PLOS Climate

RESEARCH ARTICLE

Statistical learning for climate-GDP panels: Data cleaning, flexible trend controls, and predictive validation

Christof Schötz[1,2*], **Jan Hassel**[2], **Christian Otto**[2]

1 Department of Aerospace and Geodesy, Munich Climate Center and Earth System Modelling Group, TUM School of Engineering and Design, Technical University of Munich, Munich, Germany, **2** Research Department 3: Transformation Pathways, Potsdam Institute for Climate Impact Research, Potsdam, Germany

* christof.schoetz@tum.de

OPEN ACCESS

Citation: Schötz C, Hassel J, Otto C (2026) Statistical learning for climate-GDP panels: Data cleaning, flexible trend controls, and predictive validation. PLOS Clim 5(7): e0000962. https://doi.org/10.1371/journal.pclm.0000962

Editor: María Dolores Gadea Rivas, University of Zaragoza, SPAIN

Received: January 29, 2026

Accepted: June 20, 2026

Published: July 20, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pclm.0000962

Abstract

We assess the panel-regression approach to climate econometrics—the dominant framework for estimating the effect of climate on GDP in global country–year data—using modern statistical learning techniques. Common implementations are sensitive to outliers, do not fully account for the dependence structure across countries and years, and rarely combine formal model selection with genuine out-of-sample evaluation. To address these issues, we implement knowledge-based data cleaning, nonparametric time-trend controls, and out-of-sample validation across 700 + climate variables. Our analysis reveals that widely used models and predictors—such as mean temperature—have little out-of-sample predictive power. A previously overlooked humidity-related variable emerges as the most consistent predictor, though even its performance remains limited. These findings question the robustness of common empirical practices in this literature and point toward a more data-driven approach built on data cleaning, flexible trend controls, and predictive validation.

Introduction

Climate econometrics is the study of how weather, climate, and climate change affect economic outcomes using statistical methods. Understanding these relationships is essential for informing policy, managing climate-related risks, and guiding adaptation strategies. As interest in the economic implications of climate variability and change has grown, the field has developed at the intersection of economics, climate science, and statistics [1,2].

A widely used empirical approach in climate econometrics is fixed effects panel data regression, where macroeconomic indicators—such as a country's GDP—are regressed on climate variables like average annual temperature. This method aims to quantify the extent to which economic variation can be statistically attributed to fluctuations in climatic conditions. Fixed effects account for time-invariant country

Data availability statement: All data and code necessary to reproduce the results are publicly available at https://doi.org/10.5281/zenodo.20415574.

Funding: J.H. and C.O. acknowledge support from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101081358 (ACCREU) and No. 101135481 (COMPASS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

characteristics and global shocks shared across countries, while additional controls, such as country-specific time trends, are often included to capture heterogeneous growth trajectories.

Numerous influential studies have adopted this methodology using global country-year panel datasets. For example, Dell et al. [3] estimate linear effects of temperature and precipitation on economic output, finding that higher temperatures negatively affect economic growth in poorer countries. Burke et al. [4] extend this framework by including quadratic terms for temperature and precipitation as well as quadratic time trends, uncovering a significant nonlinear (inverted U-shaped) relationship between temperature and GDP growth. Pretis et al. [5] incorporate measures of monthly variance and extreme temperatures, along with dummy variables for outlier events, to assess differences in projected economic damages under 1.5° C and 2° C warming scenarios. Other contributions have expanded the analysis by increasing the spatial resolution of the panel. Kalkuhl and Wenz [6] use economic data of subnational regions to explore temperature effects. Continuing the analysis of subnational data, Kotz et al. [7] investigate the role of intra-annual temperature variability, while Kotz et al. [8] examine the effects of diverse precipitation-related indicators. Newell et al. [9] assess model performance through out-of-sample testing, finding no significant predictive power of temperature on GDP growth, though more robust effects are observed for GDP levels. Kahn et al. [10] study deviations from long-term temperature norms, and Krichene et al. [11] incorporate extreme weather events to examine longer-term macroeconomic consequences.

Several meta-analyses and review articles provide a broader synthesis of this literature and the methodologies it employs [1,2,12–17]. A related strain of literature emphasizes spatial heterogeneity in climate impacts and the role of adaptation through migration, trade, and innovation, as reviewed by Desmet and Rossi-Hansberg [18]. A separate tradition within climate econometrics, surveyed by Castle and Hendry [19], works in a time-series rather than panel-regression framework and emphasizes explicit modeling of non-stationarity, structural breaks, and the underlying data generating process. The present paper contributes to the panel-regression strand with control variables, while engaging formally with time series stationarity in Section A.2 in S1 Appendix and with the temporal stability of estimated relationships in Stationarity.

While fixed effects panel regression is a standard tool in applied econometrics [20,21], its limitations have received increasing attention. Hill et al. [22] and Imai and Kim [23] emphasize that the validity of such models hinges on strong assumptions about functional form, treatment homogeneity, and independence, which are often untested in practice. Violations of these assumptions can create unreliable and misleading findings. Although techniques such as out-of-sample testing and formal variable selection are standard in other empirical disciplines, they are rarely applied in the current climate econometrics literature. This limits the robustness and generalizability of existing findings. In this article, we address several of these methodological gaps and present results based on more rigorous, data-driven approaches.

While outliers and structural breaks have been treated within climate econometrics using model-based detection methods such as indicator saturation [5,19,24], we take a complementary and more granular approach. We distinguish between data errors, genuine non-climate shocks, and imputation artifacts, screening each using external, source-based knowledge together with correlation- and segment-based diagnostics. This lets us remove spurious observations while retaining the extreme-climate events we wish to study, and we provide a correspondingly cleaned dataset that enhances the robustness of our results. To address temporal confounding, we introduce nonparametric time trend controls that flexibly account for slowly evolving, country-specific characteristics. These controls avoid the arbitrary specification of polynomial time trends and their induced long-range serial correlation biases. In addition to examining temporal dependencies, we assess spatial correlation and its implications for standard error estimation.

In our main analysis, we apply various statistical learning methods for variable selection and evaluate models in an out-of-sample framework to identify the most predictive climate indicators of GDP growth and levels. Our approach considers orders of magnitude more variables and models than any prior study in climate econometrics: we select from a pool of over 700 variables, resulting in roughly 10^{213} possibilities of predictor combinations. Among the methods, penalized regression techniques—LASSO, Ridge regression, and Elastic Net [25]—consistently yield the best predictive accuracy. The most robust predictor we identify is the squared maximum specific humidity of the previous year, which may influence economic activity through channels such as wet-bulb temperatures or extreme precipitation events. In contrast, commonly used variables such as mean annual temperature and its square show no predictive power, calling into question some earlier findings in the literature. Overall, even the best-performing models exhibit limited predictive power. We identify a lack of temporal stationarity as a possible reason for the limited out-of-sample predictive performance.

Setup

We first provide a description of the basic statistical framework, the panel data regression, and the data and variables used in this framework.

Panel data regression with fixed effects

Our goal is to obtain empirical evidence of the influence of climate variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ on a target economic variable $\boldsymbol{Y}$. All variables are indexed by two dimensions: One is spatial (here: countries), denoted by $s \in \mathcal{S}$. The other one is temporal (here: years), denoted by $t \in \mathcal{T}$, where $\mathcal{T} \subseteq \mathbb{Z}$ is a set of consecutive time points. Thus, the data forms a panel and panel data analysis methods can be applied. We do not assume that the panel is full, i.e., we may only observe certain region–time combinations $(s, t) \in \mathcal{I}$ for some index set $\mathcal{I} \subseteq \mathcal{S} \times \mathcal{T}$. Denote the total number of observations as $n = \#\mathcal{I}$.

To achieve our goal, we apply a linear regression model of the form

$$Y_{s,t} = Z_{1,s,t}\beta_1 + \cdots + Z_{p,s,t}\beta_p + w_{1,s,t}\alpha_1 + \cdots + w_{q,s,t}\alpha_q + \varepsilon_{s,t}, \quad (s,t) \in \mathcal{I}. \tag{1}$$

Here $Z_{k,s,t}$ are features derived from the climate variables. These can, e.g., be transformations like $\log(X_{j,s,t})$ or $X^2_{j,s,t}$; or time lags like $X_{j,s,t-1}$, $X_{j,s,t-2}$, ...; or interaction terms like $X_{j,s,t}X_{j',s,t}$. The terms $w_{k,s,t}$ are fixed effect variables, e.g., dummy variables $w_{k,s,t} = 1_{t=t_0}$ for a given time point $t_0 \in \mathcal{T}$ or time trends $w_{k,s,t} = t1_{s=s_0}$ for a given region $s_0 \in \mathcal{S}$. The parameters α_j, β_j are unknown. They differ in that we are interested in statistical inference on β_j, whereas α_j are nuisance parameters that model unobserved effects that we want to control for but have no further interest in. The errors $\varepsilon_{s,t}$ are modeled as centered random variables, which are normally distributed or fulfill moment conditions that allow to use asymptotic normality in later analysis. In matrix notation, we have

$$\boldsymbol{Y} = \boldsymbol{Z}\beta + \boldsymbol{w}\alpha + \varepsilon \tag{2}$$

where

$$\boldsymbol{Y} = (Y_{s,t})_{(s,t)\in\mathcal{I}} \in \mathbb{R}^n, \qquad \beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p, \tag{3}$$

$$\boldsymbol{Z} = (Z_{j,s,t})_{(s,t)\in\mathcal{I},j=1,\ldots,p} \in \mathbb{R}^{n\times p}, \qquad \alpha = (\alpha_1, \ldots, \alpha_q) \in \mathbb{R}^q, \tag{4}$$

$$\boldsymbol{w} = (w_{k,s,t})_{(s,t)\in\mathcal{I},k=1,\ldots,q} \in \mathbb{R}^{n\times q}, \qquad \varepsilon = (\varepsilon_{s,t})_{(s,t)\in\mathcal{I}} \in \mathbb{R}^n. \tag{5}$$

Aside from statistical inference on β, we are also interested in variable selection to identify models with high predictive power for unseen observations. While our primary metric is predictive accuracy, this framework supports causal interpretation due to the strict exogeneity of weather shocks: short-term economic fluctuations do not influence concurrent climatic conditions. Consequently, a model that robustly predicts out-of-sample is plausibly capturing stable structural relationships rather than spurious correlations, thereby validating the causal links necessary for future projections. In order to find such a model, we need the set of selected features to be as small as possible while explaining as much as possible of the variation of the target $Y_{\bullet,\bullet}$. Model selection is explored in Step 3: Model Selection via Out-of-Sample Test.

Year fixed effects cannot be estimated for unseen years, and time trends extrapolated forward are not guaranteed to remain predictive. With such time controls, the model is not intended to predict the absolute value of $Y_{s,t}$; the goal is to identify the time-persistent influence of the climate features $Z_{j,s,t}$ on $Y_{s,t}$.

Data

Our primary data source is the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP). Specifically, we utilize climate variables from the GSWP3-W5E5 dataset [26], the tropical cyclone data [27], as well as GDP (PPP) [28] and country-level population data [29]. In addition, we incorporate gridded population data from the History Database of the Global Environment (HYDE), version 3.3 [30]. To assess the robustness of our findings with respect to economic data sources, we also employ World Development Indicator GDP data [31]. All datasets are restricted to the period 1960–2018, and our analysis covers a total of 203 countries.

Climate econometric model

Our target variable in the regression is the economic growth rate $Y_{s,t} = \log(\mathrm{GDPpc}_{s,t}/\mathrm{GDPpc}_{s,t-1})$, where $\mathrm{GDPpc}_{s,t}$ is the gross domestic product per capita in country s and year t either using the ISIMIP or the World Bank data.

Most of the predictor variables we use, are based on the 11 climate variables given in Table 1. All variables are first given as daily gridded ($0.5° \times 0.5°$) values. For each grid cell, they are aggregated to yearly values using the mean, standard deviation, minimum, and maximum. This results in 44 variables indexed by grid cell and year. We have additional 12 variables derived from tas and pr (Table 1) by calculating the deviation from a rolling window of 11, 31, and 91 days (prDevi11, prDevi31, prDevi91, tasDevi11, tasDevi31, tasDevi91) and the rate of extreme events larger than $1-2^{-8} \approx 99.6\%$, $1-2^{-10} \approx 99.9\%$, $1-2^{-12} \approx 99.98\%$ of the whole observation time distribution (prExtr08, prExtr10, prExtr12, tasExtr08, tasExtr10, tasExtr12). All yearly gridded values are aggregated to country means using population as grid weights. Furthermore, we add the number of people affected by tropical cyclones of at least 34 knots, 48 knots and 64 knots (tc34kt, tc48kt, tc64kt) to our mix of variables. We arrive at $44+12+3=59$ variables.

For each of the 59 base variables, we create linear and squared terms and up to 5 lags. This leaves us at $59 \cdot 2 \cdot 6 = 708$ predictors to choose from. We do not include lagged growth-rate terms on the right-hand side, following

Table 1. Climate-related base variables from GSWP3-W5E5 and their descriptions.

Variable	Description
hurs	Near-surface relative humidity
huss	Near-surface specific humidity
pr	Total precipitation
prsn	Snowfall
ps	Surface air pressure
rlds	Long wave downwelling radiation
rsds	Short wave downwelling radiation
sfcwind	Near-surface wind speed
tas	Daily mean temperature
tasmax	Daily maximum temperature
tasmin	Daily minimum temperature

https://doi.org/10.1371/journal.pclm.0000962.t001

the standard specification in the climate-econometrics literature [4,3]. Instead, we rely on weather variables and time-dependent controls to capture dynamic effects.

The country and year fixed effects together with the time-trend controls absorb a large part of the systematic economic variation in growth—persistent cross-country differences in development and growth regimes, globally common shocks, and slowly evolving country-specific trajectories—so that climate is evaluated against the residual, higher-frequency variation. Supporting a causal interpretation of the influence of climate on GDP would rest on the assumption that year-to-year climate anomalies are exogenous to contemporaneous economic activity. This assumption is plausible in our setting: the long-run channel through which economic activity affects climate via emissions operates at low frequency and is absorbed by our nonparametric time-trend controls.

We have all variables in a base form ($Z_{k,s,t}$) and a difference form ($\Delta Z_{k,s,t} = Z_{k,s,t} - Z_{k,s,t-1}$), leading to two different kinds of regression models. We follow the terminology of Newell et al. [9] and distinguish a *growth* and a *level* regression, where the labels refer to whether the climate *level* is related to GDP *growth* or to the GDP *level*. In the growth regression, the climate predictors enter in original (level) form $Z_{k,s,t}$ and relate directly to the growth-rate target $Y_{s,t} = \Delta \log(\mathrm{GDPpc}_{s,t})$. In the level regression, the predictors enter in differenced form $\Delta Z_{k,s,t}$; since the target is also a difference, this is the differenced form of a relationship between the climate level $Z_{k,s,t}$ and the GDP level $\log(\mathrm{GDPpc}_{s,t})$, which is the source of the "level" label, even though the climate level itself does not appear explicitly in the estimated equation. We do not mix original and differenced predictors within the same model and do not delve into the interpretation of the two modeling strategies here. Instead, we refer readers to Newell et al. [9] for a detailed discussion and present results for both approaches.

Summary statistics of all variables are given in Section A.1 in S1 Appendix. We verify in Section A.2 in S1 Appendix that the GDP growth-rate target and the differenced climate variables are stationary, and that the original climate variables show no detectable non-stationarity once our sufficient time-trend controls are included. Both modeling strategies therefore yield balanced regressions, so the choice between them is not pinned down by stationarity considerations and is best interpreted in economic terms, as discussed in Newell et al. [9].

Step 1: Cleaning the data

Ensuring the quality of the data set is a critical first step before performing regression analysis. This involves identifying and addressing outliers and out-of-distribution data that could bias results or compromise the validity of statistical inference. By systematically removing problematic data points, we aim to create a robust foundation for accurate modeling and reliable inference.

Outliers

Outliers in the data can substantially distort results, potentially overshadowing true effects or introducing spurious findings that lack causal basis. Broadly speaking, outliers are data points that appear highly atypical or unexpected relative to the statistical model. We categorize outliers into two primary types here: mistakes and nuisance events.

Mistakes refer to incorrect data points, often due to issues such as instrument failure or transcription errors. As an example, version 2 of the DOSE data set [32] (MCC-PIK **D**atabase **O**f **S**ub-national **E**conomic Output) showed a decline in economic production per capita in Madrid, Spain, of almost 90% from 1994 to 1995. In version 4, there is instead an increase by 5%. The steep decline in the old version might be due to a wrongly placed decimal point. See Table K and L in S1 Appendix for more details on this example. Another mistake is the GDPpc in Oman in 1965 in the World Bank data, where it is reported to grow by a factor of 10^6 (by many orders of magnitude the largest value in the dataset and clearly wrong). We reported this issue on February 14, 2025, which led to a correction of this mistake in a subsequent version of the data.

In contrast, nuisance events are data points that accurately reflect reality but arise from rare, extreme non-climate events, which are not captured by the model and yet can be understood through external knowledge sources. For example, the ISIMIP economic data shows a decline of economic production per capita in Iraq in the year 1991 by 65%. This is an extreme event in the data compared to 90% of the data having growth values between -5.0% and 7.8%. Furthermore, it can be attributed to a non-climate event: the Gulf War. See Table M in S1 Appendix for more details on this example.

Off-distribution data

Another type of data that can disrupt a statistical analysis is *Off-Distribution Data*, which refers to data that does not originate from the distribution under study. This type of data often enters a dataset through processes like imputation, where missing values are filled with proxy estimates derived from other available information instead of direct measurements. While such data can be valuable for maintaining dataset completeness, it may also substantially skew the results of a statistical analysis. Unlike outliers, Off-Distribution data points typically have values that align with the general range of the target distribution, making them less conspicuous. However, they often exhibit distinct statistical properties, such as reduced variance or artificially inflated correlations with other variables. These discrepancies can introduce bias or lead to invalid conclusions in the analysis.

For instance, consider the ISIMIP GDP data for post-Soviet states, which spans from 1960 to 2021, including periods before these states existed independently of the Soviet Union. In the dataset, economic growth data for the years 1960–1985 is imputed with a linear method, see Fig A in S1 Appendix. This imputation introduces an artifact where the correlation coefficient between any two post-Soviet states for the economic growth rate equals 1. Consequently, there is no independent statistical information about economic changes of individual states during this period beyond what is reflected in the aggregated output of the Soviet Union. See Finding and Treating Off-Distribution Data for a more comprehensive correlation analysis.

Example of the impact of outliers

As an example, we use the following model to find the economic impact of tropical cyclone and changes in mean annual surface temperature:

$$Y_{s,t} = \beta_{\text{tas}} X_{\text{tas},s,t} + \beta_{\text{tas}^2} X^2_{\text{tas},s,t} + \sum_{\ell=0}^{5} \beta_{\text{tc},\ell} X_{\text{tc},s,t-\ell} + \alpha_t + \alpha_{s,0} + \alpha_{s,1} t + \alpha_{s,2} t^2 + \varepsilon_{s,t}, \tag{6}$$

where, for region s and year t, $Y_{s,t} = \log(\text{GDPpc}_{s,t}/\text{GDPpc}_{s,t-1})$, $X_{\text{tas},s,t}$ is the mean (population weighted) annual near-surface air temperature (tas), and $X_{\text{tc},s,t-\ell}$ is the mean ratio of people affected by a tropical cyclone (TC) with ℓ lag years.

Our initial results indicate that lags $\ell = 1, 3, 5$ of tropical cyclone events are significant, while temperature predictors are not. Through influence analysis, we identify the 2002 observation from the U.S. Virgin Islands (VIR) as exerting the greatest influence on the lag terms with a growth value of +48.7%, which ranks as the 11th highest out of 10752 observations. This value seems to be an artifact of the dataset; likely a change in source data between 2001 and 2002 (the World Bank GDP time series for U.S. Virgin Islands starts at 2002) rather than a real event: According to the US Virgin Islands Bureau of Economic Research, *there was some weakening in the performance of the US Virgin Islands economy during 2002*, see https://usviber.org/wp-content/uploads/2022/07/EconReviewJan2003.pdf. Removing this single influential observation changes the statistical significance of the lagged effects strongly, and with all outliers and off-distribution data excluded, surface air temperature and its quadratic term become significant. See Table 2 for details.

Treating outliers

A straightforward approach to detecting outliers involves inspecting the largest and smallest values of each variable; unusually high or low values may indicate potential outliers. In connection with further knowledge about the data, we can identify outliers. For example Table 3 shows the largest and smallest GDP growth values in the ISIMIP dataset. The largest decline in GDP is observed in Iraq in 1991. This extreme value can be attributed to the Gulf War and thus be counted as an outlier (it is extreme and not climate related).

In this study, we manually review potential outliers characterized by extreme growth events and investigate their causes through online searches. If a non-climate-related explanation for the extreme growth is identified, we remove the corresponding data point from the dataset. The resulting dataset is described in A Cleaned Dataset.

It is crucial to emphasize that we do not begin by performing a regression analysis to identify points with high residuals or those that might contradict our hypothesis. Such an approach could risk (either intentionally or unintentionally) introducing *p*-hacking. Instead, we base the exclusion of data points on external, independent information, ensuring that the removal decision is not influenced by its effect on the regression results.

Alternative treatments of outliers

Additionally, external datasets can be utilized to supplement information when a given data point is deemed unreliable. For instance, one might use a conflict dataset, such as the *Uppsala Conflict Data Program* [33], to exclude data points associated with armed conflicts of a certain scale. While this approach has the merits of a more formal and less laborious procedure than checking extreme values by hand, it is not without limitations. In particular, the time frame or the

Table 2. Coefficients and *p*-values for applying model (6) to different subsets of the ISIMIP dataset. *Raw* is the whole dataset; for *Clean 1*, VIR 2002 is removed; for *Clean All*, the cleaned economic data as described in A Cleaned Dataset is used.

Level Regression for TC Impacts for Different Cleaning						
	Raw		Clean 1		Clean All	
Predictor	coeff.	*p*-val.	coeff.	*p*-val.	coeff.	*p*-val.
Δtas	0.9776	0.059	0.9810	0.058	1.4773	0.001
Δtas^2	-0.9657	0.062	-0.9691	0.060	-1.4723	0.001
Δtc64kt	-0.0016	0.875	0.0016	0.870	-0.0064	0.672
Δtc64kt$_{-1}$	0.0343	0.047	0.0407	0.015	0.0536	0.043
Δtc64kt$_{-2}$	0.0311	0.080	0.0404	0.008	0.0559	0.040
Δtc64kt$_{-3}$	0.0256	0.050	0.0272	0.036	0.0509	0.051
Δtc64kt$_{-4}$	0.0247	0.065	0.0190	0.125	0.0364	0.132
Δtc64kt$_{-5}$	0.0188	0.042	0.0159	0.077	0.0258	0.123

https://doi.org/10.1371/journal.pclm.0000962.t002

Table 3. For the ISIMIP economic data, the most extreme changes in GDPpc as identified by $\Delta GDPpc_{s,t} = (GDPpc_{s,t} - GDPpc_{s,t-1})/GDPpc_{s,t-1}$. For a translation of the ISO country code to a country name, see Tables N to P in S1 Appendix. Less than 1% of all 15,189 growth values in the dataset are below −12% and less than 1% are above 14%.

10 Highest and 10 Lowest GDPpc Growth Values in ISIMIP						
Minimum				Maximum		
Year	ISO	ΔGDPpc	Rank	ΔGDPpc	ISO	Year
1991	IRQ	-65%	1	140%	GNQ	1997
2011	LBY	-62%	2	122%	LBY	2012
1994	RWA	-48%	3	92%	BIH	1996
1992	GEO	-45%	4	90%	VIR	1990
1992	ARM	-41%	5	60%	GNQ	1996
1976	LBN	-40%	6	57%	GNQ	2001
2003	IRQ	-38%	7	53%	GUY	2020
1994	KHM	-37%	8	53%	OMN	1968
2013	CAF	-37%	9	51%	LBR	1997
1989	LBN	-34%	10	49%	IRQ	2004

https://doi.org/10.1371/journal.pclm.0000962.t003

magnitude of the conflict and the economic signal too often do not align. Overall, although such datasets can provide valuable insights, the results lack sufficient precision, leading us to forgo this approach.

Another alternative approach is robust regression, which automatically down-weights data points that deviate substantially from the model. However, this method has a key drawback: extreme weather events, which often produce such outlier data points, are precisely the phenomena we aim to analyze. Excluding or down-weighting these points would mean losing valuable information, even if retaining them increases variance and uncertainty in the results. Additionally, applying robust regression to a large dataset with many predictors can be computationally expensive and numerically unstable. See Section C in S1 Appendix for an example.

Instead of removing outliers, one could introduce a dummy variable, such as one indicating the presence of armed conflict. However, because the economic consequences of armed conflict vary substantially across cases, this approach would not fully address the outlier characteristics of the affected data points.

Alternatively, one might choose to ignore outliers altogether, assuming they are neither numerous nor severe enough to substantially influence the results. However, this approach is only valid if an analysis demonstrates that the impact of outliers is indeed negligible. As shown above, outliers can have a substantial effect on the results, making this a risky assumption.

Finding and treating off-distribution data

We focus on identifying off-distribution data in the economic target variable of the ISIMIP dataset. To achieve this, we perform two key analyses: a correlation-based clustering of time series to detect unusual patterns and a search for constant segments within the data. Based on these analyses, we identify and remove off-distribution observations.

Specifically, we compute the correlation matrix between each pair of economic time series of the form $((Y_{s,t})_{t=t_{\text{from}},\ldots,t_{\text{to}}}, (Y_{s',t})_{t=t_{\text{from}},\ldots,t_{\text{to}}})$, $s \neq s'$, where $Y_{s,t}$ is the economic growth rate for country s in year t. Using these correlations, we perform hierarchical clustering, specifically the unweighted pair group method with arithmetic mean (UPGMA), to identify clusters of highly correlated countries.

Since imputation is often applied to segments of time series data, we compute correlations for each time interval $[t_{\text{from}}, t_{\text{to}}]$ within the dataset's full time range, ensuring a minimum interval length of 7 years ($t_{\text{to}} - t_{\text{from}} \geq 7$) to obtain

reliable results. Additionally, to minimize false positives due to spurious correlations, we only consider clusters with extremely high mean correlations. The resulting clusters, with mean correlations close to 1, are shown in Table Q in S1 Appendix.

The ISIMIP dataset also includes time series with segments of exactly constant economic output, which are likely the result of imputation. These segments, having zero variance, produce undefined correlations with other variables, making correlation analysis inapplicable. To address this, we identify all segments where the time series remains exactly constant for at least 3 years (zero growth over at least 2 years) and mark them as off-distribution. These segments are detailed in Table R in S1 Appendix.

Using these findings, along with additional historical context, we identify and remove observations classified as off-distribution, as described in A Cleaned Dataset.

A cleaned dataset

To obtain a cleaned economic dataset from the original ISIMIP data, we have marked: 182 observations as (non-climate related) outliers by checking the most extreme growth events individually; 2260 observations as imputed because of highly correlated time series segments; 1046 observations as imputed because of constant time series segments. Additionally, we remove post-Soviet and post-Yugoslavia observations up to 1996 (756 observations), as they are partially imputed and partially reflect political changes so that we deem them uninformative on climate events. From the original 15189 values, we remove 3422 observations in total, which is less than the sum of the number of elements in the previously mentioned categories as those intersect. Thus, the final dataset has 11767 values.

Step 2: Controlling for global shocks and regional time trends

To find evidence for a causal connection between economic and climate variables, we have to control for time-point specific and region specific effects. To illustrate, let us consider the following version of (1),

$$Y_{s,t} = f_\beta(Z_{\bullet,s,t}) + \alpha_t + g_s(t) + \varepsilon_{s,t}\,. \tag{7}$$

Here f_β is a function that models the influence of the predictors on the response; α_t is the global time-specific fixed effect (growth of the global economy in year t); $g_s(t)$ is the region specific time trend control (the smooth time $t$rend of economic growth of a given region), e.g., $g_s(t) = \alpha_{s,0} + \alpha_{s,1}t$ for a linear time trend; and $\varepsilon_{s,t}$ models the remaining error of the model.

The fixed effect α_t accounts for time-specific shocks affecting all regions, such as global economic crises, while $g_s(t)$ captures region-specific long-term trends in economic growth, such as persistent differences between developing and developed countries.

The power of Frisch–Waugh–Lovell

The number of control parameters $\alpha_t, \alpha_{s,0}, \alpha_{s,1}$ for $s \in \mathcal{S}, t \in \mathcal{T}$ may be large while the total number of observations may be too small to accurately estimate their values. Fortunately, the Frisch–Waugh–Lovell (FWL) theorem demonstrates that this limitation has minimal impact on the accuracy of the estimator of the parameters of interest, β.

Recall the general regression setup introduced in Panel Data Regression with Fixed Effects. By the FWL theorem, the least squares estimate of the parameters of interest β in the model $\boldsymbol{Y} = \boldsymbol{Z}\beta + \boldsymbol{w}\alpha + \varepsilon$ is equivalent to the least squares estimate of β in the transformed model

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{Z}}\beta + \tilde{\varepsilon}, \tag{8}$$

where the transformed variables $\tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{Z}}$ are obtained by removing the components explainable by $\boldsymbol{w}\alpha$. Specifically, $\tilde{\boldsymbol{Y}} = M\boldsymbol{Y}$ and $\tilde{\boldsymbol{Z}} = M\boldsymbol{Z}$, with M being the linear projection onto the complement of the column space of $\boldsymbol{w}$:

$$M = I - \boldsymbol{w}(\boldsymbol{w}^\top \boldsymbol{w})^{-1}\boldsymbol{w}^\top \tag{9}$$

where $I \in \mathbb{R}^{n \times n}$ is the identity matrix.

This result not only allows an accurate estimation of β without precise estimation of α, but also offers substantial computational advantages when analyzing panel data with many control variables. Given the raw data $(X_{\bullet,s,t}, Y_{s,t})_{(s,t)\in\mathcal{I}}$, we can streamline the computation by first creating the features $Z_{\bullet,s,t}$ to analyze. Next, we apply a control variable design and transform the target and features to obtain the controlled dataset $(\tilde{Z}_{\bullet,s,t}, \tilde{Y}_{s,t})_{(s,t)\in\mathcal{I}}$. This preprocessing step decouples the estimation of β from the high-dimensional controls, substantially reducing computational complexity.

This approach is especially advantageous when the dimension q of α is large relative to the dimension p of β. For instance, in the example described in (6), we have 225 countries and 61 years, requiring $q = 733$ control parameters if quadratic time trends are included. In contrast, the parameter of interest, β, has dimension $p = 8$. While q may still be manageable for a single regression, scenarios involving bootstrap or cross-validation (as discussed in Step 3: Model Selection via Out-of-Sample Test) often require thousands of regressions with the same controls. In such cases, the two-step procedure enabled by the FWL theorem provides immense computational benefits. Furthermore, covariance matrices needed for confidence intervals and hypothesis tests can be computed in the $p \times p$ dimensions of the transformed model instead of $(q + p) \times (q + p)$ dimensions of the original model equation (if small sample corrections are adjusted correctly).

Local time trends

To account for the long-term time trend of each country, the function $g_s(t)$ in (7) must be specified. A common approach is to model $g_s(t)$ as a polynomial in t, with the degree typically ranging from 0 (constant, capturing only the country-specific mean growth) to 3 (cubic, allowing for up to two changes in direction). While low-degree polynomials capture only simple trends, higher degrees may inadvertently remove meaningful information that should be attributed to the parameters of interest.

However, polynomial time trends have notable drawbacks. First, practitioners must choose the polynomial degree, and there is no universally accepted rule for making this decision. This may (involuntarily) lead to *p*-hacking when the choice of polynomial degree is influenced by the desired claim of the research article. Second, even the best-fitting polynomial may fail to adequately approximate the time trend for all regions. Third, polynomials can also introduce artificial long-distance correlations, which may distort the results (see Fig 1).

Rather than relying on polynomials, we propose a more flexible nonparametric approach that directly embodies the idea of a *typical value* for a given time period—longer than a single time point but shorter than the entire time interval. Specifically, we use a local linear estimator for the time trend. We refer to this method as *Window* and *Kernel*, respectively, depending on whether all observations are weighted equally or according to a quadratic kernel. These local linear methods resemble a rolling window approach (local constant) but offer statistical advantages, particularly near the edges of the time interval [34]. An alternative nonparametric time trend approach using basis functions instead of local regression can be found in Kneip et al. [35].

As with the rolling window method, the local linear estimator requires a parameter to define how much of the surrounding data influences the estimate at each time point. This parameter, known as the bandwidth, determines the threshold for considering neighboring time points when calculating the trend. To make the method fully automatic, we propose using leave-one-out cross-validation with the economic response variable as target to select the optimal bandwidth. The automatically selected values are such that 10–12 years before and after a given year are used to calculate the local time trend.

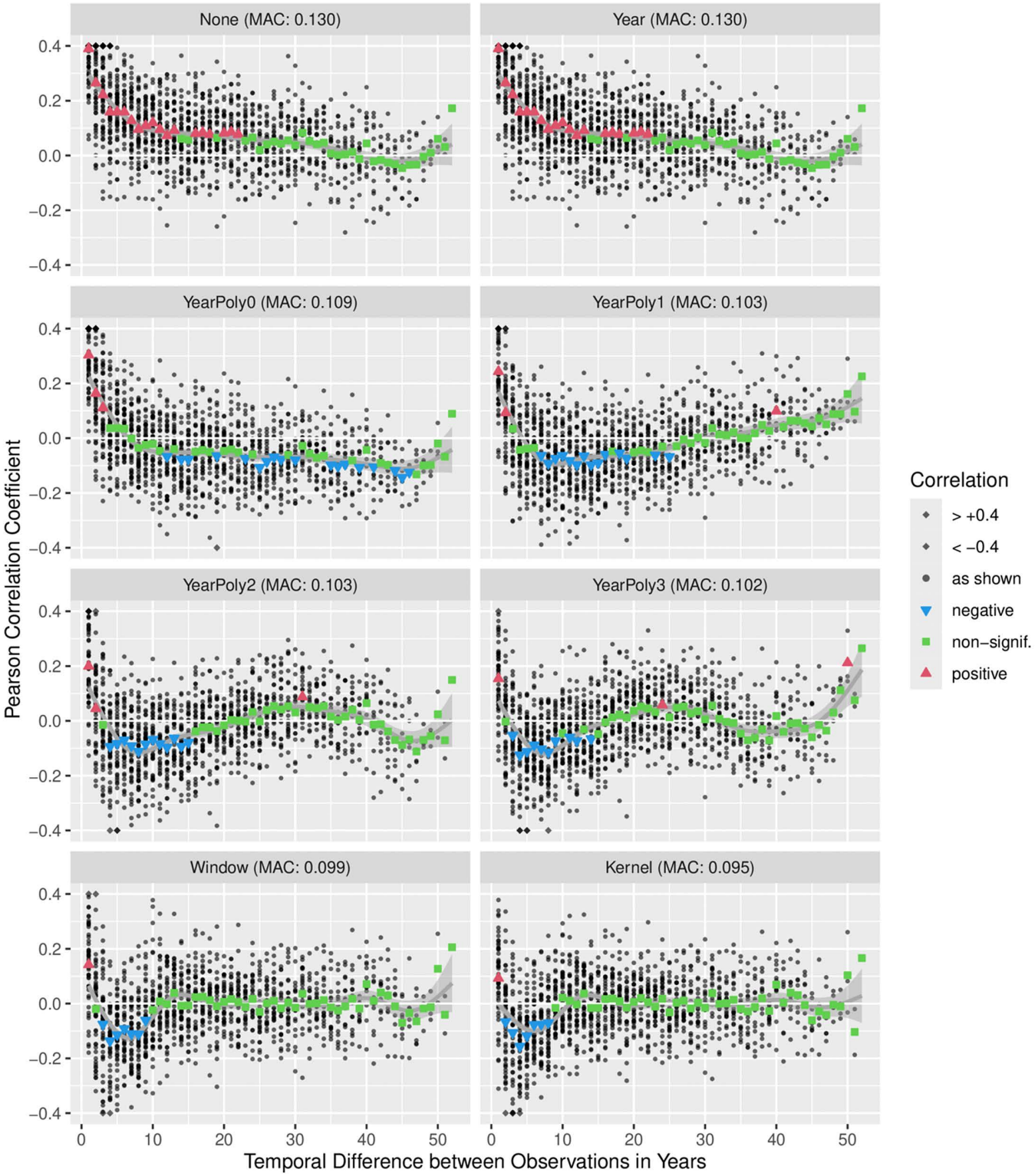

Fig 1. Effect of different control variable designs on temporal correlation of residual economic growth after controlling. Each black dot is an empirical correlation value between two different years calculated from the ISIMIP economic data. The gray curve shows a Loess smoother of these

correlations. For each fixed temporal difference, we test whether the correlations values for pairs of years with this given difference are significantly different from zero via a conservative permutation test, see Section D in S1 Appendix. Furthermore, we apply the Benjamini-Hochberg correction (with false discovery rate level of 0.05) to account for the multiple testing issue. Blue and red triangles show correlations significantly different from zero, whereas green squares indicate that non-zero correlation cannot be detected. MAC is the mean absolute correlation. Without controls (*None*), we observe a high positive correlation for more than 20 years. If only year fixed effects are applied, the result is exactly the same as this only changes the means of the correlated variables. All polynomial time trends are able to reduce the positive correlation of small temporal difference, but introduce negative correlations of temporal differences between 3 and 14 (and more) years. Both local control designs remove correlations in later years. Only (mostly negative) correlations remain for temporal differences of six to seven years. Overall the Kernel-based method seems to show the best results.

https://doi.org/10.1371/journal.pclm.0000962.g001

Figs 1 and 2 contrast the temporal and spatial correlations, respectively, resulting from different control variable designs. The local linear kernel approach yields the least systematic temporal and spatial correlations and does not introduce spurious correlation artifacts over extended time periods. Furthermore, using the automatic bandwidth choice, we remove further control design decisions ensuring objectivity of results.

Correlations and standard errors

Our investigation of correlations focuses primarily on the economic growth variable and its residual values after controlling for global shocks and regional time trends. This approach is a justified heuristic approximation: after controlling, the estimated residuals are nearly identical to the residual of a full regression model with climate variables. This holds because, in our application, no climate variable explains a large portion of the variation in economic growth. Consequently, a preliminary correlation analysis for the climate variables is not crucial. Nonetheless, should we identify a highly predictive variable, it would be appropriate to redo the analysis of correlations based on the residuals from the final regression model.

Different standard error designs. After controlling, some correlation in the error term typically remains. In order to ensure valid inference, we must choose a standard error estimator that properly accounts for the remaining correlations. In our work, we have considered several approaches, each of which deals with spatial and/or temporal dependence in distinct ways. In what follows we briefly summarize these methods and their key properties.

HAC (Newey–West) Standard Errors. The Newey–West estimator [36] is specifically designed to address auto-correlation in time series data. It applies a kernel (typically the Bartlett kernel) that weights observations based on their temporal distance, thereby correcting for serial correlation and heteroskedasticity. However, Newey–West standard errors do not address cross-sectional (spatial) dependence.

Clustered Standard Errors by Year. Clustering standard errors [37] by year allows for arbitrary correlation among cross-sectional units (e.g., countries) within the same time period. This approach assumes that, within each year, the error terms may be correlated in an unrestricted manner across units, but that there is no serial (temporal) correlation across different years.

Clustered Standard Errors by Country. In contrast, clustering by country accommodates arbitrary temporal correlation within each country. Here, it is assumed that errors within a given country over time may be arbitrarily correlated, while errors across different countries are treated as independent. In contrast to estimators that deal with temporal auto-correlation where the correlation is determined by the lag (difference between two years), here any two years can be arbitrarily correlated.

Two-Way Clustered Standard Errors by Country and by Year. Multi-way clustered standard errors [38,39] require uncorrelated error terms only for observations that do not share a common cluster in any clustering dimension. In our case, we require an observation from country s and year t to be uncorrelated with an observation from country s' and year t' for $s \neq s'$ and $t \neq t'$. But we account for correlations between observations at (s, t) and (s, t') as well as for correlations between (s, t) and (s', t). To have a reliable estimate of two-way clustered standard errors, we need to have enough clusters for each clustering dimension (enough different years, and enough different countries).

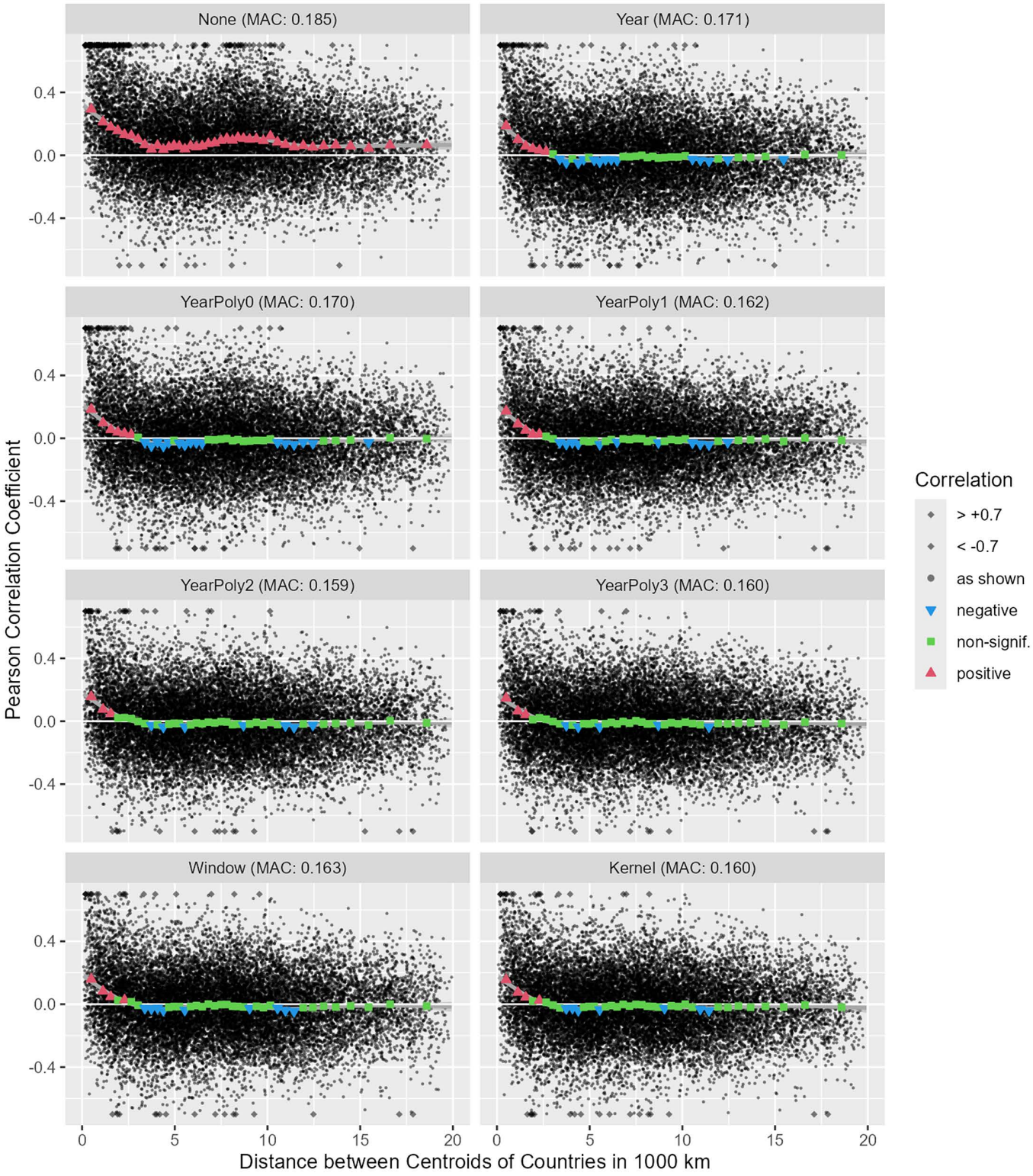

Fig 2. Effect of different control variable designs on spatial correlation of residual economic growth. Plotting details and significance testing are the same as in Fig 1, except that the horizontal axis shows the distance between country centroids. Significance tests are performed on 40 bins of

spatial distances with roughly equal numbers of pairs. Results: Without controls (*None*), we observe a high positive correlation for any spatial distance. As soon as year fixed effects are applied, only low distance positive correlations remain, but some negative correlations for larger distances get newly introduced. Adding polynomial or local time trends to the controls reduces the negative correlations. From this perspective, local time trends or polynomial time trends of degree at least 1 should be preferred.

https://doi.org/10.1371/journal.pclm.0000962.g002

Conley Standard Errors. While Newey–West account for temporal correlation based on temporal distance, Conley standard errors [40] account for spatial correlation based on spatial distance. A kernel is applied that downweights correlations as the distance increases. The distance is typically taken as the physical distance, but if other notions of distance exist, they can also be used. Some implementations can additionally deal with temporal auto-correlation (in R, the package `conleyreg`).

Driscoll–Kraay Standard Errors. Driscoll–Kraay standard errors [41] extend the HAC idea by first averaging the estimating equations over the cross-sectional units and then applying a Newey–West type correction. This method is robust to both arbitrary cross-sectional (spatial) correlation (not necessarily based on distance) and temporal auto-correlation (based on temporal distance). A potential drawback is that this estimator is more data hungry (i.e., it requires a sufficiently long time dimension) and tends to be less precise when the number of time periods is small.

Correlations in the economic growth rate. We emphasize that spatial correlation in economic data is often not driven solely by the physical distance between countries. While neighboring countries frequently exhibit strong trade links (which is an argument in favor of distance based spatial correlation), being members of the same economic union or sharing similar economic policies can induce strong correlations even if the countries are not immediate neighbors or geographically close.

To demonstrate this point, we compare correlations between countries of the European Union (EU) with correlations of countries that are geographically close. We use the countries of the EU that were members in the year 1996 (EU membership does not change thereafter until 2004) so that the effect of being part of the EU shows in the observation time series (1967–2018). The mean correlation between non-neighboring EU countries is 0.36 whereas neighboring countries globally have an average correlation of 0.15. This difference is highly significant ($p < 10^{-5}$ according to a bootstrap test). Similarly, the mean correlation between EU countries with a centroid distance larger than the median EU centroid distance (1534km) is 0.33, which is significantly ($p < 10^{-5}$ according to a bootstrap test) larger than 0.07, the global mean correlation of countries with less centroid distance than the aforementioned threshold. Thus, being part of a common economic zone such as the EU is more relevant than geographical proximity.

Furthermore, our residual analysis reveals that there are some instances of long-distance (negative) spatial correlations (Fig 2), while long temporal correlations are absent (Fig 1) when using *Kernel* controls. In view of these features, we advocate for the use of Driscoll–Kraay standard errors. Despite their higher data requirements, they offer a robust solution that simultaneously addresses both arbitrary spatial correlation and temporal auto-correlation, without suffering from the systematic bias observed in alternative estimators. This also speaks to the concern that omitted economic determinants of growth—such as shared policies or common economic shocks—may induce residual correlation: such correlation is exactly what the Driscoll–Kraay estimator is designed to absorb, leaving the consistency of our climate estimates unaffected under the exogeneity of short-run anomalies.

Small sample correction. To account for the degrees of freedom already removed in the controlled data set, the small sample correction in the calculation of standard errors has to be adjusted appropriately. In the case of local linear time trends, we use the effective number of parameters, i.e., the trace of the smoother matrix S, which fulfills $\hat{\boldsymbol{Y}} = S\boldsymbol{Y}$, where $\hat{\boldsymbol{Y}}$ is the temporally smoothed economic data, so that the controlled data is $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$.

Step 3: Model selection via out-of-sample test

Traditionally, studies have relied on annual averages of temperature (`tas`) and precipitation (`pr`) to predict GDP growth. However, such models may be overly simplistic. More recent research has introduced complex variables—often derived from `tas` and `pr`—such as measures of extreme rainfall or temperature fluctuations. In addition, studies examining specific

channels through which weather impacts the economy often incorporate variables like the number of people affected by tropical cyclones.

When accounting for all plausible combinations and transformations of these variables, the number of potential predictors grows dramatically. Selecting a model based solely on its fit, without acknowledging the many alternatives considered and discarded, risks *p*-hacking and fosters unwarranted confidence in the results. This underscores the necessity of a formal, transparent variable selection procedure.

In the analysis that follows, we implement a suite of formal model selection procedures. Crucially, model performance is assessed using out-of-sample metrics, rather than in-sample fit, which can be misleading due to reliance on assumptions that may not hold in practice. This approach provides a more robust evaluation of predictive accuracy.

Variable selection and testing are carried out on the controlled dataset, allowing us to isolate the effect of the predictors from that of the control variables. To assess the robustness of the results, we repeat the analysis across multiple datasets, each defined by a different control specification. However, it is important to emphasize that test-set performance cannot be directly compared across different control variants, as each defines a distinct prediction objective.

Methods

The controlled data is split into training (pre-2007) and testing (2007 onward) subsets, yielding 41 years of training and 11 years of testing data. To assess robustness, we also evaluate alternative splits using 1997 and 2002 as threshold years. We avoid shorter test periods, as they reduce the reliability of evaluation statistics. The split at a threshold year maintains spatial dependencies and reduces the influence of temporal auto-correlation. Furthermore, it exactly mimics the main use case of the model results in which economic damages are projected into the future.

This design reduces the influence of temporal auto-correlation but does not eliminate it: after residualizing against the controls, the data may retain some short-range dynamic dependence, and the earliest test years in particular share information with the adjacent training years. Such residual dependence would tend to make the test error slightly optimistic. We do not introduce a temporal gap ("padding") between the training and test sets to suppress this further, for two reasons. First, with only about 50 annual observations per country, the test set must remain large enough to detect an effect in this highly noisy setting; padding would consume scarce observations. Second, padding introduces an additional hyperparameter with its own trade-offs. We retain the simpler split because our central results are negative: most models show little to no predictive power even in this potentially favorable setting, so any optimistic bias from residual dependence works against our conclusions rather than inflating them.

Most selection methods involve tuning hyperparameters, which we optimize via five-fold cross-validation on the training set. Folds are composed of consecutive years (for the 2007 train-test split: 1966–1974, 1975–1982, 1983–1990, 1991–1998, 1999–2006), ensuring that all data from a given year remains in the same fold. This design again respects both spatial correlation and reduces the influence of temporal auto-correlation.

We emphasize that cross-validation here serves only to tune hyperparameters; our reported conclusions rest on the held-out test set, evaluated in temporal order. The fold structure thus affects only model tuning and not the validity of the final out-of-sample assessment.

Once optimal hyperparameters are identified, each model is retrained on the full training set and used to predict the (controlled) economic response in the test set. We assess the predictive pseudo-significance of each model by benchmarking its performance against a random prediction independent of any predictor variables and modeled by an isotropic Gaussian random vector with optimally chosen variance. See Section F in S1 Appendix for details.

To assess the stability of predictor selection, we employ a block bootstrap approach with yearly blocks, resampling the training data 1000 times. Each resample includes a number of years equal to the original training period, drawn with replacement. When a year is drawn multiple times, all data from this year appears multiple times in the resampled dataset. For each bootstrapped sample, we reapply the selection procedure (but not the hyperparameter tuning) and track

how often the originally selected predictors are re-selected. This *re-selection rate* serves as an indicator of the robustness of the variable selection process.

Note that all of these calculations happen on the controlled data. This is necessary, as we cannot estimate the controls of the test data using the training data, and all analysis has to be carried out orthogonally to the controls. We expand on this issue and compare our framework with that of Newell et al. [9] in Section G in S1 Appendix.

Information criteria. A common way of model selection is using information criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Unfortunately, these are not applicable in our setting as they require independent and identically distributed variables [42]. We instead have a complex dependence structure with temporal and spatial correlation and possible heteroskedasticity. Extending the information criteria to such dependence structures is highly non-trivial. Therefore, we do not use these criteria in our experiments.

Subset, pair, and forward selection. In *Subset Selection*, we fix the number of predictors $0 \leq k \leq p$. For each subset of k predictors, we fit the model using ordinary least squares (OLS) and compute the residual sum of squares (RSS), selecting the subset that yields the lowest RSS. The value of k is a hyperparameter to be chosen via cross-validation. However, because there are $\binom{p}{k}$ possible subsets of size k from p predictors, this approach is computationally demanding for large k: For our case with $p = 708$, the number of subsets for $k = 2, 3, 4$ is approximately 2.5×10^5, 5.9×10^7, and 1.0×10^{10}, respectively. While specialized algorithms can perform subset selection more efficiently than brute-force enumeration [43], the computational cost remains prohibitive in our high-dimensional setting.

To address this, we instead employ a more computationally efficient selection method variant known as *Forward Selection* [25]. This method builds the model sequentially by adding, at each step, the single predictor that most reduces the RSS. While this approach drastically reduces computation time, it has a key limitation: It may fail to select predictor combinations that are jointly informative but not individually strong. For instance, a variable like tas and its squared term tas^2 might only be useful when included together, and Forward Selection might overlook them individually. In contrast, Subset Selection could capture such combinations.

To assess the impact of this limitation, we also implement *Pair Selection*, i.e., Subset Selection restricted to $k = 2$, to evaluate whether jointly informative pairs are being missed by Forward Selection.

LASSO, ridge, and elastic net. The *LASSO* and *Ridge Regression* penalize the absolute values of the regression coefficients so that small absolute values are preferred. For a hyperparameter $\lambda \geq 0$, the respective coefficient estimates can be written as

$$\hat{\beta}_{\mathrm{LASSO}} = \arg\min_{\beta} \left(\frac{1}{n} \|\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{Z}}\beta\|_2^2 + \lambda \|\beta\|_1 \right), \tag{10}$$

$$\hat{\beta}_{\mathrm{Ridge}} = \arg\min_{\beta} \left(\frac{1}{n} \|\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{Z}}\beta\|_2^2 + \frac{\lambda}{2} \|\beta\|_2^2 \right). \tag{11}$$

The penalties allow these methods to be applied in settings with a large number of coefficients compared to the sample size. Meaningful estimates are possible even in settings with more coefficients than observations. The different penalties between LASSO (L^1) and Ridge (L^2) yield different coefficient estimates: LASSO prefers to set many coefficients to exactly 0 (thus implying a selection of variables), whereas Ridge typically does not estimate any coefficient to be exactly 0 but all coefficients will be shrunken sufficiently close to 0. Note that both methods tend to estimate coefficients with lower absolute values than OLS, i.e., they imply a form of shrinkage. The *Elastic Net* combines both LASSO and Ridge Regression:

$$\hat{\beta}_{\mathrm{ElasticNet}} = \arg\min_{\beta} \left(\frac{1}{n} \|\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{Z}}\beta\|_2^2 + \lambda \left(\frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right) \right). \tag{12}$$

The penalty parameter $\lambda \geq 0$ and the trade-off parameter $\alpha \in \{0.1, 0.2, \ldots, 1\}$ for ElasticNet are chosen using 5-fold cross validation on the training set with the folds consisting of all data from mutually exclusive consecutive years. Ensuring $\alpha > 0$ promotes ElasticNet to set some coefficients to exactly 0 as happens in LASSO implying variable selection.

Fordge. As pair and forward selection yield much worse results than the shrinkage methods LASSO, Ridge, and Elastic Net, we hypothesize that the missing shrinkage is the problem. Thus, we additionally test the usage of Ridge regression on the variables chosen by forward selection. We call this procedure *Fordge* (a portmanteau of forward and ridge).

Results

Here we present the results of all six different model selection methods in different settings. Detailed results can be found in Section J in S1 Appendix.

Test error

All experiments. First, we consider the mean squared error (MSE) on the test set and its pseudo-significance (see Section F in S1 Appendix) for our six methods: Ridge Regression, LASSO, Elastic Net, Pair Selection, Forward Selection and Fordge. Apart from different methods, we also compare raw vs cleaned data, ISIMIP vs World Bank GDP, 8 different control designs, level vs growth regression, and three different options for the threshold year that splits our data into train and test set. In total, this makes 1152 experiments.

For binary attributes the results are summarized in Table 4. Generally, the cleaned data is clearly easier to predict (in 197 cases with 337 draws and 42 opposite cases). This is to be expected as outliers are difficult or even impossible to forecast. The World Bank data seems somewhat more predictable than the ISIMIP data (in 136 cases with 369 draws and 71 opposite cases), but the difference is not large. Level and growth regression have similar predictability (359 draws, 93 cases where level is more predictable, 124 growth).

Predictive power slightly improves when increasing the train-test split year from 1997 to 2007: The mean test error ranks among the three threshold choices are 2.07 (split in 1997), 1.97 (split in 2002), 1.95 (split in 2007), respectively. Furthermore, the number of experiments with any predictive power (test error lower than for the trivial model without predictors) increases with later split year. Out of 384 experiments nonzero predictive power is shown in 82 (split in 1997), 96 (split in 2002), and 112 (split in 2007) experiments, respectively. Both results seem to be a consequence of better estimates due to more training data for later split years.

Table 4. Summary of experiments along different attributes. We evaluate the binary attributes *preprocessing*, *regression design*, and *GDP data source* by comparing the *p*-value on the test set for both instances of the attribute and counting how often each instance of the respective attribute has a lower value than the other instance. The high number of draws (both versions perform equally) is due to the *p*-value being set to 1 if the method's test error is larger than the variance of the test data, i.e., the test error of the trivial constant prediction.

Experiment Results Summarized for Different Attributes							
	Attribute Instance		# Lower Error Comparisons			Total Number of	
Attribute	A	B	A	B	Draws	Comparisons	Experiments
Main Experiments							
GDP Data Source	ISIMIP	World Bank	5 (42%)	2 (17%)	5 (42%)	12	24
Regression Design	growth	level	2 (17%)	7 (58%)	3 (25%)	12	24
All Experiments							
Preprocessing	clean	raw	198 (34%)	41 (7%)	337 (59%)	576	1152
GDP Data Source	ISIMIP	World Bank	70 (12%)	136 (24%)	370 (64%)	576	1152
Regression Design	growth	level	124 (22%)	93 (16%)	359 (62%)	576	1152

https://doi.org/10.1371/journal.pclm.0000962.t004

The average predictability of data with different control does not differ a lot with *Window* having the highest predictability. Note that high predictability is not necessarily a quality criterion for the control design.

The test error results for different model selection methods are collected in Section H in S1 Appendix, which we further summarize in Table 5. Ridge regression shows the highest predictive power followed by Elastic Net and LASSO, while Pair Selection, Fordge and Forward Selection show worse performances.

Main experiments. To ensure that this evaluation is not diluted by averaging over different controls and mixing cleaned and uncleaned data, we next focus only on our preferred settings with cleaned data, Kernel controls, and train-test split in 2007—we deem those to be the most relevant and reliable. This leaves 24 experiments (6 methods, ISIMIP vs World Bank, and level vs growth regression).

For the two binary attributes (GDP data source and regression design), the results are summarized in Table 4. We find the ISIMIP data to be slightly more predictable compared to the World Bank data (in 5 cases with 5 draws and 2 opposite cases). This is opposite to the evaluation of all experiments and indicates that the differences are not a strong structural property of the datasets but rather random fluctuations. In 7 out of 12 cases (with 3 draws), level regression models have higher predictive power than growth regression models. This pattern does not show in the common evaluation of all experiments. So, it is unclear how far this statement generalizes.

The three penalized regression methods—Ridge, LASSO, and Elastic Net—again outperform the other approaches (see Table 5). Therefore, we conclude that Ridge, LASSO, and Elastic Net should be preferred. Among them, Ridge consistently achieves the best performance. Nonetheless, one might still favor LASSO or Elastic Net due to their greater interpretability, as they perform explicit variable selection—unlike Ridge Regression.

Best test error. The most predictive setup of the main experiments (level regression, Elastic Net, ISIMIP data) is only able to explain 0.19% of the variance in the test data. This is equivalent to the amount achieved in 2% of cases using a prediction vector independent of predictor variables with a randomly chosen direction and optimal length (i.e., pseudo *p*-value of $p = 0.02$, see Section F in S1 Appendix). Consequently, despite the inclusion of 708 variables (in both original and differenced forms), a substantial proportion of which have been previously associated with GDP in other studies, there remains a small possibility that the predictor dataset contains no substantial information regarding GDP.

Selected variables

To evaluate which among the 708 variables in our regression are most important, we focus on the main experiments (cleaned data, Kernel controls) considering level and growth regression and ISIMIP as well as World Bank data sources for GDP. Furthermore, as we here do not need a test set for this evaluation, we use all data for training, i.e., until 2018,

Table 5. Summary of performance of different methods according to their test error rank (1 to 6) averaged over different experimental settings. All comparisons are made from all 1152 experiments. For the main comparisons we restrict to cleaned data, kernel controls, and 2007 as the split year and vary only between level and growth regression designs and ISIMIP and World Bank as GDP data sources.

Summarized Rankings of Model Selection Methods				
	All 192 Comparisons		Main 4 Comparisons	
Method	Median Rank	Mean Rank	Median Rank	Mean Rank
Ridge	3.5	2.67	1.0	2.25
LASSO	3.5	3.21	3.5	3.25
Elastic Net	3.5	3.05	3.5	3.25
Pair	4.0	3.78	4.0	4.25
Forward	4.0	4.20	4.5	4.50
Fordge	4.0	4.09	4.0	3.50

https://doi.org/10.1371/journal.pclm.0000962.t005

to get the most accurate results. We compare the four out of six model selection methods that explicitly select variables (Forward, Pair, LASSO, Elastic). The results are summarized in Section I in S1 Appendix.

There is only a single variable that gets chosen systematically (except by Forward on growth regression, which does not select any variable): The square of the maximal specific humidity of the previous year, $\text{huss}^2_{\text{max},-1}$ (or $\Delta\text{huss}^2_{\text{max},-1}$ in the level regression case). This variable also has the highest bootstrap re-selection rate among all chosen variables in all settings. Reasons for the relatively high predictive power could be the correlation of high specific humidity to heat stress, storms, and flooding, as well as correlation of low specific humidity with droughts, water shortages, wild fires, respiratory issues and airborne pollutants. Moreover, the squared term may capture the non-linear influence arising from threshold effects where impacts on human health or agriculture escalate disproportionately.

No other variable is chosen robustly. In particular, classical variables such as tas, tas^2, pr, pr^2 are never chosen by any algorithm but some transformations of tasmin and tasmax do show up for some settings without strong consistency.

Shrinkage

LASSO, Ridge Regression, and Elastic Net typically perform better than the other three methods and Fordge is often slightly better than Forward Selection. This suggests that penalized regression in general has an advantage in our setting. To investigate this hypothesis further, we select specific variables either classical tas, tas^2, pr, pr^2 or preferred by our variables selection methods ($\text{huss}^2_{\text{max},-1}$, $\text{sfcwind}^2_{\text{max}}$) and use OLS as well as LASSO and Ridge fits for these variables. 6 shows the results. In all cases the shrinkage of coefficients implied by the penalty of LASSO and Ridge helps to achieve lower test errors.

Stationarity

In our regression analysis, we model the proportional influence of shocks in the climate variables on the economic growth rate to be constant over time and space (temporal and spatial stationarity). This, of course, is an approximation. Even if this does not perfectly reflect the real world, such a model may still be useful, as it allows us to estimate an average effect. But this approximative nature of our analysis has to be kept in mind when interpreting the results or using them for projections into the future.

To ensure validity of future projections based on such regression results, we do require temporal stationarity: The weather–economy relationship should be stable over the observation time to reasonably assume that this relationship also exists in the future. We investigate this property for models involving the classical temperature parabola [4] and models involving the humidity-variable favored by our variable selection analysis. With an F-test, we check whether the data before a threshold year exhibits a similar regression coefficient to the data thereafter, see Fig 3. Most models show significantly different effects in the first time period compared to the later time period, if the threshold year is chosen between 2000 and 2010 indicating a change point within these years. Only the pure specific humidity model (huss A) for level regression is stable over the whole training period. Adding maximal wind surface speeds to the level regression humidity model (huss B), seems to exhibit a change point before 1980 but stability thereafter. These two models show the greatest temporal stationarity, which is consistent with their low test errors in Table 6.

Investigating the change of coefficients in the temperature parabola models further reveals that the quadratic influence of temperature on the economic growth seems to strongly reduce after 1992, see Fig B in S1 Appendix. Thus, when measuring predictive power with a test set of some years after 1993, the quadratic relationship learned from previous years is not predictive. This instability is reason for concern when projecting economic influence of temperature to the future. Coefficient values of the level regression humidity model (huss A), seem slightly more stable, in particular when estimated over time periods of at least 15 years.

Spatial stationarity appears less critical than temporal stationarity, as we are typically projecting outcomes for existing countries rather than entirely unseen ones. If the goal is merely to estimate the average effect of weather on the economy,

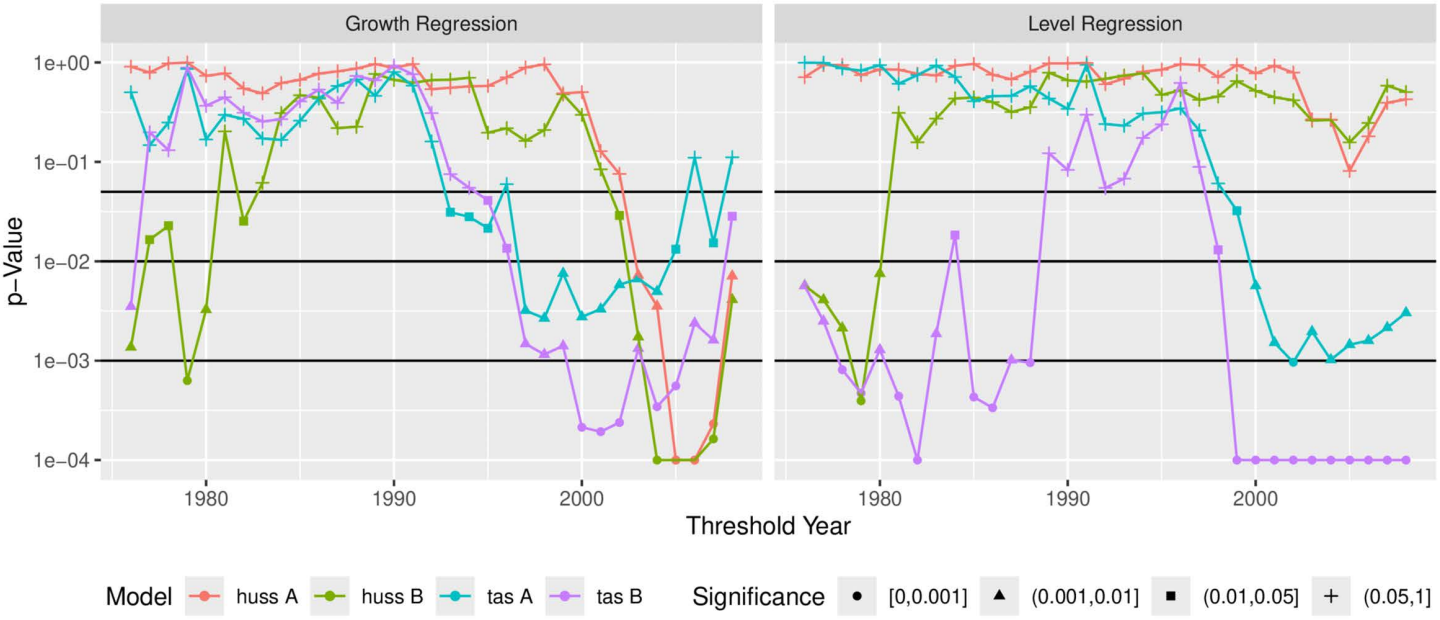

Fig 3. *p*-values of *F*-test for stationarity of effects. We split the data (Kernel controls, cleaned ISIMIP economic data) into two parts: up until the year indicated on the horizontal axis and afterward. Then, for both parts, we fit the OLS models indicated by the color of the displayed curves, see Table 6. We test the Null-hypothesis of equal coefficient vectors for both parts $\beta_{\text{before}} = \beta_{\text{after}}$ using an *F*-test for linear hypothesis and the Driscoll-Kraay covariance matrix with 8 lag years. The *p*-value of the test is shown via the vertical axis, which is on a log-scale and cut at 10^{-4}. Using different symbols, we indicate different significance levels (their threshold is displayed by black horizontal lines). We only show threshold years from 1977 to 2007, even though the data ranges from 1967 to 2018, to ensure that there is enough data on each side of the threshold year to make the results reliable. Note that there is a shift in the source data of ISIMIP between 1978 and 1979, which may explain the low *p*-values for some models around that time.

https://doi.org/10.1371/journal.pclm.0000962.g003

Table 6. For a selected number of different sets of predictor variables, we calculate the OLS, LASSO and Ridge fit, using 5-fold cross validation to determine the penalty parameter for the latter two. We use the cleaned ISIMIP data with Kernel controls and split it into train and test data in 2007. The columns MSE% show the respective change in percent in mean squared error on the test set compared to the trivial null-model. The columns vs OLS show the change of MSE% from OLS to LASSO and Ridge, respectively.

Effect of Shrinkage on Different Models

		OLS	LASSO		Ridge	
Model	**Predictors**	**MSE%**	**MSE%**	**vs OLS**	**MSE%**	**vs OLS**
Level Regression						
huss A	$\Delta\text{huss}^2_{\text{max},-1}$	-0.009	-0.012	-29%	-0.027	-187%
huss B	$\Delta\text{huss}^2_{\text{max},-1}$, $\Delta\text{sfcwind}^2_{\text{max}}$	-0.054	-0.059	-9%	-0.193	-256%
tas A	Δtas, Δtas^2	0.412	0.306	-26%	0.217	-47%
tas B	Δtas, Δtas^2, Δpr, Δpr^2	0.489	0.398	-19%	0.353	-28%
Growth Regression						
huss A	$\text{huss}^2_{\text{max},-1}$	0.856	0.847	-1%	0.745	-13%
huss B	$\text{huss}^2_{\text{max},-1}$, $\text{sfcwind}^2_{\text{max}}$	0.830	0.821	-1%	0.496	-40%
tas A	tas, tas^2	0.160	0.064	-60%	-0.022	-114%
tas B	tas, tas^2, pr, pr^2	0.412	0.338	-18%	0.302	-27%

https://doi.org/10.1371/journal.pclm.0000962.t006

spatial stationarity is not required. However, to make country-specific projections, spatial stationarity either must be demonstrated or it must be clearly stated that the projections do not reflect country-specific economic structures. Instead, they represent average outcomes for a hypothetical country with similar climatic conditions but globally averaged economic characteristics—ignoring factors such as the level of industrialization or the share of agriculture in economic output of the specific country.

It is also important to note that calculating global GDP impacts represents a form of averaging over countries, though with a different weighting scheme than that used in model estimation. While the model gives equal weight to each country—regardless of economic size, population, or geography—global damage estimates are typically dominated by the losses of wealthier countries due to their larger contributions to global GDP.

Discussion

This paper advances global panel data regression in climate econometrics by addressing key methodological shortcomings in existing research and proposing data-driven solutions. Our enhanced empirical strategies challenge earlier findings and underscore the need for more robust modeling practices. To improve future research, we provide methodological recommendations for future climate econometric research.

Outlier treatment and data integrity. We document the substantial influence of outliers on empirical results, present procedures for their identification and treatment, and provide a cleaned dataset to facilitate more reliable analyses. Only few previous works [5] treat outliers at all.

Flexible time-trend controls and correlations. We replace arbitrary, parametric time-trend specifications with a nonparametric, fully automatic procedure. This reduces practitioner bias and more effectively captures complex temporal dynamics in panel data. Furthermore, we show the influence of different control designs revealing undesirable long term temporal correlations introduced by previously common polynomial time trends.

High-dimensional variable selection with out-of-sample validation. To navigate over 700 candidate predictors, we employ formal variable-selection methods—most notably LASSO, Ridge, and Elastic Net—which guard against overfitting, *p*-hacking, and cherry picking. Crucially, we benchmark all models via out-of-sample testing, providing a realistic assessment of predictive power under unknown dependency structures and avoiding untested assumptions inherent in classical inference methods.

Empirical findings. Strikingly, almost none of the tested variables—including those emphasized in prior studies—retain meaningful predictive power out-of-sample. A key reason appears to be the violation of temporal stationarity, an assumption often made but rarely tested. Among the penalized methods, a humidity-related variable consistently emerges as the strongest predictor of economic damage. Investigating this variable—and its interactions with temperature—from both statistical and mechanistic perspectives is a promising avenue for future work.

Implications. By integrating outlier robustness, adaptive trend controls, rigorous variable selection, and out-of-sample validation, our approach reveals the fragility of many conventional findings and calls for a paradigm shift toward methodologies that prioritize predictive accuracy and generalizability in climate econometrics.

Recommendations. From our finding, we derive the following recommendations for future work in climate econometrics to improve robustness and avoid wrong conclusions.

1. Check the data for outliers and analyze their influence.
2. Check temporal and spatial correlations in controlled data to be able to choose a suitable standard error and cross-validation design, and to be able to compare different control designs.
3. Use a formal model selection procedure to avoid a subjective model choice or cherry picking (e.g., use LASSO, Ridge Regression, or Elastic Net).

4. Use out-of-sample testing that is robust against potential dependence structures in the data to verify that the chosen model generalizes.
5. Check temporal stationarity if results are used for future projections.
6. Check spatial stationarity if results are interpreted as country specific or interpret results as global averages only.
7. Try shrinking estimated coefficients, e.g., using penalized regression methods (such as LASSO, Ridge Regression, or Elastic Net) to improve predictive power and mitigate against problems with data dependency and a lack of stationarity.

Supporting information

S1 Appendix. Appendix.
(PDF)

Acknowledgments

The authors gratefully acknowledge Simon Philippssohn for his valuable contributions to the classification of economic outliers.

Author contributions

Conceptualization: Christof Schötz, Christian Otto.

Data curation: Christof Schötz, Jan Hassel.

Formal analysis: Christof Schötz.

Funding acquisition: Christian Otto.

Methodology: Christof Schötz.

Software: Christof Schötz.

Supervision: Christian Otto.

Validation: Christof Schötz, Jan Hassel.

Visualization: Christof Schötz.

Writing – original draft: Christof Schötz.

Writing – review & editing: Christof Schötz, Jan Hassel, Christian Otto.

References

1. Dell M, Jones BF, Olken BA. What Do We Learn from the Weather? The New Climate-Economy Literature. Journal of Economic Literature. 2014;52(3):740–98. https://doi.org/10.1257/jel.52.3.740
2. Hsiang S. Climate Econometrics. Annual Review of Resource Economics. 2016;8(1):43–75. https://doi.org/10.1146/annurev-resource-100815-095343
3. Dell M, Jones BF, Olken BA. Temperature Shocks and Economic Growth: Evidence from the Last Half Century. American Economic Journal: Macroeconomics. 2012;4(3):66–95. https://doi.org/10.1257/mac.4.3.66
4. Burke M, Hsiang SM, Miguel E. Global non-linear effect of temperature on economic production. Nature. 2015;527(7577):235–9. https://doi.org/10.1038/nature15725 PMID: 26503051
5. Pretis F, Schwarz M, Tang K, Haustein K, Allen MR. Uncertain impacts on economic growth when stabilizing global temperatures at 1.5°C or 2°C warming. Philos Trans A Math Phys Eng Sci. 2018;376(2119):20160460. https://doi.org/10.1098/rsta.2016.0460 PMID: 29610370
6. Kalkuhl M, Wenz L. The impact of climate conditions on economic production. Evidence from a global panel of regions. Journal of Environmental Economics and Management. 2020;103:102360. https://doi.org/10.1016/j.jeem.2020.102360

7. Kotz M, Wenz L, Stechemesser A, Kalkuhl M, Levermann A. Day-to-day temperature variability reduces economic growth. Nat Clim Chang. 2021;11(4):319–25. https://doi.org/10.1038/s41558-020-00985-5
8. Kotz M, Levermann A, Wenz L. The effect of rainfall changes on economic production. Nature. 2022;601(7892):223–7. https://doi.org/10.1038/s41586-021-04283-8 PMID: 35022593
9. Newell RG, Prest BC, Sexton SE. The GDP-Temperature relationship: Implications for climate change damages. Journal of Environmental Economics and Management. 2021;108:102445. https://doi.org/10.1016/j.jeem.2021.102445
10. Kahn ME, Mohaddes K, Ng RNC, Pesaran MH, Raissi M, Yang J-C. Long-term macroeconomic effects of climate change: A cross-country analysis. Energy Economics. 2021;104:105624. https://doi.org/10.1016/j.eneco.2021.105624
11. Krichene H, Vogt T, Piontek F, Geiger T, Schötz C, Otto C. The social costs of tropical cyclones. Nat Commun. 2023;14(1):7294. https://doi.org/10.1038/s41467-023-43114-4 PMID: 37996428
12. Howard PH, Sterner T. Few and Not So Far Between: A Meta-analysis of Climate Damage Estimates. Environ Resource Econ. 2017;68(1):197–225. https://doi.org/10.1007/s10640-017-0166-z
13. Nordhaus W, Moffat A. A Survey of Global Impacts of Climate Change: Replication, Survey Methods, and a Statistical Analysis. Cambridge, MA: National Bureau of Economic Research. 2017.
14. Kolstad CD, Moore FC. Estimating the Economic Impacts of Climate Change Using Weather Observations. Review of Environmental Economics and Policy. 2020;14(1):1–24. https://doi.org/10.1093/reep/rez024
15. Chang J-J, Mi Z, Wei Y-M. Temperature and GDP: A review of climate econometrics analysis. Structural Change and Economic Dynamics. 2023;66:383–92. https://doi.org/10.1016/j.strueco.2023.05.009
16. Intergovernmental Panel on Climate Change IPCC. Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. 1 ed. Cambridge University Press. 2023. https://doi.org/10.1017/9781009325844
17. Tol RSJ. A meta-analysis of the total economic impact of climate change. Energy Policy. 2024;185:113922. https://doi.org/10.1016/j.enpol.2023.113922
18. Desmet K, Rossi-Hansberg E. Climate Change Economics over Time and Space. Annual Review of Economics. 2024;16(1):271–304. https://doi.org/10.1146/annurev-economics-072123-044449
19. Castle JL, Hendry DF. Climate Econometrics: An Overview. Foundations and Trends® in Econometrics. 2020;10(3–4):145–322. https://doi.org/10.1561/0800000037
20. Arellano M. Panel data econometrics. repr ed. Oxford: Oxford Univ. Press. 2010.
21. Wooldridge JM. Econometric analysis of cross section and panel data. 2nd ed ed. Cambridge, Mass: MIT Press. 2010.
22. Hill TD, Davis AP, Roos JM, French MT. Limitations of Fixed-Effects Models for Panel Data. Sociological Perspectives. 2019;63(3):357–69. https://doi.org/10.1177/0731121419863785
23. Imai K, Kim IS. On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data. Polit Anal. 2020;29(3):405–15. https://doi.org/10.1017/pan.2020.33
24. Pretis F, Reade JJ, Sucarrat G. Automated General-to-Specific (GETS) Regression Modeling and Indicator Saturation for Outliers and Structural Breaks. J Stat Soft. 2018;86(3). https://doi.org/10.18637/jss.v086.i03
25. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2 ed. New York, NY: Springer. 2009.
26. Lange S, Mengel M, Treu S, Büchner M. ISIMIP3a atmospheric climate input data. 2022. https://doi.org/10.48364/ISIMIP.982724.1
27. Vogt T. ISIMIP3a tropical cyclone tracks and windfields. 2024. https://doi.org/10.48364/ISIMIP.605192
28. Sauer I, Koch J, Otto C. ISIMIP3a GDP input data. 2022. https://doi.org/10.48364/ISIMIP.824555
29. Volkholz J, Lange S, Geiger T. ISIMIP3a population input data. 2022. https://doi.org/10.48364/ISIMIP.822480.2
30. Klein Goldewijk CGM. History database of the global environment (hyde) 3.3. 2023. https://doi.org/10.17026/dans-25g-gez3
31. World Bank. GDP per capita growth (annual %) (ny.gdp.pcap.kd.zg). 2025. https://data.worldbank.org/indicator/NY.GDP.PCAP.KD.ZG
32. Wenz L, Carr RD, Kögel N, Kotz M, Kalkuhl M. DOSE - Global data set of reported sub-national economic output. Sci Data. 2023;10(1):425. https://doi.org/10.1038/s41597-023-02323-8 PMID: 37400486
33. Davies S, Engström G, Pettersson T, Öberg M. Organized violence 1989–2023, and the prevalence of organized crime groups. Journal of Peace Research. 2024;61(4):673–93. https://doi.org/10.1177/00223433241262912
34. Fan J, Gijbels I. Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66. Taylor & Francis: Chapman & Hall/CRC. 1996.
35. Kneip A, Sickles RC, Song W. A new panel data treatment for heterogeneity in time trends. Econom Theory. 2012;28(3):590–628. https://doi.org/10.1017/s026646661100034x
36. Newey WK, West KD. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica. 1987;55(3):703. https://doi.org/10.2307/1913610

37. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13–22. https://doi.org/10.1093/biomet/73.1.13
38. Cameron AC, Gelbach JB, Miller DL. Robust Inference With Multiway Clustering. Journal of Business & Economic Statistics. 2011;29(2):238–49. https://doi.org/10.1198/jbes.2010.07136
39. Thompson SB. Simple formulas for standard errors that cluster by both firm and time. Journal of Financial Economics. 2011;99(1):1–10. https://doi.org/10.1016/j.jfineco.2010.08.016
40. Conley TG. GMM estimation with cross sectional dependence. Journal of Econometrics. 1999;92(1):1–45. https://doi.org/10.1016/s0304-4076(98)00084-0
41. Driscoll JC, Kraay AC. Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data. Review of Economics and Statistics. 1998;80(4):549–60. https://doi.org/10.1162/003465398557825
42. Schötz C. Spatial correlation in economic analysis of climate change. Nature. 2025;644(8076):E27–30. https://doi.org/10.1038/s41586-025-09206-5 PMID: 40804155
43. Hanck C. I just ran two trillion regressions. Economics Bulletin. 2016;36:2037–42.