533 | 887.027 | 2.852 | 574.446 | 1,758.374 | 6,263.758 |
| R10mm | 6,536 | 38.691 | 33.066 | 0.000 | 14.700 | 55.315 | 251.752 |
| R1mm | 6,536 | 150.340 | 78.396 | 0.716 | 92.720 | 215.402 | 337.184 |
| R20mm | 6,536 | 8.917 | 10.363 | 0.000 | 2.195 | 11.931 | 106.953 |
| R95p | 6,536 | 236.716 | 205.363 | 0.000 | 98.779 | 316.589 | 1,756.307 |
| R99p | 6,536 | 74.899 | 77.687 | 0.000 | 24.518 | 98.428 | 956.574 |
| Rx1day | 6,536 | 41.502 | 19.955 | 1.603 | 26.395 | 53.463 | 137.613 |
| Rx5day | 6,536 | 94.704 | 49.498 | 2.456 | 58.375 | 122.139 | 358.305 |
| SDII | 6,536 | 7.423 | 2.336 | 0.952 | 5.729 | 8.691 | 21.146 |
| SU | 6,536 | 200.014 | 124.220 | 0.000 | 78.027 | 321.855 | 366.000 |
| TN10p | 6,536 | 9.797 | 4.023 | 0.216 | 7.088 | 11.857 | 53.207 |
| TN90p | 6,536 | 12.041 | 6.376 | 0.126 | 8.110 | 14.740 | 72.559 |
| TNn | 6,536 | 3.289 | 14.615 | -42.462 | -6.850 | 15.815 | 25.335 |
| TNx | 6,536 | 23.965 | 4.146 | 9.722 | 21.010 | 27.012 | 36.104 |
| TR | 6,536 | 150.340 | 134.160 | 0.000 | 25.618 | 281.301 | 366.000 |
| TX10p | 6,536 | 9.703 | 4.435 | 0.000 | 6.733 | 11.816 | 51.491 |
| TX90p | 6,536 | 12.083 | 6.142 | 0.056 | 8.048 | 14.879 | 76.309 |
| TXn | 6,536 | 11.674 | 12.967 | -31.970 | 1.081 | 22.719 | 27.008 |
| TXx | 6,536 | 34.722 | 5.573 | 12.775 | 30.712 | 38.034 | 50.700 |
| WSDI | 6,536 | 7.960 | 11.807 | -0.026 | 1.433 | 10.284 | 223.540 |

## 2.2 Climate Data Processing Information

All climate variables, apart from annual average temperature and precipitation, used in the estimation were obtained from an average of four reanalysis products (ERA-Interim, ERA40, NCEP-Reanalysis 1 and NCEP-Reanalysis 2) produced by Sillmann *et al.* (2013a). All spatial datasets were bilinearly resampled to a 2.5° by 2.5° grid using the *resample* function of the R "raster" package. All four reanalysis products were averaged for each variable and then a 2015-population (CIESIN, 2016) weighted average over a countries spatial area was taken. Temperature variables, where needed, were converted from Kelvin to °C by subtracting 273.15 from each value. Annual average temperatures and precipitation were obtained from Matsuura and Willmott (2018) and processed in the same way.

A similar methodology, with regard to resampling and taking a weighted average, was employed for the projected climate variables over the period of 2007 to 2099 obtained from Sillmann *et al.* (2013b), although each CMIP5 model and ensemble combination was treated independently of its forcing scenario. Each CMIP5 model/ensemble combination was categorised by its global mean surface temperature anomaly to pre-industrial temperatures in the period of 2090-2099. Ensemble specific global mean surface temperature anomalies were calculated and verified with data obtained from the KNMI Climate Explorer. The verification data can be downloaded here: https://climexp.knmi.nl/CMIP5/Tglobal/.

The SSP base projections on population and GDP used in this paper were obtained from the IIASA (2013) database (see Riahi *et al.*, 2017 for an overview), with the GDP data produced by the OECD (Dellink *et al.*, 2017) and population projected by KC and Lutz (2017).

# 3 Full Estimation results

## 3.1 Standard Model Results

Table A3. Standard Regression Model Results for gets, LASSO (re-estimated post-selection) and AATPsq

| **Standard Model Results** | | | |
|---|---|---|---|
| | *Dependent variable:* | | |
| | Δlog GDP per capita | | |
| | gets | LASSO | AATPsq |
| | (1) | (2) | (3) |
| $\Delta\log(\text{GDP per capita})_{i,t-1}$ | $0.120^{***}$ | $0.121^{***}$ | $0.121^{***}$ |
| | (0.011) | (0.011) | (0.011) |
| Temp | $0.006^{**}$ | | $0.009^{***}$ |
| | (0.003) | | (0.003) |
| $\text{Temp}^2$ | $-0.0003^{***}$ | $-0.0002^{***}$ | $-0.0003^{***}$ |
| | (0.0001) | (0.00005) | (0.0001) |
| R1mm | $0.0001^{**}$ | | |
| | (0.00004) | | |
| Rx1day | $-0.0002^{**}$ | $-0.0001^{**}$ | |
| | (0.0001) | (0.0001) | |
| TNx | $0.002^{***}$ | $0.002^{**}$ | |
| | (0.001) | (0.001) | |
| TR | $-0.0002^{**}$ | | |
| | (0.0001) | | |
| TX10p | $-0.0004^{**}$ | $-0.0003^{**}$ | |
| | (0.0002) | (0.0002) | |
| Prcp. | | | 0.001 |
| | | | (0.001) |
| $\text{Prcp}^2$ | | | -0.00004 |
| | | | (0.00003) |
| Time Fixed Effects | Yes | Yes | Yes |
| Country Fixed Effects | Yes | Yes | Yes |
| Country-Time Trends | Yes | Yes | Yes |
| Squared Country-Time Trends | Yes | Yes | Yes |
| Impulse Indicator Saturation | Yes | Yes | Yes |
| Observations | 6,536 | 6,536 | 6,536 |
| Residual Std. Error | 0.039 (df = 5876) | 0.039 (df = 5879) | 0.039 (df = 5879) |
| F Statistic | $17.873^{***}$ (df = 660; 5876) | $17.886^{***}$ (df = 657; 5879) | $17.881^{***}$ (df = 657; 5879) |
| *Note:* | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table A4. Standard Regression Model Results for Bayesian Model Selection with a Uniform Prior. Output generated using the *coef.bma()* command in the BMS R-package (Zeugner and Feldkircher, 2015). Here, 'PIP' refers to the posterior inclusion probability, 'Post Mean' and 'Post SD' refer to the posterior expected value of coefficients and standard deviation and 'cond. Pos. Sign' refers to the ratio of how often the coefficients' expected values were positive conditional on inclusion.

| **Bayesian Model Selection - Uniform Prior** | | | | |
|---|---|---|---|---|
| | PIP | Post Mean | Post SD | cond. Pos. Sign |
| $\Delta\log(\text{GDP per capita})_{i,t-1}$ | 1 | 0.121 | 0.010 | 1 |
| $\text{Temp}^2$ | 0.133 | -0.148 | 0.434 | 0 |
| Temp | 0.089 | 0.108 | 0.360 | 1 |
| TR | 0.076 | -0.032 | 0.121 | 0 |
| ID | 0.031 | 0.0002 | 0.011 | 0.138 |
| TN90p | 0.024 | -0.001 | 0.004 | 0 |
| R95p | 0.023 | -0.001 | 0.005 | 0 |
| $\text{TN90p}^2$ | 0.021 | -0.001 | 0.005 | 0 |
| R99p | 0.019 | -0.001 | 0.004 | 0 |
| $\text{TX90p}^2$ | 0.018 | -0.001 | 0.004 | 0 |
| GSL | 0.018 | 0.001 | 0.013 | 1 |
| WSDI | 0.017 | -0.0003 | 0.003 | 0 |
| TXx | 0.013 | 0.001 | 0.009 | 1 |
| Rx1day | 0.011 | -0.001 | 0.006 | 0 |
| TNx | 0.005 | 0.001 | 0.013 | 1 |
| TX90p | 0.004 | -0.0001 | 0.002 | 0 |
| SU | 0.002 | -0.0003 | 0.009 | 0 |

Table A5. Standard Regression Model Results for Bayesian Model Selection with a Fixed Prior. Output generated using the *coef.bma()* command in the BMS R-package (Zeugner and Feldkircher, 2015). Here, 'PIP' refers to the posterior inclusion probability, 'Post Mean' and 'Post SD' refer to the posterior expected value of coefficients and standard deviation and 'cond. Pos. Sign' refers to the ratio of how often the coefficients' expected values were positive conditional on inclusion.

| **Bayesian Model Selection - Fixed Prior** | | | | |
|---|---|---|---|---|
| | PIP | Post Mean | Post SD | cond. Pos. Sign |
| $\Delta\log(\text{GDP per capita})_{i,t-1}$ | 1 | 0.121 | 0.010 | 1 |
| $\text{Rx1day}^2$ | 0.301 | -0.017 | 0.028 | 0 |
| $\text{TN90p}^2$ | 0.198 | -0.007 | 0.015 | 0 |
| TX10p | 0.047 | -0.001 | 0.007 | 0 |
| $\text{Temp}^2$ | 0.043 | -0.037 | 0.205 | 0 |
| TR | 0.039 | -0.018 | 0.093 | 0 |
| $\text{Prcp}^2$ | 0.030 | -0.001 | 0.008 | 0 |
| Rx1day | 0.020 | -0.001 | 0.007 | 0 |
| PRCPTOT | 0.016 | 0.001 | 0.011 | 1 |
| Temp | 0.014 | 0.014 | 0.125 | 1 |
| $\text{Rx5day}^2$ | 0.010 | -0.0003 | 0.004 | 0 |
| TNx | 0.009 | 0.001 | 0.017 | 1 |
| R10mm | 0.006 | 0.0001 | 0.004 | 1 |
| $\text{PRCPTOT}^2$ | 0.003 | -0.0001 | 0.002 | 0 |
| SDII | 0.002 | 0.0001 | 0.003 | 1 |
| TN90p | 0.002 | -0.0001 | 0.001 | 0 |
| TX90p | 0.001 | 0.00001 | 0.001 | 1 |

Table A6. Standard Regression Model Results for Bayesian Model Selection with a PIP Prior. Output generated using the *coef.bma()* command in the BMS R-package (Zeugner and Feldkircher, 2015). Here, 'PIP' refers to the posterior inclusion probability, 'Post Mean' and 'Post SD' refer to the posterior expected value of coefficients and standard deviation and 'cond. Pos. Sign' refers to the ratio of how often the coefficients' expected values were positive conditional on inclusion.

| **Bayesian Model Selection - PIP Prior** | | | | |
|---|---|---|---|---|
| | PIP | Post Mean | Post SD | cond. Pos. Sign |
| $\Delta\log(\text{GDP per capita})_{i,t-1}$ | 1 | 0.121 | 0.010 | 1 |
| $Rx1day^2$ | 0.114 | -0.007 | 0.024 | 0 |
| TR | 0.072 | -0.031 | 0.116 | 0 |
| Rx1day | 0.052 | -0.001 | 0.017 | 0.179 |
| SU | 0.047 | -0.008 | 0.042 | 0 |
| TNx | 0.034 | 0.004 | 0.026 | 1 |
| WSDI | 0.027 | -0.001 | 0.005 | 0 |
| TX90p | 0.019 | -0.0003 | 0.003 | 0 |
| $TX10p^2$ | 0.019 | -0.0003 | 0.004 | 0 |
| $CWD^2$ | 0.008 | 0.0001 | 0.002 | 1 |
| $Temp^2$ | 0.007 | -0.003 | 0.041 | 0 |
| $R20mm^2$ | 0.007 | -0.0001 | 0.002 | 0 |

## 3.2 Adaptation Model Results

Table A7. Regression Model Results for gets and LASSO models incorporating Adaptation.

| **Adaptation Model Results** | | | | | | |
|---|---|---|---|---|---|---|
| | *Dependent variable:* | | | | | |
| | delta log GDP per capita | | | | | |
| | gets First | LASSO First | gets Last | LASSO Last | gets Income Adaptation | LASSO Income Adaptation |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| L1.diff.ln_gdp_cap | 0.069*** | 0.069*** | 0.089*** | 0.091*** | 0.080*** | 0.081*** |
| | (0.016) | (0.016) | (0.012) | (0.012) | (0.011) | (0.011) |
| temp | 0.002 | | 0.007** | | -0.005 | |
| | (0.005) | | (0.003) | | (0.013) | |
| temp_2 | -0.0003** | -0.0003*** | -0.0003*** | -0.0002*** | -0.001* | -0.001*** |
| | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0004) | (0.0002) |
| R1mm | 0.0001** | | 0.0001** | | 0.0004* | |
| | (0.0001) | | (0.0001) | | (0.0002) | |
| Rx1day | -0.0003** | -0.0002** | -0.0002*** | -0.0002** | -0.0004 | -0.0003 |
| | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0004) | (0.0004) |
| TNx | 0.001 | -0.0002 | 0.002** | 0.002** | -0.013*** | -0.016*** |
| | (0.002) | (0.002) | (0.001) | (0.001) | (0.005) | (0.003) |
| TR | -0.0001 | | -0.0001 | | 0.001 | |
| | (0.0001) | | (0.0001) | | (0.0003) | |
| TX10p | -0.001*** | -0.001*** | -0.0002 | -0.0003 | -0.003*** | -0.001 |
| | (0.0003) | (0.0003) | (0.0002) | (0.0002) | (0.001) | (0.001) |
| temp_gdppc_int | | | | | 0.001 | |
| | | | | | (0.001) | |
| temp_2_gdppc_int | | | | | 0.0001 | 0.0001*** |
| | | | | | (0.00005) | (0.00002) |
| R1mm_gdppc_int | | | | | -0.00004 | |
| | | | | | (0.00003) | |
| Rx1day_gdppc_int | | | | | 0.00003 | 0.00001 |
| | | | | | (0.00005) | (0.00005) |
| TNx_gdppc_int | | | | | 0.002*** | 0.002*** |
| | | | | | (0.001) | (0.0004) |
| TR_gdppc_int | | | | | -0.0001** | |
| | | | | | (0.00004) | |
| TX10p_gdppc_int | | | | | 0.0003** | 0.0001 |
| | | | | | (0.0001) | (0.0001) |
| Time Fixed Effects | Yes | Yes | Yes | Yes | Yes | Yes |
| Country Fixed Effects | Yes | Yes | Yes | Yes | Yes | Yes |
| Country-Time Trends | Yes | Yes | Yes | Yes | Yes | Yes |
| Squared Country-Time Trends | Yes | Yes | Yes | Yes | Yes | Yes |
| Impulse Indicator Saturation | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 3,444 | 3,444 | 4,644 | 4,644 | 6,536 | 6,536 |
| Residual Std. Error | 0.042 (df = 2937) | 0.042 (df = 2940) | 0.037 (df = 4030) | 0.037 (df = 4033) | 0.039 (df = 5869) | 0.039 (df = 5875) |
| F Statistic | 10.196*** (df = 507; 2937) | 10.232*** (df = 504; 2940) | 17.038*** (df = 614; 4030) | 17.075*** (df = 611; 4033) | 18.524*** (df = 667; 5869) | 18.604*** (df = 661; 5875) |
| *Note:* | | | | | | *p<0.1; **p<0.05; ***p<0.01 |

# 4 CMIP5 Climate Models Overview

Table A8. CMIP5 Model runs used in the projections (Part 1).

| CMIP5 Models Considered and Selected | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | RCP | Ensemble | End of Century Temperature | gets | LASSO | AATPsq | BMS-Uniform | BMS-Fixed | BMS-PIP |
| MRI-CGCM3 | 2.6 | r1i1p1 | 1.355 | Y | Y | Y | Y | Y | Y |
| NorESM1-M | 2.6 | r1i1p1 | 1.378 | N | N | N | N | N | N |
| MIROC5 | 2.6 | r2i1p1 | 1.441 | Y | Y | Y | Y | Y | Y |
| MIROC5 | 2.6 | r1i1p1 | 1.491 | Y | Y | Y | Y | Y | Y |
| MPI-ESM-LR | 2.6 | r1i1p1 | 1.619 | Y | Y | Y | Y | Y | Y |
| MPI-ESM-LR | 2.6 | r3i1p1 | 1.637 | Y | Y | Y | Y | Y | Y |
| MPI-ESM-MR | 2.6 | r1i1p1 | 1.653 | Y | Y | Y | Y | Y | Y |
| MIROC5 | 2.6 | r3i1p1 | 1.669 | Y | Y | Y | Y | Y | Y |
| CNRM-CM5 | 2.6 | r1i1p1 | 1.679 | Y | Y | Y | Y | Y | Y |
| MPI-ESM-LR | 2.6 | r2i1p1 | 1.697 | Y | Y | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 2.6 | r4i1p1 | 1.737 | N | N | Y | Y | Y | Y |
| bcc-csm1-1-m | 2.6 | r1i1p1 | 1.78 | Y | Y | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 2.6 | r8i1p1 | 1.804 | N | N | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r4i1p2 | 1.809 | N | N | Y | Y | Y | Y |
| CCSM4 | 2.6 | r1i1p1 | 1.817 | N | N | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r1i1p1 | 1.817 | N | N | Y | Y | Y | Y |
| CCSM4 | 2.6 | r3i1p1 | 1.818 | N | N | Y | Y | Y | Y |
| inmcm4 | 4.5 | r1i1p1 | 1.832 | Y | Y | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r5i1p2 | 1.837 | N | N | Y | Y | Y | Y |
| EC-EARTH | 4.5 | r7i1p1 | 1.838 | N | N | Y | Y | Y | Y |
| CCSM4 | 2.6 | r5i1p1 | 1.847 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 2.6 | r3i1p1 | 1.849 | N | N | Y | Y | Y | Y |
| bcc-csm1-1 | 2.6 | r1i1p1 | 1.858 | N | N | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r3i1p2 | 1.859 | N | N | Y | Y | Y | Y |
| CCSM4 | 2.6 | r4i1p1 | 1.86 | N | N | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r1i1p2 | 1.862 | N | N | Y | Y | Y | Y |
| EC-EARTH | 4.5 | r14i1p1 | 1.865 | N | N | Y | Y | Y | Y |
| EC-EARTH | 4.5 | r5i1p1 | 1.874 | N | N | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r2i1p1 | 1.874 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 2.6 | r2i1p1 | 1.877 | N | N | Y | Y | Y | Y |
| CCSM4 | 2.6 | r2i1p1 | 1.878 | Y | Y | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r4i1p1 | 1.885 | N | N | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r6i1p1 | 1.887 | Y | Y | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r3i1p1 | 1.914 | N | N | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r2i1p2 | 1.937 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 2.6 | r5i1p1 | 1.938 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 2.6 | r10i1p1 | 1.94 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 2.6 | r7i1p1 | 1.94 | N | N | Y | Y | Y | Y |
| IPSL-CM5A-MR | 2.6 | r1i1p1 | 1.967 | Y | Y | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 2.6 | r1i1p1 | 1.977 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 2.6 | r9i1p1 | 1.984 | N | N | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r5i1p1 | 1.993 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 2.6 | r6i1p1 | 2.054 | N | N | Y | Y | Y | Y |
| MRI-CGCM3 | 4.5 | r1i1p1 | 2.12 | Y | Y | Y | Y | Y | Y |
| MIROC5 | 4.5 | r1i1p1 | 2.214 | Y | Y | Y | Y | Y | Y |
| IPSL-CM5A-LR | 2.6 | r2i1p1 | 2.215 | Y | Y | Y | Y | Y | Y |
| IPSL-CM5A-LR | 2.6 | r3i1p1 | 2.233 | Y | Y | Y | Y | Y | Y |
| MIROC5 | 4.5 | r2i1p1 | 2.239 | Y | Y | Y | Y | Y | Y |
| IPSL-CM5A-LR | 2.6 | r4i1p1 | 2.248 | Y | Y | Y | Y | Y | Y |

Table A9. CMIP5 Model runs used in the projections (Part 2).

| CMIP5 Models Considered and Selected | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | RCP | Ensemble | End of Century Temperature | gets | LASSO | AATPsq | BMS-Uniform | BMS-Fixed | BMS-PIP |
| IPSL-CM5A-LR | 2.6 | r1i1p1 | 2.279 | Y | Y | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r1i1p3 | 2.335 | N | N | Y | Y | Y | Y |
| MIROC-ESM-CHEM | 2.6 | r1i1p1 | 2.351 | N | N | Y | Y | Y | Y |
| MRI-CGCM3 | 6 | r1i1p1 | 2.351 | Y | Y | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r4i1p3 | 2.355 | N | N | Y | Y | Y | Y |
| MIROC5 | 4.5 | r3i1p1 | 2.357 | Y | Y | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r6i1p3 | 2.407 | Y | Y | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r2i1p3 | 2.414 | N | N | Y | Y | Y | Y |
| MIROC-ESM | 2.6 | r1i1p1 | 2.455 | Y | Y | Y | Y | Y | Y |
| GISS-E2-R | 4.5 | r3i1p3 | 2.474 | N | N | Y | Y | Y | Y |
| bcc-csm1-1 | 4.5 | r1i1p1 | 2.478 | N | N | Y | Y | Y | Y |
| bcc-csm1-1-m | 4.5 | r1i1p1 | 2.483 | Y | Y | Y | Y | Y | Y |
| MPI-ESM-LR | 4.5 | r1i1p1 | 2.501 | Y | Y | Y | Y | Y | Y |
| MPI-ESM-MR | 4.5 | r3i1p1 | 2.516 | Y | Y | Y | Y | Y | Y |
| MIROC5 | 6 | r1i1p1 | 2.524 | Y | Y | Y | Y | Y | Y |
| IPSL-CM5B-LR | 4.5 | r1i1p1 | 2.546 | Y | Y | Y | Y | Y | Y |
| ACCESS1-3 | 4.5 | r1i1p1 | 2.568 | N | N | N | N | N | N |
| CNRM-CM5 | 4.5 | r1i1p1 | 2.579 | Y | Y | Y | Y | Y | Y |
| MPI-ESM-MR | 4.5 | r1i1p1 | 2.586 | Y | Y | Y | Y | Y | Y |
| CCSM4 | 4.5 | r3i1p1 | 2.588 | N | N | Y | Y | Y | Y |
| MPI-ESM-LR | 4.5 | r3i1p1 | 2.613 | Y | Y | Y | Y | Y | Y |
| MPI-ESM-MR | 4.5 | r2i1p1 | 2.616 | Y | Y | Y | Y | Y | Y |
| CCSM4 | 4.5 | r5i1p1 | 2.625 | N | N | Y | Y | Y | Y |
| CCSM4 | 4.5 | r1i1p1 | 2.639 | Y | Y | Y | Y | Y | Y |
| MIROC5 | 6 | r3i1p1 | 2.657 | N | N | Y | Y | Y | Y |
| MPI-ESM-LR | 4.5 | r2i1p1 | 2.676 | Y | Y | Y | Y | Y | Y |
| MIROC5 | 6 | r2i1p1 | 2.678 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 4.5 | r8i1p1 | 2.692 | N | N | Y | Y | Y | Y |
| CCSM4 | 4.5 | r2i1p1 | 2.699 | Y | Y | Y | Y | Y | Y |
| CCSM4 | 4.5 | r4i1p1 | 2.701 | N | N | Y | Y | Y | Y |
| EC-EARTH | 4.5 | r9i1p1 | 2.705 | N | N | Y | Y | Y | Y |
| EC-EARTH | 4.5 | r1i1p1 | 2.714 | N | N | N | N | N | N |
| CMCC-CMS | 4.5 | r1i1p1 | 2.717 | Y | Y | Y | Y | Y | Y |
| EC-EARTH | 4.5 | r12i1p1 | 2.733 | N | N | Y | Y | Y | Y |
| ACCESS1-0 | 4.5 | r1i1p1 | 2.755 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 4.5 | r1i1p1 | 2.756 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 4.5 | r9i1p1 | 2.756 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 4.5 | r2i1p1 | 2.767 | N | N | Y | Y | Y | Y |
| EC-EARTH | 4.5 | r2i1p1 | 2.791 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 4.5 | r7i1p1 | 2.792 | N | N | Y | Y | Y | Y |
| EC-EARTH | 4.5 | r8i1p1 | 2.813 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 4.5 | r4i1p1 | 2.814 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 4.5 | r3i1p1 | 2.837 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 6 | r2i1p1 | 2.856 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 6 | r7i1p1 | 2.892 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 6 | r9i1p1 | 2.901 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 4.5 | r6i1p1 | 2.908 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 4.5 | r5i1p1 | 2.928 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 6 | r3i1p1 | 2.976 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 4.5 | r10i1p1 | 2.984 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 6 | r5i1p1 | 2.985 | N | N | Y | Y | Y | Y |

Table A10. CMIP5 Model runs used in the projections (Part 3).

| CMIP5 Models Considered and Selected | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | RCP | Ensemble | End of Century Temperature | gets | LASSO | AATPsq | BMS-Uniform | BMS-Fixed | BMS-PIP |
| bcc-csm1-1-m | 6 | r1i1p1 | 3.001 | Y | Y | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 6 | r8i1p1 | 3.01 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 6 | r1i1p1 | 3.028 | N | N | Y | Y | Y | Y |
| bcc-csm1-1 | 6 | r1i1p1 | 3.032 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 6 | r4i1p1 | 3.039 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 6 | r6i1p1 | 3.044 | N | N | Y | Y | Y | Y |
| CCSM4 | 6 | r3i1p1 | 3.12 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 6 | r10i1p1 | 3.12 | N | N | Y | Y | Y | Y |
| CCSM4 | 6 | r4i1p1 | 3.164 | N | N | Y | Y | Y | Y |
| CCSM4 | 6 | r1i1p1 | 3.169 | Y | Y | Y | Y | Y | Y |
| MIROC-ESM-CHEM | 4.5 | r1i1p1 | 3.203 | N | N | Y | Y | Y | Y |
| IPSL-CM5A-LR | 4.5 | r2i1p1 | 3.212 | Y | Y | Y | Y | Y | Y |
| IPSL-CM5A-LR | 4.5 | r4i1p1 | 3.214 | Y | Y | Y | Y | Y | Y |
| CCSM4 | 6 | r2i1p1 | 3.224 | Y | Y | Y | Y | Y | Y |
| MIROC-ESM | 4.5 | r1i1p1 | 3.229 | Y | Y | Y | Y | Y | Y |
| BNU-ESM | 4.5 | r1i1p1 | 3.238 | Y | Y | Y | Y | Y | Y |
| IPSL-CM5A-LR | 4.5 | r1i1p1 | 3.273 | Y | Y | Y | Y | Y | Y |
| IPSL-CM5A-LR | 4.5 | r3i1p1 | 3.303 | Y | Y | Y | Y | Y | Y |
| inmcm4 | 8.5 | r1i1p1 | 3.358 | Y | Y | Y | Y | Y | Y |
| IPSL-CM5A-MR | 6 | r1i1p1 | 3.603 | Y | Y | Y | Y | Y | Y |
| IPSL-CM5A-LR | 6 | r1i1p1 | 3.743 | Y | Y | Y | Y | Y | Y |
| MIROC-ESM | 6 | r1i1p1 | 3.796 | Y | Y | Y | Y | Y | Y |
| MIROC-ESM-CHEM | 6 | r1i1p1 | 3.852 | N | N | Y | Y | Y | Y |
| MRI-CGCM3 | 8.5 | r1i1p1 | 3.867 | Y | Y | Y | Y | Y | Y |
| MIROC5 | 8.5 | r1i1p1 | 3.986 | Y | Y | Y | Y | Y | Y |
| MIROC5 | 8.5 | r3i1p1 | 4.047 | Y | Y | Y | Y | Y | Y |
| MIROC5 | 8.5 | r2i1p1 | 4.124 | Y | Y | Y | Y | Y | Y |
| CNRM-CM5 | 8.5 | r2i1p1 | 4.3 | N | N | Y | Y | Y | Y |
| CNRM-CM5 | 8.5 | r4i1p1 | 4.312 | N | N | Y | Y | Y | Y |
| CNRM-CM5 | 8.5 | r1i1p1 | 4.354 | Y | Y | Y | Y | Y | Y |
| CNRM-CM5 | 8.5 | r6i1p1 | 4.424 | N | N | Y | Y | Y | Y |
| IPSL-CM5B-LR | 8.5 | r1i1p1 | 4.473 | Y | Y | Y | Y | Y | Y |
| bcc-csm1-1 | 8.5 | r1i1p1 | 4.548 | N | N | N | N | N | N |
| CCSM4 | 8.5 | r3i1p1 | 4.555 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 8.5 | r2i1p1 | 4.611 | N | N | Y | Y | Y | Y |
| ACCESS1-3 | 8.5 | r1i1p1 | 4.634 | N | N | Y | Y | Y | Y |
| MPI-ESM-MR | 8.5 | r1i1p1 | 4.65 | Y | Y | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 8.5 | r8i1p1 | 4.666 | N | N | Y | Y | Y | Y |
| CCSM4 | 8.5 | r6i1p1 | 4.678 | Y | Y | Y | Y | Y | Y |
| MPI-ESM-LR | 8.5 | r1i1p1 | 4.694 | Y | Y | Y | Y | Y | Y |
| CCSM4 | 8.5 | r5i1p1 | 4.695 | N | N | Y | Y | Y | Y |
| CCSM4 | 8.5 | r4i1p1 | 4.713 | N | N | Y | Y | Y | Y |
| MPI-ESM-LR | 8.5 | r3i1p1 | 4.715 | Y | Y | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 8.5 | r9i1p1 | 4.717 | N | N | Y | Y | Y | Y |
| CCSM4 | 8.5 | r2i1p1 | 4.725 | Y | Y | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 8.5 | r4i1p1 | 4.742 | N | N | Y | Y | Y | Y |

Table A11. CMIP5 Model runs used in the projections (Part 4).

| CMIP5 Models Considered and Selected | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | RCP | Ensemble | End of Century Temperature | gets | LASSO | AATPsq | BMS-Uniform | BMS-Fixed | BMS-PIP |
| CSIRO-Mk3-6-0 | 8.5 | r7i1p1 | 4.806 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 8.5 | r1i1p1 | 4.808 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 8.5 | r5i1p1 | 4.827 | N | N | Y | Y | Y | Y |
| CMCC-CESM | 8.5 | r1i1p1 | 4.832 | Y | Y | Y | Y | Y | Y |
| MPI-ESM-LR | 8.5 | r2i1p1 | 4.833 | Y | Y | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 8.5 | r3i1p1 | 4.84 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 8.5 | r10i1p1 | 4.847 | N | N | Y | Y | Y | Y |
| CSIRO-Mk3-6-0 | 8.5 | r6i1p1 | 4.872 | N | N | Y | Y | Y | Y |
| ACCESS1-0 | 8.5 | r1i1p1 | 4.924 | N | N | Y | Y | Y | Y |
| CMCC-CMS | 8.5 | r1i1p1 | 5.011 | Y | Y | Y | Y | Y | Y |
| IPSL-CM5A-MR | 8.5 | r1i1p1 | 5.501 | Y | Y | Y | Y | Y | Y |
| CanESM2 | 8.5 | r1i1p1 | 5.506 | Y | Y | Y | Y | Y | Y |
| CanESM2 | 8.5 | r4i1p1 | 5.514 | Y | Y | Y | Y | Y | Y |
| bcc-csm1-1-m | 8.5 | r1i1p1 | 5.55 | Y | Y | Y | Y | Y | Y |
| CanESM2 | 8.5 | r3i1p1 | 5.552 | Y | Y | Y | Y | Y | Y |
| CanESM2 | 8.5 | r2i1p1 | 5.56 | Y | Y | Y | Y | Y | Y |
| BNU-ESM | 8.5 | r1i1p1 | 5.589 | Y | Y | Y | Y | Y | Y |
| CanESM2 | 8.5 | r5i1p1 | 5.628 | Y | Y | Y | Y | Y | Y |
| MIROC-ESM | 8.5 | r1i1p1 | 5.641 | Y | Y | Y | Y | Y | Y |
| IPSL-CM5A-LR | 8.5 | r3i1p1 | 5.668 | Y | Y | Y | Y | Y | Y |
| IPSL-CM5A-LR | 8.5 | r2i1p1 | 5.689 | Y | Y | Y | Y | Y | Y |
| IPSL-CM5A-LR | 8.5 | r4i1p1 | 5.713 | Y | Y | Y | Y | Y | Y |
| IPSL-CM5A-LR | 8.5 | r1i1p1 | 5.803 | Y | Y | Y | Y | Y | Y |
| MIROC-ESM-CHEM | 8.5 | r1i1p1 | 5.912 | N | N | Y | Y | Y | Y |

# 5 Modelling Strategy Overview

Our modelling strategy is outlined in Table 1 and described in detail in the following sections.

| Table A12. Modelling Strategy to express economic impacts as a function of cumulative carbon emissions. | | |
|---|---|---|
| Link economic impacts to climate variables ($Climate$). | $\Delta\log\left(GDPpc\right)_{i,t} = f(Climate_{i,t}, t)$ | Estimate *f()* using model selection and machine learning, account for adaptation over time $t$ using in-sample rolling window estimates and out-of-sample exponential smoothing forecasts. |
| Link climate variables to Global Mean Surface Temperature (*GMST*) | $Climate_{i,t} = h(GMST_t)$ | Estimate *h()* using CMIP5Projections. |
| Link GMST to cumulative emissions (*CCe*). | $GMST_t = g(CCe_t)$: | Using TCRE estimates from Goodwin et al. (2015) |

# 6 Detailed Projection Results

The end-of-century median GDP per capita damage function can be shown for several functional forms. In the main paper, we show a quadratic formulation. In equation A1 and A2 we present a linear and a cubic specification as well.

$$\frac{GDPpc_{Climate} - GDPpc_{Baseline}}{GDPpc_{Baseline}} = -0.062 - -0.113\,Temp, \text{ for } Temp \in [1.35°C, 5.9°C] \quad \text{(A1)}$$

$$\frac{GDPpc_{Climate} - GDPpc_{Baseline}}{GDPpc_{Baseline}} = -0.182 - 0.024\,Temp - 0.047\,Temp^2 + 0.005 * Temp^3, \text{ for } Temp \in [1.35°C, 5.9°C] \quad \text{(A2)}$$

Quantile regression results, which underlie the construction of equations 1 as well as A1 and A2, can be found in Table A13 for gets selected projections and in Table A14 in LASSO selected models.

Table A13. Quantile regression results for gets selected projections without adaptation.

**gets Projection Results**

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Median Difference to Baseline | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| Temperature Anomaly | -0.113*** | -0.140*** | 0.024*** | -0.088*** | 0.387*** |
| | (0.0003) | (0.002) | (0.009) | (0.006) | (0.045) |
| Temperature Anomaly Squared | | 0.004*** | -0.047*** | -0.015*** | -0.064*** |
| | | (0.0003) | (0.003) | (0.002) | (0.015) |
| Temperature Anomaly Cubed | | | 0.005*** | 0.003*** | 0.005*** |
| | | | (0.0003) | (0.0002) | (0.002) |
| Constant | -0.062*** | -0.021*** | -0.182*** | -0.468*** | -0.187*** |
| | (0.001) | (0.003) | (0.009) | (0.007) | (0.042) |
| Percentile | 50% | 50% | 50% | 5% | 95% |
| Observations | 858,000 | 858,000 | 858,000 | 858,000 | 858,000 |
| *Note:* | | | | | *p<0.1; **p<0.05; ***p<0.01 |

Table A14. Quantile regression results for LASSO selected projections without adaptation.

**LASSO Projection Results**

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Relative Difference to Baseline | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| Temperature Anomaly | -0.092*** | -0.083*** | 0.018*** | 0.028*** | -0.027** |
| | (0.0003) | (0.001) | (0.006) | (0.008) | (0.013) |
| Temperature Anomaly Squared | | -0.001*** | -0.033*** | -0.048*** | 0.007* |
| | | (0.0002) | (0.002) | (0.002) | (0.004) |
| Temperature Anomaly Cubed | | | 0.003*** | 0.005*** | -0.001** |
| | | | (0.0002) | (0.0002) | (0.0004) |
| Constant | -0.033*** | -0.047*** | -0.145*** | -0.419*** | 0.127*** |
| | (0.001) | (0.002) | (0.006) | (0.008) | (0.013) |
| Percentile | 50% | 50% | 50% | 5% | 95% |
| Observations | 858,000 | 858,000 | 858,000 | 858,000 | 858,000 |
| *Note:* | | | | | *p<0.1; **p<0.05; ***p<0.01 |

# 7 Comparison to Burke, Hsiang, Miguel 2015

The projection method in our paper differs from Burke *et al.* (2015; BHM) in some regards. Specifically, these differences include:

BHM estimate their global median as the median estimate of all bootstrap estimates where they sum global GDP divided by the sum of global population to arrive at a median GDP per capita impact. We use the $median\left(\frac{\sum_i^N GDP_{i,m}}{\sum_i^N Population_{i,m}}\right)$ where $m$ is the bootstrap replication used. In contrast, we use a median quantile regression estimate of the *country level* GDP per capita estimates $quantileRegr_{0.5}\left(\frac{GDP_i}{Population_i}\right)$ to give a more sensible estimate of the effects of a median country. This is partly due to the data from Mueller, Stock and Watson not featuring population projections but only GDP per capita projections, not allowing us to disentangle population effects. We carry out a BHM compatible calculation for SSP projections below, however.

Furthermore, BHM project their damage function by imposing a 0% difference at a 0°C anomaly. While this of course makes theoretical sense (with no warming there would be no impacts), the inclusion of this data point imposes a functional form of the damage function to go through 0. In reality though, under no conceivable scenario would the temperature anomaly at the end of the century be able to be 0°C. In fact, no CMIP5 model is capable of producing an anomaly even close to that. We therefore only estimate our damage function across the observable CMIP5 range (starting at 1.35°C). This allows the quantile regression, which we use to estimate the median through our models at every point on the temperature projection, to be functionally more flexible.

In their projections, BHM base their GDP projections on the mean GDP per capita between 1980 and 2010. In contrast, we use GDP per capita until the end of our sample 2011. This change does not affect the relative differences for impacts affects the level of our estimates. BHM also do not project estimates out of sample (any future temperature that would be higher than 30°C is set as 30°C) while we do not impose a similar constraint.

In Figure A6, we show our estimates compared to BHM results. In this figure, we use BHM's method of aggregating global estimates but only include our damage function for the range observed in CMIP5 models.

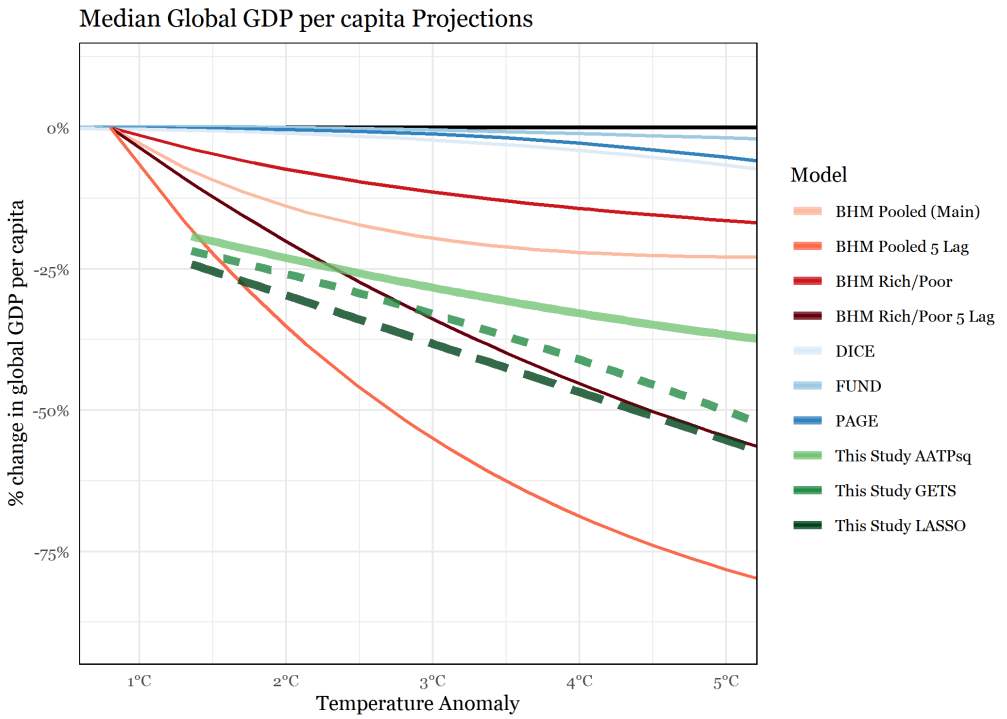


**Figure A6:** Comparison of impacts to earlier estimates. BHM models are taken from Burke *et al.* (2015) and correspond to their main model using a pooled estimator (Pooled Main), to a model with a pooled estimator but including 5-time lags to approximate a long-term response (Pooled 5 Lag). The Rich/Poor BHM models allows rich/poor countries to respond differently to temperature shocks. The DICE, FUND, and PAGE curves are taken from each of their models as processed by Revesz *et al.* (2014). The three green lines are taken from this study and correspond to the naïve model estimating using Annual Averages of Temperature and Precipitation squared (AATPsq), to the model estimated using general-to-specific model selection (GETS) and using the least absolute shrinkage and selection operator (LASSO) model selection method. Estimates from this study are only displayed across the plausible CMIP5 range, while all other curves are forced to cross 0°C temperature change, although this is not plausible in 2100.

# 8 Alternative Estimation of the Adaptation Model

We also consider two alternative forms of estimation to further investigate the robustness of our results.

We first consider whether adding the income interactions *before* selection rather than post-selection considerably changes our estimates. We therefore interact all climate variables with income before re-selecting a gets model and the project the resulting estimates again.

Second, to alleviate possible concerns about endogeneity due to interacting climate with contemporaneous income, we also consider a model where we interact the selected climate variables with the lagged incomes. Subsequently we again project the results to the end of the century.

Both alternative estimation results are contained in Figure A7 below and results show that the estimates are remarkably similar to the standard Adaptation Model Projections.

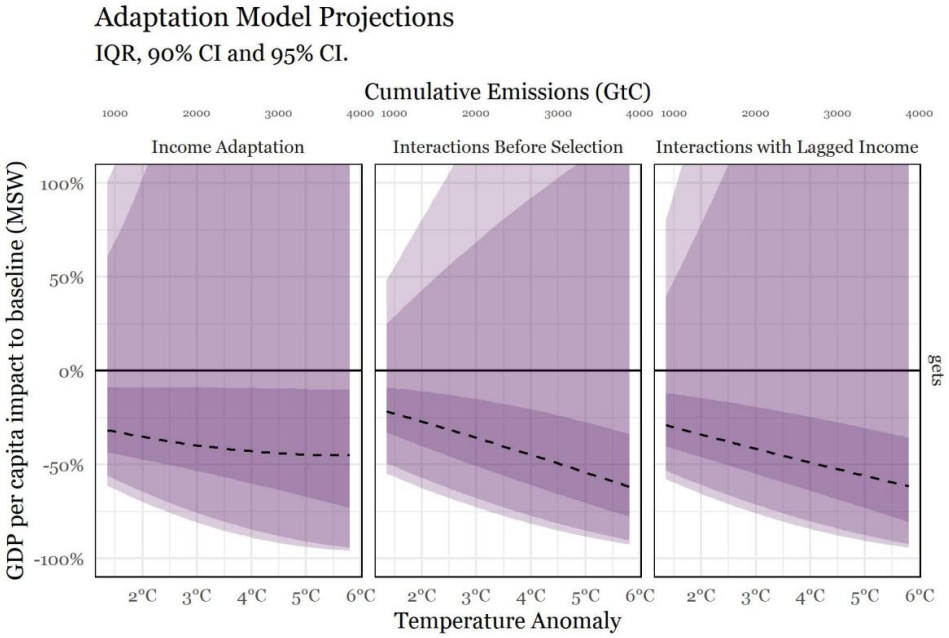


**Figure A7:** Comparison of impacts under different adaptation models.

# 9 Reference List


Burke, Marshall; Hsiang, Solomon M. and Miguel, Edward (2015), Global non-linear effect of temperature on economic production, *Nature*, Vol. 527 No. 7577, pp. 235–239.

CIESIN (2016), Gridded Population of the World, Version 4 (GPWv4): Population Count: *http://dx.doi.org/10.7927/H4X63JVC. Accessed DAY MONTH YEAR.,* Center for International Earth Science Information Network - Columbia University - NASA Socioeconomic Data and Applications Center (SEDAC), Series: Palisades, NY.

Dellink, Rob; Chateau, Jean; Lanzi, Elisa and Magné, Bertrand (2017), Long-term economic growth projections in the Shared Socioeconomic Pathways, *Global Environmental Change*, Vol. 42, pp. 200–214.

Donat, M. G; Alexander, L. V; Yang, H; Durre, I; Vose, R; Dunn, R. J.H; Willett, K. M; Aguilar, E; Brunet, M; Caesar, J; Hewitson, B; Jack, C; Klein Tank, A. M.G; Kruger, A. C; Marengo, J; Peterson, T. C; Renom, M; Oria Rojas, C; Rusticucci, M; Salinger, J; Elrayah, A. S; Sekele, S. S; Srivastava, A. K; Trewin, B; Villarroel, C; Vincent, L. A; Zhai, P; Zhang, X. and Kitching, S. (2013), Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century. The HadEX2 dataset, *Journal of Geophysical Research: Atmospheres*, Vol. 118 No. 5, pp. 2098–2118.

IIASA (2013), SSP Basic Elemensts data, IIASA, Series: 2013rd ed., available at: https://tntcat.iiasa.ac.at/SspDb/download/basic_elements/SspDb_country_data_2013-06-12.csv.zip (accessed 21 March 2019).

KC, Samir and Lutz, Wolfgang (2017), The human core of the shared socioeconomic pathways. Population scenarios by age, sex and level of education for all countries to 2100, *Global Environmental Change*, Vol. 42, pp. 181–192.

Matsuura, Kenji and Willmott, Cort J. (2018), Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1901 - 2017): *Version 5. Data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA,* University of Delaware, Series: available at: https://www.esrl.noaa.gov/psd/ (accessed 26 February 2019).

Revesz, Richard L; Howard, Peter H; Arrow, Kenneth; Goulder, Lawrence H; Kopp, Robert E; Livermore, Michael A; Oppenheimer, Michael and Sterner, Thomas (2014), Global warming. Improve economic models of climate change, *Nature*, Vol. 508 No. 7495, pp. 173–175.

Riahi, Keywan; van Vuuren, Detlef P; Kriegler, Elmar; Edmonds, Jae; O'Neill, Brian C; Fujimori, Shinichiro; Bauer, Nico; Calvin, Katherine; Dellink, Rob; Fricko, Oliver; Lutz, Wolfgang; Popp, Alexander; Cuaresma, Jesus C; KC, Samir; Leimbach, Marian; Jiang, Leiwen; Kram, Tom; Rao, Shilpa; Emmerling, Johannes; Ebi, Kristie; Hasegawa, Tomoko; Havlik, Petr; Humpenöder, Florian; Da Silva, Lara A; Smith, Steve; Stehfest, Elke; Bosetti, Valentina; Eom, Jiyong; Gernaat, David; Masui, Toshihiko; Rogelj, Joeri; Strefler, Jessica; Drouet, Laurent; Krey, Volker; Luderer, Gunnar; Harmsen, Mathijs; Takahashi, Kiyoshi; Baumstark, Lavinia; Doelman, Jonathan C; Kainuma, Mikiko; Klimont, Zbigniew; Marangoni, Giacomo; Lotze-Campen, Hermann; Obersteiner, Michael; Tabeau, Andrzej and Tavoni, Massimo (2017), The

Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications. An overview, *Global Environmental Change*, Vol. 42, pp. 153–168.

Sillmann, J; Kharin, V. V; Zhang, X; Zwiers, F. W. and Bronaugh, D. (2013a), Climate extremes indices in the CMIP5 multimodel ensemble. Part 1. Model evaluation in the present climate, *Journal of Geophysical Research: Atmospheres*, Vol. 118 No. 4, pp. 1716–1733.

Sillmann, J; Kharin, V. V; Zwiers, F. W; Zhang, X. and Bronaugh, D. (2013b), Climate extremes indices in the CMIP5 multimodel ensemble. Part 2. Future climate projections, *Journal of Geophysical Research: Atmospheres*, Vol. 118 No. 6, pp. 2473–2493.

Zeugner, Stefan and Feldkircher, Martin (2015), Bayesian Model Averaging Employing Fixed and Flexible Priors. The BMS Package for R, *Journal of Statistical Software*, Vol. 68 No. 4, pp. 1–37.

# Appendix for Paper 3

# Appendix

The appendix contains three distinct sections that aim to contextualise and validate results presented in the main paper.

Firstly in Section A, further summary statistics are presented for the data used. This is done both for the estimation input data as well as for the data that is subsequently used to put these results into context i.e. for the data considered from them EM-DAT database.

Secondly in Section B, the performance of the algorithms that were considered are presented.

Lastly in Section C, Variable Importance figures are presented. Those figures are meant to illustrate the relative importance of the different variables used in the model.

## A Further Summary Statistics

In this section, I present further summary statistics. In Table 3, I show the data that is primarily used in the estimation of the machine-learning modes. In Table 4, I show the summary data of the EM-DAT data used. This data is the basis for quantitatively contextualising the magnitude of the identified events that were previously missed.

| | Mean | SD | Min | Max | N |
|---|---|---|---|---|---|
| Drought | 0.053 | 0.224 | 0.000 | 1.000 | 266868 |
| Landslide | 0.007 | 0.085 | 0.000 | 1.000 | 266868 |
| Flood | 0.100 | 0.300 | 0.000 | 1.000 | 266868 |
| Storm | 0.058 | 0.234 | 0.000 | 1.000 | 266868 |
| Extreme Temperature | 0.045 | 0.208 | 0.000 | 1.000 | 266868 |
| Monthly Mean of Daily Mean Temperature | 9.410 | 17 | -46.013 | 41 | 266868 |
| Monthly Minimum of Daily Minimum Temperature | -1.302 | 20 | -60.046 | 29 | 266868 |
| Monthly Maximum of Daily Maximum Temperature | 21 | 16 | -34.275 | 50 | 266868 |
| Total Monthly Precipitation | 66 | 80 | -0.003 | 890 | 266868 |
| Maximum Daily Total Precipitation | 14 | 13 | 0.000 | 181 | 266868 |
| Minimum Daily Total Precipitation | 0.048 | 0.198 | 0.000 | 4.921 | 266868 |
| Sum of Dry Days | 7.415 | 8.654 | 0.000 | 31 | 266868 |
| Elevation | 548 | 742 | -2,271.962 | 5,083 | 266868 |
| Population | 2,801,761 | 7,338,132 | 0.000 | 88,239,594 | 247536 |

Table 3: Extensive Summary Statistics for the used data. Note that these summary statistics do not consider any area or population weights, so e.g. measures like average temperature will disproportionately be biased towards values of very high latitudes. Furthermore statistical grid resampling to convert to lower grid resolutions can introduce slight biases, such as a slightly negative Total Monthly Precipitation.

| n | Disaster Type | Measure | Total Deaths | Total Affected | Total Damages in Mio. USD |
|---|---|---|---|---|---|
| 479 | Drought | Mean | 9,727 | 19,062,602 | 3,616 Mio. US$ |
| | | Median | 24 | 2,520,000 | 1,662 Mio. US$ |
| | | Mean Across all Events | 711 | 11,660,422 | 1,827 Mio. US$ |
| | | Sum | 340,432 | 5,585,342,333 | 875,084 Mio. US$ |
| 765 | Extreme temperature | Mean | 424 | 123,921 | 983 Mio. US$ |
| | | Median | 36 | 17,000 | 323 Mio. US$ |
| | | Mean Across all Events | 394 | 58,802 | 108 Mio. US$ |
| | | Sum | 301,118 | 44,983,218 | 82,533 Mio. US$ |
| 4997 | Flood | Mean | 88 | 1,594,099 | 1,697 Mio. US$ |
| | | Median | 21 | 49,092 | 211 Mio. US$ |
| | | Mean Across all Events | 74 | 1,446,716 | 845 Mio. US$ |
| | | Sum | 367,367 | 7,229,239,919 | 4,222,560 Mio. US$ |
| 173 | Landslide | Mean | 88 | 43,135 | 368 Mio. US$ |
| | | Median | 37 | 3,091 | 219 Mio. US$ |
| | | Mean Across all Events | 84 | 32,912 | 85 Mio. US$ |
| | | Sum | 14,586 | 5,693,785 | 14,730 Mio. US$ |
| 3397 | Storm | Mean | 61 | 997,600 | 3,503 Mio. US$ |
| | | Median | 16 | 22,273 | 232 Mio. US$ |
| | | Mean Across all Events | 52 | 798,491 | 2,569 Mio. US$ |
| | | Sum | 176,506 | 2,712,474,003 | 8,727,140 Mio. US$ |

Table 4: Summarised EM-DAT data distinguished by each considered disaster type. Note that this summary does not take the spatial dimension into account and has simply formed the average across events, not across grid cells.

# B Algorithm Performance

In this section, I present the performance of the algorithms that were considered when estimating the different models. Those plots are presented here for landslide, extreme temperature, flood, and storm events. Drought events are presented in the main text.

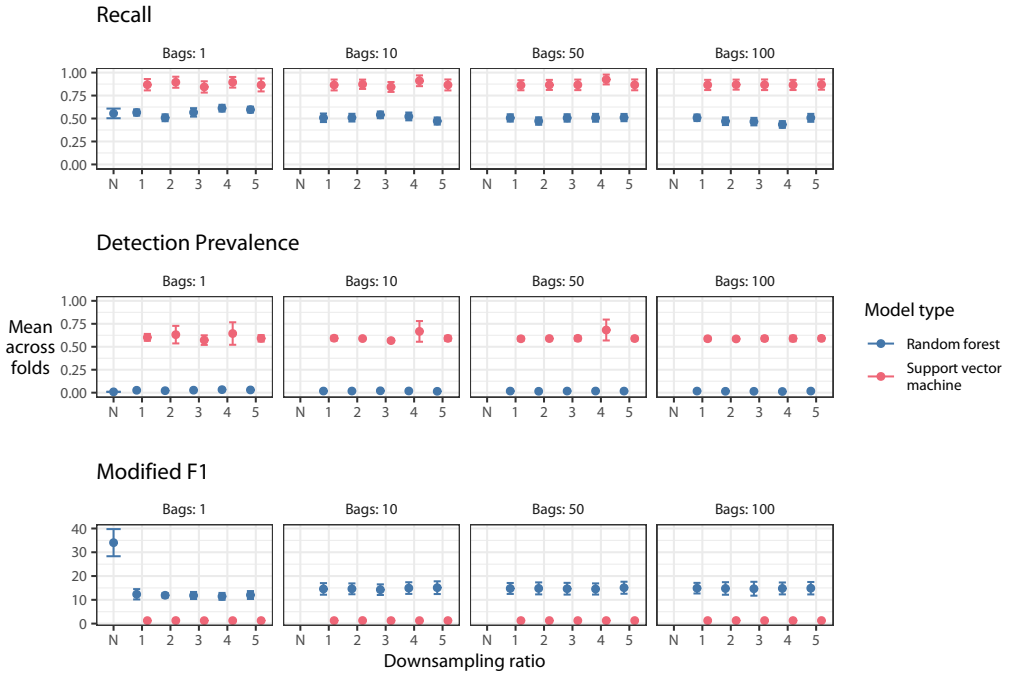


Figure B.1: Algorithm Performance for Landslide Events. N denotes the naïve classifier. Note that the y-axis scale varies for each sub-plot and is bounded from 0 to 1 for the two upper plot rows. The modified F1 combines the information of both rows 1 and 2, providing the essential information to decide which algorithm to use.

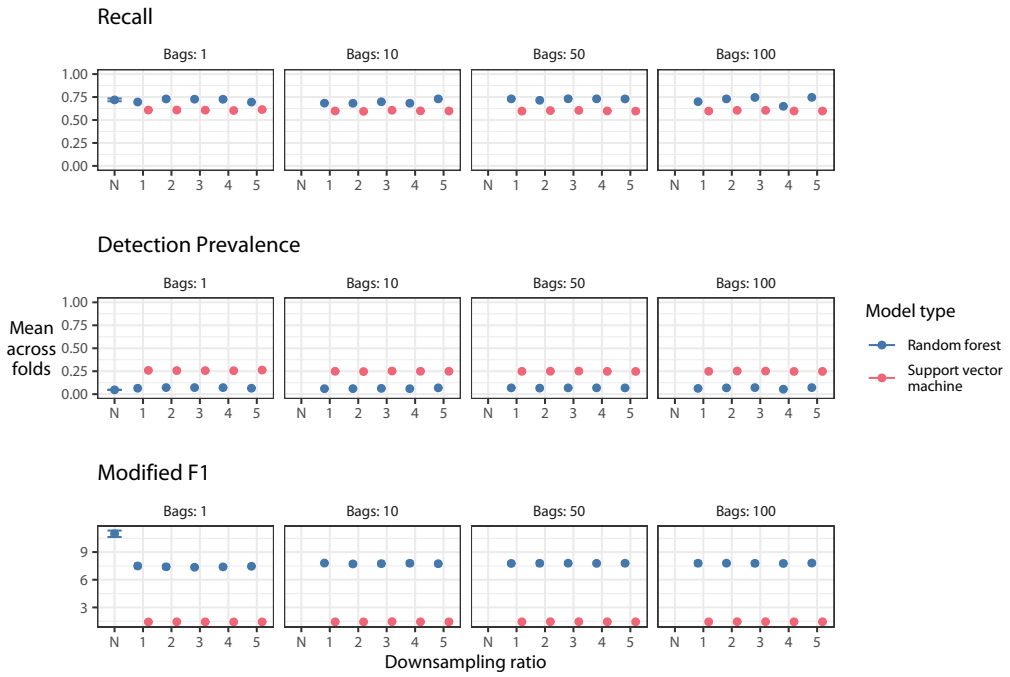


Figure B.2: Algorithm Performance for Extreme Temperature Events. N denotes the naïve classifier. Note that the y-axis scale varies for each sub-plot and is bounded from 0 to 1 for the two upper plot rows. The modified F1 combines the information of both rows 1 and 2, providing the essential information to decide which algorithm to use.

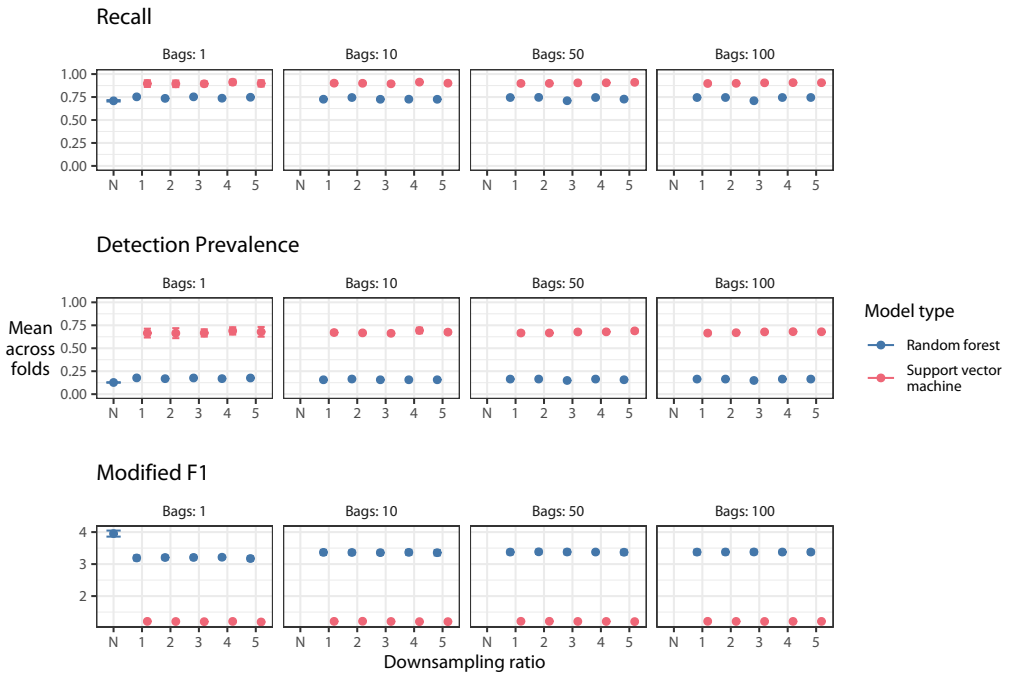


Figure B.3: Algorithm Performance for Flood Events. N denotes the naïve classifier. Note that the y-axis scale varies for each sub-plot and is bounded from 0 to 1 for the two upper plot rows. The modified F1 combines the information of both rows 1 and 2, providing the essential information to decide which algorithm to use.

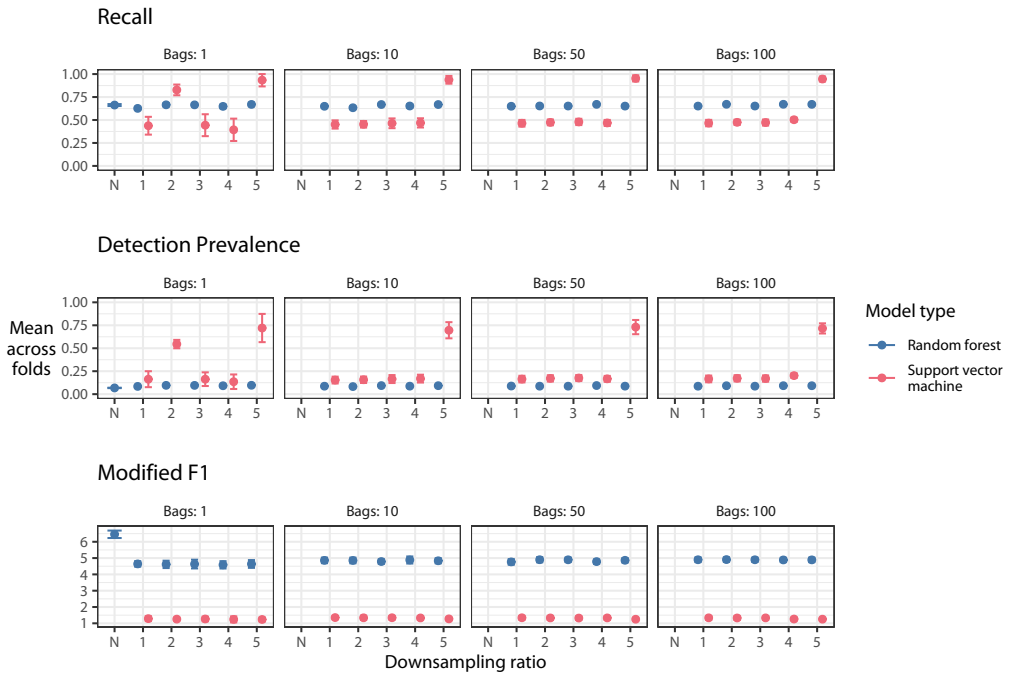


Figure B.4: Algorithm Performance for Storm Events. N denotes the naïve classifier. Note that the y-axis scale varies for each sub-plot and is bounded from 0 to 1 for the two upper plot rows. The modified F1 combines the information of both rows 1 and 2, providing the essential information to decide which algorithm to use.

# C Variable Importance

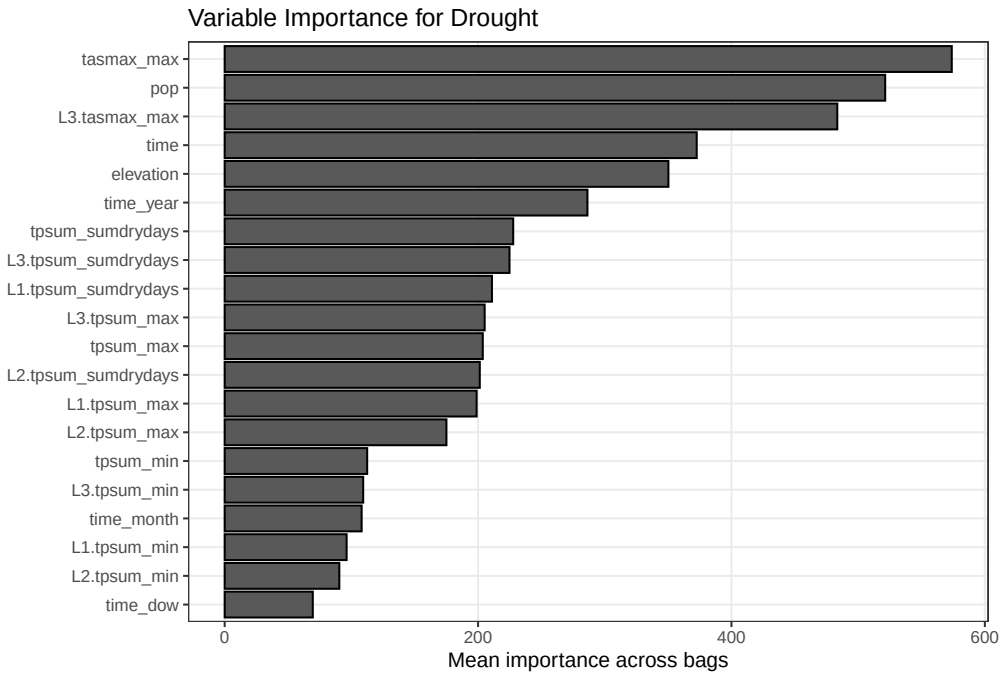


Figure C.1: Variable Importance for Drought Events. This figure is intended to show the relative importance of variables in the selected model. The figure provides information on the mean importance across the bags that are estimated. The numerical value is not a useful metric for comparison across models though – this figure is only intended to provide relative importance.

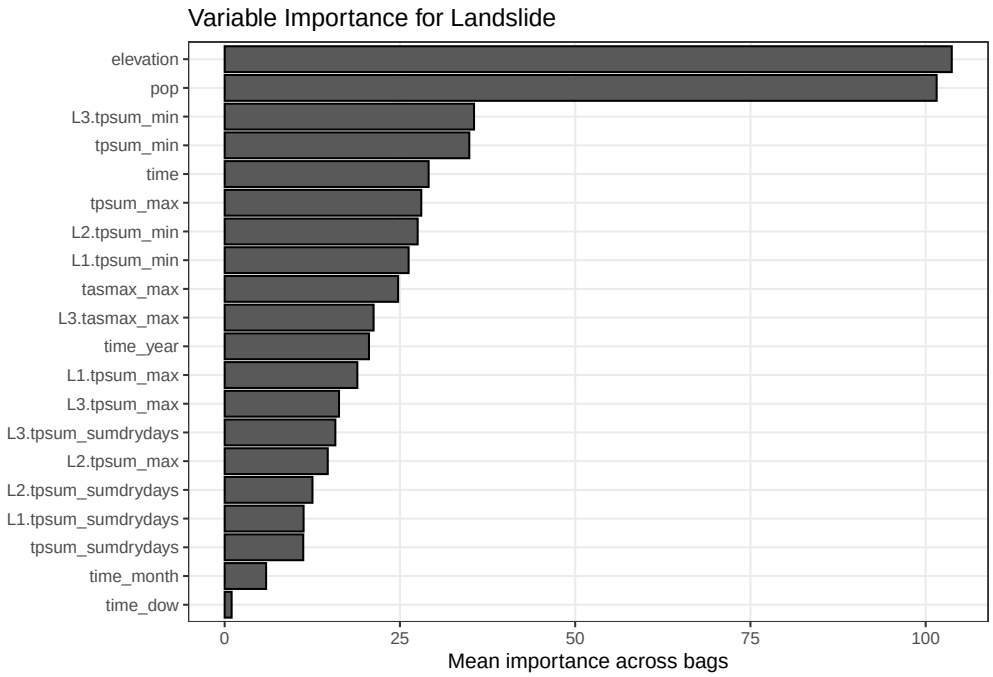


Figure C.2: Variable Importance for Landslide Events. This figure is intended to show the relative importance of variables in the selected model. The figure provides information on the mean importance across the bags that are estimated. The numerical value is not a useful metric for comparison across models though – this figure is only intended to provide relative importance.

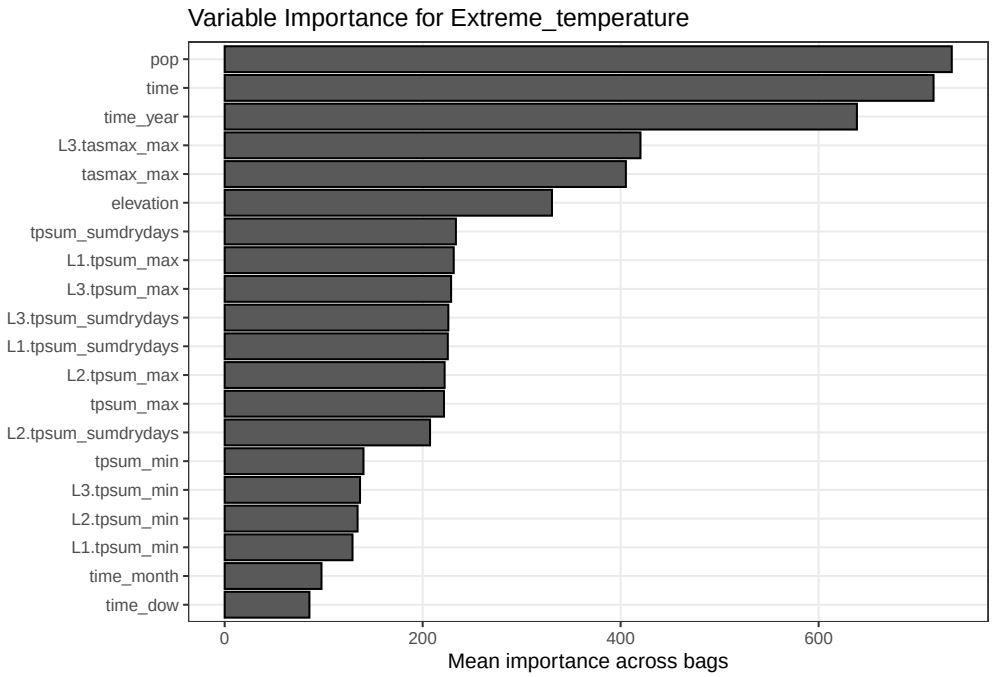


Figure C.3: Variable Importance for Extreme Temperature Events. This figure is intended to show the relative importance of variables in the selected model. The figure provides information on the mean importance across the bags that are estimated. The numerical value is not a useful metric for comparison across models though – this figure is only intended to provide relative importance.

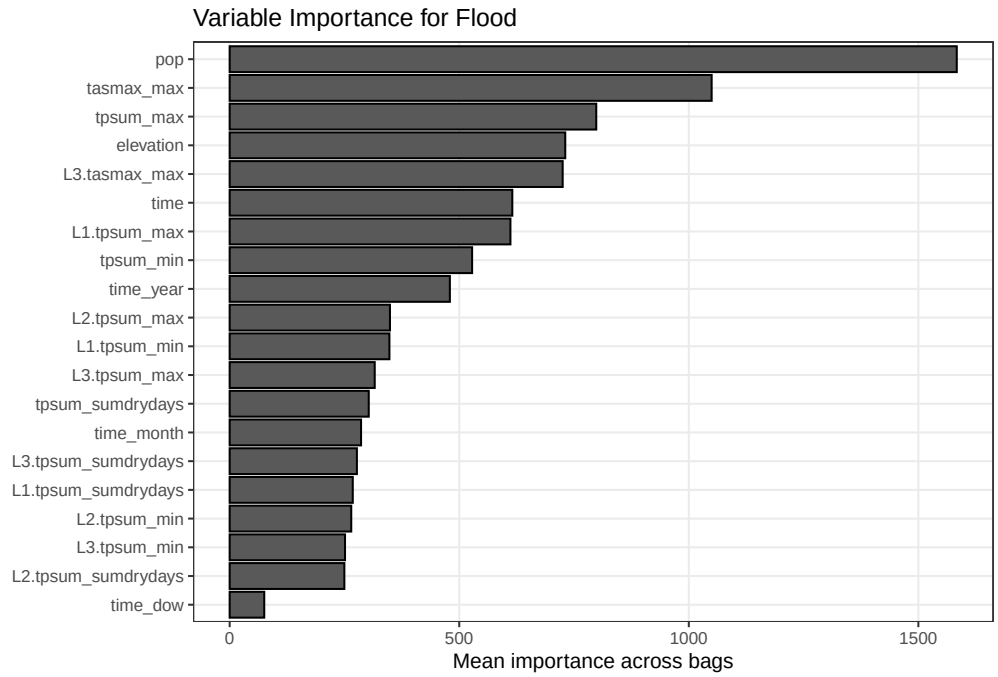


Figure C.4: Variable Importance for Flood Events. This figure is intended to show the relative importance of variables in the selected model. The figure provides information on the mean importance across the bags that are estimated. The numerical value is not a useful metric for comparison across models though – this figure is only intended to provide relative importance.

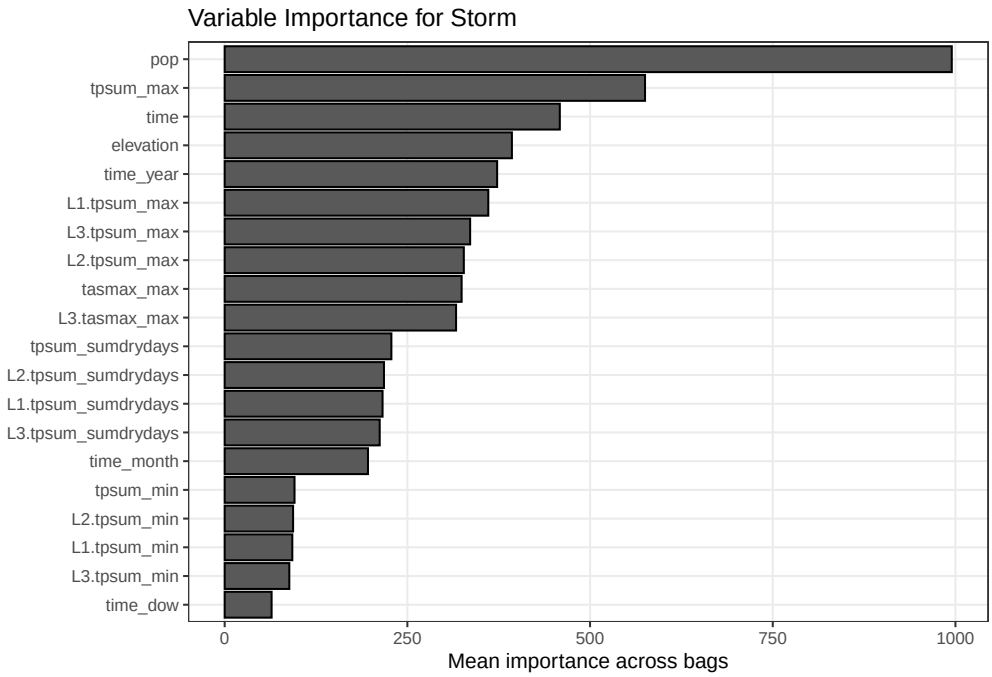


Figure C.5: Variable Importance for Storm Events. This figure is intended to show the relative importance of variables in the selected model. The figure provides information on the mean importance across the bags that are estimated. The numerical value is not a useful metric for comparison across models though – this figure is only intended to provide relative importance.

# Appendix for Paper 4

# 6 Supplementary Material

## 6.1 Identifying Heterogeneous Treatment Effects in Staggered Treatment

Here we briefly summarise the results from Wooldridge (2021), deriving equation (30) to identify treatment effects when treatment is staggered. We assume there is no anticipation of treatment for each $r = q, q+1, ..., Q$:

$$E[y_t(r) - y_t(\infty)|\mathbf{d}] = 0, \text{ for } t < r. \tag{52}$$

We also require a common trend assumption that the trend in absence of treatment is common regardless of state of treatment:

$$E[y_t(\infty) - y_1(\infty)|d_q, ...d_Q] = E[y_t(\infty) - y_1(\infty)] = \theta_t, \text{ for } t = 2, ..., T \tag{53}$$

and we assume at least one untreated group. The observed outcome in any period is given by:

$$y_t = y_t(\infty) + d_q te_t(q) + ... + d_Q te_t(Q) \tag{54}$$

where no anticipation implies that for the pre-treatment period ($t < q$):

$$E[y_t|\mathbf{d}] = E[y_t(\infty)|\mathbf{d}] \tag{55}$$

and for $t \geq q$:

$$E[y_t|\mathbf{d}] = E[y_t(\infty)|\mathbf{d}] + d_q\tau_{q,t} + ... + d_Q\tau_{Q,t} \tag{56}$$

We then write the never treated outcome $y_t(\infty)$ as an initial outcome and change relative to the initial period:

$$y_t(\infty) = y_1(\infty) + g_t(\infty) \tag{57}$$

By the common trend assumption $E[g_t(\infty)|\mathbf{d}] = \theta_t$:

$$E[y_t(\infty)|\mathbf{d}] = E[y_1(\infty)|\mathbf{d}] + E[g_t(\infty)|\mathbf{d}] = \eta + \lambda_q d_q + ... + \lambda_Q d_Q + \theta_t \tag{58}$$

which subsequently allows us to write the expected outcome as equation (30).

If we are interested in treatment effects of units treated at one point relative to those treated at a different point in time, as in Wooldridge (2021), we can define a sub-group ATT for those treated at $r$ compared to for example one period later at $r + 1$ as

$$\tau_{(r:r+1)} = E[y_t(r) - y_t(r+1)|d_r = 1] \tag{59}$$

This can be expressed as the difference in treatment effects relative to the untreated group:

$$y_t(r) - y_t(r+1) = [y_t(r) - y_t(\infty)] - [y_t(r+1) - y_t(\infty)] \tag{60}$$

Thus

$$\tau_{(r:r+1)} = \tau_{r,t} - E[y_t(r+1) - y_t(\infty)|d_r = 1] \tag{61}$$

which under no anticipation and parallel trends simplifies to:

$$\tau_{(r:r+1)} = \tau_{r,t} - \tau_{r+1,t} \tag{62}$$

and which is matched by the difference in coefficients $\hat{\tau}$ obtained post-break detection on treatment dummies (step-functions or impulses).

### 6.2 Simulation Study

Here we investigate the properties of detecting treatment in our reverse causal setting using 'gets' and the adaptive LASSO. For the simulations we focus on detecting piece-wise constant treatment in the form of step-functions. Future work will expand simulations to also include fully-time-varying effects through impulse indicators.

We vary the treatment effect size $\sigma$ as well as the number of treated units $n$. We compare the detection of unknown treatments against the 'known treatment' standard TWFE estimator for a single treated unit as well as multiple treated units. We then consider the case where we impose a known treatment while searching for additional treatment as described in section 3.2.

We simulate the DGP in (23) with errors drawn from the standard normal distribution and evaluate the performance of treatment detection as follows. For 'gets' we select over the full set of break functions using varying target levels of significance ($\gamma_c$). We use cross-validation to determine the penalty level for the adaptive LASSO. To measure the false positive rate of detection we compute the proportion of spuriously retained breaks (out of all possible spurious breaks). To measure whether we correctly identify treatment, we classify the proportion of correctly identified treated observations as those for which the detected breaks include the true treatment effect within a $(1 - \gamma_c)$ confidence interval.

Figure 5 shows the false positive rate together with the correctly classified proportion of treated observations for a single treated unit when varying the treatment magnitude (as a function of the standard deviation of the error term). Note that for a treatment effect size of 0 no treatment is present and hence no treatment should be identified – in this case therefore the rejection frequency yields a measure of the false-positive rate. Results show that treatment detection using 'gets' (red, solid) is close to the benchmark of a known treatment estimated using a conventional TWFE estimator (blue solid). The false positive rate is stable around the chosen level of significance of selection (red dashed). The adaptive LASSO (green solid) using cross-validation to choose the penalty factor also achieves a high level of accurate classification, however, is consistently lower than 'gets' for all significance levels considered. The adaptive LASSO using cross-validation also exhibits an erratic false positive rate (green dashed).

We increase the number of treated units from one to two and then five in our simulations (with identical treatment timing and homogeneous treatment effects) with results shown in Figures 6 and 7. As expected, as we increase the number of treated units, the correct classification (detection of treatment) falls relative to the known treatment case (using a single dummy variable) as our treatment detection approach has to identify a separate treatment dummy per treated unit. Nevertheless, the correct rejection frequency remains high given that no prior information about treatment assignment or timing was used.

Finally, we consider the costs of searching for additional treatment when there is a single known treatment that has been imposed from the outset (i.e. forced in the model and not selected over; see section 3.2).

Figure 8 shows the root-mean-squared error (RMSE) of the estimated treatment effect on the known treatment dummy when selecting over additional break variables relative to the simple TWFE estimator (without selection), together with the false-positive rate of detected treatment (gauge). The DGP only contains the single known treatment, with no other unknown treatment occurring. Thus this provides an assessment of the costs of searching for additional breaks when a known treatment is embedded. The results in Figure 8 show that searching for additional treatment when a known treatment is imposed, increases the RMSE on the estimated treatment effect for known treatment, however, for increasingly conservative selection significance levels this cost shrinks close to zero. This can be seen as an insurance cost – controlling for possible treatment (or breaks) that have been omitted from a standard model increases the RMSE of the known treatment indicator while providing robustness against omitted breaks. In other words, searching for additional breaks (i.e. treatment) lowers the precision on a known forced treatment somewhat, but the degree to which the RMSE increases can be easily controlled by choosing conservative levels of selection when using 'gets'. For 'gets', as Figure 8 shows, the false positive rate

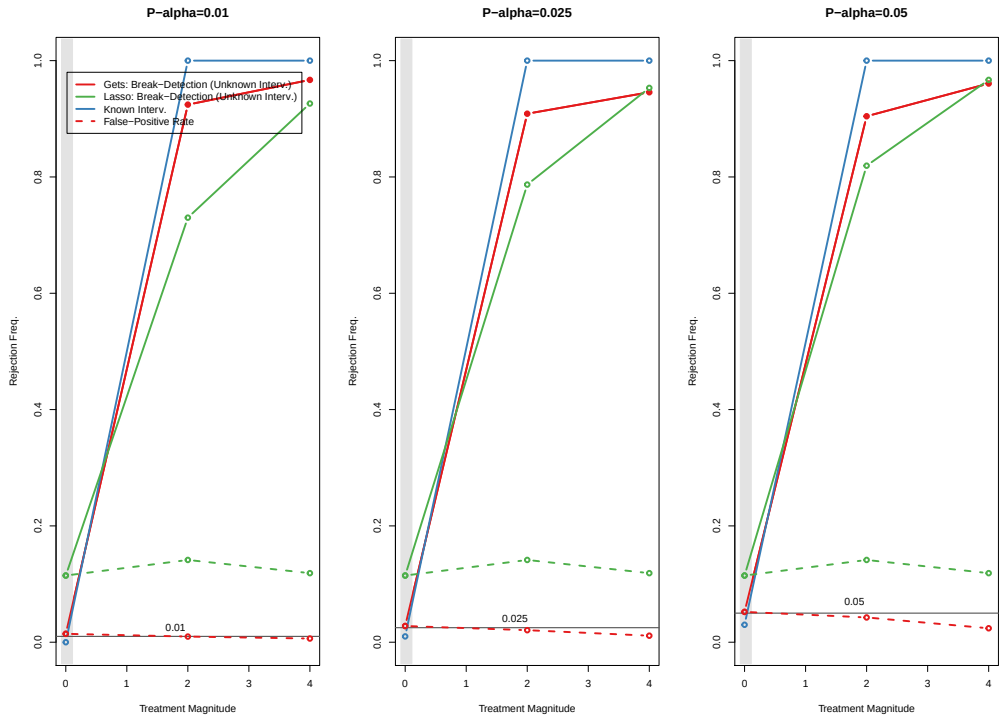


Figure 5: Simulation: Detecting treatment with **one** unknown treated units using 'gets' (SIS) and the adaptive LASSO compared to a 'known' treatment with $N = 10$

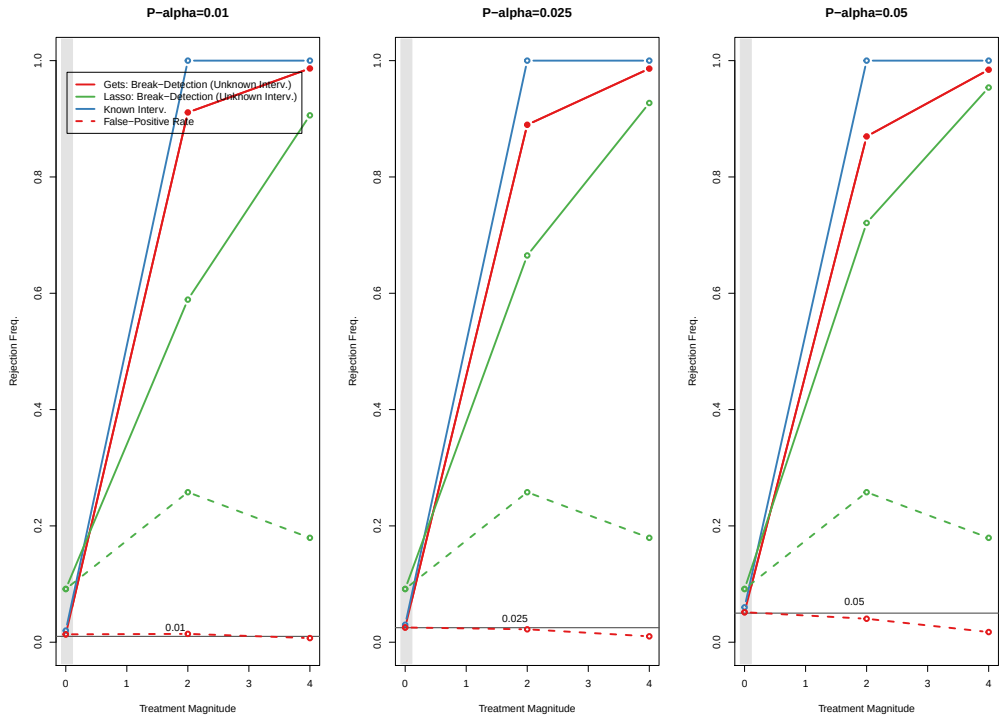


Figure 6: Simulation: Detecting treatment with **two** unknown treated units using 'gets' (SIS) and the adaptive LASSO compared to a 'known' treatment with $N = 10$

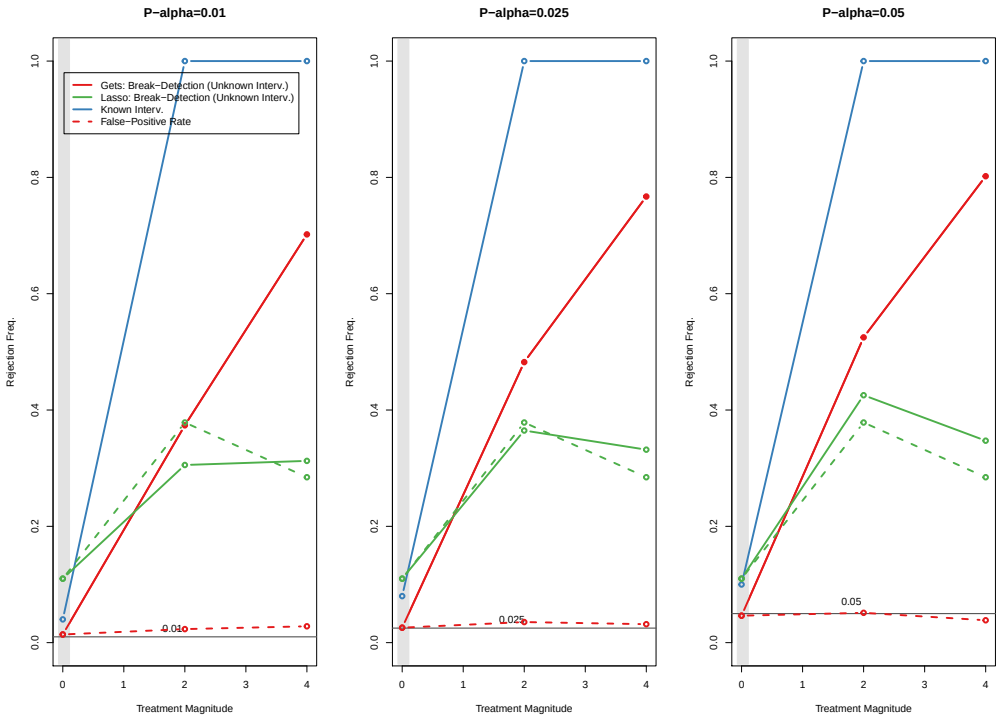


Figure 7: Simulation: Detecting treatment with **five** unknown treated units using 'gets' (SIS) and the adaptive LASSO compared to a 'known' treatment with $N = 10$

(gauge) again is stable around the specified nominal level of significance. Such control is more difficult to achieve when using the LASSO due to not targeting the false-positive rate.

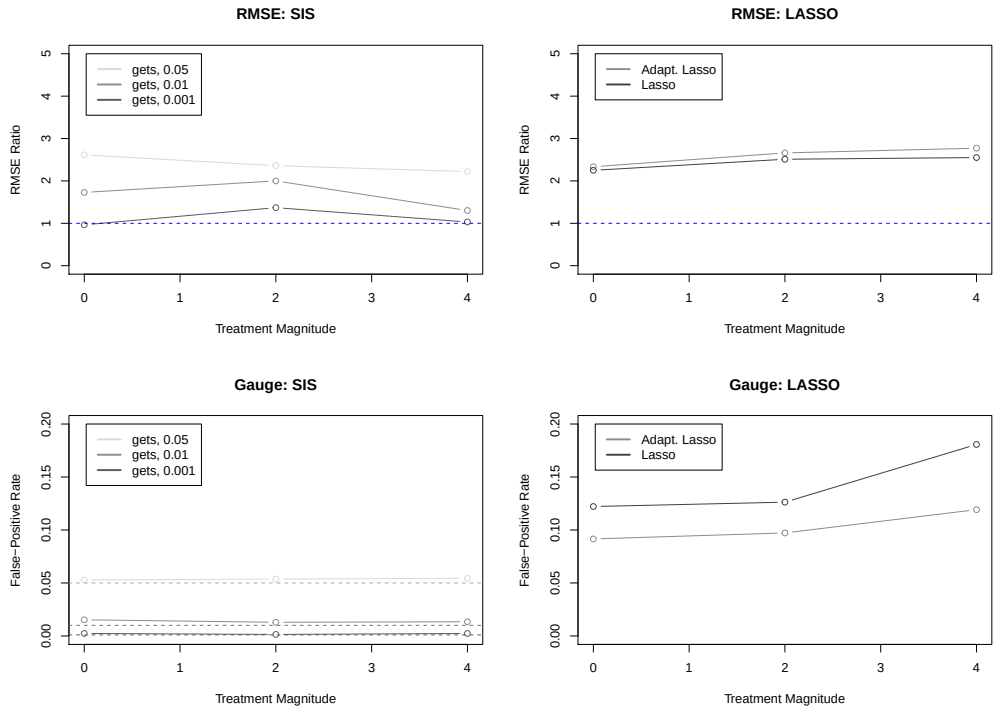


Figure 8: Top: RMSE of estimated 'known' single treatment effect when searching for additional treatment relative to known TWFE estimator. Bottom: False positive rate of detected breaks.

### 6.3 Simple Time Series Approach

We estimate a simple time series model of Basque GDP per capita in (63) and demonstrate that in absence of control groups, we are unable to detect the impact of ETA Basque terrorism *a-priori*. We model the log of GDP per capita as a function of log investment (one of the original control variables in Abadie and Gardeazabal), while searching for structural breaks in the intercept using step-indicators with 'gets' at a conservative target significance level of $\gamma_c = 0.001$:

SIS – Time Series for Basque Country only:

$$log(GDPpc_t) = \beta_0 + \beta_1 log(Inv)_t + \sum_{s=1966}^{1995} \tau_s 1_{\{t \geq s\}} + \epsilon_t \tag{63}$$

Estimation results of this time series model are shown in Table 4 and Figure 9. While multiple breaks are found, the negative impact of ETA terrorism on GDP per capita in the Basque region is not detectable due to the lack of control regions. There are no detected breaks with negative coefficients during the period that ETA was active.

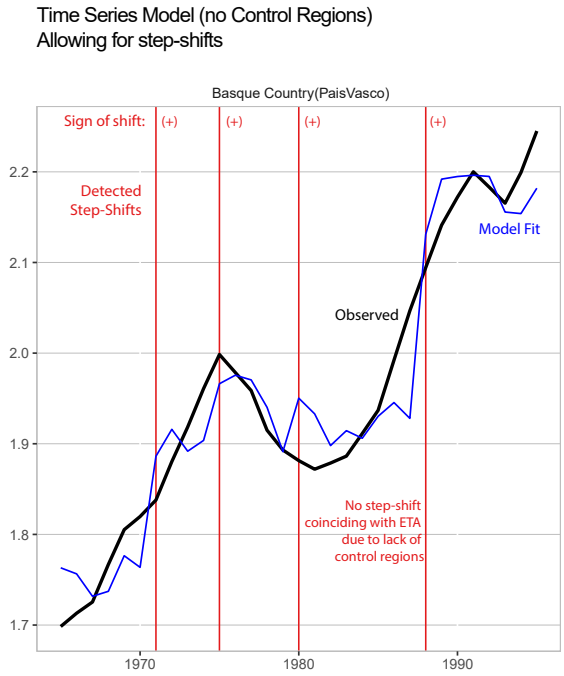


Figure 9: Simple Time Series Model of Basque GDP per Capita – Breaks detected using 'gets' and $\gamma_c = 0.001$.

Table 4: Detecting Breaks in a Simple Time Series Model (Basque Country)

| Dependent Variable: | log(GDPpc) |
|---|---|
| Model: | Time Series |
| *Variables* | |
| Constant | 0.5367 |
| | (0.5008) |
| log(Invest) | 0.3788** |
| | (0.1556) |
| Break ($i$=Basq., $t \geq 1971$) | 0.1663*** |
| | (0.0321) |
| Break ($i$=Basq., $t \geq 1975$) | 0.1308*** |
| | (0.0463) |
| Break ($i$=Basq., $t \geq 1980$) | 0.0536 |
| | (0.0417) |
| Break ($i$=Basq., $t \geq 1988$) | 0.1737*** |
| | (0.0392) |
| *Fit statistics* | |
| Observations | 31 |
| $R^2$ | 0.92 |

*Standard-errors in parentheses*
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

# Appendix for Paper 5

nature portfolio

https://doi.org/10.1038/s41560-022-01095-6

## Supplementary information

# Attributing agnostically detected large reductions in road $CO_2$ emissions to policy mixes

In the format provided by the authors and unedited

# Supplementary Information for:

Attributing agnostically-detected large reductions in road $CO_2$ emissions to policy mixes

Nicolas Koch, Lennard Naumann, Felix Pretis, Nolan Ritter, and Moritz Schwarz

## Supplementary Note 1: Further background information pertaining to the empirical approach

Here we provide additional details on our break detection approach to detect unknown treatment timing and assignment. The discussion here is based on the work in [1] which provides additional details.

### 1.1 Relating break detection to conventional two-way fixed effects policy evaluation

If there is only a single treatment in a single country, then imposing a binary treatment variable for a known policy intervention in a TWFE model is equivalent to agnostically detecting the break at time of treatment and estimating its effect. In case of multiple treated countries, our break detection approach can be seen as a heterogeneous treatment effects model in which the binary treatment variables are interacted with binary variables for each treated country to capture potential heterogeneity (as in [2] for the known-treatment case). This holds regardless of whether all treated countries receive treatment simultaneously or whether treatment is staggered because each treatment effect is estimated using binary variables for each treated country and policy [see 1]. Thus, when faced with multiple treated countries, the break detection approach deviates from the conventional difference-in-differences TWFE literature in that we do not estimate a single average treatment effect on the treated. Instead, we estimate a separate treatment effect for each detected break in each country which identifies heterogeneous treatment effects for each treated cohort. This reduces power compared to a known, homogeneous, single treatment that affects multiple countries. However, this disadvantage must be weighted against the advantage of allowing for heterogeneous treatment effects without suffering from the weighting-problem in staggered treatment designs [such as in 3, 4, 5].

### 1.2 Selection algorithm "gets"

To select over treatment variables in our model saturated with a full set of step-shifts we use the tree search "gets." It is equivalent to applying step-indicator saturation [SIS, 6] in a fixed effects panel. SIS uses a near exhaustive tree search over candidate variables based on a specified level of significance $\gamma_c$ which controls the expected false-positive rate. Pretis and Schwarz [1] further discuss the false positive rate of detection in the context of detecting treatment effects in panels. We can estimate the set of treated units $\widehat{H}$ as those that have at least one treatment indicator retained. Specifically, if we allow for possible treatment of each unit at every point in time, then – in absence of treatment – the expected number of detected breaks is $\gamma_c \times N(T-1)$ in a balanced panel. (Albeit simulation results show a higher false-positive rate for SIS in small samples, warranting perhaps a more conservative choice of $\gamma_c$.) Which translates into a probability of a single country (i.e. unit) being classified as treated as:

$$P(i \in \widehat{H} | d_i = 0) = 1 - (1 - \gamma_c)^{(T-1)} \tag{1}$$

which increases with $T$ because for larger samples (and fixed $\gamma_c$) the probability of retaining an indicator spuriously increases as the number of indicators increases with $T$. Let $\hat{M}$ denote

the estimated number of treated units. Under the null of no treatment (when in fact no unit is treated), the expected number of falsely detected treated units (countries) is then given by:

$$E[\hat{M}] = P(i \in \widehat{H} | d_i = 0) \times N = (1 - (1 - \gamma_c)^{T-1})N \tag{2}$$

Thus, if we are concerned about the false positive rate of treatment classification, then treatment detection in a panel warrants tighter target significance levels $\gamma_c$ than conventionally used in the selection/break detection literature. We therefore report results in the main text using both looser and more conservative significance levels of detection.

Variations of "gets" have been widely applied in the time series literature, with applications ranging from detecting heterogeneity in behaviour in medical practise [7] to assessing misreporting in pollution monitoring [8]. The properties of "gets" transfer from time series to panel models when interpreted as a least-squares dummy variable estimator.

Alternatives to "gets" include the algorithm *Autometrics* in the wider general-to-specific family of model selection – see [9], or shrinkage-based methods such as the LASSO and variants thereof, though these do not target the false positive rate – see [10, 11, 12]. See [1] for a more detailed discussion.

### 1.3 Combining break detection with conventional policy evaluation approach

There are two ways to implement break detection in policy evaluation. First, it can be applied as a purely agnostic data-driven approach that identifies interventions without any prior knowledge of their occurrence. We take this approach in our paper. One drawback is that estimated effects may not be generalizable in the sense of providing benchmark estimates for the effectiveness of a particular policy instrument. Another potential drawback is a loss in power if treatment assignment and timing is known and there are multiple treated units with a homogeneous treatment effects.

The second approach can partly address these issues by inserting known policy interventions into the break detection model. If a policy's treatment assignment and timing are known, we can include it in the model without having to rely on the selection step to find it and estimate its impact, all the while searching for additional treatments. In this case, we force the known treatment indicator into the model and select over additional treatment indicators. This allows for the detection of additional, unknown treatments while the coefficient estimate on the forced break variable provides an estimate of the conventional treatment effect in a difference-in-differences TWFE estimator. We hope that future research can apply this second implementation strategy to make progress on improving our understanding of the workings of specific policy instruments.

### 1.4 Role of subject-specific knowledge in ex-post attribution

In a post-estimation analysis, we manually attribute the detected breaks to policy interventions. This requires subject-specific knowledge, e.g. from the literature or policy databases. This is similar to the need for subject-specific knowledge in the conventional difference-in-differences TWFE literature to justify the assumption that a specific known intervention is exogenous or as-if randomly assigned. In fact, once a potential cause for a detected break has been identified, we could have simply estimated a conventional difference-in-differences TWFE model using the "known" intervention. Thus, in the proposed reverse causal approach, we must argue that a particular detected break is associated with a particular policy intervention. Naturally, there may be many such policy interventions. While such confounding treatments pose a challenge for the conventional difference-in-differences TWFE setting, there is no need for us to argue that a given policy intervention was unique. In fact, we pursue a reverse causal approach that agnostically identifies effective, emission reducing interventions because we are particularly interested in the effectiveness of interacting policies. So while the search for causes is different than arguing that a cause was unique, subject-specific knowledge will be necessary in both settings.

In our application, we find at least one policy intervention for every detected break. However, should there be a break that cannot seemingly be attributed to a policy intervention, researchers may want to consider widening their search or consider that their model may be misspecified and in need of adaptation.

## Supplementary Note 2: Summary statistics

Supplementary Tables 1 through 3 summarize our dependent variable and our control variables by country in terms of their means, their standard deviations, as well as their minimum and maximum values for the time period between 1995 and 2018. Although we log the variables for our analyses, we present them in levels here to facilitate understanding.

| Country | mean | sd | min | max |
|---|---|---|---|---|
| Austria | 20.7 | 2.6 | 15.0 | 23.9 |
| Belgium | 24.5 | 1.5 | 21.4 | 26.7 |
| Bulgaria | 6.6 | 1.6 | 3.9 | 9.0 |
| Croatia | 5.0 | 1.0 | 3.0 | 6.3 |
| Cyprus | 1.9 | 0.2 | 1.5 | 2.2 |
| Czech Republic | 14.8 | 3.6 | 7.2 | 21.1 |
| Denmark | 11.3 | 0.8 | 10.3 | 13.1 |
| Estonia | 1.9 | 0.3 | 1.3 | 2.4 |
| Finland | 11.1 | 0.6 | 10.1 | 12.2 |
| France | 121.6 | 4.6 | 116.4 | 129.1 |
| Germany | 154.4 | 9.4 | 141.7 | 172.1 |
| Greece | 16.4 | 2.1 | 13.7 | 21.0 |
| Hungary | 10.4 | 2.0 | 6.8 | 13.5 |
| Iceland | 0.7 | 0.1 | 0.5 | 1.0 |
| Ireland | 10.5 | 2.1 | 5.6 | 14.0 |
| Italy | 106.3 | 7.8 | 94.8 | 117.7 |
| Latvia | 2.5 | 0.5 | 1.7 | 3.5 |
| Lithuania | 3.9 | 0.8 | 2.7 | 5.5 |
| Luxembourg | 5.7 | 1.2 | 3.3 | 7.2 |
| Malta | 0.5 | 0.1 | 0.3 | 0.7 |
| Netherlands | 31.2 | 1.9 | 28.0 | 34.3 |
| Norway | 9.9 | 0.6 | 8.8 | 10.9 |
| Poland | 37.4 | 10.8 | 21.4 | 58.0 |
| Portugal | 16.3 | 1.9 | 12.4 | 19.3 |
| Romania | 12.3 | 2.7 | 6.9 | 16.8 |
| Slovak Republic | 5.2 | 1.0 | 3.5 | 7.0 |
| Slovenia | 4.7 | 0.8 | 3.5 | 6.0 |
| Spain | 81.9 | 10.6 | 61.9 | 101.7 |
| Sweden | 19.9 | 0.8 | 18.6 | 21.4 |
| Switzerland | 16.0 | 0.9 | 14.0 | 17.1 |
| United Kingdom | 114.2 | 3.7 | 107.8 | 120.5 |

Supplementary Table 1: $CO_2$ emissions from road transport by country in millions of metric tons. This table summarizes $CO_2$ emissions by country for the years 1995 through 2018. All values are in millions of metric tons. sd is for standard deviation, min is for minimum, and max is for maximum.

| Country | mean | sd | min | max |
|---|---|---|---|---|
| Austria | 371.5 | 44.4 | 290.4 | 442.5 |
| Belgium | 451.2 | 56.2 | 352.4 | 538.4 |
| Bulgaria | 45.1 | 9.5 | 31.1 | 60.9 |
| Croatia | 55.2 | 7.8 | 39.6 | 65.5 |
| Cyprus | 22.0 | 3.9 | 15.1 | 27.4 |
| Czech Republic | 188.4 | 34.5 | 139.0 | 247.9 |
| Denmark | 315.8 | 29.5 | 257.1 | 370.3 |
| Estonia | 18.9 | 4.8 | 10.4 | 26.4 |
| Finland | 230.9 | 31.6 | 162.9 | 269.2 |
| France | 2525.2 | 264.2 | 2019.5 | 2927.8 |
| Germany | 3335.1 | 316.0 | 2841.0 | 3937.2 |
| Greece | 267.3 | 36.2 | 210.3 | 332.1 |
| Hungary | 126.0 | 19.9 | 92.5 | 162.6 |
| Iceland | 13.0 | 2.8 | 8.2 | 18.2 |
| Ireland | 216.3 | 68.4 | 107.2 | 373.1 |
| Italy | 2083.1 | 96.7 | 1868.1 | 2236.6 |
| Latvia | 22.9 | 6.0 | 12.8 | 31.3 |
| Lithuania | 34.7 | 9.4 | 19.3 | 49.4 |
| Luxembourg | 48.8 | 10.8 | 30.6 | 67.3 |
| Malta | 8.6 | 2.2 | 5.6 | 13.9 |
| Netherlands | 796.4 | 94.8 | 597.9 | 948.1 |
| Norway | 409.1 | 51.5 | 306.9 | 489.3 |
| Poland | 423.6 | 113.2 | 252.4 | 633.9 |
| Portugal | 224.3 | 16.9 | 181.1 | 247.2 |
| Romania | 154.0 | 37.1 | 107.3 | 225.6 |
| Slovak Republic | 77.9 | 20.9 | 46.8 | 112.1 |
| Slovenia | 43.6 | 7.4 | 30.1 | 55.3 |
| Spain | 1298.6 | 178.3 | 942.9 | 1539.5 |
| Sweden | 461.2 | 76.4 | 334.3 | 589.3 |
| Switzerland | 547.2 | 74.0 | 434.9 | 674.6 |
| United Kingdom | 2377.7 | 321.3 | 1780.0 | 2879.3 |

Supplementary Table 2: GDP by country in billions of US dollars of 2010. This table summarizes GDP by country for the years 1995 through 2018. All values are in billions of US dollars of 2010. sd is for standard deviation, min is for minimum, and max is for maximum.

| Country | mean | sd | min | max |
|---|---|---|---|---|
| Austria | 8.3 | 0.3 | 7.9 | 8.8 |
| Belgium | 10.7 | 0.4 | 10.1 | 11.4 |
| Bulgaria | 7.7 | 0.4 | 7.0 | 8.4 |
| Croatia | 4.3 | 0.1 | 4.1 | 4.6 |
| Cyprus | 1.0 | 0.1 | 0.9 | 1.2 |
| Czech Republic | 10.4 | 0.1 | 10.2 | 10.6 |
| Denmark | 5.5 | 0.2 | 5.2 | 5.8 |
| Estonia | 1.4 | 0.0 | 1.3 | 1.4 |
| Finland | 5.3 | 0.1 | 5.1 | 5.5 |
| France | 63.5 | 2.5 | 59.5 | 67.0 |
| Germany | 81.9 | 0.7 | 80.3 | 82.9 |
| Greece | 10.9 | 0.2 | 10.6 | 11.1 |
| Hungary | 10.1 | 0.2 | 9.8 | 10.3 |
| Iceland | 0.3 | 0.0 | 0.3 | 0.4 |
| Ireland | 4.3 | 0.4 | 3.6 | 4.9 |
| Italy | 58.5 | 1.5 | 56.8 | 60.8 |
| Latvia | 2.2 | 0.2 | 1.9 | 2.5 |
| Lithuania | 3.2 | 0.3 | 2.8 | 3.6 |
| Luxembourg | 0.5 | 0.1 | 0.4 | 0.6 |
| Malta | 0.4 | 0.0 | 0.4 | 0.5 |
| Netherlands | 16.4 | 0.5 | 15.5 | 17.2 |
| Norway | 4.8 | 0.3 | 4.4 | 5.3 |
| Poland | 38.2 | 0.2 | 38.0 | 38.7 |
| Portugal | 10.4 | 0.2 | 10.0 | 10.6 |
| Romania | 21.1 | 1.1 | 19.5 | 22.7 |
| Slovak Republic | 5.4 | 0.0 | 5.4 | 5.4 |
| Slovenia | 2.0 | 0.0 | 2.0 | 2.1 |
| Spain | 43.9 | 2.8 | 39.7 | 46.8 |
| Sweden | 9.3 | 0.4 | 8.8 | 10.2 |
| Switzerland | 7.6 | 0.5 | 7.0 | 8.5 |
| United Kingdom | 61.5 | 2.8 | 58.0 | 66.5 |

Supplementary Table 3: Population by country in millions. This table summarizes the population size by country for the years 1995 through 2018. All values are in millions. sd is for standard deviation, min is for minimum, and max is for maximum.

# Supplementary Note 3: Robustness checks

## 3.1 Alternative emissions data

Our main estimates are based on $CO_2$ emissions data for road transport as defined by NRF code 1A3b. It covers aggregate emissions from passenger Cars (1A3bi), light duty vehicles (1A3bii), heavy duty vehicles and buses (1A3biii), and mopeds & motorcycles (1A3biv). Because policy efforts may have a particular focus on reducing emissions from passenger vehicles, we re-estimate our model based on more disaggregated 1A3bi data only.

| Country | | Model | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| | | EU-15 | EU-15 | EU-15 | EU-31 | EU-31 | EU-31 |
| significance level in break detection | | 5% | 1% | 0.1% | 5% | 1% | 0.1% |
| Denmark | effect | | | | | | |
| | se | | | | | | |
| | year | | | | | | |
| Finland | effect | −0.054 | | | −0.073 | | |
| | se | (0.020) | | | (0.025) | | |
| | year | 2000 | | | 2000 | | |
| Germany | effect | −0.135 | −0.125 | −0.140 | −0.125 | −0.142 | −0.139 |
| | se | (0.023) | (0.023) | (0.025) | (0.021) | (0.021) | (0.024) |
| | year | 2003 | 2003 | 2003 | 2005 | 2005 | 2005 |
| Ireland (1st break) | effect | | | | | | |
| | se | | | | | | |
| | year | | | | | | |
| Ireland (2nd break) | effect | | | | −0.208 | −0.166 | |
| | se | | | | (0.031) | (0.032) | |
| | year | | | | 2015 | 2015 | |
| Luxembourg (1st break) | effect | −0.066 | −0.073 | | −0.107 | | |
| | se | (0.024) | (0.025) | | (0.029) | | |
| | year | 2006 | 2006 | | 2006 | | |
| Luxembourg (2nd break) | effect | | | −0.133 | −0.171 | −0.141 | −0.152 |
| | se | | | (0.028) | (0.032) | (0.033) | (0.036) |
| | year | | | 2015 | 2015 | 2015 | 2015 |
| Portugal | effect | | | | | | |
| | se | | | | | | |
| | year | | | | | | |
| Sweden (1st break) | effect | −0.076 | −0.081 | −0.123 | −0.107 | −0.158 | −0.152 |
| | se | (0.019) | (0.020) | (0.018) | (0.024) | (0.021) | (0.023) |
| | year | 2004 | 2004 | 2004 | 2003 | 2004 | 2004 |
| Sweden (2nd break) | effect | −0.084 | −0.089 | | −0.124 | | |
| | se | (0.019) | (0.020) | | (0.023) | | |
| | year | 2011 | 2011 | | 2010 | | |

Supplementary Table 4: Results for passenger vehicle emissions only. This table shows the relative impact of treatments on $CO_2$ emissions, their standard errors, as well as the year of the break.

Supplementary Table 4 shows the emissions breaks detected in the alternative data set that is limited to passenger vehicles. With the exemption of the breaks in Portugal in 2011 and in Denmark in 2012, which both were detected only when holding the false discover rate in break testing at the 5% level of significance, and the break in Ireland in 2011, we find the same emission breaks as in our baseline analysis (Table 1). Notably, the timing of the breaks detected in Germany and Sweden slightly shifts, which can be explained by the fact the policies under the "Ecological Tax Reform" in Germany and the "Green Tax Shift" in Sweden were implemented in a staggered manner over time.

### 3.2 Clustered standard errors

One concern is that our outcome variable, $CO_2$ emissions, may be persistent, and, thus, the error term in our model may be serially correlated. This might affect our inference. In Supplementary Table 5 we, therefore, report our main results with HAC standard errors and HAC standard errors that are clustered at the county level. Neither of these alternatives change our inferences.

| Country | | Model | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| | | EU-15 | EU-15 | EU-15 | EU-31 | EU-31 | EU-31 |
| significance level in break detection | | 5% | 1% | 0.1% | 5% | 1% | 0.1% |
| Denmark | effect | | | | −0.080 | | |
| | HAC se | | | | (0.010) | | |
| | clustered HAC se | | | | [0.013] | | |
| Finland | effect | −0.103 | −0.123 | −0.128 | −0.156 | −0.171 | |
| | HAC se | (0.006) | (0.014) | (0.018) | (0.018) | (0.021) | |
| | clustered HAC se | [0.016] | [0.017] | [0.023] | [0.025] | [0.027] | |
| Germany | effect | −0.105 | −0.131 | −0.108 | −0.112 | −0.112 | |
| | HAC se | (0.008) | (0.011) | (0.020) | (0.015) | (0.014) | |
| | clustered HAC se | [0.012] | [0.013] | [0.033] | [0.024] | [0.022] | |
| Ireland | effect | −0.087 | | −0.127 | | | |
| (1st break) | HAC se | (0.012) | | (0.016) | | | |
| | clustered HAC se | [0.017] | | [0.019] | | | |
| Ireland | effect | −0.148 | −0.192 | | −0.247 | −0.244 | −0.229 |
| (2nd break) | HAC se | (0.031) | (0.032) | | (0.031) | (0.046) | (0.051) |
| | clustered HAC se | [0.032] | [0.036] | | [0.034] | [0.043] | [0.047] |
| Luxembourg | effect | −0.136 | | | −0.108 | | |
| (1st break) | HAC se | (0.012) | | | (0.018) | | |
| | clustered HAC se | [0.019] | | | [0.026] | | |
| Luxembourg | effect | | | −0.214 | −0.193 | −0.227 | −0.262 |
| (2nd break) | HAC se | | | (0.034) | (0.025) | (0.029) | (0.024) |
| | clustered HAC se | | | [0.033] | [0.026] | [0.030] | [0.034] |
| Portugal | effect | | | | −0.094 | | |
| | HAC se | | | | (0.010) | | |
| | clustered HAC se | | | | [0.016] | | |
| Sweden | effect | −0.095 | −0.103 | −0.110 | | | |
| (1st break) | HAC se | (0.008) | (0.010) | (0.015) | | | |
| | clustered HAC se | [0.021] | [0.023] | [0.024] | | | |
| Sweden | effect | | | | −0.108 | −0.115 | |
| (2nd break) | HAC se | | | | (0.013) | (0.015) | |
| | clustered HAC se | | | | [0.035] | [0.037] | |

Supplementary Table 5: Results with (cluster) robust standard errors. This table shows the relative impact of treatments on $CO_2$ emissions, their HAC standard errors, and their HAC standard errors clustered at the country level.


51

### 3.3 Bias adjusted estimates

Another concern is that our estimated effects (Table 1) may not be generalizable because we are more likely to detect large breaks. To assess the extent to which our model selection approach introduces bias, we estimate one and two-step bias-corrected effect estimates following [13]. Overall, Supplementary Table 6 provides strong evidence that model selection did not introduce any substantial biases in the estimated effect sizes.

| Country | | Model | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| | | EU-15 | EU-15 | EU-15 | EU-31 | EU-31 | EU-31 |
| significance level in break detection | | 5% | 1% | 0.1% | 5% | 1% | 0.1% |
| Denmark | uncorrected | | | | −0.080 | | |
| | one-step | | | | (−0.079) | | |
| | two-step | | | | [−0.078] | | |
| Finland | uncorrected | −0.103 | −0.123 | −0.128 | −0.156 | −0.171 | |
| | one-step | (−0.103) | (−0.123) | (−0.126) | (−0.156) | (−0.171) | |
| | two-step | [−0.103] | [−0.123] | [−0.126] | [−0.156] | [−0.171] | |
| Germany | uncorrected | −0.105 | −0.131 | | −0.112 | −0.112 | |
| | one-step | (−0.105) | (−0.131) | | (−0.112) | (−0.111) | |
| | two-step | [−0.105] | [−0.131] | | [−0.112] | [−0.111] | |
| Ireland | uncorrected | −0.087 | | −0.127 | | | |
| (1st break) | one-step | (−0.086) | | (−0.126) | | | |
| | two-step | [−0.086] | | [−0.126] | | | |
| Ireland | uncorrected | −0.148 | −0.192 | | −0.247 | −0.244 | −0.229 |
| (2nd break) | one-step | (−0.148) | (−0.192) | | (−0.247) | (−0.244) | (−0.229) |
| | two-step | [−0.148] | [−0.192] | | [−0.247] | [−0.244] | [−0.229] |
| Luxembourg | uncorrected | −0.136 | | | −0.108 | | |
| (1st break) | one-step | (−0.136) | | | (−0.103) | | |
| | two-step | [−0.136] | | | [−0.102] | | |
| Luxembourg | uncorrected | | | −0.214 | −0.193 | −0.227 | −0.262 |
| (2nd break) | one-step | | | (−0.214) | (−0.193) | (−0.227) | (−0.262) |
| | two-step | | | [−0.214] | [−0.193] | [−0.227] | [−0.262] |
| Portugal | uncorrected | | | | −0.094 | | |
| | one-step | | | | (−0.094) | | |
| | two-step | | | | [−0.094] | | |
| Sweden | uncorrected | −0.095 | −0.103 | −0.110 | | | |
| (1st break) | one-step | (−0.095) | (−0.102) | (−0.108) | | | |
| | two-step | [−0.095] | [−1.020] | [−0.107] | | | |
| Sweden | uncorrected | | | | −0.108 | −0.115 | |
| (2nd break) | one-step | | | | (−0.108) | (−0.114) | |
| | two-step | | | | [−0.108] | [−0.114] | |

Supplementary Table 6: Bias corrected estimates. This table shows the relative impact of treatments on $CO_2$ emissions. For each detected break, the table holds three coefficients one below the other. The topmost is the coefficient from the models introduced in the main body of the text. The second and third are from applying either a one-step or two-step bias correction following [13]. The similarity of the coefficients provides strong evidence that model selection did not introduce any substantial biases in effect sizes.

### 3.4 Estimates without Austria and Luxembourg

Policy interventions in one country may also affect neighboring countries. In particular, a fuel or carbon tax increase in a given country may cause fuel tourism in private consumers in the border region or cause firms to reroute their trucks to refuel in neighboring countries. To evaluate the potential bias from such spillovers, we exclude Austria and Luxembourg, two major transit countries with a low fuel tax regime, from our sample to estimate our final model. Supplementary Table 10 shows that we obtain very similar results in the restricted sample. Overall, any differences in coefficients are negligible.

| Country | | Model | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| | | EU-15 | EU-15 | EU-15 | EU-31 | EU-31 | EU-31 |
| significance level in break detection | | 5% | 1% | 0.1% | 5% | 1% | 0.1% |
| Denmark | effect | | | | −0.070 | | |
| | se | | | | (0.020) | | |
| | year | | | | 2012 | | |
| Finland | effect | −0.101 | −0.115 | −0.121 | −0.138 | −0.173 | |
| | se | (0.020) | (0.022) | (0.025) | (0.025) | (0.029) | |
| | year | 2000 | 2000 | 2000 | 2000 | 2000 | |
| Germany | effect | −0.105 | −0.134 | −0.107 | −0.127 | −0.109 | |
| | se | (0.017) | (0.020) | (0.022) | (0.021) | (0.025) | |
| | year | 2002 | 2002 | 2002 | 2003 | 2003 | |
| Ireland | effect | −0.081 | | −0.127 | | | |
| (1st break) | se | (0.020) | | (0.024) | | | |
| | year | 2011 | | 2011 | | | |
| Ireland | effect | −0.152 | −0.186 | | −0.219 | −0.250 | −0.208 |
| (2nd break) | se | (0.027) | (0.028) | | (0.031) | (0.035) | (0.038) |
| | year | 2015 | 2015 | | 2015 | 2015 | 2015 |
| Portugal | effect | | | | −0.097 | | |
| | se | | | | (0.021) | | |
| | year | | | | 2011 | | |
| Sweden | effect | −0.093 | −0.099 | −0.107 | | | |
| (1st break) | se | (0.016) | (0.019) | (0.022) | | | |
| | year | 2001 | 2001 | 2001 | | | |
| Sweden | effect | | | | −0.101 | −0.118 | |
| (2nd break) | se | | | | (0.019) | (0.022) | |
| | year | | | | 2006 | 2006 | |

Supplementary Table 7: Robustness Check: Models excluding Austria and Luxembourg. This table shows the relative impact of treatments on $CO_2$ emissions, their standard errors, as well as the year of the break.

### 3.5 Alternative base model specifications

| Country | | Model | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| | | EU-15 | EU-15 | EU-15 | EU-31 | EU-31 | EU-31 |
| significance level in break detection | | 5% | 1% | 0.1% | 5% | 1% | 0.1% |
| Denmark | effect | | | | −0.085 | | |
| | se | | | | (0.036) | | |
| Finland | effect | −0.136 | −0.136 | −0.127 | −0.148 | −0.144 | |
| | se | (0.022) | (0.023) | (0.024) | (0.040) | (0.035) | |
| Germany | effect | −0.112 | −0.122 | −0.109 | −0.284 | −0.231 | |
| | se | (0.021) | (0.022) | (0.023) | (0.037) | (0.031) | |
| Ireland | effect | −0.046 | | −0.127 | | | |
| (1st break) | se | (0.024) | | (0.023) | | | |
| Ireland | effect | −0.155 | −0.169 | | −0.288 | −0.245 | −0.261 |
| (2nd break) | se | (0.033) | (0.030) | | (0.048) | (0.042) | (0.043) |
| Luxembourg | effect | −0.090 | | | −0.119 | | |
| (1st break) | se | (0.031) | | | (0.055) | | |
| Luxembourg | effect | | | −0.216 | −0.193 | −0.251 | −0.204 |
| (2nd break) | se | | | (0.037) | (0.051) | (0.044) | (0.045) |
| Portugal | effect | | | | −0.114 | | |
| | se | | | | (0.038) | | |
| Sweden | effect | −0.093 | −0.103 | −0.110 | | | |
| (1st break) | se | (0.020) | (0.021) | (0.022) | | | |
| Sweden | effect | | | | −0.158 | −0.145 | |
| (2nd break) | se | | | | (0.033) | (0.029) | |

Supplementary Table 8: Robustness Check: Squared log(population size). This table shows the relative impact of treatments on $CO_2$ emissions and their standard errors for models with the covariates $log(GDP)$, $log(GDP)^2$, $log(population)$, and $log(population)^2$.

| Country | | Model | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| | | EU-15 | EU-15 | EU-15 | EU-31 | EU-31 | EU-31 |
| significance level in break detection | | 5% | 1% | 0.1% | 5% | 1% | 0.1% |
| Denmark | effect | | | | −0.088 | | |
| | se | | | | (0.037) | | |
| Finland | effect | −0.134 | −0.130 | −0.131 | −0.130 | −0.138 | |
| | se | (0.022) | (0.024) | (0.024) | (0.041) | (0.036) | |
| Germany | effect | −0.144 | −0.165 | −0.112 | −0.234 | −0.195 | |
| | se | (0.021) | (0.022) | (0.022) | (0.036) | (0.033) | |
| Ireland | effect | −0.047 | | −0.127 | | | |
| (1st break) | se | (0.024) | | (0.023) | | | |
| Ireland | effect | −0.145 | −0.159 | | −0.263 | −0.244 | −0.280 |
| (2nd break) | se | (0.034) | (0.031) | | (0.050) | (0.044) | (0.045) |
| Luxembourg | effect | −0.143 | | | −0.016 | | |
| (1st break) | se | (0.029) | | | (0.055) | | |
| Luxembourg | effect | | | −0.211 | −0.077 | −0.138 | −0.078 |
| (2nd break) | se | | | (0.031) | (0.051) | (0.043) | (0.044) |
| Portugal | effect | | | | −0.044 | | |
| | se | | | | (0.039) | | |
| Sweden | effect | −0.109 | −0.115 | −0.114 | | | |
| (1st break) | se | (0.020) | (0.022) | (0.022) | | | |
| Sweden | effect | | | | −0.165 | −0.153 | |
| (2nd break) | se | | | | (0.034) | (0.030) | |

Supplementary Table 9: Robustness Check: Urban population. This table shows the relative impact of treatments on $CO_2$ emissions and their standard errors for models with the covariates $log(GDP)$, $log(GDP)^2$, $log(population)$, and $Urban\ population$.

| Country | | Model | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| | | EU-15 | EU-15 | EU-15 | EU-31 | EU-31 | EU-31 |
| significance level in break detection | | 5% | 1% | 0.1% | 5% | 1% | 0.1% |
| Denmark | effect | | | | −0.066 | | |
| | se | | | | (0.036) | | |
| Finland | effect | −0.133 | −0.129 | −0.118 | −0.117 | −0.121 | |
| | se | (0.022) | (0.024) | (0.024) | (0.040) | (0.035) | |
| Germany | effect | −0.143 | −0.165 | −0.109 | −0.228 | −0.189 | |
| | se | (0.021) | (0.022) | (0.022) | (0.035) | (0.031) | |
| Ireland | effect | −0.051 | | −0.150 | | | |
| (1st break) | se | (0.025) | | (0.023) | | | |
| Ireland | effect | −0.149 | −0.164 | | −0.354 | −0.306 | −0.340 |
| (2nd break) | se | (0.034) | (0.032) | | (0.051) | (0.044) | (0.046) |
| Luxembourg | effect | −0.142 | | | −0.004 | | |
| (1st break) | se | (0.029) | | | (0.054) | | |
| Luxembourg | effect | | | −0.206 | −0.070 | −0.119 | −0.054 |
| (2nd break) | se | | | (0.031) | (0.049) | (0.042) | (0.044) |
| Portugal | effect | | | | −0.187 | | |
| | se | | | | (0.045) | | |
| Sweden | effect | −0.109 | −0.115 | −0.109 | | | |
| (1st break) | se | (0.020) | (0.022) | (0.022) | | | |
| Sweden | effect | | | | −0.156 | −0.145 | |
| (2nd break) | se | | | | (0.033) | (0.029) | |

Supplementary Table 10: Robustness Check: Urban population, squared. This table shows the relative impact of treatments on $CO_2$ emissions and their standard errors for models with the covariates $log(GDP)$, $log(GDP)^2$, $log(population)$, $Urban\ population$ and $(Urban\ population)^2$.

| Country | | Model | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| | | EU-15 | EU-15 | EU-15 | EU-31 | EU-31 | EU-31 |
| significance level in break detection | | 5% | 1% | 0.1% | 5% | 1% | 0.1% |
| Denmark | effect | | | | −0.124 | | |
| | se | | | | (0.049) | | |
| Finland | effect | −0.176 | −0.162 | −0.145 | −0.105 | −0.124 | |
| | se | (0.028) | (0.030) | (0.029) | (0.048) | (0.044) | |
| Germany | effect | −0.115 | −0.106 | −0.222 | −0.131 | −0.153 | |
| | se | (0.034) | (0.037) | (0.031) | (0.052) | (0.048) | |
| Ireland | effect | −0.105 | | −0.090 | | | |
| (1st break) | se | (0.034) | | (0.035) | | | |
| Ireland | effect | −0.193 | −0.197 | | −0.260 | −0.306 | −0.313 |
| (2nd break) | se | (0.038) | (0.039) | | (0.059) | (0.048) | (0.045) |
| Luxembourg | effect | −0.087 | | | −0.232 | | |
| (1st break) | se | (0.032) | | | (0.058) | | |
| Luxembourg | effect | | | −0.242 | −0.338 | −0.255 | −0.371 |
| (2nd break) | se | | | (0.034) | (0.058) | (0.051) | (0.042) |
| Portugal | effect | | | | −0.062 | | |
| | se | | | | (0.059) | | |
| Sweden | effect | −0.060 | −0.037 | −0.013 | | | |
| (1st break) | se | (0.028) | (0.030) | (0.030) | | | |
| Sweden | effect | | | | −0.019 | −0.027 | |
| (2nd break) | se | | | | (0.055) | (0.051) | |

Supplementary Table 11: Robustness Check: Individual Country Trends. This table shows the relative impact of treatments on $CO_2$ emissions and their standard errors for models with the covariates $log(GDP)$, $log(GDP)^2$, $log(population)$, and country-specific time trends.

### 3.6 Oster test

Break detection methods evaluate the presence of breaks relative to a specified model, in our case a model that controls for log(GDP), log(GDP)$^2$ and log(population size). Misspecification of this model may lead to the detection of spurious breaks. Therefore, we use the specification test suggested by Oster [14] to evaluate the robustness of our results with respect to omitted variable bias. Supplementary Table 12 reports Oster's *delta*. This proportional selection coefficient estimates how much more important selection on unobservables has to be relative to the selection on observables to reduce our estimated effects to zero. For $\delta > 1$, Oster [14] considers coefficients as stable. The delta values corresponding to our main results reported in Table 12 are either far above 1 or negative, which suggests that the model we use to detect breaks is highly robust. For instance, the unobservables underlying the coefficient for the break in Germany in the EU-15 sample at the 5% false discovery rate (column 1) would need to increase about 28-fold relative to the included control variables to nullify our finding. Negative test statistics indicate trivially robust coefficients because adding additional control variables would only reduce the importance of selection on unobservables relative to observables.

| Country | Model | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| | EU-15 | EU-15 | EU-15 | EU-31 | EU-31 | EU-31 |
| significance level in break detection | 5% | 1% | 0.1% | 5% | 1% | 0.1% |
| Denmark | | | | −57.4 | | |
| Finland | 110.6 | 125.4 | 110.1 | −16.8 | −13.2 | |
| Germany | 28.2 | 27.7 | 19.2 | −45.3 | −32.0 | |
| Ireland (1st break) | −94.0 | | −70.1 | | | |
| Ireland (2nd break) | −96.1 | −88.3 | | −43.8 | −31.6 | −25.8 |
| Luxembourg (1st break) | −2.9 | | | −1.5 | | |
| Luxembourg (2nd break) | | | −7.6 | −8.6 | −7.3 | −6.8 |
| Portugal | | | | 104.9 | | |
| Sweden (1st break) | 374.1 | 344.1 | 359.5 | | | |
| Sweden (2nd break) | | | | −27.7 | −20.6 | |

Supplementary Table 12: Specification test by Oster [14]. $\delta$ measures how much more important selection on unobservables has to be relative to the selection on observables to reduce our estimated effects to zero. For $\delta > 1$, Oster [14] considers coefficients as stable. Negative test statistics indicate trivially robust coefficients because adding additional control variables would only reduce the importance of selection on unobservables relative to observables.

## Supplementary References


[1] Pretis F, Schwarz M. Discovering What Mattered: Answering Reverse Causal Questions by Detecting Unknown Treatment Assignment and Timing as Breaks in Panel Models. Working Paper. 2022;.

[2] Wooldridge J. Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators. Working Paper Available at SSRN 3906345. 2021;.

[3] Goodman-Bacon A. Difference-in-differences with variation in treatment timing. Journal of Econometrics. 2021;225(2):254–277.

[4] Callaway B, Sant'Anna PH. Difference-in-differences with multiple time periods. Journal of Econometrics. 2021;225(2):200–230.

[5] Baker A, Larcker DF, Wang CC. How Much Should We Trust Staggered Difference-In-Differences Estimates? Available at SSRN 3794018. 2021;.

[6] Castle JL, Doornik JA, Hendry DF, Pretis F. Detecting location shifts during model selection by step-indicator saturation. Econometrics. 2015;3(2):240–264.

[7] Walker AJ, Pretis F, Powell-Smith A, Goldacre B. Variation in responsiveness to warranted behaviour change among NHS clinicians: Novel implementation of change detection methods in longitudinal prescribing data. BMJ. 2019;367.

[8] Turiel JS, Kaufmann RK. Evidence of air quality data misreporting in China: An impulse indicator saturation model comparison of local government-reported and US embassy-reported PM2. 5 concentrations (2015–2017). PLOS ONE. 2021;16(4):e0249063.

[9] Doornik JA. Autometrics. In: Castle JL, Shephard N, editors. The Methodology and Practice of Econometrics. Oxford: Oxford University Press; 2009. p. 88–121.

[10] Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267–288.

[11] Zou H. The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association. 2006;101(476):1418–1429.

[12] Okui R, Wang W. Heterogeneous structural breaks in panel data models. Journal of Econometrics. 2021;220(2):447–473.

[13] Hendry DF, Krolzig HM. The properties of automatic Gets modelling. The Economic Journal. 2005;115(502):C32–C61.

[14] Oster E. Unobservable selection and coefficient stability: Theory and evidence. Journal of Business & Economic Statistics. 2019;37(2):187–204.

# Works Cited in Part I and III


Barreca, Alan et al. (2016). "Adapting to Climate Change: The Remarkable Decline in the US Temperature-Mortality Relationship over the Twentieth Century". In: *Journal of Political Economy* 124.1, pp. 105–159.

Berlemann, Michael and Daniela Wenzel (2018). "Hurricanes, Economic Growth and Transmission Channels". In: *World Development* 105.C, pp. 231–247.

Boyd, Emily et al. (2017). "A Typology of Loss and Damage Perspectives". In: *Nature Climate Change* 7.10, pp. 723–729.

Burke, Marshall, W. Matthew Davis, and Noah S. Diffenbaugh (2018). "Large Potential Reduction in Economic Damages under UN Mitigation Targets". In: *Nature* 557.7706, pp. 549–553. DOI: 10.1038/s41586-018-0071-9.

Burke, Marshall and Kyle Emerick (2016). "Adaptation to Climate Change: Evidence from US Agriculture". In: *American Economic Journal: Economic Policy* 8.3, pp. 106–140.

Burke, Marshall, Solomon M. Hsiang, and Edward Miguel (2015). "Global Non-Linear Effect of Temperature on Economic Production". In: *Nature* 527.7577, pp. 235–239. DOI: 10.1038/nature15725.

CarbonBrief (2022). *COP27: Key Outcomes Agreed at the UN Climate Talks in Sharm El-Sheikh - Carbon Brief*. https://www.carbonbrief.org/cop27-key-outcomes-agreed-at-the-un-climate-talks-in-sharm-el-sheikh/.

Carleton, Tamma A et al. (2020). *Valuing the Global Mortality Consequences of Climate Change Accounting for Adaptation Costs and Benefits*. Tech. rep. National Bureau of Economic Research.

Cohen, Francois et al. (2020). "The Challenge of Using Epidemiological Case Count Data: The Example of Confirmed COVID-19 Cases and the Weather". In: 76.4, pp. 1187–1213. DOI: 10.1101/2020.05.21.20108803.

Davis, Steven J. et al. (May 2022). "Emissions Rebound from the COVID-19 Pandemic". In: *Nature Climate Change* 12.5, pp. 412–414. DOI: 10.1038/s41558-022-01332-6.

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken (2012). "Temperature Shocks and Economic Growth: Evidence from the Last Half Century". In: *American Economic Journal: Macroeconomics* 4.3, pp. 66–95. DOI: 10.1257/mac.4.3.66.

Dietz, Simon et al. (2021). "Economic Impacts of Tipping Points in the Climate System". In: *Proceedings of the National Academy of Sciences* 118.34, e2103081118.

Donat, M. G. et al. (Mar. 2013). "Updated Analyses of Temperature and Precipitation Extreme Indices since the Beginning of the Twentieth Century: The HadEX2 Dataset: HADEX2-GLOBAL GRIDDED CLIMATE EXTREMES". In: *Journal of Geophysical Research: Atmospheres* 118.5, pp. 2098–2118. DOI: 10.1002/jgrd.50150.

Dow, Kirstin et al. (Apr. 2013). "Limits to Adaptation". In: *Nature Climate Change* 3.4, pp. 305–307. DOI: 10.1038/nclimate1847.

Edmonds, Christopher and Ilan Noy (Jan. 2018). "The Economics of Disaster Risks and Impacts in the Pacific". In: *Disaster Prevention and Management: An International Journal* 27.5, pp. 478–494. DOI: 10.1108/DPM-02-2018-0057.

Engels, Anita et al. (Feb. 2023). *Hamburg Climate Futures Outlook: The Plausibility of a 1.5°C Limit to Global Warming - Social Drivers and Physical Processes*. Tech. rep. Universität Hamburg. DOI: 10.25592/UHHFDM.11230.

Eskander, Shaikh M. S. U. and Sam Fankhauser (Aug. 2020). "Reduction in Greenhouse Gas Emissions from National Climate Legislation". In: *Nature Climate Change* 10.8, pp. 750–756. DOI: 10.1038/s41558-020-0831-z.

Farmer, J. et al. (2015). "A Third Wave in the Economics of Climate Change". In: *Environmental & Resource Economics* 62.2, pp. 329–357.

Felbermayr, Gabriel and Jasmin Gröschl (Nov. 2014). "Naturally Negative: The Growth Effects of Natural Disasters". In: *Journal of Development Economics*. Special Issue: Imbalances in Economic Development 111, pp. 92–106. DOI: 10.1016/j.jdeveco.2014.07.004.

Fomby, Thomas, Yuki Ikeda, and Norman V. Loayza (2013). "The Growth Aftermath of Natural Disasters". In: *Journal of applied econometrics* 28.3, pp. 412–434.

Gall, Melanie, Kevin A. Borden, and Susan L. Cutter (June 2009). "When Do Losses Count?: Six Fallacies of Natural Hazards Loss Data". In: *Bulletin of the American Meteorological Society* 90.6, pp. 799–810. DOI: 10.1175/2008BAMS2721.1.

Gelman, Andrew (2011). *Causality and Statistical Learning*.

Gelman, Andrew and Guido Imbens (Nov. 2013). *Why Ask Why? Forward Causal Inference and Reverse Causal Questions*. Working Paper 19614. National Bureau of Economic Research. DOI: 10.3386/w19614.

Grubb, Michael, Claudia Wieners, and Pu Yang (May 2021). "Modeling Myths: On DICE and Dynamic Realism in Integrated Assessment Models of Climate Change Mitigation". In: *WIREs Climate Change* 12.3. DOI: 10.1002/wcc.698.

Hänsel, Martin C et al. (2020). "Climate Economics Support for the UN Climate Targets". In: *Nature Climate Change* 10.8, pp. 781–789.

Harrington, Luke J and Friederike EL Otto (2020). "Reconciling Theory with the Reality of African Heatwaves". In: *Nature Climate Change* 10.9, pp. 796–798.

Hendry, David F (2020). *First in, First out: Econometric Modelling of UK Annual CO2 Emissions, 1860–2017*. Tech. rep. Oxford: Economics Group, Nuffield College, University of Oxford.

Hendry, David F, Soren Johansen, and Carlos Santos (2008). "Automatic Selection of Indicators in a Fully Saturated Regression". In: *Computational Statistics* 23.2, pp. 317–335.

Hinkel, Jochen et al. (2014). "Coastal Flood Damage and Adaptation Costs under 21st Century Sea-Level Rise". In: *Proceedings of the National Academy of Sciences* 111.9, pp. 3292–3297.

Howard, Peter H. and Thomas Sterner (Sept. 2017). "Few and Not So Far Between: A Meta-analysis of Climate Damage Estimates". In: *Environmental and Resource Economics* 68.1, pp. 197–225. DOI: 10.1007/s10640-017-0166-z.

Hsiang, Solomon (2016). "Climate Econometrics". In: *Annual Review of Resource Economics* 8.1, pp. 43–75. DOI: 10.1146/annurev-resource-100815-095343.

Hsiang, Solomon et al. (June 2017). "Estimating Economic Damage from Climate Change in the United States". In: *Science* 356.6345, pp. 1362–1369. DOI: 10.1126/science.aal4369.

Hsiang, Solomon M. and Amir S. Jina (July 2014). *The Causal Effect of Environmental Catastrophe on Long-Run Economic Growth: Evidence From 6,700 Cyclones*. Working Paper. DOI: 10.3386/w20352.

IPCC (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of IPCC the Intergovernmental Panel on Climate Change*. Ed. by Thomas F Stocker et al. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

— (2014). *Climate Change 2014: Mitigation of Climate Change Working Group III Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. New York NY: Cambridge University Press.

— (2022a). *Climate Change 2022: Impacts, Adaptation and Vulnerability*. Summary for Policymakers. Cambridge, UK and New York, USA: Cambridge University Press, pp. 3–33.

— (2022b). "Summary for Policymakers". In: *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by H. O. Pörtner et al. Cambridge, UK: Cambridge University Press, In Press.

Kalkuhl, Matthias and Leonie Wenz (2020). "The Impact of Climate Conditions on Economic Production. Evidence from a Global Panel of Regions". In: *Journal of Environmental Economics and Management* 103, p. 102360.

Kotz, Maximilian, Anders Levermann, and Leonie Wenz (Jan. 2022). "The Effect of Rainfall Changes on Economic Production". In: *Nature* 601.7892, pp. 223–227. DOI: 10.1038/s41586-021-04283-8.

Lenton, Timothy M et al. (2008). "Tipping Elements in the Earth's Climate System". In: *Proceedings of the national Academy of Sciences* 105.6, pp. 1786–1793.

Loayza, Norman V. et al. (2012). "Natural Disasters and Growth: Going beyond the Averages". In: *World Development* 40.7, pp. 1317–1336.

Majda, Andrew J. and Boris Gershgorin (2010). "Quantifying Uncertainty in Climate Change Science through Empirical Information Theory". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.34, pp. 14958–14963. DOI: 10.1073/pnas.1007009107.

Mark, Ebba, Ryan Rafaty, and Moritz Schwarz (2022). "Spatial-Temporal Dynamics of Employment Shocks in Declining Coal Mining Regions and Potentialities of the 'Just Transition'". In: DOI: 10.48550/ARXIV.2211.12619.

Martinez, Andrew B. (2020a). "Forecast Accuracy Matters for Hurricane Damage". In: *Econometrics* 8.2, p. 18.

— (2020b). "Improving Normalized Hurricane Damages". In: *Nature Sustainability* 3.7, pp. 517–518.

Matsuura, Kenji and Cort J. Willmott (2018). "Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1901 - 2017): Version 5. Data Provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA". In.

Menne, Matthew J. et al. (July 2012). "An Overview of the Global Historical Climatology Network-Daily Database". In: *Journal of Atmospheric and Oceanic Technology* 29.7, pp. 897–910. DOI: 10.1175/JTECH-D-11-00103.1.

Mill, John Stuart (1843). *A System of Logic*. Vol. 1. London: Parker.

Newbold, Stephen and Alex Marten (2014). "The Value of Information for Integrated Assessment Models of Climate Change". In: *Journal of Environmental Economics and Management* 68.1, pp. 111–123.

Newell, Richard G, Brian C Prest, and Steven E Sexton (2021). "The GDP-temperature Relationship: Implications for Climate Change Damages". In: *Journal of Environmental Economics and Management* 108, p. 102445.

Nordhaus, William (2018). "Evolution of Modeling of the Economics of Global Warming: Changes in the DICE Model, 1992–2017". In: *Climatic Change* 148.4, pp. 623–640. DOI: 10.1007/s10584-018-2218-y.

Nordhaus, William and Paul Sztorc (2013). "DICE 2013R: Introduction and User's Manual". In: *Yale University and the National Bureau of Economic Research, USA*.

Nordhaus, William D. (1992). "The 'DICE'Model: Background and Structure of a Dynamic Integrated Climate-Economy Model of the Economics of Global Warming". In.

Otto, Friederike EL et al. (2016). "The Attribution Question". In: *Nature Climate Change* 6.9, pp. 813–816.

Panwar, Vikrant and Subir Sen (2020). "Disaster Damage Records of EM-DAT and DesInventar: A Systematic Comparison". In: *Economics of Disasters and Climate Change* 4.2, pp. 295–317.

Perron, Pierre (1989). "The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis". In: *Econometrica: journal of the Econometric Society*, pp. 1361–1401.

Pindyck, Robert S. (2013). "Climate Change Policy: What Do the Models Tell Us?" In: *Journal of Economic Literature* 51.3, pp. 860–872. DOI: 10.1257/jel.51.3.860.

Pretis, Felix et al. (2018). "Uncertain Impacts on Economic Growth When Stabilizing Global Temperatures at 1.5°C or 2°C Warming". In: *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 376.2119. DOI: 10.1098/rsta.2016.0460.

Rafaty, Ryan, Geoffroy Dolphin, and Felix Pretis (2022). "Carbon Pricing and the Elasticity of Co2 Emissions". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.4188459.

Riahi, Keywan et al. (2017). "The Shared Socioeconomic Pathways and Their Energy, Land Use, and Greenhouse Gas Emissions Implications: An Overview". In: *Global Environmental Change* 42, pp. 153–168. DOI: 10.1016/j.gloenvcha.2016.05.009.

Roughgarden, Tim and Stephen H. Schneider (1999). "Climate Change Policy: Quantifying Uncertainties for Damages and Optimal Carbon Taxes". In: *Energy Policy* 27.7, pp. 415–429. DOI: 10.1016/S0301-4215(99)00030-0.

Ruppert, David and Raymond J Carroll (1980). "Trimmed Least Squares Estimation in the Linear Model". In: *Journal of the American Statistical Association* 75.372, pp. 828–838.

Scheer, Antonina et al. (Sept. 2022). "Whose Jobs Face Transition Risk in Alberta? Understanding Sectoral Employment Precarity in an Oil-Rich Canadian Province". In: *Climate Policy* 22.8, pp. 1016–1032. DOI: 10.1080/14693062.2022.2086843.

Schlenker, Wolfram and Michael J Roberts (2009). "Nonlinear Temperature Effects Indicate Severe Damages to US Crop Yields under Climate Change". In: *Proceedings of the National Academy of sciences* 106.37, pp. 15594–15598.

Stern, Nicholas (2008). "The Economics of Climate Change". In: *American Economic Review* 98.2, pp. 1–37. DOI: 10.1257/aer.98.2.1.

Sterner, Thomas (2015). "Higher Costs of Climate Change". In: *Nature* 527.7577, pp. 177–178.

Stock, James H. and Mark W. Watson (2015). *Introduction to Econometrics*. Updated third edition, global edition. The Pearson Series in Economics. Harlow: Pearson Education Limited / Askews and Holts, distributor.

Strobl, Eric (2011). "The Economic Growth Impact of Hurricanes: Evidence from US Coastal Counties". In: *Review of Economics and Statistics* 93.2, pp. 575–589.

— (2012). "The Economic Growth Impact of Natural Disasters in Developing Countries: Evidence from Hurricane Strikes in the Central American and Caribbean Regions". In: *Journal of Development economics* 97.1, pp. 130–141.

Thalheimer, Lisa, Moritz P. Schwarz, and Felix Pretis (Mar. 2023). "Large Weather and Conflict Effects on Internal Displacement in Somalia with Little Evidence of Feedback onto Conflict". In: *Global Environmental Change* 79, p. 102641. DOI: 10.1016/j.gloenvcha.2023.102641.

Tol, Richard S. J. (June 2009). "The Economic Effects of Climate Change". In: *Journal of Economic Perspectives* 23.2, pp. 29–51. DOI: 10.1257/jep.23.2.29.

Tol, Richard S.J., Samuel Fankhauser, and Joel B. Smith (June 1998). "The Scope for Adaptation to Climate Change: What Can We Learn from the Impact Literature?" In: *Global Environmental Change* 8.2, pp. 109–123. DOI: 10.1016/S0959-3780(98)00004-1.

Wilson, Charlie et al. (2017). "Evaluating Process-Based Integrated Assessment Models of Climate Change Mitigation". In: *IIASA Working Paper WP-17-007*.