Underestimation of extreme heat damages in climate financial risk assessment: evidence from bias-corrected climate projections

Olivia Vashti Ayim[1], Matt Goldklang[2,3,4], Jane Smyth[3] & Myles R. Allen[1]

[1]Atmospheric, Oceanic and Planetary Physics, University of Oxford, Clarendon Laboratory, Sherrington Rd, Oxford OX1 3PU, United Kingdom

[2]School of Geography and the Environment, University of Oxford, Oxford, UK

[3]Man Group, London, UK

[4]Columbia University Climate School, New York, USA

Abstract

Standard climate–economy models rely on mean temperature projections, yet near-term economic damages from heat operate primarily through extremes. Using IFS reforecast data as a benchmark, we show that widely used bias-corrected climate products produce sensitivity estimates up to 33% lower for how strongly regional warming amplifies local extreme heat frequency. We estimate a global subnational damage function in which EHD frequency independently reduces GDP growth, and find that this extremes channel is both economically significant and operates in parallel with the conventional mean-temperature effect. At the near-term stress-testing horizon ($\Delta T \approx 0.85\,^{\circ}\mathrm{C}$ by 2040), mean warming is mildly beneficial through the temperature channel, since global mean temperatures remain below the estimated economic optimum. Net climate damage therefore comes entirely from the EHD channel. Consequently, approximating EHD frequency correctly is crucial: projections using IFS-derived sensitivities yield 2040 damages roughly 65% higher than those based on conventional downscaled inputs. These results reveal a structural downward bias in climate damage projections. Accurately characterizing how regional warming amplifies local extreme heat is not a refinement but a prerequisite for scenario analysis, macroprudential stress testing, and climate risk pricing.

Main

Climate change is increasingly recognized as an important factor influencing economic growth and financial stability. Global average temperatures have risen by over 1°C since the preindustrial period, and each year between 2015 and 2024 ranks among the ten warmest ever recorded (1). Extreme heat events have also increased in frequency, intensity, and duration (2; 3; 4), causing direct impacts on productivity, health, agriculture, and infrastructure (5; 6). It is critical to understand how the rising prevalence of extreme heat translates into economic damages. There is evidence that markets do not efficiently price climate risk (7), and more research is needed to clarify and monitor the systemic stress that climate change poses to the economy and the financial system (8; 9).

Central banks and regulators now conduct climate stress tests and scenario analyses to assess the resilience of the financial system to climate change (10; 11). These exercises, along with climate disclosures under the ISSB (12), depend on downscaled climate projections that translate coarse global climate model (GCM) output into high-resolution hazard data. The damage function—the mapping from climate hazard to economic loss—is the critical bridge between these projections and assessments of systemic economic stress. Existing damage functions relate mean temperature to GDP loss (13; 14; 15), with recent extensions incorporating

temperature variability and precipitation (16; 17). However, the most recent NGFS damage function (16) has been withdrawn from *Nature* pending revision, with an updated methodology not expected until late 2026 (10). Meanwhile, Bilal and Känzig (18) estimate that 1°C of warming reduces world GDP by over 20%—approximately six times larger than prior estimates—arguing that global temperature correlates more strongly with extreme climatic events than local temperature used in previous work. These developments expose both the fragility of current damage functions and the underrepresentation of the extremes channel in assessing climate-driven economic stress.

Despite the relevance for pricing climate risk and stress testing, there is limited research on how the *climate data pipeline* itself affects damage estimates. Bias-corrected products such as the Global Downscaled Projections for Climate Impacts Research (GDPCIR) (19) apply Quantile Delta Mapping (QDM) (20) to preserve quantile-specific trends while mapping GCM distributions to ERA5 reanalysis (21). Lafferty et al. (22) showed that downscaling and bias-correction contribute substantial uncertainty to local climate projections, but the implications for projected economic damages have not been tested. Developing a fuller understanding of how the data pipeline affects projected damages would enable more accurate pricing of climate risk and more robust stress testing of economic resilience.

We present a new empirical approach to examining how extreme heat exposure affects economic output and how the choice of climate data source biases the result. First, we analyze the relationship between regional mean temperatures and the frequency of exceeding the local extreme temperature threshold (maximum temperature above the 95th percentile) using IFS reforecast data globally. This provides a physically grounded alternative estimate of how the risk of extreme heat days responds to mean warming. We then compare this IFS-derived relationship to the equivalent relationship in the QDM-downscaled GDPCIR product and show that the downscaled slopes—the change in extreme heat day frequency per degree of warming—are systematically flatter. Second, we estimate a panel damage function at the admin2 level (second-order administrative units such as counties, districts, and prefectures) in which EHD frequency enters linearly alongside mean temperature controls, demonstrating that the extremes channel operates independently of and amplifies the conventional mean-temperature effect, and that at near-term warming horizons (to 2040) it accounts for virtually all projected GDP damage. Third, we project damages to 2040 under SSP2-4.5 via two pathways—one using the downscaled slopes to derive EHD frequencies from warming, the other using the IFS slopes—and quantify the resulting gap in damage estimates. This work conveys the sensitivity of climate damage projections to the underlying data pipeline and can inform more accurate stress testing and pricing of climate risk.

Estimating Heat Sensitivity from Reforecast Data

This study takes a novel approach, using ensemble reforecast data from the ECMWF Integrated Forecasting System (IFS) to quantify the sensitivity of extreme-heat day frequency to regional mean warming. A core challenge in studying rare temperature extremes is that the observational record, limited to a single realization of past weather, rarely provides enough independent samples to estimate changes in the tails of the temperature distribution with confidence (23; 24). The rarity of such events means that even multi-decadal datasets can misrepresent long-term trends when internal variability dominates, and short records have been shown to systematically underestimate return levels of high-impact heat events (23; 25; 26). Climate models have also been shown to underestimate positive trends in the upper tail of extreme surface temperature distributions by a factor of four compared to reanalysis data across global land areas (27), further motivating the use of a reforecast-based approach. The IFS ensemble addresses this directly by producing multiple physically

consistent simulations of past atmospheric states, substantially expanding the sample available for analyzing rare exceedance events at the global scale (28; 29). Although the reforecast dataset was originally designed for model calibration and performance evaluation, it is repurposed here to estimate how the probability of exceeding local climatological temperature thresholds responds to regional warming across all land areas. This provides a physically grounded alternative to relying on bias-corrected climate model output, whose limitations in representing tail behavior are precisely what this study seeks to quantify.

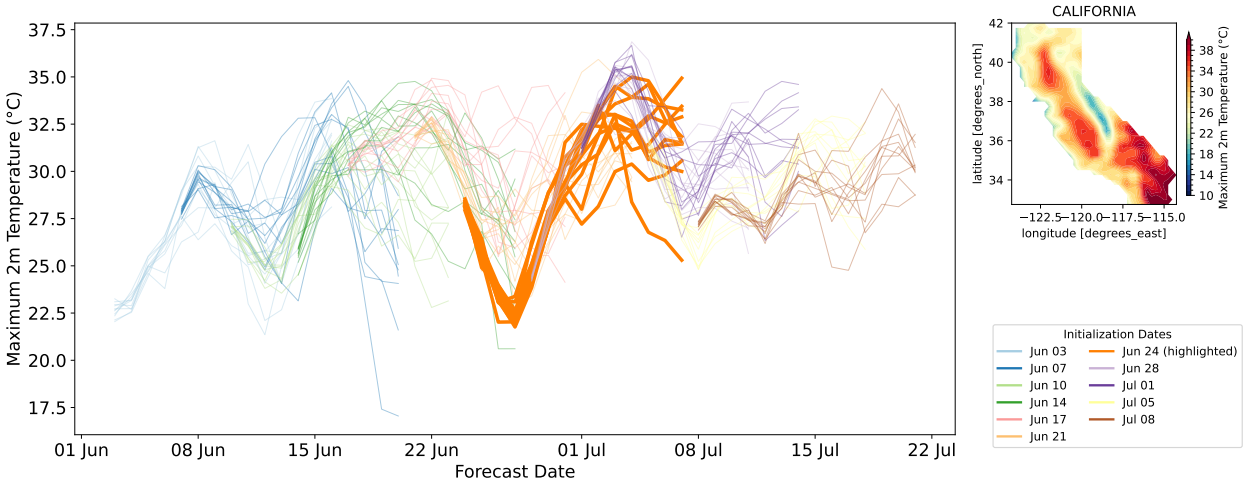


Figure 1: Evolution of daily maximum 2m temperature (°C) in the IFS reforecast dataset for the hindcast year 2001, showing ensemble members across different initialization dates over California (inset). Each line represents a forecast trajectory from lead day 0 to lead day 12. At lead day 0, all ensemble members begin from identical initial conditions and therefore show no spread. As lead time increases, the spread widens progressively, reflecting reduced predictability from initial conditions and greater dependence on large-scale boundary forcing and internal model dynamics. The highlighted time series (orange) corresponds to the forecast initialized on 24 June 2001, and is used to illustrate the ensemble characteristics at the lead time adopted in this study.

At each grid point, we estimate how the probability of exceeding the local climatological 95th percentile temperature threshold changes with regional warming, averaged over a spatial scale of 200 km. Since the increased frequency of hot extremes is largely explained by a shift in the surface temperature distribution driven by mean warming (30; 31; 32), the key question is what temperature predictor best captures this shift. Regional temperature anomalies are preferred over local ones because they are governed by large-scale atmospheric and oceanic processes that are both more stable over time and more consistently reproduced across climate models (33; 34; 35). Local temperatures, by contrast, are sensitive to microscale surface conditions and microclimatic factors that introduce noise and reduce model agreement. Regional changes also scale approximately linearly with global mean warming (33; 34), making them a reliable basis for translating GCM warming signals into local exceedance probabilities.

Linking local temperature extremes to regional warming in this way establishes a physically grounded framework for estimating how extreme heat risk evolves with continued warming. To assess whether the IFS reforecast reliably captures this relationship, Figure 2 compares the estimated slopes against those derived directly from ERA5 observations. Both datasets show broadly consistent spatial patterns, with elevated sensitivity concentrated in tropical and subtropical regions. Differences between the two are moderate at the 200 km smoothed scale (Figure 2, panels e,f), supporting the use of the IFS reforecast as a reliable source of

extreme heat sensitivity estimates.

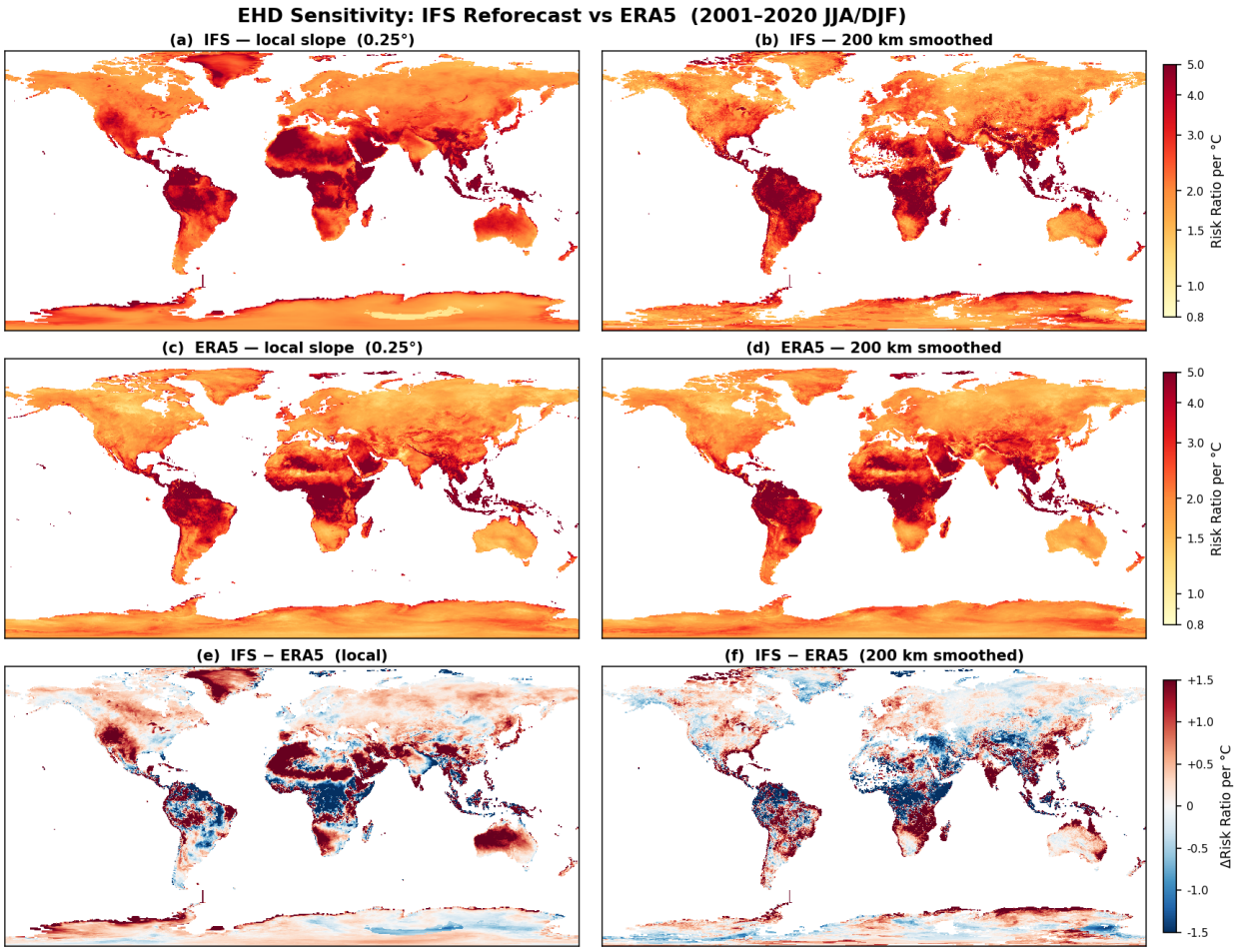


Figure 2: Global comparison of EHD sensitivity slopes (risk ratio per °C warming, log scale) for 2001–2020 (JJA/DJF). Panels (a, b): IFS ECMWF reforecast slopes at local (0.25°) and 200 km smoothed resolutions. Panels (c, d): ERA5 observed slopes at local (0.25°) and 200 km smoothed resolutions. Panels (e, f): Difference (IFS minus ERA5) at local and 200 km smoothed scales. Both datasets show broadly consistent spatial patterns of higher sensitivity in tropical and subtropical land areas; differences are moderate, validating the IFS reforecast as a reliable source of extreme heat sensitivity estimates.

Divergence Between IFS and Downscaled Slopes

We compare the IFS-derived slopes to slopes estimated from the GDPCIR dataset (19), which applies QDM (20) for bias adjustment to ERA5 and Quantile-Preserving Localised-Analog Downscaling (QPLAD) for spatial disaggregation at 0.25° resolution.

The GDPCIR-derived slopes are systematically flatter than the IFS-derived slopes (Figure 2). That is, for a given increase in mean tasmax, the downscaled product implies a smaller increase in the risk of extreme heat days than the IFS relationship. The sensitivity of EHD frequency to warming is underestimated, not the baseline frequency itself. This divergence arises because QDM anchors the bias-corrected distribution to ERA5. QDM faithfully preserves the GCM's quantile-specific *trend*—the relative change at each quantile between historical and future periods—but the absolute distribution reflects ERA5. ERA5 is produced

at approximately 31 km resolution with parameterized convection and boundary layer physics (21), which smooth local temperature extremes, cap peak temperatures in convective environments, and omit urban heat island effects (36; 37; 38). The result is a reference distribution whose tails are too thin: the warm-side tail is truncated, so a given mean shift produces a smaller increase in threshold exceedances than in the IFS, where the tail structure is resolved by the model's own physics. The GCM trend is preserved, but it is applied to a distribution that is less sensitive to warming in the extremes.

To quantify this divergence, we compute land-masked global mean slopes for the IFS (local and 200 km smoothed) and for all 15 available QDM CMIP6 models, applying a one-sample t-test with df = 14 (Table 1). Ocean cells are excluded using the IFS ERA5-Land footprint.

Table 1: Global one-tailed t-test: IFS vs. QDM CMIP6 ensemble (15 models, df = 14). Land-masked global means only; ocean cells excluded via IFS ERA5-Land footprint.

Scale	IFS mean	GCM μ	GCM σ	t	p	
Local (0.25°)	2.830	2.132	0.069	−38.96	< 0.0001	***
200 km smoothed	2.808	2.320	0.096	−19.79	< 0.0001	***

RR per °C. *** $p < 0.001$.

On land, IFS slopes are significantly *higher* than CIL CMIP6 slopes at both spatial scales ($p < 0.0001$). The IFS local mean (2.83 RR/°C) exceeds the GCM land mean (2.13) by 33%, and the 200 km means differ by 21% (2.81 vs. 2.32). Supplementary Figure S2 illustrates this divergence at the city level, where the IFS (blue diamond) and GCM ensemble (coloured circles) are plotted on a common local-vs-200 km axis for nine major cities.

This finding is not specific to GDPCIR or QDM. Any bias-correction method that maps to a reanalysis inherits the reanalysis tail biases. ISIMIP3BASD (39) uses W5E5—itself derived from ERA5—presenting the same fundamental problem. Alternative downscaling methods such as BCSD (40) face analogous limitations. The implication is structural: the pipeline from GCM to high-resolution hazard data contains a systematic downward bias in the sensitivity of extreme heat risk to warming, with direct consequences for any damage projection or stress test built on these products.

Extreme Heat Days as a Damage Driver

We estimate an admin2-level panel model linking economic output to extreme heat day frequency:

$$D_{i,t} = \beta \cdot R_{95,i,t} + \mathbf{X}'_{i,t}\boldsymbol{\delta} + \mu_i + \tau_t + \epsilon_{i,t} \tag{1}$$

where $D_{i,t}$ is GDP growth per capita in admin2 region i and year t, $R_{95,i,t}$ is the annual extreme heat day frequency (settlement-weighted, ERA5), $\mathbf{X}_{i,t}$ contains annual mean temperature and total precipitation (ERA5 0.25°), μ_i and τ_t are region and year fixed effects, and standard errors use the Conley (41) spatial HAC estimator with a 500 km distance cutoff. Country-specific quadratic time trends are removed prior to two-way demeaning via the Frisch–Waugh–Lovell theorem.

Much of the existing literature focuses on aggregate economic impacts of temperature, modeling damages as a function of annual mean temperature and its square (13; 14). To assess whether EHD frequency captures a distinct damage channel, we estimate a nested specification:

$$D_{i,t} = \beta_1 \bar{T}_{i,t} + \beta_2 \bar{T}^2_{i,t} + \beta_3 R_{95,i,t} + \mathbf{X}'_{i,t}\boldsymbol{\delta} + \mu_i + \tau_t + \epsilon_{i,t} \tag{2}$$

Estimated on 1,531,392 observations spanning 47,856 admin2 entities across 237 countries (1991–2022) using the Kummu et al. (42) downscaled subnational GDP per capita dataset, the baseline coefficient is:

$$\hat{\beta} = -0.0280 \quad (SE_{\text{Conley 500km}} = 0.0085,\ t = -3.30,\ p = 0.001) \tag{3}$$

A one percentage point increase in the share of summer days exceeding the local 95th-percentile threshold is associated with a 0.028 percentage point reduction in same-year GDP growth per capita. In the fully nested specification (EHD, $\bar{T}$, $\bar{T}^2$, $\bar{P}$), annual mean temperature remains positive and highly significant ($\hat{\beta}_T = +0.0080$, $t = +4.58$, $p < 0.001$), with the quadratic term negative and significant ($\hat{\beta}_{T^2} = -2.0 \times 10^{-4}$, $t = -3.98$, $p < 0.001$), implying an optimal temperature of $T^* \approx 20\,^{\circ}\text{C}$. Precipitation enters negatively but is not statistically significant ($t = -0.36$). These results indicate that the extremes channel (EHD frequency) and the mean-temperature channel operate independently and in parallel. Critically, at the near-term warming horizon relevant to stress testing ($\Delta T \approx 0.85\,^{\circ}\text{C}$ by 2040), the EHD channel accounts for **more than 100%** of projected global GDP damages: because the GDP-weighted mean baseline temperature ($T_0 = 17.4\,^{\circ}\text{C}$) lies below the estimated optimum ($T^* \approx 20\,^{\circ}\text{C}$), the mean-temperature channel provides a slight offsetting benefit (+0.06 pp/yr), so the EHD channel alone (−0.34 pp/yr, IFS local) exceeds the total combined damage (−0.27 pp/yr). This decomposition is a back-of-the-envelope illustration that applies the GDP-weighted panel mean T_0 uniformly; a rigorous assessment would aggregate entity-level EHD and temperature channel contributions using time-varying GDP weights and would require combining the climate drag with an explicit baseline GDP growth scenario. The T channel only becomes a net drag globally once warming exceeds approximately 4.4 °C above the 2020 baseline—beyond any near-term stress-testing horizon (Supplementary Figure S3). The non-linearity in climate damages therefore arises primarily from the non-linear relationship between mean warming and EHD frequency, consistent with Bilal and Känzig (18) and the standard quadratic specification of Burke et al. (13). The linearity of the EHD effect implies that the economic damage per additional extreme heat day is approximately constant across the EHD distribution.

To confirm that the estimate is not driven by any particular set of admin2 units, we implement a bootstrap in which 10% of entities are dropped at random across 1,000 replications (Supplementary Table S1). The bootstrap distribution is tightly centered on the full-sample estimate ($\bar{\beta}_{\text{boot}} = -0.0280$, sd = 0.0006), and the coefficient is statistically significant at the 5% level in all 1,000 replications, confirming robustness to any particular subset of admin2 units.

Supplementary Table S2 presents a comprehensive set of robustness checks. *Panel A* shows that the estimate is significant across all seven standard error specifications—from entity-clustered ($t = -16.6$) to the most conservative Conley 1000 km ($t = -2.7$, $p = 0.008$)—confirming that the inference is not an artifact of any particular approach to spatial or serial dependence. *Panel B* shows that the coefficient is stable when simpler $T + P$ controls are used instead ($\hat{\beta} = -0.032$, significant at $p < 0.001$). A natural cubic spline in EHD frequency (4 degrees of freedom) produces a joint Wald statistic of 127.0 ($p < 0.001$), and the quintile Q5 estimate (Q5 vs Q1: $\hat{\beta} = -0.0033$, $p < 0.05$) confirms that damage is concentrated in the highest extreme heat exposure group. *Panel C* shows that the lead coefficient is negligible and insignificant ($\hat{\beta} = -0.0075$, $p = 0.32$), ruling out reverse causality or anticipation effects. *Panel D* shows negative EHD effects across all income groups—low- (−0.046) and middle-income (−0.031) regions are significant at $p < 0.01$; high-income regions show a negative but marginally significant effect (−0.018, $p = 0.07$)—consistent with better adaptive capacity at higher income levels. *Panel E* shows stability across spatial aggregation levels: the effect remains negative and significant when data are aggregated to admin1 boundaries (first-order administrative units such as states and provinces; −0.048), consistent with measurement-error attenuation at finer aggregation.

Damage Underestimation to 2040

We apply the damage function (Eq. 1) under three pathways:

Pathway A (QDM): GCM mean tasmax anomaly under SSP2-4.5 to 2040, converted to EHD frequency via the GDPCIR-derived slope → damage function.

Pathway B (IFS): GCM mean tasmax anomaly under SSP2-4.5 to 2040, converted to EHD frequency via the IFS-derived slope → damage function.

Both pathways use the same GCM warming trajectories and the same $\hat{\beta}$; the comparison isolates the effect of the warming–extremes slope on damage estimates. Results are reported for the 2035–2040 window under SSP2-4.5, using ECS-weighted CMIP6 trajectories ($\Delta T_{\text{global}} \approx +0.85\,^{\circ}\text{C}$ above 2020 in that window). This horizon aligns with typical stress-test windows while minimizing stationarity concerns: the IFS- and ERA5-derived slopes are calibrated on 2001–2020 data and their validity diminishes as warming diverges further from the historical record. These projections report the *incremental* GDP growth drag attributable to the climate signal, holding all else constant; they are a stylized illustration of the damage gap and should not be interpreted as GDP growth forecasts. Translating the climate drag into projected GDP growth rates would require combining these estimates with a baseline growth scenario — for example from IMF or World Bank projections — and we flag this as a priority for future work.

IFS-based methods yield a GDP-weighted $\Delta\overline{\text{EHD}}$ roughly 65% larger than QDM-based methods (0.12–0.13 vs. 0.07–0.08), translating into projected annual growth drags of -0.34 to -0.36 pp (IFS) versus -0.20 to -0.23 pp (CIL). The gap—approximately 0.13 pp per year—is already material within the 2025–2040 window directly relevant to macroprudential stress testing. Total uncertainty is wider for IFS-based methods ($\pm$0.30–0.40 pp, 1 SE) than for QDM-based methods ($\pm$0.09–0.13 pp), reflecting large cross-model variation in projected warming; even at the lower bound of the IFS confidence interval the damage exceeds the CIL central estimate. Critically, the bias compounds with warming: as mean temperatures rise further from the historical baseline, the flatter downscaled slope diverges further from the IFS slope, amplifying the damage gap over time even within our conservative projection horizon. Figure 3 illustrates the annual growth drag and compounded GDP loss under each pathway.

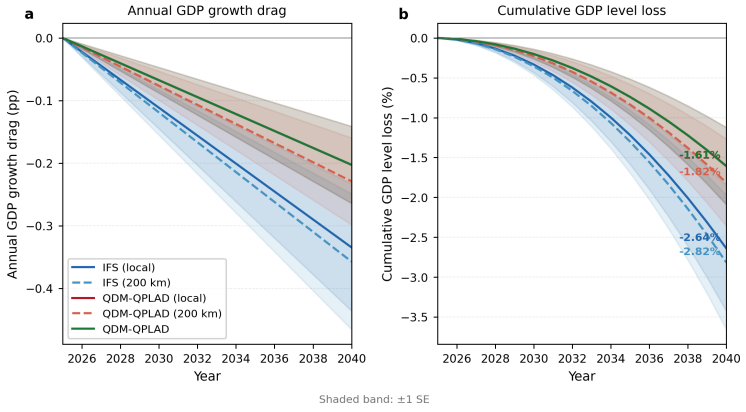


Figure 3: Left: annual GDP growth drag (pp) by projection method. Right: compounded GDP-level loss relative to a no-warming counterfactual. Shaded bands show $\pm$1 SE combining regression uncertainty (Conley 500 km) and GCM ensemble spread in quadrature. IFS-based pathways consistently project larger losses than QDM-based pathways.

Table 2: Projected annual GDP growth impact by 2040 (SSP2-4.5, $\Delta T_{\text{global}} \approx +0.85\,^{\circ}\text{C}$ above 2020). Reg. SE is parametric uncertainty from the damage-function coefficient (Conley 500 km); GCM SE is the ECS-weighted cross-model standard deviation of the projected impact across the CMIP6 ensemble.

Method	$\Delta\overline{\text{EHD}}$	Impact (pp)	Reg. SE (pp)	GCM SE (pp)
EHD channel only ($\hat{\beta}_{EHD} \times \Delta\overline{EHD}$)				
IFS local (0.25°)	0.120	−0.335	0.102	0.23
IFS 200 km smoothed	0.128	−0.358	0.109	0.25
QDM CMIP6 local	0.072	−0.201	0.061	0.07
QDM CMIP6 200 km smoothed	0.082	−0.229	0.070	0.04
GDPCIR QDM-QPLAD (direct)	0.073	−0.204	0.062	0.09
EHD + temperature channels ($\hat{\beta}_{EHD} \times \Delta\overline{EHD} + \hat{\beta}_T \Delta T + \hat{\beta}_{T^2}(\Delta T)^2$, $\Delta T = +0.85\,^{\circ}C$ at $T_0 = 17.4\,^{\circ}C$)				
IFS local (0.25°)	0.120	−0.274	0.238	0.23 §
IFS 200 km smoothed	0.128	−0.296	0.241	0.25 §
QDM CMIP6 local	0.072	−0.140	0.223	0.07 §
QDM CMIP6 200 km smoothed	0.082	−0.168	0.226	0.04 §
GDPCIR QDM-QPLAD (direct)	0.073	−0.143	0.224	0.09 §
Literature benchmarks under the same SSP2-4.5 trajectory				
Burke, Hsiang & Miguel (2015) (13)	—	+0.08 †	0.02	0.03
Kotz et al. (2024) (16)	—	−0.31	0.16	0.14
Bilal & Känzig (2024) (18)	—	−0.23 ‡	0.06	0.11

Reg. SE: propagated as $|\Delta\overline{\text{EHD}}| \times \text{SE}_{\hat{\beta}}$ (Conley 500 km) for our EHD-only rows; includes additional parametric uncertainty from $\hat{\beta}_T$ and $\hat{\beta}_{T^2}$ for EHD + temperature rows (≈0.22 pp), which is higher because $T_0 < T^* \approx 20\,^{\circ}\text{C}$ so the temperature channel partially offsets EHD damages; propagated from published confidence intervals for the literature. *GCM SE*: ECS-weighted standard deviation of projected impact across 15 CMIP6 models (the same ensemble used in the slope t-test); for literature studies, computed as $|\partial\text{damage}/\partial(\Delta T)| \times \text{SD}(\Delta T_{\text{GCM}})$ using each study's central damage function and the same ensemble ($\text{SD}(\Delta T) = 0.38\,^{\circ}\text{C}$ in the 2035–2040 window). §EHD + temperature rows report GCM SE from the EHD channel only; the temperature channel adds $\lesssim$0.03 pp in quadrature and partially cancels with the EHD channel. †GDP-weighted baseline temperature (12.0 °C) falls below the BHM optimum ($T^* = 12.9\,^{\circ}\text{C}$); warming is marginally beneficial. ‡BK 2024 estimates a GDP-*level* effect (−5.6% per °C); −0.23 pp/yr is the annualised rate of the implied cumulative loss at $\Delta T \approx 0.85\,^{\circ}\text{C}$.

Our EHD-based damage estimates under the IFS pathway (−0.34 to −0.36 pp/yr by 2040) are consistent with mid-range estimates from mean-temperature specifications (13; 17), and broadly comparable to literature benchmarks evaluated under the same SSP2-4.5 trajectory (Table 2), though direct comparison requires care given differences in damage function structure and aggregation. Mean-based approaches miss the extreme heat channel entirely and do not isolate whether damage operates via extremes or mean temperature. Kotz et al. (16) incorporate temperature variability but do not isolate EHD frequency. Bilal and Känzig (18) capture the extremes channel indirectly through global temperature; our approach identifies the same mechanism at higher spatial resolution through a directly observable metric. Importantly, our results show that even using moderate damage function coefficients, the *climate input data* biases the result downward—damages are underestimated regardless of which damage function is used, so long as it depends on extreme heat frequency derived from bias-corrected projections.

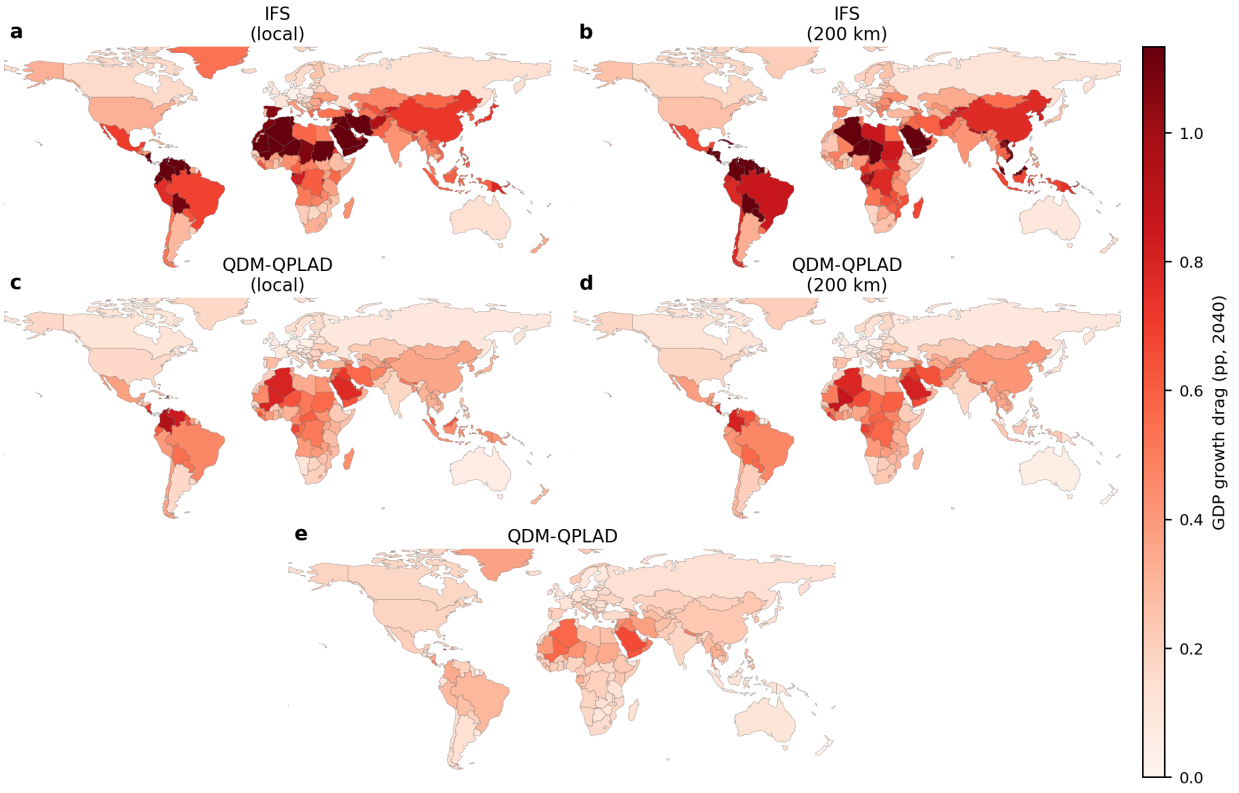


Figure 4: Country-level annual GDP growth drag (pp) in the 2035–2040 window under SSP2-4.5 ($\Delta T_{\text{global}} \approx +0.85\,^{\circ}\text{C}$ above 2020) for all five ΔEHD methods. Tropical and subtropical regions face the largest projected losses across all methods; IFS-based methods (top two panels) show materially larger damages than QDM-based methods (bottom three panels) in South Asia, sub-Saharan Africa, and Central America.

Discussion

This study reveals that bias-corrected climate projections systematically underestimate the sensitivity of extreme heat day frequency to mean warming, with material consequences for economic damage estimates. The IFS-derived relationship between warming and EHD frequency is steeper than the equivalent relationship in the QDM-downscaled product, because the downscaling pipeline inherits ERA5's underrepresentation of the warm-side tail of the temperature distribution. This is a bias in the *slope*, how rapidly extreme heat risk increases with warming, rather than in the historical baseline frequency. When this flatter slope is used to project EHD frequencies forward and the result is propagated through our damage function, the consequence is a downward bias of approximately 0.15 pp per year in projected GDP growth drag that is already significant under moderate warming at a 2040 horizon.

The finding that EHD frequency enters the damage function linearly, alongside a statistically significant quadratic mean-temperature effect ($T^{*} \approx 20\,^{\circ}\text{C}$), has broader implications for how climate-driven economic stress is modeled. Rather than displacing the mean-temperature channel, EHD frequency identifies an independent extremes channel that operates even after controlling for the full quadratic temperature response. This suggests that the standard quadratic temperature specification (13; 14) captures only part of the climate damage story, and that directly modeling the extremes channel adds economically significant explanatory power beyond mean temperature alone. This is consistent with Bilal and Känzig (18), who

find larger damages when using global temperature, which may correlate with extreme events, rather than local temperature. Our results indicate that local EHD frequency provides a directly observable, spatially resolved mechanism for capturing this extremes component, identifying it at the admin2 level rather than through global temperature variation. Local EHD frequency increases disproportionately as the temperature distribution shifts rightward (31; 43), and regions with short warm-side tails exhibit amplified increases in threshold exceedances—precisely the regions where the downscaled product most severely underestimates the warming–extremes slope.

The non-linearity in climate damages, often attributed to the quadratic temperature term in conventional specifications, may in fact arise from the non-linear relationship between mean warming and EHD frequency rather than from the economic response to temperature itself. Our linear EHD specification accommodates this: the damage per extreme heat day is constant, but the number of extreme heat days accelerates non-linearly with warming (4). This decomposition has practical value for stress testing, because it separates the climate science question (how does warming translate into extremes?) from the economic question (what is the damage per extreme heat day?), allowing each to be assessed and updated independently.

For the economy and financial system, the implications are direct. Climate stress tests conducted by central banks under NGFS scenarios (10; 11) use downscaled climate projections to estimate GDP losses, sectoral disruptions, and counterparty risks under adverse warming pathways. If the climate data underpinning these exercises underestimates how rapidly extreme heat risk increases with warming, the resulting stress scenarios are insufficiently severe. This matters not only for macroprudential oversight but for any application that prices climate risk, from sovereign credit assessment to infrastructure investment to agricultural lending. The damage gap we identify is not a tail scenario; it is present under moderate warming (SSP2-4.5) within a 15-year horizon, well within the planning horizons of governments, insurers, and long-duration asset owners.

Battiston et al. (44) showed that neglecting tail acute risks leads to underestimation of investor losses by up to 82%. Our findings provide a complementary mechanism: even when extreme heat is explicitly modelled, the underlying climate data understates how much extreme heat frequency *changes* as the climate warms. The combination of these two effects, missing tail risks entirely, and underestimating the rate of change when they are included, suggests that current assessments of climate-driven economic stress are conservative on multiple fronts.

The withdrawal of Kotz et al. (16) has left the NGFS without a peer-reviewed damage function for its long-term scenarios. Our EHD-based damage function offers a complementary approach: simpler than the multi-variable specification of Kotz et al., more granular than the global temperature approach of Bilal and Känzig (18), and directly connected to an observable climate metric whose projection biases we quantify. By linking climate science (the warming–extremes slope) to economics (the damage per EHD) through a transparent and decomposable framework, this approach can inform the next generation of climate scenario analysis and stress testing.

By demonstrating the sensitivity of damage estimates to the slope of the warming–extremes relationship in the climate data pipeline, this study motivates further research to explore how reanalysis tail biases propagate through other hazard types and damage channels. As climate change continues to intensify, understanding how the data pipeline shapes the projected *rate of change* in climate risk is crucial for accurate pricing of climate-exposed assets, robust macroprudential assessment, and sound long-term economic planning. This study highlights that better representation of the sensitivity of tail climate events to warming, here through the use of reforecast data, better informs the economic implications of continued warming.

Methods

Estimating extreme heat sensitivity

Reforecast configuration

The ECMWF reforecast dataset used in this study is based on IFS model cycle CY47R2, comprising an 11-member ensemble initialized twice weekly at 00:00 UTC and covering the period 2001 to 2020. Each forecast extends to a 14-day lead time at a horizontal resolution of $0.25° \times 0.25°$. The atmospheric component runs at TCo639 resolution with 137 vertical levels and is coupled to NEMO v3.4, which uses 75 vertical levels and a horizontal resolution of 0.25°. Forecasts are extracted at day 12 of lead time, the point at which ensemble variance approaches observed climatology (Supplementary Figure S1) while retaining physically consistent circulation patterns (28). Across 25 initialization dates per year and 11 ensemble members, this yields 275 data points annually and 5,500 over the 20-year record, providing a broad statistical sample for extreme value analysis.

Threshold definition

For each grid point, a climatological 95th-percentile temperature threshold is computed from all available daily maximum 2m temperature values pooled across years. The threshold is defined separately for the local summer season, taken as June to August (JJA) in the Northern Hemisphere and December to February (DJF) in the Southern Hemisphere. The 95th percentile was selected to isolate the upper tail of the temperature distribution while maintaining a sufficient sample size for robust statistical estimation.

Exceedance probability and regression

For each year, the exceedance probability is defined as the fraction of summer days on which daily maximum temperature exceeds the fixed climatological threshold. The relationship between exceedance probability and temperature is analyzed in logarithmic probability space, which linearizes tail behavior and enables consistent comparison across grid points with differing variances. The log of the yearly exceedance probability is regressed onto the corresponding annual value of the Regional temperature indicator at each grid point:

$$\beta = \frac{d \log P(T_{95})}{d(T_{\text{reg}})} \tag{4}$$

where $P(T_{95})$ is the annual probability of exceeding the local climatological 95th-percentile threshold and T_{reg} is the regional temperature anomaly. The slope coefficient β is exponentiated to express the relationship as the multiplicative change in exceedance probability per 1°C of regional warming, referred to throughout as the risk ratio slope.

Regional temperature and Gaussian filtering

Regional temperature indicators are constructed by applying a Gaussian filter to the daily maximum temperature field, defined in physical distance with a smoothing radius of 200 km. The filter is applied separately in the latitude and longitude directions, with the longitude smoothing scale adjusted by $\cos(\phi)$, where ϕ is latitude, to maintain a consistent physical distance at all grid points. Although the Gaussian weighting formally extends beyond the chosen radius, contributions from distances greater than approximately three standard deviations are negligible, so the effective influence is limited to the selected scale (35). The 200 km scale

is selected because it captures large-scale atmospheric and oceanic forcing that drives regional temperature variability while filtering out microscale surface effects that reduce model agreement across datasets (33; 34).

Uncertainty estimation

Uncertainty in the estimated slopes is assessed through bootstrap resampling with 1,000 iterations. In each iteration, data are sampled with replacement and a linear regression is fitted between the regional temperature anomaly and the logarithm of exceedance probability. The resulting distribution of slope coefficients is used to derive 95% confidence intervals. Identical procedures are applied across the IFS reforecast, ERA5, and GDPCIR datasets to ensure that differences in estimated sensitivities reflect intrinsic dataset characteristics rather than methodological inconsistencies.

Downscaled climate data

We use the GDPCIR dataset (19), which provides global daily tasmax at 0.25° resolution for 25 CMIP6 GCMs (45) under SSP2-4.5 (46) over 2015–2040. GDPCIR applies QDM (20) for bias adjustment to ERA5 (21) and QPLAD for spatial downscaling. We extract daily tasmax and compute EHD frequencies at the admin2 level. To estimate the QDM-derived EHD sensitivity slopes for comparison with the IFS, we use the full GCM historical runs extending back to 1950, regressing log annual EHD frequency against summer-mean regional tasmax anomalies across the entire available historical record. Using the full historical run maximizes the sample of interannual variability used to estimate the slope and ensures the QDM comparison is not period-specific. The global one-tailed t-test (Table 1) uses the 15 models for which both slope files and EHD projection data are available, with land masking applied via the IFS ERA5-Land footprint to exclude ocean cells that inflate QDM global means.

Panel damage function

Panel structure. We estimate Eqs. 1 and 2 on an unbalanced admin2-level panel spanning 237 countries over 1991–2022. After merging the climate and GDP growth data and dropping observations with missing controls, the estimation sample contains 1,531,392 observations across 47,856 admin2 entities. The panel is unbalanced: coverage ranges from 1 to 32 years per entity depending on GDP data availability. All specifications include entity fixed effects μ_i (absorbing time-invariant heterogeneity such as geography, institutional quality, and long-run income level) and year fixed effects τ_t (absorbing global shocks including world business cycles, commodity price movements, and aggregate climate trends). To control for country-specific long-run growth trends that could confound the climate signal, we additionally include country-specific quadratic time trends, estimated and removed prior to two-way demeaning via the Frisch–Waugh–Lovell (FWL) theorem (47). The within-transformation (two-way demeaning) is applied iteratively to convergence. Full panel descriptive statistics are provided in Supplementary Table S3.

Dependent variable. The dependent variable is the log first difference of GDP per capita at the admin2 level (annualised growth rate). GDP per capita data at admin2 resolution are drawn from the Kummu et al. subnational GDP database (42), extended where possible with national accounts data disaggregated using population-weighted interpolation. Observations with implausible growth rates ($|D_{i,t}| > 50\%$) are excluded.

EHD frequency. Extreme heat day frequency $R_{95,i,t}$ is constructed from ERA5 daily maximum 2 m temperature at 0.25° resolution. For each calendar day and grid cell, a climatological 95th-percentile threshold is computed over the 1990–2020 reference period using a ±7-day rolling window to smooth the seasonal cycle,

yielding a day-of-year-specific threshold at each location. EHD frequency in year t is the fraction of summer days (JJA in the Northern Hemisphere, DJF in the Southern Hemisphere) on which the threshold is exceeded. Grid-cell values are aggregated to admin2 polygons using settlement-area weights from the Global Human Settlement Layer (GHS-BUILT-S, R2023A) (48), time-varying across six GHS epochs (1975, 1980, 1990, 2000, 2005, 2010, 2015, 2020) mapped annually. This settlement weighting ensures that EHD exposure reflects where people actually live rather than administrative-area averages driven by uninhabited terrain.

Climate controls. Annual mean 2 m temperature (°C) and annual total precipitation (mm) are derived from ERA5 reanalysis at 0.25° resolution and aggregated to admin2 polygons via area-weighted mean. These controls absorb the slow-moving component of the climate signal that may be correlated with EHD frequency, ensuring that $\hat{\beta}$ captures variation in extreme heat frequency conditional on mean temperature and precipitation levels. Their inclusion does not qualitatively affect the EHD coefficient but reduces residual variance.

Standard error estimation. The baseline standard error is a Conley (41) spatial HAC estimator with a Bartlett kernel and a 500 km great-circle distance cutoff. The implementation computes entity-level score vectors $S_i = \sum_t \tilde{X}_{it}\hat{e}_{it}$ (where tildes denote within-transformed variables and $\hat{e}$ are OLS residuals from the demeaned regression), then forms the sandwich estimator $\hat{V} = (X'X)^{-1}\big(\sum_{i,j} w_{ij} S_i S_j'\big)(X'X)^{-1}$ where $w_{ij} = \max(0,\, 1 - d_{ij}/500)$ is the Bartlett weight for entity pair (i, j) with great-circle distance d_{ij}. Entity centroids are the representative points of the admin2 polygons. The Conley estimator accounts for spatial autocorrelation in both the climate regressors and the GDP growth residuals; entity and time fixed effects are fully absorbed prior to its application.

Robustness checks

Standard error robustness. We report the EHD coefficient under seven standard error specifications (Panel A of Supplementary Table S2): (i) entity-clustered, (ii) time-clustered (by year), (iii) two-way clustered (entity and year), (iv) Driscoll–Kraay (49) with a 5-year bandwidth (robust to both cross-sectional and serial dependence), and Conley spatial HAC at (v) 250 km, (vi) 500 km (baseline), and (vii) 1000 km cutoffs. The point estimate is identical across all seven since it derives from the same OLS regression; only the variance estimator changes. Significance is maintained at the 5% level under all seven approaches, including the most conservative Conley 1000 km specification.

Functional form. We assess sensitivity to three aspects of functional form (Panel B). First, we report the simpler $T + P$ controls specification as a robustness check against the baseline $T + T^2 + P$ specification; the EHD coefficient is stable ($\hat{\beta} = -0.032$, $p < 0.001$). Second, we replace the linear EHD term with a natural cubic spline (4 degrees of freedom with knots at the 25th, 50th, and 75th percentiles of the EHD distribution); the joint Wald statistic is 127.0 ($p < 0.001$), confirming that EHD frequency is jointly significant. We also estimate a quintile bin specification in which EHD frequency is grouped into five quintiles (Q1 as reference); the Q5 coefficient ($\hat{\beta} = -0.003$, $p < 0.05$) confirms that economic damage is concentrated in the highest extreme heat exposure group, with intermediate quintiles near zero.

Lag and lead structure. We estimate the contemporaneous effect ($t \to t$) and a one-period lead ($t \to t + 1$, Panel C). The lead coefficient is negligible ($\hat{\beta} = -0.007$, $p = 0.32$), ruling out anticipation effects or reverse causality whereby anticipated future GDP shocks drive current EHD measurement. The contemporaneous dominance of the effect is consistent with a supply-side mechanism (labor productivity, agricultural output) that operates within the calendar year.

Income heterogeneity. We partition the sample into World Bank income groups (low, lower-middle and

upper-middle, high) and estimate separately (Panel D). Low- and middle-income groups show negative and statistically significant EHD effects. The gradient from low (-0.046) to middle income (-0.031) is consistent with better adaptive capacity and greater service-sector shares in higher-income economies reducing heat sensitivity at the margin. High-income regions also show a negative effect (-0.018) that is marginally significant ($p = 0.07$), though differences across income groups are not statistically significant from one another at conventional levels.

Spatial aggregation. The baseline admin2 panel (47,856 entities, 1,531,392 observations) is compared against an alternative spatial definition (Panel E). The *admin1* panel ($N = 10{,}994$ entities, 342,700 observations) aggregates all entities to admin1 boundaries. The EHD coefficient remains negative and significant at this coarser level (-0.048), consistent with measurement-error attenuation at finer aggregation levels.

Entity-dropout bootstrap. To assess sensitivity to any particular set of admin2 units, we implement a block bootstrap that drops 10% of entities at random in each of 1,000 replications. Country-specific quadratic trends are pre-estimated on the full panel (minor approximation: removing 10% of entities negligibly affects country-level trend estimates). For each subsample, two-way demeaning and Conley SEs are recomputed from scratch. The bootstrap distribution of $\hat{\beta}$ has mean -0.0280 and standard deviation 0.0006, and the coefficient is statistically significant at the 5% level in all 1,000 replications.

Damage projection pathways

Pathway A (QDM). For each GCM in the GDPCIR ensemble, we extract the mean tasmax anomaly under SSP2-4.5 and convert to $R_{95,i,t}$ using the GDPCIR-derived slope; the estimated $\hat{\beta}$ from Eq. 1 translates EHD into a GDP growth impact. Results are reported as rolling-average annual impacts over 2035–2040 ($\Delta T_{\text{global}} \approx +0.85\,^{\circ}\text{C}$ above 2020).

Pathway B (IFS). For each GCM, we extract the same mean tasmax anomaly and convert to $R_{95,i,t}$ using the IFS-derived slope (local at 0.25° and 200 km smoothed variants) and apply the same $\hat{\beta}$.

The multi-model mean and spread across the GCM ensemble are reported for both pathways.

Aggregation from admin2 to country and global. Projected ΔEHD is first computed at the IFS 0.25° grid, then interpolated to 0.1° and aggregated to admin2 entities via settlement-weighted bincount (the same weighting scheme used for the regression). This gives a vector of entity-level ΔEHD_i for the ∼47,000 entities in the damage function sample. Entity-level GDP impacts are computed as $\hat{\beta} \times \Delta\text{EHD}_i$.

For global aggregation, entity-level impacts are averaged using GDP per capita weights (time-averaged over the sample period from the Kummu et al. (42) dataset), yielding the GDP-weighted mean annual growth drag $\Delta\overline{\text{EHD}} \times \hat{\beta}$ reported in Table 2. For country-level maps, admin2 entities are assigned to their ISO3 country code and GDP-per-capita-weighted means are computed within each country, producing the country-level damage estimates shown in Figure 4. Total uncertainty on the projected impact has two independent components. The first is regression uncertainty, propagated as $\text{SE}_{\text{reg}} = |\Delta\overline{\text{EHD}}| \times \text{SE}_{\hat{\beta}}$, where $\text{SE}_{\hat{\beta}}$ is the Conley 500 km standard error. The second is GCM ensemble spread: for each method, we compute $\Delta\overline{\text{EHD}}$ individually for the same 15 CMIP6 models used in the slope t-test and take the cross-model standard deviation SD_{GCM}, converted to impact units as $\text{SE}_{\text{GCM}} = |\hat{\beta}| \times \text{SD}_{\text{GCM}}$. The two components are combined in quadrature as $\text{SE}_{\text{total}} = \sqrt{\text{SE}_{\text{reg}}^2 + \text{SE}_{\text{GCM}}^2}$, and reported as the SE column in Table 2 and the shaded bands in Figure 3. For IFS-based methods the ensemble spread is the dominant uncertainty source, driven by wide variation in projected regional warming across models (T at the grid cell level ranging from 0.7°C to 2.6°C for the 2005-baseline to 2040-projection window under SSP2-4.5).

Data availability

The ECMWF IFS reforecast dataset is available from the ECMWF data portal (https://apps.ecmwf.int/mars-catalogue/?year=2021&type=cf&class=od&stream=enfh&expver=1). The GDPCIR dataset is publicly available on the Microsoft Planetary Computer (https://planetarycomputer.microsoft.com/dataset/group/cil-gdpcir). ERA5 data are available from the Copernicus Climate Data Store (https://cds.climate.copernicus.eu). Economic panel data sources are described in Supplementary Information.

Code availability

Code to reproduce the analysis is available at [repository URL].

Acknowledgements

The authors are funded by Man Group and the University of Oxford Department of Physics.

Author contributions

M.G. and O.V.A. conducted the analysis and wrote the manuscript. J.S. contributed to the writing. M.R.A. advised on the research and reviewed the manuscript.

Competing interests

M.G. and J.S. are employed by Man Group, which manages investment portfolios that may be affected by climate risk. The remaining authors declare no competing interests.

References

[1] Copernicus Climate Change Service. Copernicus: 2024 first year to exceed 1.5°c above pre-industrial level. https://climate.copernicus.eu/copernicus-2024-first-year-exceed-15degc-above-pre-industrial-level (2024). Accessed: April 24, 2025.

[2] Perkins, S. E., Alexander, L. V. & Nairn, J. R. Increasing frequency, intensity and duration of observed global heatwaves and warm spells. *Geophys. Res. Lett.* **39** (2012).

[3] Perkins-Kirkpatrick, S. E. & Lewis, S. C. Increasing trends in regional heatwaves. *Nat. Commun.* **11**, 3357 (2020).

[4] Liu, Q., Fu, C., Xu, Z. *et al.* Global warming intensifies extreme day-to-day temperature changes in mid–low latitudes. *Nat. Clim. Change* **16**, 69–76 (2026).

[5] Hsiang, S. *et al.* Estimating economic damage from climate change in the United States. *Science* **356**, 1362–1369 (2017).

[6] Abatzoglou, J. T. & Williams, A. P. Impact of anthropogenic climate change on wildfire across western US forests. *Proc. Natl Acad. Sci. USA* **113**, 11770–11775 (2016).

[7] Hong, H., Li, F. W. & Xu, J. Climate risks and market efficiency. *J. Econom.* **208**, 265–281 (2019).

[8] Brunetti, C. *et al.* Climate change and financial stability. *FEDS Notes* (2021).

[9] Bolton, P., Despres, M., Pereira da Silva, L. A., Samama, F. & Svartzman, R. The green swan: central banking and financial stability in the age of climate change. Tech. Rep., Bank for International Settlements (2020).

[10] Network for Greening the Financial System. NGFS climate scenarios for central banks and supervisors—phase V. Tech. Rep., NGFS (2024).

[11] Basel Committee on Banking Supervision. Climate-related financial risks—measurement methodologies. Tech. Rep., Bank for International Settlements (2021).

[12] International Sustainability Standards Board. IFRS S2 climate-related disclosures. Tech. Rep., IFRS Foundation (2023).

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

[14] Dell, M., Jones, B. F. & Olken, B. A. Temperature shocks and economic growth: evidence from the last half century. *Am. Econ. J. Macroecon.* **4**, 66–95 (2012).

[15] Nordhaus, W. D. Revisiting the social cost of carbon. *Proc. Natl Acad. Sci. USA* **114**, 1518–1523 (2017).

[16] Kotz, M., Levermann, A. & Wenz, L. The economic commitment of climate change. *Nature* **628**, 551–557 (2024). Retracted.

[17] Kalkuhl, M. & Wenz, L. The impact of climate conditions on economic production. Evidence from a global panel of regions. *J. Assoc. Environ. Resour. Econ.* **7**, 731–772 (2020).

[18] Bilal, A. & Känzig, D. R. The macroeconomic impact of climate change: global vs. local temperature. *Q. J. Econ.* (2024). NBER Working Paper No. 32450.

[19] Gergel, D. R. *et al.* Global Downscaled Projections for Climate Impacts Research (GDPCIR): preserving quantile trends for modeling future climate impacts. *Geosci. Model Dev.* **17**, 191–227 (2024).

[20] Cannon, A. J., Sobie, S. R. & Murdock, T. Q. Bias correction of GCM precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes? *J. Clim.* **28**, 6938–6959 (2015).

[21] Hersbach, H. *et al.* The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).

[22] Lafferty, D. C. *et al.* Downscaling and bias-correction contribute considerable uncertainty to local climate projections in CMIP6. *npj Clim. Atmos. Sci.* **6**, 158 (2023).

[23] Zeder, J., Fischer, E. M. & Knutti, R. On the effect of short observational records on the statistics of temperature extremes. *Environmental Research Letters* **18**, 024014 (2023).

[24] Vautard, R. *et al.* Human contribution to the record-breaking june and july 2019 heatwaves in western europe. *Environmental Research Letters* **15**, 094077 (2020).

[25] Fischer, E. M., Sippel, S. & Knutti, R. Increasing probability of record-shattering climate extremes. *Nature Climate Change* **11**, 689–695 (2021).

[26] Diffenbaugh, N. S. *et al.* Quantifying the influence of global warming on unprecedented extreme climate events. *Proceedings of the National Academy of Sciences* **114**, 4881–4886 (2017).

[27] Kornhuber, K., Bartusek, S., Seager, R., Schellnhuber, H. J. & Ting, M. Global emergence of regional heatwave hotspots outpaces climate model simulations. *Proceedings of the National Academy of Sciences* **121**, e2411258121 (2024).

[28] Hamill, T. M., Whitaker, J. S. & Mullen, S. L. Reforecasts: An important dataset for improving weather predictions. *Bull. Am. Meteorol. Soc.* **87**, 33–46 (2006).

[29] Kelder, T., Fischer, E. M., Pfahl, S., Lehner, F. & Sippel, S. Using large climate model ensembles to evaluate rare event probabilities. *Earth System Dynamics* **13**, 895–910 (2022).

[30] Patel, R. N., Bonan, D. B. & Schneider, T. Changes in the frequency of observed temperature extremes largely driven by a distribution shift. *Geophys. Res. Lett.* **51**, e2024GL110707 (2024).

[31] Loikith, P. C. & Neelin, J. D. Short-tailed temperature distributions over North America and implications for future changes in extremes. *Geophys. Res. Lett.* **42**, 8577–8585 (2015).

[32] Rhines, A. & Huybers, P. Frequent summer temperature extremes reflect changes in the mean, not the variance. *Proc. Natl Acad. Sci. USA* **110**, E546 (2013).

[33] Diez-Sierra, J. *et al.* Consistency of the regional response to global warming levels from CMIP5 and CORDEX projections. *Clim. Dyn.* **61**, 4047–4060 (2023).

[34] Santer, B. D. *et al.* On the linearity of local and regional temperature changes from pattern scaling. *J. Clim.* **31**, 7131–7145 (2018).

[35] Liang, X., Schaller, N. & Sillmann, J. Pitfalls in diagnosing temperature extremes. *Nat. Commun.* **15**, 1234–1245 (2024).

[36] Decker, M. *et al.* Evaluation of reanalysis products for near-surface temperature. *J. Clim.* **25**, 5283–5301 (2012).

[37] Parker, W. S. Reanalyses and observations: what's the difference? *Bull. Am. Meteorol. Soc.* **97**, 1565–1572 (2016).

[38] Hassler, B. & Lauer, A. Comparison of reanalysis and observational precipitation datasets including ERA5 and WFDE5. *Atmosphere* **12**, 1462 (2021).

[39] Lange, S. Trend-preserving bias adjustment and statistical downscaling with ISIMIP3BASD (v1.0). *Geosci. Model Dev.* **12**, 3055–3070 (2019).

[40] Thrasher, B. *et al.* NASA global daily downscaled projections, CMIP6. *Sci. Data* **9**, 262 (2022).

[41] Conley, T. G. GMM estimation with cross sectional dependence. *J. Econom.* **92**, 1–45 (1999).

[42] Kummu, M., Kosonen, M. & Masoumzadeh Sayyar, S. Downscaled gridded global dataset for gross domestic product (GDP) per capita PPP over 1990–2022. *Scientific Data* **12**, 178 (2025).

[43] Schär, C. *et al.* Short warm-side temperature distribution tails drive hot spots of extreme future warming. *J. Clim.* **30**, 7033–7046 (2017).

[44] Battiston, S., Mandel, A., Monasterolo, I., Schütze, F. & Visentin, G. Asset-level assessment of climate physical risk matters for adaptation finance. *Nat. Commun.* **15**, 5371 (2024).

[45] O'Neill, B. C. *et al.* The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. *Geosci. Model Dev.* **9**, 3461–3482 (2016).

[46] Riahi, K. *et al.* The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: an overview. *Glob. Environ. Change* **42**, 153–168 (2017).

[47] Davidson, R. & MacKinnon, J. G. *Econometric Theory and Methods* (Oxford University Press, New York, 2004).

[48] Schiavina, M., Melchiorri, M. & Corbane, C. GHS-BUILT-S R2023A — GHS built-up surface grid, derived from Sentinel-2 composite and Landsat, multitemporal (1975–2030). http://data.europa.eu/89h/9f06f36f-4b11-47ec-abb0-4f8b7b1d72ea (2023).

[49] Driscoll, J. C. & Kraay, A. C. Consistent covariance matrix estimation with spatially dependent panel data. *Rev. Econ. Stat.* **80**, 549–560 (1998).

Supplementary Material

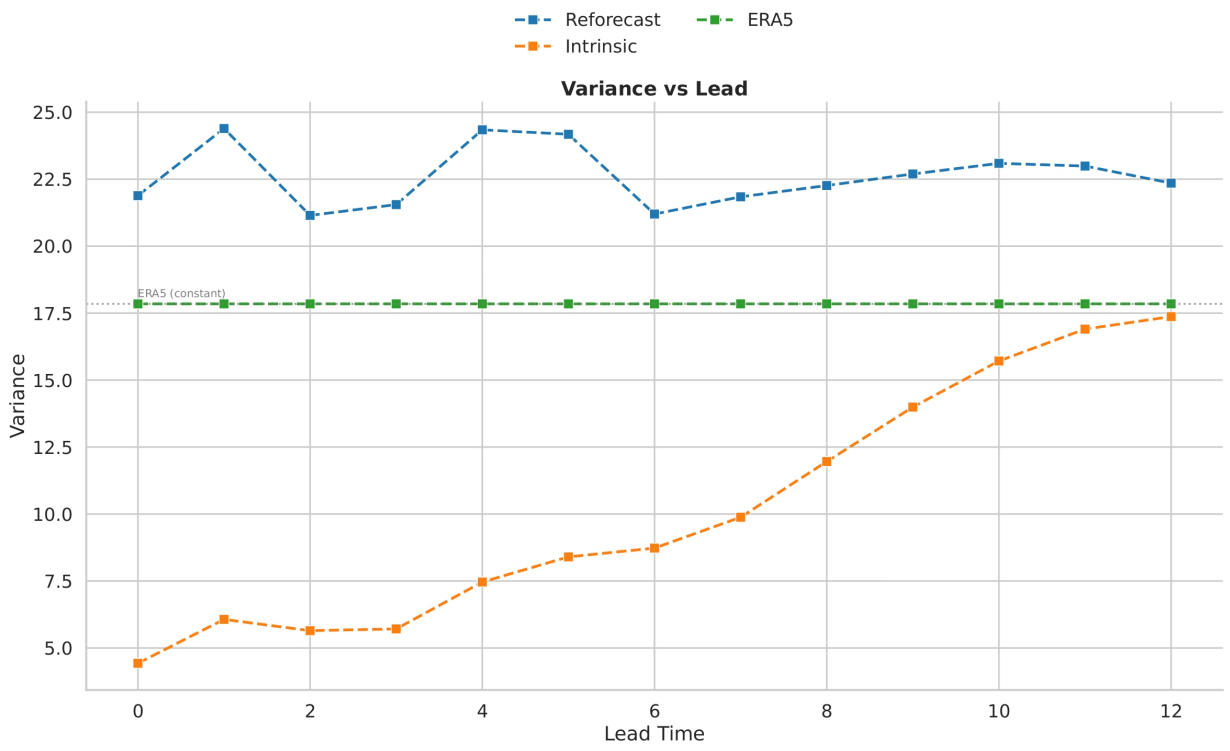


Figure S1: Evolution of temperature variance across forecast lead times for the reforecast ensemble, intrinsic variability, and ERA5 reference (constant). Ensemble variance initially inflates at short leads due to the dominance of initial-condition uncertainty, then gradually stabilizes as forecasts lose memory of initialization. The intrinsic variance increases steadily with lead time, approaching the ERA5 reference value by day 12, indicating that forecast spread converges toward the climatological variability represented in ERA5. This convergence supports the use of 12-day lead reforecasts to assess climate-driven variability.

Table S1: Entity-dropout bootstrap (1,000 replications, 10% of admin2 entities dropped per draw, ≈43,070 entities retained per iteration).

	$\hat{\beta}$	Conley 500 km SE
Full sample (baseline)	−0.0279	0.0084
Bootstrap mean	−0.0280	—
Bootstrap std	0.0006	—
Bootstrap 5th pct	−0.0289	—
Bootstrap 95th pct	−0.0271	—
Fraction $\lvert t\rvert > 1.96$		1.000

Table S2: Robustness of the EHD damage estimate. Dependent variable: log GDP growth per capita. All specifications include region and year fixed effects and country-specific quadratic time trends. Conley 500 km SEs unless otherwise noted in Panel A. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Specification	N	$\hat{\beta}$	SE	t
Panel A: Standard error specification ($\hat{\beta} = -0.0280$ throughout)				
Entity-clustered	1,531,392	−0.0280	0.0017	-16.56^{***}
Time-clustered			0.0127	-2.20^{*}
Two-way clustered			0.0127	-2.20^{*}
Driscoll–Kraay			0.0088	-3.16^{***}
Conley 250 km			0.0065	-4.31^{***}
Conley 500 km (baseline)			**0.0085**	$\mathbf{-3.30}^{***}$
Conley 1000 km			0.0105	-2.67^{**}
Panel B: Alternative controls and functional forms				
$T + T^2 + P$ controls **(baseline)**	1,531,392	−0.0280	0.0085	-3.30^{***}
$T + P$ controls	1,531,392	−0.0321	0.0085	-3.78^{***}
Natural spline (4 df), joint Wald F	1,531,392	$F = 127.0$, $p < 0.001$		
Quintile Q5 vs Q1 (top 20% EHD)	1,531,392	−0.0033	0.0016	-2.03^{*}
Panel C: Lag structure				
Contemporaneous EHD (baseline)	1,531,392	−0.0280	0.0085	-3.30^{***}
Lead EHD $t \to t{+}1$ (placebo)	1,483,536	−0.0075	0.0075	−0.99
Panel D: Income subsamples				
Low income	510,464	−0.0457	0.0142	-3.21^{**}
Middle income	510,432	−0.0306	0.0114	-2.69^{**}
High income	510,496	−0.0181	0.0100	$-1.81^{\dagger}$
Panel E: Alternative spatial aggregation levels				
Admin2 (baseline)	1,531,392	−0.0280	0.0085	-3.30^{***}
Admin1	342,700	−0.0483	0.0122	-3.95^{***}
Panel F: Climate control variable coefficients (baseline spec, Conley 500 km)				
Annual temp. (°C)	1,531,392	+0.0080	0.0018	$+4.58^{***}$
Annual temp.^{2} (°C^2)		-2.0×10^{-4}	5.0×10^{-5}	-3.98^{***}
Annual precip. (mm)		-1.2×10^{-6}	2.0×10^{-6}	−0.36

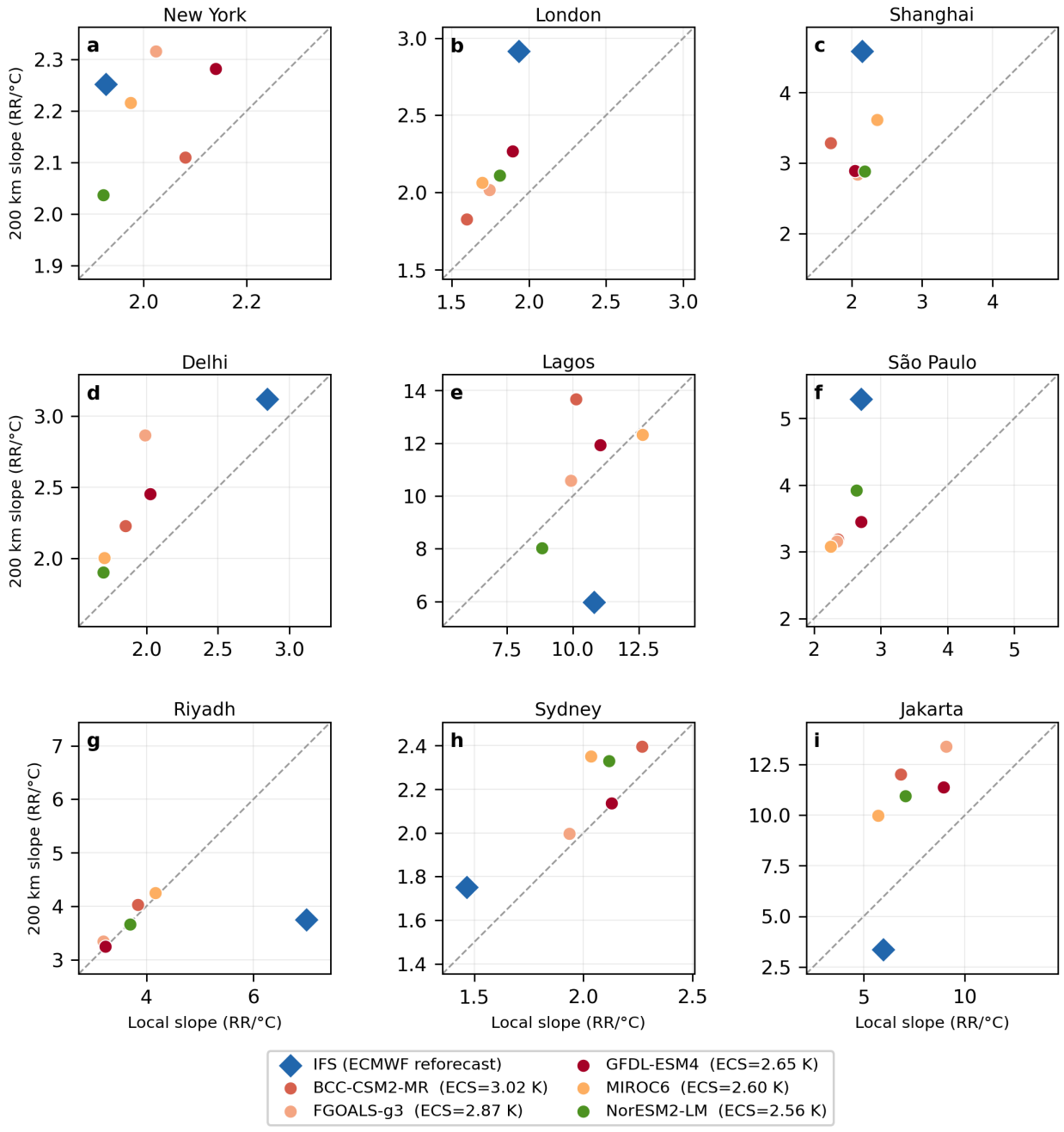


Figure S2: Local (x-axis) vs. 200 km smoothed (y-axis) EHD sensitivity slopes (RR per °C) at nine major cities. Blue diamonds: IFS ECMWF reforecast. Colored circles: five highest-ECS-weighted QDM CMIP6 models (BCC-CSM2-MR, GFDL-ESM4, MIROC6, NorESM2-LM, NorESM2-MM; ECS-weighted Gaussian kernel, center 3.0 K, $\sigma = 0.75$ K). Dashed line: $y = x$ (smoothing has no effect).

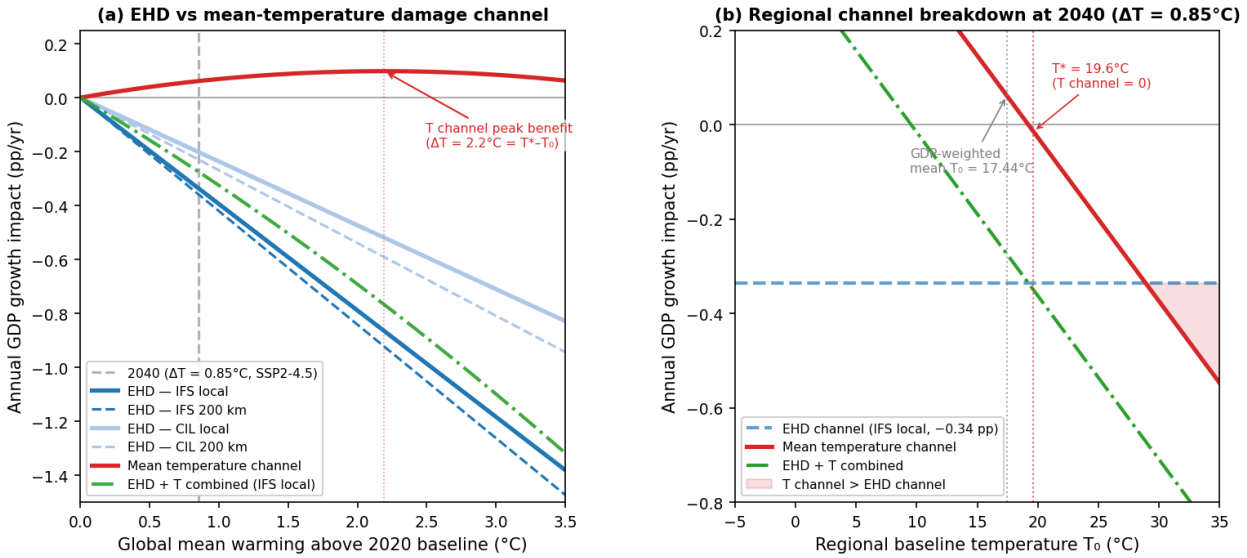


Figure S3: Decomposition of GDP damage into the EHD and mean-temperature channels. *Left*: annual GDP growth impact (pp/yr) as a function of global mean warming above the 2020 baseline, for the IFS-local and QDM-local EHD pathways and for the mean-temperature channel estimated from the baseline regression ($T_0 = 17.4\,^{\circ}\mathrm{C}$). The temperature channel is beneficial until warming reaches approximately $4.4\,^{\circ}\mathrm{C}$; the 2040 horizon ($\Delta T \approx 0.85\,^{\circ}\mathrm{C}$, dashed line) lies well within the EHD-dominated regime. *Right*: regional channel breakdown at the 2040 horizon as a function of regional baseline temperature T_0. The mean-temperature channel exceeds the EHD channel in magnitude only for regions with $T_0 > 28.9\,^{\circ}\mathrm{C}$ (shaded red); for the vast majority of global economic activity the EHD channel dominates.

Table S3: Descriptive statistics for the estimation panel. Panel spans 1,531,392 admin2–year observations, 47,856 entities, 237 countries, 1991–2022. GDP growth and EHD frequency are log-differenced and share-of-days units respectively; raw (pre-demeaning) values reported here.

Variable	Mean	SD	Min	Max
EHD frequency (share of summer days)	0.049	0.057	0.000	0.837
GDP growth per capita	0.018	0.107	−3.077	2.939
Annual mean temperature (°C)	17.44	7.85	−23.02	32.38
Annual total precipitation (mm)	1,274	1,029	0	30,577
Panel dimensions				
Observations				1,531,392
Admin2 entities				47,856
Countries				237
Year range				1991–2022