Protected

MSR

Working Papers No. 29

July 2026

Did Tariff Exposure Bend National Emission Trajectories? A Pre-Trends Cautionary Tale from the US–China Trade War

Diwani, M., Hamid, M., Santangelo, G.

Publisher and Distributor

M&S Research Hub institute

Landgraf Karl Str. 11, 34131 Kassel, Germany

Telephone +49 0 15225886301

Email: info@ms-researchhub.com,

Web: www.ms-researchhub.com

Copyright © M&S Research Hub 2018-2026

All rights reserved.

No part of this publication may be reproduced in any form or by any electronic or mechanical means, without a written permission from the publisher.

The findings, interpretations and conclusions expressed in this publication are entirely those of the author(s).

.

Did Tariff Exposure Bend National Emission Trajectories? A Pre-Trends Cautionary Tale from the US–China Trade War

Dr. Mazen Diwani
Senior Lecturer in Economics, MS Research Hub, Germany
meezdin86@gmail.com

Dr. Mohamed Hamid
Lecturer in Economics, University of the People, USA
azumnile2013@gmail.com

Dr. Giacomo Santangelo
Senior Lecturer in Economics, Fordham University, USA
gsantangelo1@fordham.edu

Abstract

The 2018 US–China trade war was the sharpest escalation of trade barriers among major economies in decades, yet its consequences for carbon emissions remain little studied. We ask whether tariff exposure altered the national CO_2 emission trajectories of affected countries, using propensity score matching across 171 countries over 2015–2019 (162 on common support), with treatment defined on trade-weighted tariff quartiles. A conventional matched comparison appears to show high-exposure countries — predominantly rapidly industrializing economies — growing emissions 4.5 percentage points faster than controls ($p = 0.007$), a gap stable across alternative algorithm labels and sample exclusions. This apparent robustness is illusory. The properly weighted matched ATT — the intended estimand — is never statistically distinguishable from zero (0.28 pp before a pre-trend control, 0.32 pp after). The significant 4.5 pp figure is instead an unweighted two-sample comparison, the wrong estimator for a matched design, and does not correct itself: it stays 4.4 pp ($p < 0.01$) even with the pre-trend covariate in the match. A placebo over 2015–2016, before any tariff, finds them already growing 2.4 pp faster than controls ($p = 0.009$). We find no detectable effect on national trajectories beyond pre-existing dynamics, but the design's limits are real: the confidence interval is wide (we can exclude effects above roughly 6 pp, not moderate ones), the window is short, and national aggregates miss the cross-border reallocation channel. The transferable lesson is methodological: baseline covariate balance offers no protection against selection on growth dynamics, and placebo and pre-trend diagnostics are part of identification, not optional refinements.

Keywords: Propensity score matching · Parallel trends · Pre-treatment trends · Carbon emissions · US–China trade war · Policy evaluation

JEL Classification: F18 · Q54 · Q56 · O13 · C21

1 Introduction

The 2018 US–China trade war was the sharpest escalation of trade barriers among major economies in decades. Its economic consequences — for trade volumes, employment, prices, and firm performance — have been extensively documented. Its environmental consequences have not. Would production reallocation push activity toward countries with weaker emissions standards? Could disrupted supply chains slow the diffusion of cleaner technologies? And what, if anything, does a trade conflict of this scale reveal about the link between trade policy and climate outcomes?

This paper asks a deceptively simple question: did the trade war move CO_2 emissions in countries that faced higher tariff exposure? The answer — no detectable causal effect on national emission trajectories — may be unsurprising in retrospect, but arriving there was not straightforward. The central difficulty is one that plagues observational policy evaluation broadly: separating genuine treatment effects from pre-existing trends when the same structural characteristics that attracted tariffs also drive emissions growth independently.

Matching countries on 2016 baseline emissions, GDP, GDP per capita, and population yields an apparent 4.5-percentage-point gap in emissions growth between high- and low-exposure groups over 2016–2019 ($p = 0.007$). The pattern looks solid: it appears within the first year of the trade war, barely moves when China or India is dropped, is echoed — at smaller, non-significant magnitudes — by doubly robust specifications, and looks stable across alternative matching algorithms, though as Section 4.2 notes those algorithm labels re-report a single underlying test rather than independent estimates.

The robustness is an illusion, and the way it dissolves is the heart of the paper. Two facts, established under estimators held fixed, jointly deliver the null. First, the *properly weighted* matched estimator — the average treatment effect on the treated (ATT) that propensity score matching is designed to produce — is statistically indistinguishable from zero from the outset, 0.28 pp before any pre-trend control and 0.32 pp after it. There was never a significant matched effect to "collapse." Second, the significant 4.5 pp gap comes from an *unweighted* two-sample mean comparison — the wrong estimator for a matched design — and this estimator does not correct itself: it remains 4.4 pp ($p = 0.009$) even after 2015–2016 emission growth is added as a matching covariate. A diagnostic meant to neutralize a pre-existing trend leaves the naive comparison essentially unchanged, and that stubbornness is the tell. What reveals the null is therefore not the unweighted comparison "collapsing" but the properly weighted estimand (never statistically distinguishable from zero) together with a placebo test over 2015–2016, before any tariff, which finds high-exposure countries already growing emissions 2.4 percentage points faster than matched controls ($p = 0.009$). Under the weighted estimand the matched groups grow at nearly the same rate — 7.6 percent in high-exposure countries versus 7.3 percent in controls.

This paper makes a focused contribution. Methodologically, it is a replication and extension of Daw and Hatfield (2018), who show that matching on pre-treatment outcome levels can manufacture spurious estimates when those levels correlate with treatment assignment. We reproduce that mechanism on policy-salient trade–environment data and add a further, practically important twist: an apparent effect can survive a full battery of conventional robustness checks — alternative algorithms, sample exclusions, doubly robust estimators — and still dissolve entirely once a placebo test and a pre-trend control are run. The demonstration is that unweighted post-match comparison

and level-only matching can *jointly* fabricate a "robust" effect that neither diagnostic on its own would expose. Substantively, we provide a cross-country assessment of the 2018 trade war's effect on national CO_2 trajectories and find none beyond pre-existing industrialization dynamics. We are deliberate about scope: our design speaks only to national aggregate emissions and is silent on the cross-border reallocation channel that likely carries the trade war's true environmental footprint. Our precision is moderate — the weighted estimand can rule out national emission-growth effects larger than about 6 percentage points but not smaller ones — which bounds how strongly the null should be read.

The correlation between tariff exposure and pre-existing emissions dynamics is not accidental. Countries with low labor costs, expanding manufacturing sectors, and deep integration into global supply chains attract tariffs precisely because those features make them competitive threats to domestic producers — and those same features drive rapid industrialization and rising emissions. Vietnam, Bangladesh, Cambodia, and the Philippines did not face elevated tariff exposure by chance; they did so because of structural characteristics that also fuel emissions growth during development transitions.

That trade policy leaves no detectable imprint on national emission trajectories here does not mean environmental considerations are irrelevant to trade policy design. It means something more specific: trade interventions have limited leverage over near-term national emissions in economies undergoing structural transformation, where emissions are largely determined by energy systems, production technologies, and infrastructure — things that evolve over decades and bend only weakly in response to short-run trade policy shifts.

Section 2 reviews the relevant literature on trade–environment interactions, causal inference methods, and the economic evidence on the US–China trade war. Section 3 describes our empirical strategy, including the difference-in-differences design we initially pursued and abandoned when parallel trends failed, and the propensity score matching approach we ultimately adopted. Section 4 presents the results: the apparently significant initial finding, the robustness checks that sustain it, the falsification tests that expose it, and the null estimate that survives once pre-treatment growth is controlled. Sections 5 and 6 turn to policy implications and conclusions.

2 Literature Review

2.1 Trade and Environmental Outcomes: Theoretical Frameworks

2.1.1 The Scale, Composition, and Technique Effects

The standard framework for analyzing trade's environmental consequences decomposes overall impacts into three channels (Grossman and Krueger, 1991). The scale effect captures the additional pollution that comes with higher output, holding technology and sector composition fixed. The composition effect reflects shifts in a country's sectoral structure driven by comparative advantage: trade that pushes an economy toward dirtier industries raises emissions, while trade toward cleaner ones reduces them. The technique effect refers to improvements in production processes and emissions control arising from technological advancement, higher income levels, and more stringent environmental regulation, which tend to reduce pollution intensity over time.

Formally, total emissions E can be written as the product of a scale term, a composition term, and a technique term,

$$E = Y \sum_s \frac{Y_s}{Y} \frac{E_s}{Y_s}, \tag{1}$$

where Y is total output (scale), Y_s/Y is the share of sector s in output (composition), and E_s/Y_s is emission intensity in sector s (technique). A trade war can operate through all three: reduced trade may lower Y (scale); production may shift toward more pollution-intensive sectors (composition); and disrupted technology transfer may slow efficiency improvements (technique). Early theoretical models indicated that these effects could work in opposing directions, making the net impact ambiguous (Copeland and Taylor, 2004), and empirical work disentangling them (Stern, 2004) has been a central concern of environmental economics for decades.

2.1.2 The Environmental Kuznets Curve and Development Trajectories

A related strand focuses on the Environmental Kuznets Curve (EKC), which posits that emissions first rise with income and then decline as economies grow rich enough to adopt cleaner technologies, tighten regulation, and shift toward less pollution-intensive activities (Grossman and Krueger, 1995). A common EKC specification is

$$\ln E_{it} = \alpha + \beta_1 \ln y_{it} + \beta_2 (\ln y_{it})^2 + \gamma' Z_{it} + \varepsilon_{it}, \tag{2}$$

where E_{it} is environmental degradation (here CO_2 emissions) in country i at time t, y_{it} is GDP per capita, and Z_{it} collects other controls (population, urbanization, trade openness). The EKC hypothesis requires $\beta_2 < 0$, with a turning point at $\ln y^* = -\beta_1/(2\beta_2)$.

The implication for trade is that development stage mediates the environmental consequences of policy changes: a tariff shock that hits a rapidly industrializing economy operates in a very different context than one that hits a mature, services-oriented one. Countries with high tariff exposure in our sample are concentrated on the upward-sloping segment of the EKC, in a phase of rapid industrialization; observed increases in their emissions may therefore reflect underlying development dynamics rather than the causal effect of trade policy. Recent work casts doubt on the general applicability of the EKC for CO_2, where sustained energy-consumption growth can outweigh efficiency gains even at higher incomes (Stern, 2004). Nonetheless, development stage remains a critical determinant of emission trajectories: the countries most exposed to US tariffs in our sample are predominantly middle-income economies undergoing industrial expansion, a stage at which emissions typically rise regardless of trade policy (Dinda, 2004).

2.2 Empirical Evidence on Trade Liberalization and Emissions

Empirical work on trade and the environment has produced mixed results, which is unsurprising given how much the relevant mechanisms depend on context. Cross-country panel studies find both improvements and deterioration following trade liberalization, with the direction varying by comparative advantage, regulatory capacity, industrial composition, and development stage. These inconsistencies are plausibly attributable to that heterogeneity rather than to any single stable relationship. Recent studies continue to document non-linear, context-dependent patterns. Using the Regional Comprehensive Economic Partnership as a case study, Tian et al. (2022) estimate that complete tariff

elimination would increase global CO_2 emissions from fuel combustion by roughly 3.1 percent annually — more than double the historical average growth rate — driven mainly by developing member countries, though technology spillovers from deeper integration may partially offset this over time.

Causal identification remains difficult throughout this literature. Trade policy is not exogenous; liberalization decisions correlate with economic and environmental conditions in ways that are hard to observe fully; and incomplete measurement of regulatory capacity, technology adoption, and industrial structure creates omitted-variable problems that standard panel methods do not easily resolve.

2.3 Methodological Approaches to Causal Inference in Policy Evaluation

2.3.1 Difference-in-Differences and Parallel Trends

Difference-in-differences (DiD) remains the workhorse of applied policy evaluation, constructing counterfactuals by comparing how outcomes change over time across treated and control groups. Its credibility rests on the parallel trends assumption: absent the intervention, the two groups would have followed similar trajectories. Recent work has sharpened both the tools for testing this assumption and the conceptual care it requires. Roth et al. (2023) provide an integrative overview, categorizing advances into settings with multiple periods and variation in treatment timing, departures from parallel trends, and new approaches to inference. Caetano and Callaway (2024) study settings where parallel trends holds only conditional on covariates, showing that two-way fixed-effects specifications can suffer hidden bias when covariates are necessary for identification. Roth and Sant'Anna (2023) show that parallel trends is generally not preserved under functional-form transformations unless restrictive conditions hold, so the assumption should be evaluated with respect to the chosen outcome specification.

When parallel trends fails — as it does here, with high-exposure countries exhibiting systematically different pre-treatment emission patterns — the DiD estimator conflates genuine treatment effects with pre-existing differentials. Modern practice increasingly relies on event-study designs that estimate separate parameters for multiple pre-treatment periods, facilitating both visual inspection and formal testing (de Chaisemartin and D'Haultfœuille, 2020, 2023). In our data, the pre-treatment coefficients are jointly non-zero ($F(6, 170) = 3.87$, $p = 0.0012$; see Fig. 2), a clear violation of the identifying assumption that motivated our pivot away from DiD.

2.3.2 The Power and Pitfalls of Pre-Trend Testing

A pre-trend test is only as informative as its power, and a recent literature warns against reading a passed test as licence. Roth (2022) shows that conventional pre-tests are frequently underpowered against exactly the trend violations that would most bias treatment-effect estimates, and that conditioning on having passed a pre-test can itself distort inference. Rambachan and Roth (2023) develop a "more credible" approach that, rather than testing parallel trends as a binary hypothesis, characterizes how conclusions change across a transparent range of possible pre-trend violations and reports the robust set of estimates. This frontier is directly relevant to our setting. Our concern is the mirror image of an underpowered pre-test: we *detect* a significant pre-treatment differential and must decide what it implies. As Rambachan and Roth (2023) make clear, the honest

response is to ask whether the substantive conclusion survives across plausible characterizations of the violation. In our case it does — the properly weighted matched estimand is indistinguishable from zero throughout, and the one estimator that stays significant is diagnosed as the wrong estimator for a matched design rather than as evidence of an effect (Section 4).

2.3.3 Propensity Score Matching and Falsification

Propensity score matching, developed by Rosenbaum and Rubin (1983), takes a different route to identification. Rather than exploiting time variation, it constructs treated and control groups similar on observed pre-treatment characteristics, assuming selection into treatment is driven by observables. When that conditional independence assumption holds, comparisons within the matched sample identify the ATT (Heckman et al., 1998). Effective implementation requires careful attention to model specification and diagnostics: Zhao et al. (2021) emphasize that although PSM has proven valuable across observational datasets, credible application depends on careful specification of the propensity score model and substantive knowledge of the factors driving treatment selection.

Crucially, good baseline balance does not guarantee balance on growth trajectories, and selection on unobservables can survive matching on levels (Daw and Hatfield, 2018). Placebo tests in pre-treatment periods are therefore essential in matching studies, not optional refinements. Ryan et al. (2015) test multiple estimators under parallel-trends violations and find that matching combined with DiD can add robustness but does not eliminate the need for pre-trend verification. Daw and Hatfield (2018) provide the central cautionary result: when the pre-treatment outcome level is correlated with treatment assignment, matching on that level can produce biased estimates, with bias increasing in the pre-treatment level gap and decreasing in serial correlation. Consistent with these concerns — and, importantly, without asserting that we have cleanly identified a *slope* rather than a level or anticipatory difference — our finding of a significant pre-treatment differential (Section 4.3) shows that countries destined for high tariff exposure were already on different short-run emission dynamics before treatment. We engage the competing readings of that differential, including regression to the mean and anticipation, directly in Section 4.

2.3.4 Doubly-Robust Estimation

Doubly robust estimators hedge against misspecification by combining propensity score weighting with outcome regression: consistency requires only that one of the two models is correct. As we show below, in our data these estimators return positive point estimates in the 2.0–2.3 pp range that are individually non-significant — but, critically, they are computed *without* the pre-trend covariate, and so inherit the same level-only-matching vulnerability as the unweighted baseline comparison rather than resolving it.

2.4 The US–China Trade War: Economic and Environmental Context

2.4.1 Structure and Evolution of Trade War Tariffs

By most measures, the 2018 US–China trade war was the most sweeping increase in US trade barriers since the 1930s. Tariffs on solar panels and washing machines came first,

followed by steel and aluminum, and eventually duties covering hundreds of billions of dollars in Chinese goods; China retaliated across a broad front. China was the primary target, but the tariff structure swept in many others. Rapidly developing economies across Asia, Africa, and Latin America — including Vietnam, Bangladesh, Cambodia, and the Philippines — had export-oriented manufacturing, deep supply-chain integration, and ongoing industrialization: precisely the characteristics that drew tariff scrutiny and independently drive emissions growth. That non-random selection is the identification challenge at the center of this paper.

2.4.2 Economic Impacts and Supply Chain Adjustments

Research on the trade war's economic consequences documents substantial impacts on trade flows, prices, and welfare, including reduced US export growth, lower employment, and negative effects on aggregate real income in both the US and China. Alfaro and Chor (2023) document that between 2017 and 2022 China's share of US imports fell from 21.6 to 16.3 percent — back to 2007 levels — as importers turned to new sources, particularly Vietnam and Mexico. Fajgelbaum et al. (2024), using gravity models, find that higher US tariffs significantly reduced China's exports to the US while redirecting them toward third-country markets. Freund et al. (2024), analyzing US import data through 2022, likewise find that the 2018 tariffs cut China's share of US imports from about 22 to 16 percent, with sourcing reallocated toward Mexico and Vietnam — countries that increasingly source intermediates from China, illustrating the persistence of China's role in global supply chains despite bilateral decoupling.

2.5 The Correlation Between Development Stage and Trade War Exposure

That high-exposure countries are predominantly fast-growing developing economies is not a quirk of the data; it reflects how US trade policy was designed and what global manufacturing looks like. Labor-cost advantages, expanding manufacturing capacity, and supply-chain integration made these economies attractive production platforms and competitive threats to domestic US producers, which is why they attracted tariffs. The result is a fundamental correlation between tariff exposure and development stage — and that correlation is what makes causal inference here so difficult. Recent studies on supply-chain reallocation during 2018–2023 confirm the pattern: the trade war accelerated pre-existing diversification away from China toward other emerging markets, with the most exposed countries already on steep industrialization trajectories. Alfaro and Chor (2023) note that while China dominated US imports through the 2000s, the post-2017 period saw significant reallocation, with Vietnam gaining nearly two percentage points of US import market share. The implication is that high-exposure countries would likely have experienced faster emission growth than others even absent the trade war, simply by virtue of their position along development trajectories.

2.6 BRICS Climate Ambitions and Policy Responses

Recent research documents diverse emission profiles across the BRICS economies (Brazil, Russia, India, China, and South Africa). Basso and Viola (2022) find that between 1990 and 2019 Chinese greenhouse-gas emissions rose 317 percent, with the energy sector's

share increasing from 81.64 to 88.08 percent; India's emissions grew 79.87 percent over the same period, driven by its energy sector, a pattern characteristic of industrialization; Brazil's profile differs, with land use and forestry accounting for the largest share. Sandrin and Amigo (2024), drawing on Climate Action Tracker data, find that all five BRICS countries' policies and Nationally Determined Contributions fall short of Paris Agreement targets, with ratings from "insufficient" (Brazil, South Africa) to "highly insufficient" (China, India) to "critically insufficient" (Russia).

2.7 Implications for This Study

Several threads matter directly for what follows. On the substantive question, theory offers no clear prior and the empirical literature no consensus. On methodology, the recent literature has been increasingly explicit that baseline covariate balance does not guarantee balance in growth dynamics, and that falsification and pre-trend diagnostics are part of the identification argument (Daw and Hatfield, 2018; Roth, 2022; Rambachan and Roth, 2023). On the trade war itself, studies have thoroughly documented supply-chain diversion toward Mexico and Vietnam and the resilience of Chinese export volumes (Alfaro and Chor, 2023; Fajgelbaum et al., 2024; Freund et al., 2024), while environmental consequences remain comparatively underexplored. Our contribution is to apply careful causal inference to the trade war's environmental consequences while systematically testing the identifying assumptions on which those methods depend.

3 Methodology

3.1 Empirical Strategy and Identification Challenges

Data sources. Carbon dioxide emissions data come from the Emissions Database for Global Atmospheric Research (EDGAR), which provides disaggregated emissions for fossil-fuel combustion and industrial processes. Tariff-exposure measures are constructed at the country level from product-level tariff data in the World Integrated Trade Solution (WITS) database, compiled from UNCTAD TRAINS and WTO IDB sources. Control variables — GDP, GDP per capita, population, and energy consumption per capita — come from the World Bank's World Development Indicators. Emissions are drawn from the EDGAR 2022 release; the WITS tariff extract covers 2018–2021, of which the 2018–2019 lines fall within our panel window and define treatment.

One data-quality caveat bears directly on our identification concern. EDGAR's national totals are constructed partly from modelled activity data, and coverage quality is weakest for exactly the small, data-poor, rapidly industrializing economies that populate our treated arm. Measurement error in the outcome is therefore likely concentrated in the treated group, which can inflate an apparent treated–control differential and feeds the regression-to-the-mean concern we take up below.

Tariff-exposure measure. Our treatment is the trade-weighted average US tariff rate each country faces, in percent. For each country-year it is $\sum$(tariff rate $\times$ US import value)/$\sum$(US import value) across fourteen pollution-intensive HS-2 sectors — iron and steel, chemicals, mineral fuels, textiles and apparel, and transport equipment; the full list of chapters is in Online Resource, Table S10 — from WITS tariff and US import data; we average this rate over 2018–2019 and cut it into quartiles, defining treatment as membership in the top exposure quartile — the 44 highest-exposure

countries, with mean rates of roughly 6 percent and above — of which 42 lie on common support. Restricting exposure to these emissions-relevant sectors is a deliberate design choice: it concentrates the measure on the trade most likely to bear on production and emissions, rather than diluting it across all of the country's exports to the US. Two features bear on interpretation. First, it is a tariff *rate*, not a dollar measure of exposure: because the weights are the value of US imports from that country in each sector (i.e., the country's sectoral exports to the US), a small overall trader can still face a high trade-weighted average rate when its exports to the US are concentrated in high-tariff sectors. Second, quartile membership is therefore not a function of a country's total US trade volume — within the treated quartile the correlation between the tariff rate and total mean trade volume is only -0.12. Section 4.2 confirms the results are unchanged when the smallest-trade economies are removed from the treated arm.

Our initial approach was a difference-in-differences design with an event-study specification. Countries were grouped into quartiles by the trade-weighted average US tariff rate they faced during 2018–2019, creating time-invariant treatment groups tracked from 2015 through 2019:

$$\ln E_{it} = \delta_i + \lambda_t + \sum_{q \in \{2,3,4\}} \sum_{\tau \neq 0} \beta_{q,\tau} \left(\text{Quartile}_{iq} \times \text{Year}_{\tau t} \right) + \varepsilon_{it}, \tag{3}$$

where i indexes countries, t indexes years (2015–2019), $\text{Quartile}_{iq} = 1$ if country i is in exposure quartile q (Q1 the reference), $\text{Year}_{\tau t}$ indicates year relative to 2017 with $\tau \in \{-2, -1, 0, 1, 2\}$, and δ_i, λ_t are country and year fixed effects with errors clustered by country. Under parallel trends, the pre-treatment coefficients $\beta_{q,-2}$ and $\beta_{q,-1}$ in Equation (3) should be zero. A joint test rejects decisively ($F(6, 170) = 3.87$, $p = 0.0012$), with pre-treatment coefficients ranging from -2.4 to -6.3 percentage points relative to the low-exposure reference (Fig. 2).

One qualification deserves emphasis, and it disciplines how we interpret this rejection. Parallel trends requires that *growth rates* would have evolved comparably absent the intervention, not that emission *levels* be equal in the pre-period; level differences are absorbed by country fixed effects and are not themselves a violation. The coefficients above blend level and growth differences, and the joint F-test does not isolate the growth component. We therefore do not claim to have identified a pre-existing *slope*; we claim only that treated and control countries differed in their short-run pre-treatment dynamics in a way we cannot fully decompose into level, slope, and anticipation. The decision to abandon DiD nonetheless stands on independent grounds: with only five years of data, collinearity prevents separate identification of group-specific trends alongside country and year fixed effects (Appendix A). A longer panel would permit event-study designs that separate pre-trend levels from slopes; we flag this as a priority for future work.

3.2 Propensity Score Matching Approach

We turn instead to propensity score matching. Treatment is membership in the highest-exposure quartile (Q4); the control group is drawn from Q1–Q3. We estimate a logit propensity score

$$\begin{gathered} \Pr(T_i = 1 \mid X_i) = \Lambda(X_i'\beta), \\ X_i = [\, 1,\ \ln \text{CO}_{2i,2016},\ \ln \text{GDP}_{i,2016},\ \ln \text{GDPpc}_{i,2016},\ \ln \text{Pop}_{i,2016} \,], \end{gathered} \tag{4}$$

with coefficient vector $\beta = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4]$ and $\Lambda(\cdot)$ the logistic CDF. We match on four covariates measured in 2016 — the last complete pre-war year — capturing baseline emission level, economic size, development stage, and country scale. The outcome is the change in log CO_2 emissions,

$$\Delta \ln \mathrm{CO}_{2i} = \ln \mathrm{CO}_{2i,2019} - \ln \mathrm{CO}_{2i,2016}. \tag{5}$$

We implement logistic matching with three nearest neighbors and a caliper of 0.05 standard deviations, retaining only treated units on common support. The ATT is estimated as

$$\widehat{\mathrm{ATT}} = \frac{1}{N_T} \sum_{i \in T} \left(Y_i - \sum_{j \in C(i)} w_{ij} \, Y_j \right), \tag{6}$$

where N_T is the number of treated units on common support, $C(i)$ the matched controls for treated unit i, and w_{ij} the matching weights. This weighted ATT, Equation (6), is the estimand propensity score matching is designed to deliver — as distinct from the unweighted two-sample comparison of the outcome (5) across groups — and it is the quantity we anchor our conclusions on. We report standardized bias statistics and a joint balance test; standard errors do not adjust for estimation of the propensity score, which may slightly understate them.

4 Results

4.1 Main Propensity Score Matching Results

4.1.1 Covariate Balance and Common Support

Of the 171 countries with complete emissions and tariff data, 164 had the full set of 2016 matching covariates; of these, matching placed 162 (98.8 percent) on common support, excluding two high-exposure countries with extreme propensity scores. Figure 1 shows the resulting overlap in estimated propensity scores, from the logit model in Equation (4), between treated and control units. Mean standardized bias fell from 45.3 percent before matching to 10.8 percent after — well below the conventional 25 percent threshold — and the joint balance test cannot reject balance ($\chi^2 = 0.73$, $p = 0.948$). Table 1 reports the covariate-by-covariate diagnostics. The seven countries dropped for incomplete 2016 covariates (Antigua and Barbuda, East Timor, Eritrea, Gibraltar, Marshall Islands, Romania, and Venezuela) are all low- or moderate-exposure controls; none is a treated unit, so the attrition does not selectively thin the treated arm.

4.1.2 Treatment Effect Estimates and the Estimator Question

The simple mean comparison on the matched sample shows high-exposure countries growing emissions 7.7 percent over 2016–2019 versus 3.2 percent for matched controls — a difference of 4.5 percentage points, significant at the 1 percent level ($t = -2.73$, $p = 0.007$). This is an *unweighted* two-sample comparison: it discards the matching weights and counts each retained control once. The estimand that propensity score matching is designed to deliver is the *weighted* ATT, and that estimate tells a different story. Applying the three-nearest-neighbor matching weights, the ATT is 0.28 percentage points ($t = 0.14$, $p = 0.89$) — statistically indistinguishable from zero. The two numbers are produced

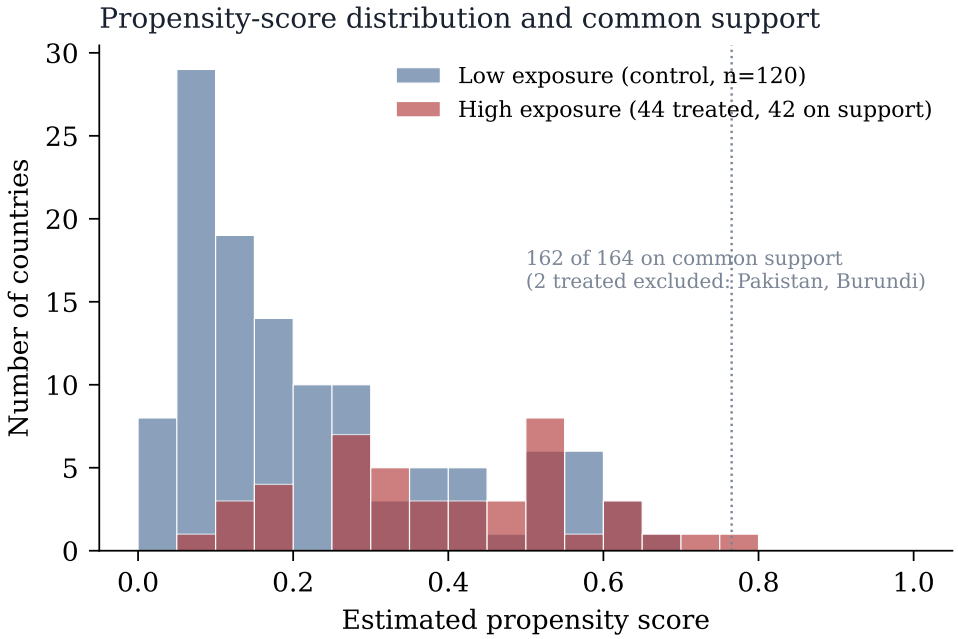

Figure 1. Common-support overlap. Distribution of estimated propensity scores for treated (high-exposure, Q4) and control (Q1–Q3) countries. The region of common support covers 162 of 164 countries (98.8 percent); two treated units with extreme scores fall off support and are excluded.

Table 1. Propensity score matching balance diagnostics

Variable	Sample	Mean (T)	Mean (C)	% bias	% bias red.	t	$p > \|t\|$
ln CO_2 2016	Unmatched	16.681	17.200	−20.2		−1.19	0.237
	Matched	16.630	16.317	12.2	39.7	0.57	0.574
ln GDP 2016	Unmatched	23.601	24.641	−43.7		−2.41	0.017
	Matched	23.578	23.213	15.3	64.9	0.73	0.466
ln GDP per capita 2016	Unmatched	7.640	9.004	−102.7		−5.58	0.000
	Matched	7.761	7.780	−1.4	98.6	−0.08	0.936
ln Population 2016	Unmatched	15.902	15.568	14.5		0.85	0.397
	Matched	15.815	15.486	14.3	1.2	0.65	0.517

Notes: Sample summary — unmatched: mean bias 45.3, LR $\chi^2 = 31.06$ ($p = 0.000$); matched: mean bias 10.8, LR $\chi^2 = 0.73$ ($p = 0.948$). Common support: 120 controls and 42 of 44 treated units retained ($N = 162$). Data sources: EDGAR, World Bank, WITS.

by two different estimators on the same matched data, and disclosing that distinction is essential to reading the results correctly.

Table 2 makes the point transparent by crossing the estimator (unweighted mean comparison vs. weighted ATT) with the presence of a pre-treatment-trend control, whose construction we describe in Section 4.3.2.

Table 2. Emission-growth differential, high- vs. low-exposure countries, by estimator and pre-trend control

Estimator	(a) Baseline match — no pre-trend control	(b) Augmented match — with pre-trend control
Unweighted two-sample mean comparison (42 treated vs. 120 controls on common support)	**4.5 pp** ($t = -2.73$, $p = 0.007$)***	**4.4 pp** ($t = -2.64$, $p = 0.009$)***
Weighted matched ATT (psmatch2, 3-NN, caliper 0.05)	0.28 pp ($t = 0.14$, $p = 0.89$)	0.32 pp ($t = 0.15$, $p = 0.88$)

Notes: pp = percentage points of 2016–2019 log-CO_2 growth. The pre-trend control adds 2015–2016 emission growth as a fifth matching covariate. Both unweighted cells (top row) are significant at 1 percent; both weighted matched ATTs (bottom row) are indistinguishable from zero. The reversal runs across the rows (estimator), not the columns: adding the pre-trend covariate changes neither estimator's verdict — only the choice of estimator does. The weighted matched ATT — the estimand propensity score matching is designed to deliver — is never statistically distinguishable from zero. The unweighted comparison is over 42 treated and 120 control units on common support; the weighted ATT applies three-nearest-neighbour weights to those same 42 treated units. Values are reproduced in the replication package. *** significant at 1 percent.

The table isolates the source of the apparent effect. The causal null does **not** depend on the pre-trend control: the properly weighted ATT is indistinguishable from zero both before (0.28 pp) and after (0.32 pp) that control enters. The significant 4.5 pp, by contrast, comes entirely from using the wrong estimator — an unweighted two-sample comparison on matched data — and that estimator does not respond to the pre-trend diagnostic: it remains 4.4 pp ($p = 0.009$) when 2015–2016 growth is added to the match. The reversal in the table runs across the rows (estimator), not the columns (pre-trend control): whichever way the pre-trend enters, the weighted estimand reads ≈0 and the unweighted comparison reads ≈4.5 pp. It is the estimator, not the covariate, that separates the spurious effect from the null.

A regression-adjusted ATT that conditions on the matching covariates within the matched sample gives 2.08 percentage points (SE = 1.89, $t = 1.10$, $p = 0.275$) — not significant, and we do not characterize it as a treatment effect. (The corresponding "≈140 percent higher" figure sometimes attached to this estimate in fact derives from the 4.5 pp unweighted mean gap, not from the non-significant regression-adjusted coefficient, and we do not describe a non-significant estimate as economically substantial.)

Because the null rests on a non-rejection, its informativeness depends on precision. The weighted matched ATT is estimated precisely enough to exclude large effects but not moderate ones: the 95 percent confidence interval is approximately $[-3.9, +4.5]$ percentage points — essentially unchanged whether the propensity-score-based standard error or the Abadie–Imbens matching standard error (which adjusts for estimation of the propen-

sity score) is used — implying a minimum detectable effect near 6 percentage points at 80 percent power (Online Resource, Table S7). We can therefore rule out an emission-growth effect larger than about 6 pp; we cannot rule out effects of moderate magnitude, and we do not claim the effect is precisely zero.

4.2 Robustness Checks

The unweighted 4.5 pp baseline comparison is remarkably stable across a conventional robustness battery, and it is precisely that stability — under an estimator that Section 4.3 shows to be compromised — that motivates our cautionary point. In brief: simple mean comparisons yield 4.5 percentage points across three alternative matching algorithms (kernel, radius, and one-to-one, all $p = 0.007$) — though we note that the source reports the identical 3-nearest-neighbour two-sample test under each algorithm header rather than independently re-estimated, algorithm-specific effects, so this agreement is not independent corroboration (Online Resource, Table S1) — with identical 98.8 percent common-support retention; the differential is already present in the first year (3.0 pp over 2016–2018, $p = 0.021$) and cumulates to 4.5 pp by 2019; it is essentially unchanged when the largest emitters are excluded; and doubly robust estimators return positive but non-significant estimates of 2.0–2.3 pp. Full detail appears in the Online Resource (Tables S1–S4). Two features of this battery deserve main-text emphasis because they bear on the reversal.

First, on the largest emitters: China — the world's largest emitter — is the single BRICS economy in the high-exposure (Q4) treatment group, so we re-estimate the effect with China removed from the treated units. India — the third-largest emitter — falls in the moderate-exposure quartile (Q3) and therefore enters as a control; we accordingly re-estimate with India removed from the control pool as a separate sensitivity check. In both cases the estimate is materially unchanged (excluding China, 4.5 pp, $p = 0.008$; excluding India, 4.6 pp, $p = 0.007$). Second, and decisively, the entire robustness battery is computed on level-only matching *without* the pre-treatment-trend control. As Daw and Hatfield (2018) show, level-only matching is exactly the design that manufactures a spurious but stable estimate when treatment correlates with pre-trends. The consistency across algorithms is therefore not evidence of a genuine effect; it is evidence that these estimators share a common vulnerability. The falsification tests in Section 4.3 identify it.

Two further checks, reported in Appendix B, reinforce this reading and remain subordinate to it. Randomization inference—2,000 random reassignments of the high-exposure label across the covariate-complete countries, re-estimating the full match each draw—places the weighted matched ATT at the centre of its placebo distribution (57th percentile, permutation $p = 0.88$), while the unweighted comparison falls in the extreme tail ($p < 0.01$). The unweighted gap is significant under randomization precisely because it captures the real selection that defines the treated group, which random reassignment destroys; the weighted estimand, which removes that selection, is unremarkable. This is the same estimator contrast obtained a third, independent way, not evidence of an effect. A leave-one-out over all 44 treated countries—beyond the China and India exclusions above—leaves both quantities intact: the unweighted comparison stays within $[+3.9, +4.9]$ pp and the weighted matched ATT within $[-0.5, +2.3]$ pp (its largest excursion, from dropping Greenland, remains far below the ≈ 6 pp minimum detectable effect), so neither the artifact nor the null hinges on any single country.

A further check addresses the composition of the treated arm directly. The high-

exposure quartile is dominated by genuine manufacturing and low-income exporters — Vietnam, Bangladesh, Cambodia, Sri Lanka, Pakistan, and Indonesia, alongside high-tariff least-developed economies such as Burundi, Comoros, and Mauritania — and only two or three units (Micronesia, Nauru, Kiribati) are genuinely small-trade economies for which the tariff-rate coding may overstate real exposure (Online Resource, Table S8). Removing these small economies from the treated arm — whether by dropping the six smallest-economy units flagged in review or by imposing a minimum-trade threshold — leaves all three findings intact (Online Resource, Table S9): the unweighted comparison stays significant and if anything strengthens (from 4.5 to as much as 5.1 pp, $p < 0.01$), the weighted matched ATT remains statistically indistinguishable from zero in every subset (largest $|t| = 1.52$), and the placebo gap stays near 2.5 pp ($p \approx 0.01$). We disclose one sensitivity honestly: the weighted ATT *point estimate* drifts upward on the trade-exposed subsets, from 0.28 pp to between 1.8 and 3.1 pp. It remains non-significant and well inside the ≈ 6 pp minimum detectable effect, so the null is unchanged as an inference, but the estimand is composition-sensitive and we do not read it as a hard zero.

4.3 Critical Falsification Tests

4.3.1 Placebo Test Using the Pre-Treatment Period

A critical test of the identification strategy is whether we observe spurious "treatment effects" in periods before any tariff. We re-run the matching design using 2015 as the baseline and 2016 as the placebo "post" year, holding the treatment assignment (based on actual 2018–2019 exposure) and the matching procedure fixed.

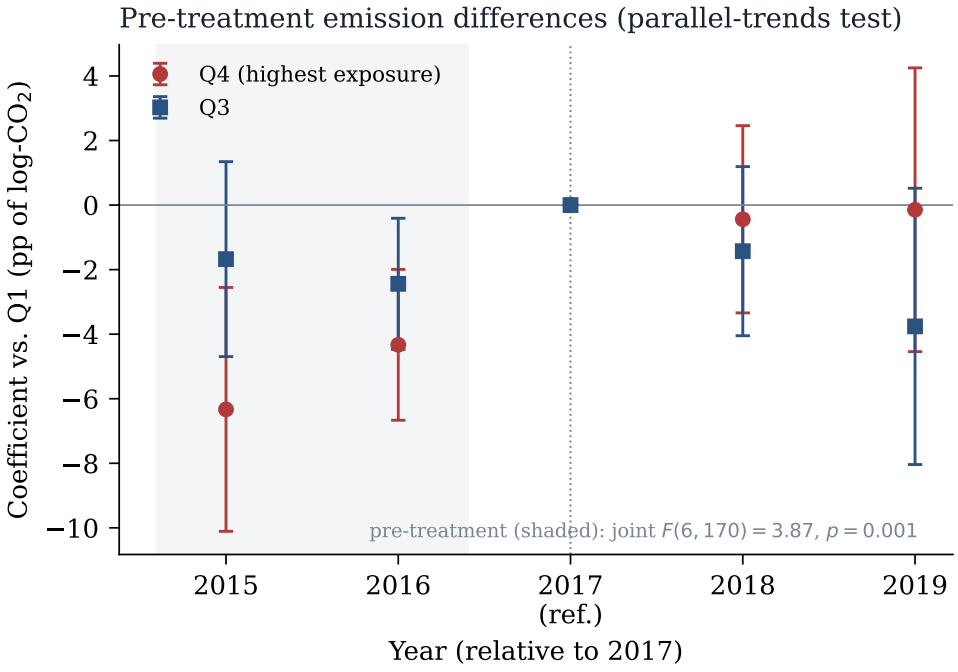

Figure 2. Event-study pre-treatment coefficients (test of parallel trends). Estimated quartile-by-year coefficients relative to the low-exposure reference group, with 95 percent confidence intervals, from the event-study difference-in-differences specification. The jointly significant pre-treatment coefficients ($F(6, 170) = 3.87$, $p = 0.0012$) reject parallel trends and motivate the pivot to matching.

Table 3. Placebo test results (pre-treatment period 2015–2016)

Group	Obs.	Mean growth	Std. err.
Low exposure (control)	120	1.5%	0.0044
High exposure (treated)	42	3.9%	0.0089
Difference		**2.4 pp**	

Notes: Two-sample t-test on the matched support, $t = -2.64$, $p = 0.009$. Data sources: EDGAR, World Bank, WITS.

During 2015–2016, before any tariff was imposed, high-exposure countries grew emissions by 3.9 percent while matched controls grew by 1.5 percent — a 2.4-percentage-point gap that is statistically significant ($t = -2.64$, $p = 0.009$). Countries destined for high tariff exposure were already on different short-run emission dynamics, which means the matching design's implicit assumption — that treated and control groups would have behaved similarly absent treatment — was not satisfied.

Two competing readings of this gap deserve to be confronted rather than assumed away, and the null survives under each. The first is *regression to the mean*, the mechanism Daw and Hatfield (2018) emphasize: if treated countries were matched on a transiently low 2016 level, a mechanical rebound could generate an apparent differential in either window. The second is *anticipation*: Section 301 investigations were announced in August 2017 and tariff threats were widely reported, so redirected investment may have begun accelerating industrialization in exposed countries before formal tariffs. Both readings imply that the observed post-2018 differential reflects something other than a causal tariff effect on national trajectories, and neither restores a treatment effect. What the placebo cannot do — with a single pre-treatment year — is adjudicate between a persistent structural trend, transient variation, and anticipation; a naïve annualized comparison of the one-year placebo against the three-year post rate should not be pushed, since single-year growth rates are noisier and the placebo confidence interval is wide enough to encompass the post-treatment rate. The placebo establishes that a pre-existing differential exists; it is not designed to characterize its precise rate of evolution. For the placebo window specifically, the timing cuts against anticipation: the 2015–2016 gap predates the August 2017 Section 301 investigations by well over a year, so it cannot reflect anticipatory reallocation triggered by the tariff threat, and is more naturally read as a pre-existing structural trend. Anticipation remains relevant for the 2017 pre-treatment year and for interpreting the pre-trend covariate, but not for the placebo gap itself.

The pre-existing gap is not an artifact of the noisiest treated economies. Repeating the placebo with each treated country dropped in turn, the 2015–2016 differential stays within $[1.79, 2.58]$ percentage points and remains significant at every drop (largest $p = 0.036$; the most influential single exclusion, Laos, lowers the gap only to 1.79 pp). Measurement error concentrated in a few small, data-poor economies therefore cannot account for the pre-treatment differential, which is robust to the removal of any single country (Online Resource, Table S9).

4.3.2 Controlling for Pre-Treatment Trends

To address the placebo finding directly, we augment the matching specification with the 2015–2016 emission growth rate as a fifth matching covariate alongside the four 2016-level variables, defining the pre-trend as $\ln \mathrm{CO}_{2i,2016} - \ln \mathrm{CO}_{2i,2015}$. This creates matched

groups similar not only in 2016 levels and economic characteristics but also in recent emission dynamics.

Table 4. Controlling for pre-treatment trends (augmented match)

Diagnostic / estimate	Value
Pre-trend covariate standardized bias, before → after matching	47.9% → 10.8% (77.3% reduction)
Weighted matched ATT	0.32 pp (SE = 2.15, $t = 0.15$, $p = 0.88$)
Regression-adjusted ATT	1.15 pp (SE = 1.76, $t = 0.65$, $p = 0.515$)
Unweighted two-sample comparison	4.4 pp (7.6% treated vs. 3.2% control; $t = -2.64$, $p = 0.009$)

Notes: Adding the pre-trend covariate cuts overall standardized bias to excellent levels (joint $\chi^2 = 0.74$, $p = 0.981$; pre-trend logit coefficient $z = 2.20$, $p = 0.028$). The weighted matched ATT and the regression-adjusted ATT are both indistinguishable from zero; the unweighted two-sample comparison, which ignores the matching weights, remains significant — confirming that the apparent effect is an estimator artifact, not a treatment effect that the pre-trend covariate removes. All values are reproduced in the replication package. Data sources: EDGAR, World Bank, WITS.

Once pre-treatment growth enters the match, the properly weighted estimate remains indistinguishable from zero: the weighted ATT is 0.32 pp ($p = 0.88$), essentially where it started (0.28 pp), with matched groups growing at 7.6 versus 7.3 percent (Online Resource, Table S6). The unweighted two-sample comparison, however, does not move — it remains 4.4 pp ($t = -2.64$, $p = 0.009$), because matching on the pre-trend reweights the estimand but leaves untouched a comparison that discards the matching weights. The regression-adjusted estimate does fall, from 2.08 pp to 1.15 pp ($p = 0.515$), but like the weighted ATT it was never significant. This is the central result, and Table 2 shows its logic: the significant 4.5 pp differential is an artifact of the unweighted estimator, not a treatment effect that the pre-trend covariate removes, and it reflects the continuation of pre-existing growth dynamics rather than anything the trade war caused. High-exposure countries — fast-growing economies in Asia, Africa, and Latin America — were emitting more rapidly before 2018 because they were industrializing, and continued doing so after 2018 for the same reason.

One more question concerns the event-study design we set aside. Rather than assume its rejected parallel-trends assumption away, we ask what it would imply about the post-treatment effect once a bounded pre-trend violation is allowed, using the Rambachan and Roth (2023) relative-magnitudes restriction on the high-exposure arm. Even permitting the post-period violation to reach twice the largest pre-period violation, the robust confidence set for the post-period effect still includes zero (Appendix B.1). Because the post-period event-study coefficients are themselves near zero, relaxing the bound only widens an already zero-centred interval and never manufactures an effect. This concerns the set-aside difference-in-differences design, not the matching null, which rests on the weighted ATT and the placebo.

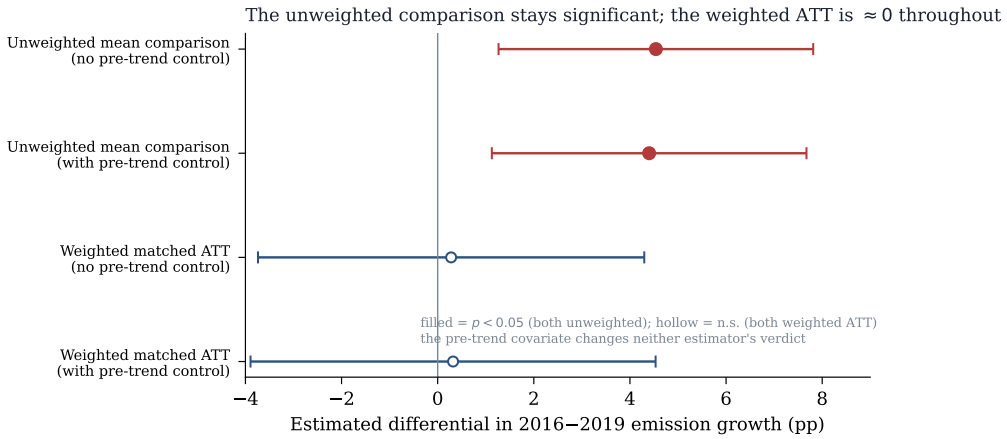

Figure 3. Estimator reversal. Emission-growth differential across specifications. The unweighted two-sample comparison is significant both before and after the pre-trend covariate enters the match (4.5 and 4.4 pp; filled markers), whereas the weighted matched ATT — the estimand matching is designed to deliver — is indistinguishable from zero throughout (0.28 pp baseline, 0.32 pp augmented; hollow markers). The reversal is across estimators, not across the pre-trend control: adding the pre-trend covariate changes neither estimator's verdict.

4.4 Heterogeneity Analyses Not Feasible Due to Data Limitations

We had intended to test whether emission patterns differed for the BRICS economies or by trade orientation. Neither was feasible, for reasons that are themselves informative. BRICS membership is near-completely separated from treatment: Brazil and Russia fall in Q1, South Africa in Q2, and India in Q3, while China is the sole BRICS economy in the high-exposure quartile (Q4). Of the 44 Q4 countries, 43 are non-BRICS. China's presence in Q4 reflects that its trade-weighted average tariff exposure, though aggregated from many moderate product-level rates, is high enough to reach the top quartile; the other four BRICS economies remain below that threshold. Because propensity score matching requires overlap within each subsample, this near-complete separation precludes estimating heterogeneous BRICS effects (Online Resource, Table S5). A second intended split, by trade orientation (trade value relative to GDP in 2016), was likewise infeasible: trade-value data were missing for a non-random subset of high-exposure countries, so treated units were systematically absent from the trade-orientation subsamples (Online Resource, Table S5). We do not attach a precise count to that missingness here; the count stated in earlier drafts did not reconcile against the estimation data and is corrected in the replication materials.

4.5 Interpretation and Limitations

4.5.1 Summary of Empirical Findings

Our findings must be read together. First, a conventional level-only matched comparison identifies a 4.5-percentage-point emission-growth differential (2016–2019, $p = 0.007$) that is robust to alternative algorithms, temporal windows, exclusion of major emitters, and doubly robust estimation — but this stability holds only under an unweighted estimator on level-only matches; the properly weighted matched ATT is never statistically

distinguishable from zero (and no larger than about 3 pp on any subsample; Section 4.2). Second, falsification testing reveals a significant 2.4-percentage-point pre-treatment differential (2015–2016, $p = 0.009$); augmenting the match to control for pre-treatment growth leaves the weighted ATT at zero (0.32 pp, $p = 0.88$) and the regression-adjusted estimate insignificant (1.15 pp, $p = 0.515$), while the unweighted two-sample comparison — the source of the headline effect — itself remains significant (4.4 pp, $p = 0.009$), because it is unaffected by the reweighting. The null is carried by the correct estimand and the placebo, not by the pre-trend covariate correcting the naive comparison. Third, intended heterogeneity analyses were infeasible: four of the five BRICS economies (Brazil, Russia, South Africa, India) fell into the low-to-moderate exposure quartiles, while China was the single BRICS economy in the high-exposure quartile — a near-complete, not complete, separation between BRICS status and treatment — and missing trade-orientation data for exposed countries precluded the trade-orientation split.

We state the causal conclusion at the weakest level the design supports. We find no evidence that tariff exposure altered national emission trajectories *beyond pre-existing dynamics*; we do not claim the effect is precisely zero. The confidence interval is wide and cannot rule out effects of moderate magnitude, and — as Section 4.5.2 explains — the pre-trend control may itself be partly endogenous.

4.5.2 Causal Interpretation, Scope, and Contribution

The falsification tests reframe the analysis. The 4.5-percentage-point differential during 2016–2019 tracks the continuation of pre-existing industrialization dynamics, not the causal footprint of tariff exposure. The appropriate summary is the weighted matched null, which the placebo and pre-trend diagnostics confirm reflects pre-existing dynamics rather than a treatment effect.

Two limitations bound the claim. First, **scope**. This paper examines CO_2 at the national aggregate level, but the most consequential environmental mechanism of a trade war operates through reallocation. If production shifts from Chinese facilities to plants elsewhere with different emissions intensities — older capital, heavier coal dependence, weaker process standards — global emissions could rise or fall even where no single country's national trajectory shows a discontinuity. Our framework cannot detect this channel. Alfaro and Chor (2023) and Freund et al. (2024) document that substantial reallocation occurred; whether it was emissions-neutral, harmful, or beneficial globally is a question this methodology cannot answer. A null on national trajectories is not evidence that the supply-chain reorganization was environmentally neutral. Second, **anticipation**. The trade war did not begin without warning in January 2018; Section 301 investigations were announced in August 2017 and firms were already redirecting investment. If anticipatory reallocation accelerated industrialization in high-exposure countries during 2015–2017, the pre-trend growth rate we use as the fifth matching covariate is not a clean confounder but partly endogenous to the treatment pathway, and controlling for it would bias the estimate toward zero. The analysis is observationally equivalent under two hypotheses — that the trade war had no effect, and that anticipatory reallocation began before formal tariffs and is absorbed by the pre-trend control. Distinguishing them would require investment-flow and production-capacity data at finer temporal resolution than annual national emissions allow.

These limitations sharpen rather than diminish the paper's contribution, which is best understood as a replication and extension of Daw and Hatfield (2018). Their result

is that matching on pre-treatment outcome levels can produce biased estimates when levels correlate with treatment; we demonstrate that mechanism on policy-salient trade–environment data and show, further, that the resulting spurious estimate can survive an entire conventional robustness battery and be exposed only by a placebo test and a pre-trend control. The methodological lesson is transferable well beyond this episode; the substantive lesson is scoped narrowly to national aggregate CO_2 over a single, short, well-anticipated policy episode.

Read through the scale–composition–technique lens (Grossman and Krueger, 1991), our null is specifically a null on the composition-plus-technique response of national aggregates: tariff exposure did not measurably shift exposed countries onto a different national emission path within the window. It is silent on the between-country composition shift that reallocation produces — the margin on which Tian et al. (2022) locate the largest trade-policy emission effects. The practical implication for trade–environment evaluation is therefore sharper than a generic call for placebo tests: because tariff exposure is assigned on the same development-stage characteristics that drive the technique effect, national-aggregate matching and DiD will systematically confound the two, and credible evaluation of a trade shock's environmental footprint must either difference out development-stage trajectories explicitly or move to the sub-national or plant level where the reallocation margin is observable.

4.5.3 Implications for Research Design

Two design lessons follow. First, analyses must test for and account for pre-existing differential dynamics between treated and control groups; placebo tests in pre-treatment periods and pre-trend controls should be standard practice in applied matching studies, because matching on levels alone is insufficient when treatment assignment correlates with growth dynamics. Second, separating trade-policy effects from underlying development trajectories favors alternative identification strategies: event studies of sharp, unanticipated policy changes; instrumental variables exploiting exposure variation unrelated to development stage; or synthetic control methods that match explicitly on pre-treatment trends as well as levels.

5 Policy Implications

Our central result is a null specific to this episode. Over 2015–2019, and once pre-existing dynamics are accounted for, we find no evidence that exposure to the 2018 US–China tariffs altered the national emission trajectories of affected countries. The most exposed economies were already on faster emission paths before the trade war — driven by industrialization, manufacturing expansion, and energy-intensive development, forces that operate on horizons far longer than a discrete tariff shock — and their trajectories did not measurably bend with it. What our single-episode, national-aggregate, 2015–2019 design **cannot** say is whether trade interventions in general, at other scales or in other settings, move emissions; we make no such claim. The narrow, defensible reading is that this particular tariff shock left no detectable imprint on exposed countries' national emission paths within a one-to-three-year window.

This conclusion should not be over-read. Our design identifies the effect of tariff exposure on countries' own national CO_2 trajectories; it is silent on the mechanism most

likely to carry the trade war's true environmental footprint — the cross-border reallocation of production. If tariffs shifted output from Chinese facilities to plants elsewhere with different emissions intensities, global emissions could have risen or fallen even where no single country's national trajectory shows a discontinuity. Measuring that channel requires plant- or sector-level intensity data beyond the national aggregates used here, and it remains the most policy-relevant open question the trade war raises. A null on national trajectories is therefore not evidence that the supply-chain reorganization was environmentally neutral.

The stronger and more transferable implication is methodological. The apparent 4.5-percentage-point effect that a conventional matching analysis produced — stable across algorithm labels and sample exclusions, and echoed at smaller, non-significant magnitudes by doubly robust estimators — dissolved once we replaced the unweighted comparison with the properly weighted matched estimand, which is never statistically distinguishable from zero; placebo and pre-trend diagnostics then confirmed that the apparent effect tracked pre-existing industrialization rather than the tariffs. Evaluations of environmental and trade policy that rely on matching or difference-in-differences should treat placebo and pre-trend diagnostics not as optional robustness but as part of the identification argument, especially when treatment assignment correlates with the structural transformation that drives the outcome. Absent such checks, evaluations risk attributing to policy what is merely the continuation of pre-existing development dynamics.

For environmental trade policy, our results give no reason to expect tariff exposure of the kind studied here to advance emissions goals; where such goals are paramount, measures that price or regulate emissions intensity, support cleaner production, and account for cross-border reallocation act on emissions more directly than tariffs imposed for other reasons. We do not extend our results into prescriptions about climate finance, differentiated responsibilities, or border-carbon adjustments — our design does not identify the effects of those policies, and a null result on this tariff episode provides no basis for them. The appropriate lesson is narrow: for the 2018 US–China trade war we find no environmental effect on national emission trajectories to speak of, while the transferable methodological point — that credible environmental policy evaluation must take pre-existing trends seriously — holds well beyond this case.

6 Conclusion

We set out to test whether the 2018 US–China trade war altered carbon dioxide emission trajectories across 171 countries differentiated by tariff exposure. A first-pass propensity score matching comparison, using 2016 baseline emissions, GDP, GDP per capita, and population, appears to show high-exposure countries — predominantly rapidly developing economies in Asia, Africa, and Latin America — growing emissions 4.5 percentage points faster than matched controls over 2016–2019 ($p = 0.007$), a differential that appears stable across alternative algorithm labels and the exclusion of China or India, and that doubly robust estimators echo at smaller, non-significant magnitudes.

Two facts unravel that picture. Under the properly weighted matched estimator — the quantity matching is designed to deliver — the ATT is never statistically distinguishable from zero, 0.28 pp before a pre-trend control and 0.32 pp after, and no larger than about 3 pp on any subsample: there was no significant matched effect to collapse. The unweighted comparison that produced the 4.5 pp figure, by contrast, is simply the wrong estimator,

and it does not repair itself — it remains 4.4 pp ($p = 0.009$) even after the pre-trend enters the match. What exposes the artifact is the falsification evidence: a placebo test finds high-exposure countries already growing 2.4 percentage points faster during 2015–2016 ($p = 0.009$), before any tariff. The null is delivered by the correct estimand and the placebo, not by the naive comparison collapsing.

Three conclusions follow. First, once pre-existing emission dynamics are accounted for, we find no detectable acceleration in exposed countries' national trajectories beyond what industrialization predicts; the confidence interval is wide, and the appropriate inference is not that the environmental consequences were precisely zero but that this episode left no measurable imprint on national trajectories. Second, the analysis is a case study in the limits of baseline covariate balance: matching achieved excellent balance on 2016 levels (mean bias 10.8 percent, joint test $p = 0.948$), yet that balance did not carry over to growth dynamics, which represent selection on unobservables that cross-sectional matching cannot detect. Third, the association between tariff exposure and emission growth reflects the architecture of global manufacturing and US trade policy — shared underlying causes — not a causal mechanism from tariffs to emissions.

For policymakers, the implication is that trade barriers, even substantial ones, are weak instruments for shifting near-term national emission trajectories when growth is driven by development processes operating over decades. For empirical researchers, the challenge is structural: when treatment assignment correlates with the economic transformations that independently drive outcome dynamics, standard matching and DiD estimators are likely to mislead unless placebo and pre-trend diagnostics are treated as part of the identification argument. The takeaway is not that trade wars carry no environmental consequences; it is that isolating those consequences demands identification strategies robust to selection on both observables and the structural dynamics that observables alone cannot capture. Future work should examine longer-term environmental impacts as supply chains stabilize, and — most importantly — measure the cross-border reallocation channel that national aggregates cannot see.

Data Availability Statement

The data and Stata code that support the findings of this study, including a runnable master script that reproduces every table and figure from the raw sources (EDGAR, World Bank WDI, and WITS), are available in an anonymized replication repository: [anonymized repository — link provided for review]. The repository will be deposited in a public, citable archive prior to peer review in accordance with the journal's data-sharing policy; the de-anonymized link will be supplied at final submission.

Appendices

A Equations and Definitions

This appendix collects, in one place, the full set of equations and definitions underlying the analysis, including the auxiliary estimators referenced in the main text and in the Online Resource. Equations reproduced here that also appear in the main text are cross-referenced to their main-text numbers.

A.1 Theoretical framework

Following the scale–composition–technique decomposition (Grossman and Krueger, 1991), total emissions can be written as

$$E = Y \sum_{s} \frac{Y_s}{Y} \frac{E_s}{Y_s}, \tag{A.1}$$

where E is total emissions, Y is total output (the scale term), Y_s/Y is the sectoral composition, and E_s/Y_s is sectoral emission intensity (the technique term). A trade war operates through all three margins: scale (reduced trade lowers Y), composition (production shifts across sectors of differing intensity), and technique (disrupted technology transfer slows efficiency gains). The environmental Kuznets curve (EKC) specification is

$$\text{Emissions}_{it} = \alpha_i + \beta_1 \ln(\text{Income})_{it} + \beta_2 \Big[\ln(\text{Income})_{it}\Big]^2 + Z'_{it}\gamma + \epsilon_{it}, \tag{A.2}$$

with the EKC hypothesis $\beta_2 < 0$. High-exposure countries in our sample sit predominantly on the rising portion of the EKC (the rapid-industrialization phase), which drives faster emission growth independently of trade policy.

A.2 Identification: event study and parallel-trends test

The event-study difference-in-differences specification (main-text Eq. (3)) is

$$\ln(\text{CO}_2)_{it} = \alpha + \sum_{q \in \{2,3,4\}} \sum_{\tau \neq 0} \beta_{q,\tau} \left(\text{Quartile}_{qi} \times \text{Year}_{\tau t}\right) + \delta_i + \lambda_t + \varepsilon_{it}, \tag{A.3}$$

with country and year fixed effects δ_i, λ_t and errors clustered by country. The parallel-trends test is the joint hypothesis on the pre-treatment coefficients,

$$H_0 : \ \beta_{2,-2} = \beta_{2,-1} = \beta_{3,-2} = \beta_{3,-1} = \beta_{4,-2} = \beta_{4,-1} = 0, \tag{A.4}$$

which is decisively rejected, $F(6, 170) = 3.87$, $p = 0.0012$, with pre-treatment coefficients ranging from -2.4 to -6.3 percentage points relative to the low-exposure reference.

A.3 Propensity score matching

The propensity score (main-text Eq. (4)) is

$$\Pr(\text{Treatment}_i = 1 \mid X_i) = \Lambda(X'_i\beta) = \frac{1}{1 + \exp(-X'_i\beta)}, \tag{A.5}$$

with $X_i = [\,1,\ \ln(\text{CO}_2)_{i,2016},\ \ln(\text{GDP})_{i,2016},\ \ln(\text{GDPpc})_{i,2016},\ \ln(\text{Pop})_{i,2016}\,]$ and $\Lambda(\cdot)$ the logistic CDF; $\text{Treatment}_i = 1$ for Q4 (highest exposure). The outcome (main-text Eq. (5)) is

$$\Delta \ln(\text{CO}_2)_i = \ln(\text{CO}_2)_{i,2019} - \ln(\text{CO}_2)_{i,2016}. \tag{A.6}$$

The common-support set retains treated units whose estimated score lies within the control score range,

$$S = \left\{ i : \ \min_{j:T_j=0} \hat{p}_j < \hat{p}_i < \max_{j:T_j=0} \hat{p}_j \right\}, \tag{A.7}$$

and caliper matching selects, for each treated unit i, the $K = 3$ nearest control neighbours j with

$$\left|\hat{p}_i - \hat{p}_j\right| < 0.05 \times \mathrm{SD}(\hat{p}). \tag{A.8}$$

The average treatment effect on the treated (main-text Eq. (6)) is

$$\mathrm{ATT} = E\Big[\Delta \ln(\mathrm{CO_2})_i \mid T_i = 1,\, i \in S\Big] - E\Big[\Delta \ln(\mathrm{CO_2})_i \mid T_i = 0,\, i \in S\Big], \tag{A.9}$$

estimated as

$$\widehat{\mathrm{ATT}} = \frac{1}{N_T} \sum_{i \in T} \Big(\Delta \ln(\mathrm{CO_2})_i - \sum_{j \in C(i)} \omega_{ij}\, \Delta \ln(\mathrm{CO_2})_j \Big), \tag{A.10}$$

where N_T is the number of treated units on common support, $C(i)$ the matched controls, and ω_{ij} the matching weights ($\omega_{ij} = 1/K$ for $j \in C(i)$, 0 otherwise). Discarding these weights and comparing the two group means directly gives the *unweighted* two-sample comparison, distinguished throughout the paper from the weighted ATT above.

A.4 Balance diagnostics

For each covariate X the standardized bias is

$$\mathrm{Bias} = 100 \times \frac{\bar{X}_T - \bar{X}_C}{\sqrt{(S^2_{X,T} + S^2_{X,C})/2}}, \tag{A.11}$$

with $\bar{X}_T, \bar{X}_C$ the treated/control means and $S^2_{X,T}, S^2_{X,C}$ their variances; bias reduction is

$$\text{Bias reduction} = 100 \times \Big(1 - \frac{|\mathrm{Bias}_{\mathrm{after}}|}{|\mathrm{Bias}_{\mathrm{before}}|}\Big). \tag{A.12}$$

Joint balance is assessed by a likelihood-ratio test of $H_0 : E[X_i \mid T_i = 1] = E[X_i \mid T_i = 0]$; the matched sample gives $\chi^2(4) = 0.73$, $p = 0.948$, so joint balance cannot be rejected.

A.5 Auxiliary estimators

The regression-adjusted ATT is $\hat{\gamma}_1$ from

$$\Delta \ln(\mathrm{CO_2})_i = \gamma_0 + \gamma_1 T_i + \gamma_2 \ln(\mathrm{CO_2})_{i,2016} + \gamma_3 \ln(\mathrm{GDP})_{i,2016} + \gamma_4 \ln(\mathrm{GDPpc})_{i,2016} + \gamma_5 \ln(\mathrm{Pop})_{i,2016} + u_i, \tag{A.13}$$

estimated on the matched sample $i \in S$ with heteroskedasticity-robust standard errors ($\hat{\gamma}_1 = 0.0208$, SE $= 0.0189$, $t = 1.10$, $p = 0.275$ in the baseline; 0.0115, $p = 0.515$ with the pre-trend covariate). The doubly-robust (AIPW) estimator is

$$\widehat{\mathrm{ATT}}_{\mathrm{DR}} = \frac{1}{N_T} \sum_{i=1}^{N} \Big[\frac{T_i \,(Y_i - \hat{\mu}_0(X_i))}{\hat{p}_i} - \frac{(1 - T_i)\,\hat{p}_i \,(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{p}_i} \Big], \tag{A.14}$$

with $Y_i = \Delta \ln(\mathrm{CO_2})_i$ and $\hat{\mu}_0(X_i)$ the control-group prediction from the regression above; it is consistent if either $\hat{p}_i$ or $\hat{\mu}_0$ is correctly specified.

A.6 Falsification and pre-trend control

The placebo re-runs the design on the pre-treatment window, holding treatment assignment fixed and defining the placebo outcome

$$\Delta \ln(\mathrm{CO}_2)_{i,\text{placebo}} = \ln(\mathrm{CO}_2)_{i,2016} - \ln(\mathrm{CO}_2)_{i,2015}. \tag{A.15}$$

The pre-trend control augments the score with the 2015–2016 growth rate, $\text{PreTrend}_i = \ln(\mathrm{CO}_2)_{i,2016} - \ln(\mathrm{CO}_2)_{i,2015}$, giving

$$\Pr(\text{Treatment}_i = 1 \mid X_i^*) = \Lambda(X_i^{*\prime}\beta^*), \qquad X_i^* = [\, X_i,\ \text{PreTrend}_i \,]. \tag{A.16}$$

A.7 Set-aside DiD: partial-identification notation

The sensitivity analysis in Appendix B uses the Rambachan and Roth (2023) relative-magnitudes restriction $\Delta^{RM}(\bar{M})$, which bounds the post-treatment violation of parallel trends by $\bar{M}$ times the largest pre-treatment violation; $\bar{M} = 1$ allows a post-period violation no larger than the biggest pre-period one. The randomization-inference exhibit reports a permutation p-value, the share of placebo estimates (under random reassignment of treatment) at least as extreme in absolute value as the observed estimate.

A.8 Alternative identification and the collinearity obstacle

We attempted to accommodate the parallel-trends violation directly by adding group-specific linear trends, $\ln(\mathrm{CO}_2)_{it} = \delta_i + \lambda_t + \sum_q \theta_q \,(\text{Quartile}_{qi} \times t) + \varepsilon_{it}$. With only five years of data and two-way fixed effects, the θ_q trends are collinear with the country and year effects, and Stata omits them for collinearity. This obstacle, not a modelling preference, is what motivated the pivot to propensity score matching with an explicit pre-trend control.

B Robustness of the Null

This appendix reports three additional checks on the central null. Each is subordinate to the main propensity-score-matching result: none is offered as a substitute for it, and none is used to claim more than the weighted matched ATT already supports. Together they show that the null is not an artifact of the event-study design that we set aside (§B.1), not an artifact of the particular treatment assignment (§B.2), and not driven by any single treated country (§B.3).

B.1 HonestDiD sensitivity for the set-aside difference-in-differences

We abandoned the event-study DiD because its pre-treatment coefficients reject parallel trends (Section 4.3.1). For completeness we ask what that set-aside design would imply about the post-treatment effect once the pre-trend violation is acknowledged rather than assumed away, using the Rambachan and Roth (2023) relative-magnitudes restriction $\Delta^{RM}(\bar{M})$ applied to the highest-exposure (Q4) arm. The restriction bounds the post-period violation of parallel trends by $\bar{M}$ times the largest pre-period violation, and returns

a robust confidence set for the average post-period effect ($l = (0.5, 0.5)$ over 2018 and 2019).

Table B.1 and Fig. B.1 report the result. The Q4 pre-period coefficients are large (-6.3 and -4.3 percentage points), but the post-period coefficients are already indistinguishable from zero, so the average post-period effect is -0.3 pp with an original (parallel-trends) confidence set of $[-3.4, +2.8]$ pp. The robust confidence set includes zero at *every* bound tested, up to $\bar{M} = 2$, i.e. even when the post-period violation is allowed to be twice the largest pre-period violation. There is no breakdown value: because the observed post-period effect is itself ≈ 0, relaxing the bound only widens an already zero-centred interval and can never manufacture a non-zero effect. This is a statement about the set-aside DiD design; it is not, and is not offered as, a validation of the matching null, which rests on the weighted ATT and the placebo. Because the post-period coefficients are already ≈ 0, a robust set that covers zero is close to tautological; the exhibit is best read as a completeness check that the set-aside design harbours no hidden post-treatment effect, not as independent corroboration of the null.

Table B.1. HonestDiD relative-magnitudes sensitivity, Q4 (high-exposure) arm

Restriction	$\bar{M}$	Robust 95% CI (pp)	Includes 0
Original (parallel trends)	—	$[-3.37, +2.78]$	Yes
$\Delta^{RM}(\bar{M})$	0.0	$[-3.30, +2.73]$	Yes
$\Delta^{RM}(\bar{M})$	0.5	$[-6.37, +5.93]$	Yes
$\Delta^{RM}(\bar{M})$	1.0	$[-10.33, +10.08]$	Yes
$\Delta^{RM}(\bar{M})$	1.5	$[-14.60, +14.34]$	Yes
$\Delta^{RM}(\bar{M})$	2.0	$[-18.99, +18.74]$	Yes

Notes: Target parameter: average post-period effect over 2018–2019 ($l = (0.5, 0.5)$), point estimate ≈ -0.3 pp. Q4 pre-period event-study coefficients: -6.3 pp (2015) and -4.3 pp (2016); post-period: ≈ 0. Robust confidence sets computed by the conditional–least-favourable method (Rambachan and Roth 2023; `HonestDiD` 0.2.8). The complementary smoothness restriction $\Delta^{SD}(M)$ was not computed (an archived-CVXR version detail in the solver, unrelated to the relative-magnitudes results); see Appendix C.

B.2 Permutation (randomization) inference

We reassign the high-exposure label at random across the 164 covariate-complete countries—holding the number of treated units fixed at 44—and re-estimate the full matching design on each of 2,000 draws (seed fixed; every draw converged). This builds the distribution of each estimate under the sharp null of random assignment. The two estimators must be read *together*, as a contrast; that contrast is the point.

Table B.2 reports it. The weighted matched ATT is completely unremarkable under randomization: the observed $+0.28$ pp sits at the 57th percentile of its placebo distribution, dead centre, with a two-sided permutation $p = 0.88$. The null is therefore not an artifact of this particular treatment assignment. The unweighted mean comparison behaves oppositely: the observed $+4.5$ pp sits in the extreme upper tail (99.9th percentile, $p < 0.01$). This is not evidence of a treatment effect—it is the estimator-artifact thesis shown a third, independent way. The unweighted comparison is significant under randomization inference precisely *because* it captures the real, systematic selection that defines the treated group (high-exposure countries are high-growth countries on a

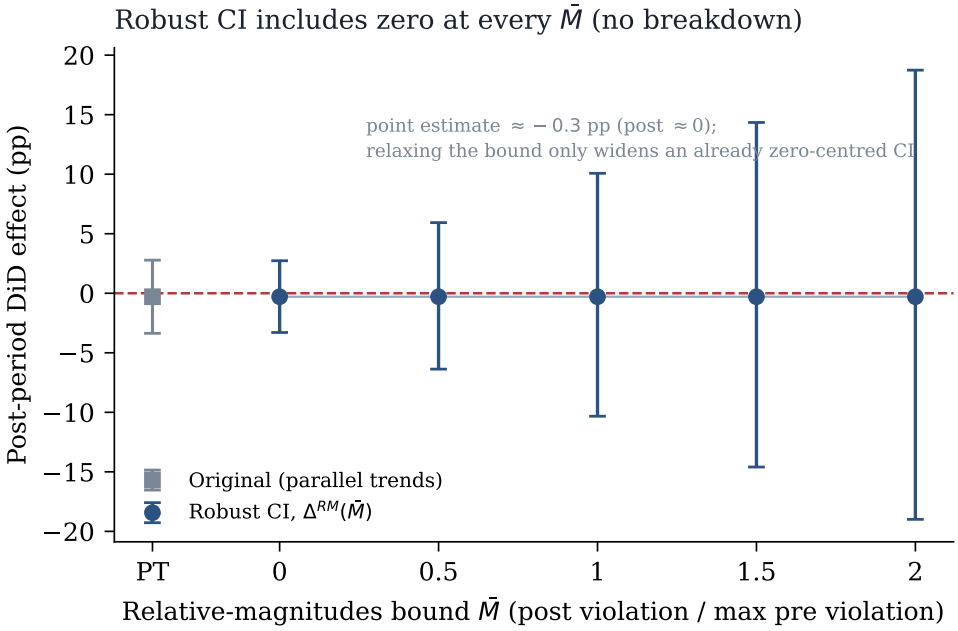

Figure B.1. HonestDiD robust confidence set for the post-period DiD effect (Q4 arm) across relative-magnitudes bounds $\bar{M}$. The set includes zero at every bound; because the post-period effect is itself ≈ 0 (pre-period coefficients $-6.3/-4.3$ pp; post ≈ 0), relaxing the bound only widens an already zero-centred interval.

pre-existing emission trajectory), which random reassignment destroys; the weighted estimand, which is designed to remove that selection, is null. Same data, same draws, opposite verdicts—delivered by the choice of estimator, exactly as in the main text and in Fig. 3.

Table B.2. Permutation inference: observed estimate vs. its placebo distribution

Estimator	Observed (pp)	Placebo mean (SD)	Percentile	Perm. p (2-sided)
Weighted matched ATT	+0.28	−0.04 (1.77)	57th	0.876
Unweighted mean comparison	+4.54	−0.05 (1.62)	99.9th	< 0.01

Notes: 2,000 random reassignments of the high-exposure label among the 164 covariate-complete countries, treated count fixed at 44; the full logit–match–estimate pipeline is re-run each draw (0 failures). Two-sided permutation p-value: the share of placebo estimates at least as extreme in absolute value as the observed estimate. The contrast—weighted null unremarkable ($p = 0.88$), unweighted gap in the tail ($p < 0.01$)—reflects estimator choice, not a treatment effect.

B.3 Leave-one-out over treated countries

We drop each of the 44 treated countries in turn and re-estimate both quantities, beyond the China and India exclusions already reported in Section 4.2. Table B.3 gives the ranges. The unweighted comparison is rock-stable (+3.9 to +4.9 pp; no single drop moves it by more than 1 pp). The weighted matched ATT is mildly more drop-sensitive: six small economies (Greenland, Micronesia, Nauru, Eswatini, Lesotho, Mongolia) each move it by more than 1 pp, with the largest single excursion to +2.27 pp when Greenland

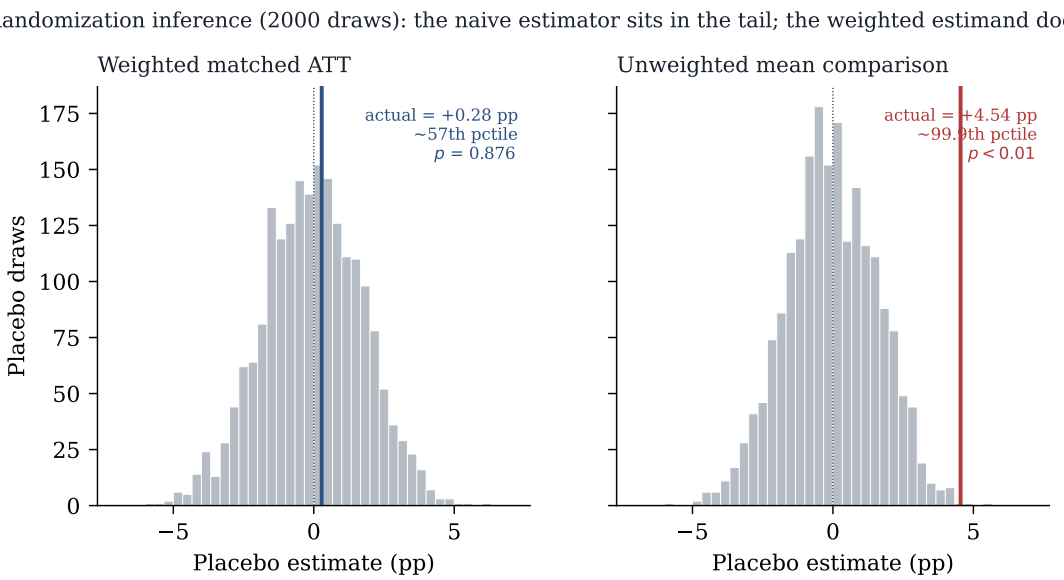

Figure B.2. Randomization-inference distributions (2,000 draws). Left: the weighted matched ATT (+0.28 pp) sits at the centre of its placebo distribution ($p = 0.88$). Right: the unweighted mean comparison (+4.54 pp) sits in the extreme tail ($p < 0.01$). The naive estimator is significant under randomization because it captures the real selection into treatment; the weighted estimand, which removes it, is null.

is dropped. We disclose this rather than smooth it over—but every excursion remains firmly null, far below the ≈ 6 pp minimum detectable effect, so no single country drives either the artifact or the null.

Table B.3. Leave-one-out over the 44 treated countries

Estimate	Full sample	LOO range	Most influential drop
Weighted matched ATT	+0.28 pp	[−0.54, +2.27] pp	Greenland (+2.27)
Unweighted mean comparison	+4.54 pp	[+3.91, +4.88] pp	Maldives (+3.91)

Notes: Each treated country dropped once; the full logit–match–estimate pipeline is re-run. Six small economies (Greenland, Micronesia, Nauru, Eswatini, Lesotho, Mongolia) move the weighted ATT by more than 1 pp; all excursions stay well below the ≈ 6 pp minimum detectable effect, so the null holds under every drop. No single drop moves the unweighted comparison by more than 1 pp.

C Replication and Computational Environment

C.1 Data and code availability

All data and code required to reproduce every number, table, and figure in this paper—from the raw sources (EDGAR emissions, World Bank WDI, and WITS tariff data) through the final outputs—are provided in an anonymized replication repository (DRYAD). A single master script runs the full pipeline; the repository will be deposited in a public, citable archive prior to peer review in accordance with the journal's data-sharing policy, and the de-anonymized link will be supplied at final submission.

C.2 Software environment

The core estimation runs in **Stata 15.1** (user commands `psmatch2`, `reghdfe`, and built-in `teffects psmatch` for Abadie–Imbens matching standard errors). Figures and the permutation summary use **Python 3** (`pandas`, `numpy`, `statsmodels`, `matplotlib`). The partial-identification sensitivity analysis (Appendix B.1) uses **R 4.3.3** with the `HonestDiD` package.

C.3 HonestDiD installation recipe

Because `HonestDiD` and its convex-optimization dependencies are version sensitive, we record the exact environment used, so a replicator can reproduce the relative-magnitudes results without encountering the dependency conflicts we did. System libraries (Debian/Ubuntu): `libgmp-dev`, `libmpfr-dev`, `libglpk-dev`. R packages: `CVXR` 1.0-11 installed from the archive via `remotes::install_version(..., upgrade="never")`, together with `TruncatedNormal` and `lpSolveAPI`, then `HonestDiD` 0.2.8. With this configuration the relative-magnitudes restriction $\Delta^{RM}(\bar{M})$ (conditional–least-favourable method) runs as reported. The complementary smoothness restriction $\Delta^{SD}(M)$, which routes through a fixed-length optimal confidence interval, failed on a namespace detail in the archived `CVXR` build and was not computed; the relative-magnitudes restriction is the standard object for this exhibit and is sufficient for it.

C.4 Shipped scripts

The replication package includes the estimation do-files (the master do-file `00_master` and the propensity-score-matching and robustness do-files), the figure builders (`make_figures.py` for Figs 1–3, `make_appendix_figures.py` for Figs B.1–B.2), and the three scripts underlying this appendix: `honestdid_Q4_sensitivity.R` (Appendix B.1), the do-file `perm_loo` (the permutation and leave-one-out estimation of Appendices B.2–B.3), and `summarize_perm_loo.py` (the permutation p-values and leave-one-out ranges). Each reads the estimation data and reproduces the corresponding table and figure.

References

Laura Alfaro and Davin Chor. Global supply chains: The looming "great reallocation". Social Science Research Network. https://ssrn.com/abstract=4567667, 2023.

Larissa Basso and Eduardo Viola. Are the BRICS engaged in the low carbon transition? *CEBRI Journal*, 1(4), 2022.

Carolina Caetano and Brantly Callaway. Difference-in-differences when parallel trends holds conditional on covariates. Working paper, 2024.

Brian R. Copeland and M. Scott Taylor. Trade, growth, and the environment. *Journal of Economic Literature*, 42(1):7–71, 2004. doi: 10.1257/002205104773558047.

Jamie R. Daw and Laura A. Hatfield. Matching and regression to the mean in difference-in-differences analysis. *Health Services Research*, 53(6):4138–4156, 2018. doi: 10.1111/1475-6773.12993.

Clément de Chaisemartin and Xavier D'Haultfœuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996, 2020. doi: 10.1257/aer.20181169.

Clément de Chaisemartin and Xavier D'Haultfœuille. Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal*, 26(3):C1–C30, 2023. doi: 10.1093/ectj/utac017.

Soumyananda Dinda. Environmental Kuznets curve hypothesis: A survey. *Ecological Economics*, 49(4):431–455, 2004. doi: 10.1016/j.ecolecon.2004.02.011.

Pablo Fajgelbaum, Pinelopi Goldberg, Patrick Kennedy, Amit Khandelwal, and Daria Taglioni. The US–China trade war and global reallocations. *American Economic Review: Insights*, 6(2):295–312, 2024. doi: 10.1257/aeri.20230094.

Caroline Freund, Aaditya Mattoo, Alen Mulabdic, and Michele Ruta. Is US trade policy reshaping global supply chains? *Journal of International Economics*, 152:104011, 2024. doi: 10.1016/j.jinteco.2024.104011.

Gene M. Grossman and Alan B. Krueger. Environmental impacts of a north american free trade agreement. NBER Working Paper 3914, National Bureau of Economic Research, 1991.

Gene M. Grossman and Alan B. Krueger. Economic growth and the environment. *The Quarterly Journal of Economics*, 110(2):353–377, 1995. doi: 10.2307/2118443.

James J. Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *Review of Economic Studies*, 65(2):261–294, 1998. doi: 10.1111/1467-937X.00044.

Ashesh Rambachan and Jonathan Roth. A more credible approach to parallel trends. *Review of Economic Studies*, 90(5):2555–2591, 2023. doi: 10.1093/restud/rdad018.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

Jonathan Roth. Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3):305–322, 2022. doi: 10.1257/aeri.20210236.

Jonathan Roth and Pedro H. C. Sant'Anna. When is parallel trends sensitive to functional form? *Econometrica*, 91(2):737–747, 2023. doi: 10.3982/ECTA19402.

Jonathan Roth, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244, 2023. doi: 10.1016/j.jeconom.2023.03.008.

Andrew M. Ryan, James F. Burgess, and Justin B. Dimick. Why we should not be indifferent to specification choices for difference-in-differences. *Health Services Research*, 50(4):1211–1235, 2015. doi: 10.1111/1475-6773.12270.

Paula Sandrin and Camila Amigo. Climate ambition of the BRICS countries. BRICS Policy Center, 2024.

David I. Stern. The rise and fall of the environmental Kuznets curve. *World Development*, 32(8):1419–1439, 2004. doi: 10.1016/j.worlddev.2004.03.004.

Kailan Tian, Yuze Zhang, Yuhan Li, Xi Ming, Shangrong Jiang, Hongbo Duan, Cuihong Yang, and Shouyang Wang. Regional trade agreement burdens global carbon emissions mitigation. *Nature Communications*, 13(1):408, 2022. doi: 10.1038/s41467-022-28004-5.

Qi-Yang Zhao, Jing-Chao Luo, Ying Su, Yi-Jun Zhang, Guo-Wei Tu, and Zhe Luo. Propensity score matching with R: Conventional methods and new features. *Annals of Translational Medicine*, 9(9):812, 2021. doi: 10.21037/atm-20-3998.