**Institute for Economic Studies, Keio University**

**Keio-IES Discussion Paper Series**

# When the Air Doesn't Move the Data: The Thermal Inversion Instrument in Developing Countries

**Hiroyuki Yamada**

**18 August, 2026**
**DP2026-018**
**https://ies.keio.ac.jp/en/publications/27955/**

Institute for Economic Studies, Keio University
2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan
ies-office-group@keio.jp
18 August, 2026

When the Air Doesn't Move the Data: The Thermal Inversion Instrument in Developing Countries

Hiroyuki Yamada

Keio-IES DP2026-018
18 August, 2026
JEL Classification: Q53, I15, O13, C26
Keywords: thermal inversion, instrumental variables, air pollution, child health, external validity, Sub-Saharan Africa, South Asia

Abstract

Thermal inversions have become the standard instrument for estimating the causal effect of air pollution, underpinning at least 81 published studies of health, labor, firm, and crime outcomes. Seventy-seven percent of those studies concern China, and none concerns Sub-Saharan Africa. This paper examines whether the design can be transported to the regions where pollution research is expanding fastest. We link 120,820 children in 26 Sub-Saharan African countries and 128,919 in India, Bangladesh, and Nepal to 0.01° satellite PM2.5 and to ERA5 temperature profiles measured at local dawn, and we estimate a single two-stage least squares specification six times, changing only the definition of the inversion instrument. Five of the six definitions come from published practice, and the sixth is an elevation-robust variant of our own. In the African sample, we find that all six first stages are strong and correctly signed (F=57-356). However, the second stages split by instrument family. Pressure-level definitions give positive but insignificant effects on child respiratory illness, whereas surface-layer definitions, the modal implementation in the literature, give significant negative ones. Over-identification tests reject the cross-family pairings of the count and continuous instruments (p<10^(-4) by the Sargan statistic and p≤0.002 with cluster-robust weighting) while accepting every pairing within the pressure-level family. In addition, the surface-layer family fails a diarrhea placebo test. In India, two instruments built from the identical temperature difference, differing only in functional form, have opposite-signed first and second stages. In Bangladesh, where the terrain is at sea level, pollution is extreme, and winter inversions are severe, all six first stages are significantly negative because inversions peak in March while transport-driven pollution peaks in January. These failures appear to be structural. An inversion day moves measured monthly PM2.5 by roughly one microgram per cubic meter, so large F statistics can coexist with an experiment too small to dominate the direct health effects of the accompanying weather. Moreover, the conditions that strengthen the first stage are the same conditions that open those direct channels. Validation against ground monitors further suggests that the fine-scale exposure contrasts many of these designs rely on are largely absent from the satellite data. We provide a nine-item diagnostic protocol and apply it to 26 applications coded in full text. None satisfies more than three of the first eight items, while the two seed papers satisfy six. The seed papers also used different definitions (balloon-sounding profiles in Arceo et al., the 1000/925 hPa pressure difference in Jans et al.), and their instruments moved exposure by 3.4 to 25 percent of its mean, compared with 0.01 to 0.07 percent in the citing studies that report comparable magnitudes. In our data, the seed papers' pressure-level definition behaves coherently, whereas the unvalidated surface-layer definition fails.

Hiroyuki Yamada
Keio University
hyamada@keio.jp

Acknowledgement: This work was supported by an academic fund of Keio University.

# When the Air Doesn't Move the Data:
# The Thermal Inversion Instrument in Developing Countries

Hiroyuki Yamada

Faculty of Economics, Keio University

2-15-45 Mita, Minato–ku, Tokyo 108-8345, Japan

Tel (Direct): +81-(0)3-5427-1271

E-mail: hyamada@keio.jp

https://orcid.org/0000-0003-3709-1914

**Acknowledgements**: This work was supported by an academic fund of Keio University.

**Conflict of interest**: The author declares no conflicts of interest.

# When the Air Doesn't Move the Data:
# The Thermal Inversion Instrument in Developing Countries


## Abstract

Thermal inversions have become the standard instrument for estimating the causal effect of air pollution, underpinning at least 81 published studies of health, labor, firm, and crime outcomes. Seventy-seven percent of those studies concern China, and none concerns Sub-Saharan Africa. This paper examines whether the design can be transported to the regions where pollution research is expanding fastest. We link 120,820 children in 26 Sub-Saharan African countries and 128,919 in India, Bangladesh, and Nepal to 0.01° satellite PM2.5 and to ERA5 temperature profiles measured at local dawn, and we estimate a single two-stage least squares specification six times, changing only the definition of the inversion instrument. Five of the six definitions come from published practice, and the sixth is an elevation-robust variant of our own. In the African sample, we find that all six first stages are strong and correctly signed ($F = 57 - 356$). However, the second stages split by instrument family. Pressure-level definitions give positive but insignificant effects on child respiratory illness, whereas surface-layer definitions, the modal implementation in the literature, give significant negative ones. Over-identification tests reject the cross-family pairings of the count and continuous instruments ($p < 10^{-4}$ by the Sargan statistic and $p \leq 0.002$ with cluster-robust weighting) while accepting every pairing within the pressure-level family. In addition, the surface-layer family fails a diarrhea placebo test. In India, two instruments built from the identical temperature difference, differing only in functional form, have opposite-signed first and second stages. In Bangladesh, where the terrain is at sea level, pollution is extreme, and winter inversions are severe, all six first stages are significantly negative because inversions peak in March while transport-driven pollution peaks in January. These failures appear to be structural. An inversion day moves measured monthly PM2.5 by roughly one microgram per cubic meter, so large $F$ statistics can coexist with an experiment too small to dominate the direct health effects of the accompanying weather. Moreover, the conditions that strengthen the first stage are the same conditions that open those direct channels. Validation against ground monitors further suggests that the fine-scale exposure contrasts many of these designs rely on are largely absent from the satellite data. We provide a nine-item diagnostic protocol and apply it to 26 applications coded in full text. None satisfies more than three of the first eight items, while the two seed papers satisfy six. The seed papers also used different definitions (balloon-sounding profiles in Arceo et al., the 1000/925 hPa pressure difference in Jans et al.), and their instruments moved exposure by 3.4 to 25 percent of its mean, compared with 0.01 to 0.07 percent in the citing studies that report comparable magnitudes. In our data, the seed papers'

pressure-level definition behaves coherently, whereas the unvalidated surface-layer definition fails.

**JEL:** Q53, I15, O13, C26

**Keywords:** thermal inversion, instrumental variables, air pollution, child health, external validity, Sub-Saharan Africa, South Asia

## 1. Introduction

Sub-Saharan Africa is urbanizing faster than any region in recorded history, and its cities are polluting as they grow. Whether, and by how much, that pollution damages children's health is a first-order question for development policy. The region will add nearly a billion urban residents by mid-century, and the investments that would abate urban air pollution compete for funds against every other use. Therefore, answering the question requires causal estimates, which in turn require confronting the endogeneity of pollution, because pollution concentrates exactly where economic activity, congestion, and poverty do. This paper began as an attempt to produce such estimates.

The standard tool for the job is the thermal-inversion instrument. Introduced by Arceo, Hanna and Oliva (2016) for Mexico City, the design instruments pollution with episodes in which warm air overlies cool surface air, suppressing vertical mixing and trapping emissions near the ground. An inversion is a meteorological event, plausibly excludable conditional on weather, and reanalysis data make it observable worldwide at no cost. As a result, the design has become very widely used. We document below that its four canonical papers have accumulated 1,032 distinct citing works, that at least 81 published studies use the instrument, that adoption grew from two studies in 2020 to nineteen in 2023 and has continued at fourteen to sixteen per year since, and that applications now span infant mortality, firm productivity, patenting, crime, fertility, and migration.

We also document a fact that, to our knowledge, the literature has not remarked upon. Of those studies, 77 percent concern China, the original validation is Mexico City, and not one concerns Sub-Saharan Africa. In other words, the instrument's evidentiary base consists almost entirely of high-emission basin cities where inversions visibly drive pollution episodes, which are the settings where its first stage is mechanically strongest. However, whether the design carries over to the regions where pollution research is expanding fastest has not been tested in the published record.

This paper provides that test. We assemble a quarter of a million geocoded child-health observations: 120,820 children across 26 Sub-Saharan African countries and 80

Demographic and Health Survey (DHS) rounds, plus 128,919 across India, Bangladesh, and Nepal. We link them to 0.01° satellite PM2.5 and construct, from ERA5 temperature profiles at local dawn, six versions of the inversion instrument spanning pressure-level and surface-layer temperature differences in count, strength, and continuous form. Five of the six come from published practice, and the sixth is an elevation-robust variant of our own. We then estimate the same two-stage least squares specification six times, changing only the instrument definition.

In the African sample, we find that all six instruments have strong first stages of the theoretically correct sign, with $F$ statistics between 57 and 356, so each would pass review in isolation. However, the two-stage estimates split by family. Pressure-level definitions, the implementation the seed papers validated, give positive but insignificant effects of PM2.5 on child respiratory illness, whereas surface-layer definitions, the China standard that dominates the applied literature, give significant negative ones. Over-identification tests reject the count and continuous cross-family pairings decisively, at $p < 10^{-4}$ by the Sargan statistic and $p \leq 0.002$ with cluster-robust weighting, and accept every pairing within the pressure-level family. The one strength-based cross-family pair rejects under the Sargan statistic ($p = 0.044$) but not under clustering. In addition, the surface-layer count measure fails a diarrhea placebo test that the pressure-level family passes. India sharpens the contradiction further. Two instruments built from the identical physical quantity, one counting the days on which the surface-layer difference is positive and one averaging its magnitude, have first stages of opposite sign and second stages of opposite sign, both of them "significant." In Bangladesh, the setting that looks ideal on paper because of its sea-level terrain, extreme pollution, and severe winter inversions, all six first stages are significantly negative, because inversions peak in March while transport-driven PM2.5 peaks in January. In Nepal, the highest first-stage $F$ across the six definitions is 8.0.

Section 6 traces these failures to their sources. First, the identifying variation is physically very small. An inversion day moves measured monthly PM2.5 by roughly one microgram per cubic meter in our samples, a few percent of the mean, so enormous $F$ statistics can coexist with an experiment too small to dominate the direct effects of the weather that accompanies it. Second, the obvious remedies appear to make matters worse, because the conditions that strengthen the first stage, such as dust seasons and dry seasons, are the same conditions that open direct meteorological channels into health. Specifically, in our data, the dust tercile where the instrument is strongest is the tercile where the placebo fails, and the tercile where the exclusion logic is cleanest is the tercile where $F$ collapses to 5.6. Moreover, alternative instruments proposed for such settings, including upwind biomass-fire radiative power and wind direction interacted with local population, reproduce similar pathologies.

In addition, a third consideration underlies the other two: the satellite exposure product itself does not resolve the finest contrasts on which pollution-health designs rely. Specifically, the 0.01° surface is a smooth interpolation of the 0.10° field, and against ground monitors Yamada (2026) reports a reliability of 0.0097 for within-city contrasts, against 0.748 for between-place contrasts and 0.786 for the temporal contrast that our design uses (Section 6.6). Classical measurement error in the endogenous variable leaves two-stage least squares consistent, so this finding does not overturn our estimates. However, it removes the ability to verify the size and the independence of the experiment the instrument runs. This is a limitation of the data available for the region rather than of any author's choices, and it applies to every pollution-health design that relies on satellite exposure in Sub-Saharan Africa, not only to the inversion instrument.

Our study makes four contributions to the literature. First, to our knowledge it provides the first systematic, multi-region, multi-definition validation exercise for a design now used in more than eighty published papers, and shows that the design fails outside its home settings, with sign reversals rather than marginal shifts. Second, it identifies the mechanism: the identifying perturbation is too small, and too seasonally entangled, to carry causal weight, even though the instruments are strong in the conventional statistical sense. Third, it documents that the literature's implementations, nominally a single method, are physically distinct instruments that our data show to be mutually contradictory, which suggests that part of the cross-study variation in published estimates reflects definition choice rather than context. Reading the seed papers against the studies that cite them also shows that none of these implementations reproduces the measurement on which the design was validated. Fourth, it distills the findings into a nine-item diagnostic protocol (Table 8) that any researcher can run before committing to the design, and applies the protocol retrospectively to 26 applications coded in full text, eighteen of which rely on a single untested definition and none of which has compared definitions across the two families.

Two caveats delimit the scope of our claims. First, this paper is not an ex-post critique of any individual study. None of the 81 studies concerns the region where we show the design fails most comprehensively, and we deliberately frame our African evidence as an ex-ante warning for a literature that is heading there. Second, we do not claim that the instrument is invalid everywhere. Where inversions move pollution by a large fraction of its mean (25 percent in the Swedish setting of Jans et al. and 3.4 to 5.7 percent in Arceo et al.'s Mexico City, compared with 0.01 to 0.07 percent in the applications that inherited the design), the trade-offs we document may resolve favorably. What we show is that strong first stages, verbal exogeneity arguments, and even passing placebo tests are jointly insufficient to establish that they do. Therefore, the substantive question that motivated this paper, the health cost of Africa's urban air, remains open for want of a credible design. We return to both points in the conclusion.

The remainder of this paper is organized as follows. Section 2 quantifies the instrument's use, codes its implementations, and compares them against the seed papers' own. Section 3 describes the data, and Section 4 describes the six instruments and the specification. Section 5 presents the central results, and Section 6 examines the mechanisms. Section 7 states the diagnostic protocol, and Section 8 concludes the paper.

## 2. The Thermal Inversion Instrument in Practice

### 2.1 Logic and origin

A thermal inversion is a departure from the atmosphere's normal state in which a layer of warmer air overlies cooler air near the surface, suppressing vertical mixing and trapping locally emitted pollutants in the boundary layer. Because inversions form through radiative and synoptic processes that no economic actor controls, they are a natural candidate for instrumenting ambient pollution. Conditional on weather, the argument runs, inversions affect outcomes only by raising pollutant concentrations. Arceo, Hanna and Oliva (2016) introduced the design to economics, instrumenting particulates and carbon monoxide in Mexico City with weekly inversion counts to estimate effects on infant mortality. The appeal of the design is easy to understand. It requires no policy discontinuity, no monitor network, and no natural experiment beyond the weather itself, and reanalysis products supply the required temperature profiles globally at no cost. As a result, it has become a standard tool.

### 2.2 How widely, and where, the instrument is used

To quantify that standardization, we conducted a systematic forward-citation search from four seed papers that applied studies routinely cite as methodological authority: Arceo et al. (2016), Jans, Johansson and Nilsson (2018), Deschênes, Wang, Wang and Zhang (2020), and Fu, Viard and Zhang (2021). The four seeds have accumulated 1,032 distinct citing works (OpenAlex), and the flow of citations to them roughly tripled between 2020 and 2024. Screening the union of Scopus and Web of Science forward citations (245 unique records after deduplication) on title, abstract, and keywords identifies 82 records that use thermal inversions as an instrumental variable for air pollution. An independent OpenAlex full-text query returns 85, confirming the count from a second source. In the course of reference checking we verified one of the 82 to be a false positive, an unrelated article that had entered the export under a mis-assigned title, leaving 81 confirmed studies. We report both figures because we checked only the records we examined in full and cannot rule out further errors of either sign (Appendix E.5). Because abstract screening also misses studies that mention the instrument only in the body, these figures are lower bounds.

Table 1 and Figure 1 summarize the corpus. Three features matter for this paper. First, adoption accelerated from two studies in 2020 to nineteen in 2023 and has continued at fourteen to sixteen per year since, with nine more already by mid-2026. Second, the outcome space has expanded far beyond the design's epidemiological origins, to firm productivity, patenting, exports, robot adoption, crime, fertility, analyst forecasts, and migration, so that the instrument now underpins estimates across essentially every applied field that touches pollution. Third, and most consequential for us, the evidentiary base is geographically concentrated to a degree that has, to our knowledge, gone unremarked. Of the 81 confirmed studies, 62 (77 percent) concern China, the original validation is Mexico City, five each concern Europe and the United States, and India appears twice. Ten records identify no study region in their titles or abstracts, among them the Taiwan-based study we code in full text. In contrast, Sub-Saharan Africa appears zero times. The settings in which the instrument was validated, high-emission cities in basins and valleys where inversions visibly drive pollution episodes, are the settings in which its first stage is mechanically strongest. Therefore, whether the design can be transported to the regions where pollution research is now expanding fastest is, on the published record, an open question. The journals most exposed are those most likely to receive such applications: the Journal of Environmental Economics and Management alone has published six of the 81 studies, the International Journal of Environmental Research and Public Health four, and the Journal of Development Economics, the Journal of the Association of Environmental and Resource Economists, Environmental and Resource Economics, and the Journal of Cleaner Production three each.

### 2.3 One name, many instruments

We then coded 26 studies in full text (10 must-read, 16 secondary; Appendix E lists the protocol and one departure from the citation-ranked selection rule). Table 2, Panel A reports the result. Behind the shared label "thermal inversion instrument" lies substantial and essentially undiscussed implementation heterogeneity.

Twenty of the 26 studies draw temperatures from MERRA-2, but they compare different layers. Seventeen, the China standard, which cites Arceo et al. as its authority, compare the product's two lowest model layers at 110 m and 320 m above sea level. Two compare the 1000 hPa and 925 hPa pressure surfaces, two (the UK studies) compare surface air temperature against 925 hPa, one compares 110 m against 540 m, one cites a pressure surface and a geometric layer together, and three do not describe the comparison precisely enough to classify. Functional forms differ as well, and often within a single paper: counts of inversion days or episodes appear in fourteen studies, continuous strength measures in eleven, and daily binary indicators in six. Timing is almost never discussed. Nineteen studies aggregate the product's fixed six-hourly snapshots as given, two do not state an acquisition time at all, and only four deliberately restrict to local nighttime or dawn hours, when surface inversions actually exist. However, these differences are consequential. The 110/320 m comparison measures stability of the lowest few hundred meters of the atmosphere, the surface layer, while the 1000−925 hPa comparison measures stability of a slab centered several hundred meters higher. As Sections 5 and 6 demonstrate, these two quantities can carry opposite-signed information in the same data. Panel C of Table 2 maps the lineages onto the two instrument families we take to the data.

The literature is more careful than a first pass suggests. Eighteen of the 26 studies report some placebo or falsification exercise, eight report results under two or more inversion definitions, two others add a second instrument of an entirely different kind, and four report an over-identification test. However, none of them varies the family. All eight robustness exercises stay inside one family, substituting the 540 m layer for the 320 m layer or a binary indicator for a count, and no study in the corpus has placed a pressure-level and a surface-layer definition side by side, either as alternative specifications or jointly in an over-identified instrument set. The same holds for the placebos. No study in the corpus applies one separately to two definitions, and none reports a placebo that fails.

As a consequence, the literature's own robustness checks appear to have been recording the instability we document without recognizing it. When Zhang et al. (2022) replace the 320 m layer with the 540 m layer, they report that "only the regression coefficient was different" and conclude that the results "remain robust and reliable." When Chen et al. (2020) make the same substitution, they describe it as "a more stringent definition." In both cases, a coefficient that moves when the layer changes is read as confirmation rather than as evidence that the two layers measure different things. In addition, two items are missing from the record entirely. No study reports the share of observations at which a count-type instrument equals zero at the aggregation used in estimation, and only four convert the first stage into physical units per unit of instrument for comparison against the exposure mean.

Panel B documents that the technical problems we study are already visible in the published record. The most-cited China application acknowledges in a footnote that its MERRA-2 layers are referenced to sea level, so that grid cells above roughly 320 m of elevation have no valid first or second layer, and states that it substitutes "the two closest non-missing layers," an adjustment under which the instrument's physical definition varies with the elevation of the observation. Four studies describe the layer geometry incorrectly or in language that mixes pressure with height. One places the first layer "approximately 110 m above ground level," a second locates the inversion "between ground level and the next higher level of 320 m," a third calls the model levels "barometric pressure layers" whose altitude is measured "from the ground," and a fourth calls them "sea-level pressure layers." Two state the geometry correctly. The remaining twenty do not describe it at all.

## 2.4 What the seed papers did, and what the literature kept

The comparison that organizes this section is between the literature's practice and that of the two papers it cites as methodological authority, rather than between our practice and the literature's. We read Arceo, Hanna and Oliva (2016) and Jans, Johansson and Nilsson (2018) in full alongside the 26 citing studies. We find that the divergence is larger than the shared label suggests, and that it runs in a specific direction.

Arceo et al. (2016) did not difference two layers of a gridded product. Their inversion data come from balloon soundings conducted almost daily by the Mexico City environment ministry, which measure temperature at a range of altitudes. An inversion is declared "upon finding non-monotonic temperature gradients," wherever in the profile the non-monotonicity occurs, and the records carry the duration and the thickness of the inverted layer as well. A profile-based rule detects an inversion at whatever height it forms, whereas a two-layer difference detects it only if it happens to straddle the two levels chosen. Therefore, these are different measurements.

Arceo et al. also recognized the altitude problem explicitly. Observing that particulate concentrations are lower at higher altitudes "presumably because higher areas are likely to be above the thermal inversion layer," they let the instrument's content depend on elevation. To instrument two pollutants at once, they use the inversion count, the count interacted with municipal altitude, and the inverted layer's thickness interacted with altitude. Three further features matter. They absorb seasonality at fine granularity, with bimonthly × municipality fixed effects that "control for common factors in a given two month block that could affect both pollution levels and infant mortality within a municipality." They control for weather with a fourth-order polynomial in mean temperature, third-order polynomials in the week's minimum and maximum, and second-order polynomials in precipitation, cloud cover and humidity. They also run a health-outcome placebo, re-estimating the model for deaths from internal and from externally caused sources, stating that "this can be viewed as a placebo test," and finding the effect confined to internal-cause mortality, while noting candidly that external deaths are rare enough to limit the test's power. Finally, they name the threat this paper documents: "Controlling for temperature is important for the exclusion restriction to hold, since inversions have a clear seasonal pattern and temperature may independently affect infant mortality."

Jans et al. (2018) supplied the reanalysis form of the design, and it is our family A. They take vertical temperature profiles from NASA's Atmospheric Infrared Sounder and use "the temperatures for the two pressure levels closest to the ground (1000 hPa and 925 hPa)," noting that the 1000 hPa layer "corresponds to the surface conditions" and the 925 hPa layer "measures conditions at approximately 600 m above sea level." That is precisely the pressure-level difference from which we build inv_days and inv_str, and Jans et al. construct both forms, an episode count and a strength measure. They restrict to nighttime episodes and justify the restriction. They report that inversions occur on 25 percent of days in their sample, so the instrument's frequency is on the face of the paper, and they report the first stage in interpretable units, in the abstract, as a 25 percent increase in PM10. Their injury analysis, sometimes read as a placebo, is in the published version a test for avoidance behavior. The outcome is the share of injuries occurring indoors, and the question is whether households change children's activity patterns during inversions. They find no significant effect, but they note that rain and wind do predict that share, a reminder that the weather accompanying the instrument moves behavior.

The modal implementation in the literature descends from neither. Seventeen of the 26 citing studies use the MERRA-2 110 m/320 m layer pair, a surface-layer measure and our family B, at fixed six-hourly snapshots with no local-time restriction. That is a different layer pair from the one Jans et al. validated, taken at a different time of day, and a different measurement principle from the one Arceo et al. validated. In other words, the citation trail is unbroken, but the measurement to which it points back was not carried forward.

This distinction matters because of where our diagnostics land. Family A, the Jans definition, is the family that in our data is internally coherent, passes the placebo, and finds nothing. In contrast, family B, the definition with no seed-paper warrant, is the family that produces significant estimates, fails the diarrhea placebo, rejects against family A in the count and continuous cross-family over-identification tests, and in South Asia contradicts itself outright. To our knowledge, the literature's migration from the validated definition to the unvalidated one was never argued for, presumably because it was never noticed.

The divergence extends past the definition to the diagnostics. Measured against the first eight items of Section 7's protocol, the two seed papers between them satisfy six items. The instrument is allowed to vary with elevation (Arceo), it is aligned to local night (Jans), its frequency and coverage are reported (both), the first stage is reported in units interpretable against the exposure mean (both), a health-outcome placebo is run (Arceo), and seasonality is absorbed at bimonth-by-municipality granularity and named as the threat (Arceo). Of the 26 citing studies, none satisfies more than three, and the modal study satisfies one. The two items the seed papers do not satisfy are items 4 and 6, the cross-definition and cross-family comparisons, which could not have occurred to them because at the time there was only one definition.

We draw no inference about whether either seed paper's own estimates are right. Arceo et al. study a single basin city with dense monitoring and a sounding program. Jans et al. study Sweden, at night, with an instrument that moves PM10 by a quarter of its mean. Nothing in our data speaks to either, and Section 6.1 shows that both instruments are far stronger, in the sense that matters, than the ones their descendants use. Our claim is narrower: the design as practiced is not the design that was validated.

### 2.5 The first stage in the published record

The clearest illustration of the first-stage problem also comes from the published record rather than from our own estimates. Zhang et al. (2022) report a first stage of 0.0066 μg/m³ per inversion day and observe, correctly, that this is 0.01 percent of the sample mean. Alongside it they report a Kleibergen–Paap $F$ of 668 and a second-stage effect of 2.45 standard deviations of the outcome per standard deviation of PM2.5. Within that single paper the $F$ statistic ranges from 2.0 to 668 across subsamples, and coefficients are reported and interpreted at $F = 2.0$. Nothing in the paper's own diagnostics flags any of this, because the diagnostic on which the field relies, the first-stage $F$, is the one that a physically negligible but statistically precise instrument passes most easily.

### 2.6 Three hypotheses

This audit of practice yields three testable propositions that organize the empirical work to follow.

H1 (Definition dependence). If the different implementations captured a common mechanism, namely trapped surface emissions, then estimates should be invariant to the choice among them. If instead each definition loads on layer-specific meteorology, estimates will differ across definitions in the same data, potentially in sign, while every definition maintains a strong first stage.

H2 (Geography dependence). The instrument's validated settings share high local emissions and topography that makes stability the binding constraint on concentrations. In settings where concentrations are governed by transported dust or smoke (the Sahel), by monsoon washout (the Bengal delta), or where reference layers sit below ground (highlands), the first stage should weaken, reverse, or degenerate. Where it remains strong, it should do so for reasons that compromise the exclusion restriction.

H3 (Implementation fragility). Because inversions are a dawn phenomenon measured at fixed UTC hours, and because count-type definitions are zero-inflated, mechanical implementation choices such as the snapshot hour and the functional form should be capable of producing large, "significant," and entirely artifactual results.

Sections 5 and 6 test these propositions on a quarter of a million children across two continents. Section 6.6 adds a fourth consideration, not hypothesized in advance, about the data on which all three tests stand: the satellite exposure product does not resolve the within-city contrasts on which many designs in this literature rely.

## 3. Data

### 3.1 Child health: Demographic and Health Surveys

Child health outcomes come from the Demographic and Health Surveys (DHS), the standard microdata source for health research in low- and middle-income countries and the natural test bed for any instrument intended for such settings. We assemble every geocoded standard DHS in Sub-Saharan Africa, 80 surveys across 26 countries fielded 1998–2024, together with all geocoded surveys for India (NFHS-4 2015–16 and NFHS-5 2019–21), Bangladesh (seven surveys, 1999–2022), and Nepal (five surveys, 2001–2022).

Our primary outcome is acute respiratory infection (ARI) in the two weeks preceding the interview, defined in the standard way as a reported cough accompanied by short, rapid breathing (variables h31 and h31b of the children's recode). Our placebo outcome is diarrhea over the same recall window (h11). It shares the module, the respondent, the recall period, and a similar prevalence, but ambient PM2.5 has no plausible short-run pathway to it. The design has a precedent at the root of this literature, since Arceo et al. (2016) split infant mortality into internal and externally caused deaths and note that "this can be viewed as a placebo test." Among the 26 citing studies we code, however, only Liu and Ao (2021), who use outpatient spending on fractures, dislocations, and sprains as an outcome that "should not be affected by air pollution levels," run such a test. Section 6.3 adds a second placebo, antimalarial treatment, which shares the seasonality of respiratory illness without sharing its pathway. Interview timing is taken from the century-month code of the actual interview (v008), so that children interviewed in different calendar years of a multi-year survey are matched to exposure in the correct year. Nepali and Ethiopian surveys recorded in the Bikram Sambat and Ethiopian calendars are converted to the Gregorian calendar and validated against each survey's documented field dates. Cluster coordinates are the DHS public-release locations, which are randomly displaced by up to 2 km in urban areas, a feature our exposure construction accommodates directly.

### 3.2 Satellite-based PM2.5

Exposure is the Atmospheric Composition Analysis Group's V6GL03 hybrid product (Shen et al. 2024), which combines satellite aerosol optical depth, chemical-transport modeling, and ground monitors into monthly PM2.5 surfaces at 0.01° resolution (≈1.1 km). We sample the surface at each cluster as the mean within a 2 km radius, matching the maximum urban displacement, for the month and year of each child's interview. Results are insensitive to using the grid-point value, a 5 km buffer, or the coarser 0.1° release (Appendix D.2). However, we note at the outset what that insensitivity does and does not establish. The 0.01° and 0.1° products are correlated 0.99 at our clusters because the fine product is a smooth interpolation of the coarse field, not an independent measurement. Only 1.4 percent of the product's spatial variance lies within a single 0.1° cell, and the effective resolution of the African field is about 55 km rather than the nominal 1.1 km (Yamada 2026). Exchanging the two products is therefore not a test of resolution. The direct test, a validation against ground monitors, is reported in Section 6.6.

### 3.3 Meteorology and the timing of the instrument

Temperature profiles come from ERA5. We extract daily snapshots of temperature at the 1000 and 925 hPa pressure surfaces, 2-meter air temperature, and 10-meter winds, and aggregate them over each child's interview month. Because surface inversions are a nocturnal phenomenon that erodes after sunrise, the snapshot hour matters. We use 06 UTC in Africa (06:00–09:00 local across our sample) and 00 UTC in South Asia (05:30–06:00 local), in both cases local dawn. This timing is a deliberate design choice. Section 6.4 shows that importing the African convention into South Asia, where 06 UTC is local noon, degenerates the instrument to near-constancy, and that the degenerate instrument nevertheless produces "significant" estimates. Two dust measures from the CAMS EAC4 reanalysis appear below and should not be confused. The dust share of aerosol optical depth at each cluster defines the sample screen of Section 3.4, and the mineral-dust share of PM2.5 mass (the speciation also used in Appendix D.3) defines the tercile split of Section 6.2.

### 3.4 Samples and descriptive statistics

The Sub-Saharan African analysis sample applies three restrictions, each motivated by the instrument's own logic as articulated in the literature: urban clusters, where trapped local emissions are the relevant margin; elevation below 1,000 m, where the 1000 hPa reference surface, at roughly 110 m altitude, exists above ground for most of the sample (49 percent of clusters still lie above it, a fact we exploit); and a CAMS aerosol-optical-depth dust share below 0.30, which excludes Saharan-dust-dominated clusters where the trapped-emissions logic does not apply. The resulting sample contains 120,820 children in 8,456 clusters, 26 countries, and 80 surveys. The South Asian sample applies no elevation or dust filter, because the region's elevation gradient is part of the test, and contains 128,919 urban children in 17,799 clusters (India 104,676; Bangladesh 15,953; Nepal 8,290).

Table 3 reports descriptive statistics by country. Three features anticipate the results. First, exposure levels differ sharply between the two regions. Mean PM2.5 is 23 μg/m³ in the African sample (country means from 6 to 43) against 51 μg/m³ in South Asia (India 49, Bangladesh 64, Nepal 46). If pollution levels were the binding constraint on the design, South Asia would be its showcase. Second, ARI prevalence, 7.3 percent in Africa and 6.0 percent in South Asia, is elicited by the same questionnaire items everywhere, so cross-country comparisons are not contaminated by differences in how the outcome is measured.

Third, and most consequential for our purposes, the identifying variation available to the instrument is very thin. The median cluster-month records zero inversion days in 24 of our 26 African countries, and zero shares run from 55 percent (Bangladesh) to 85 percent (the African sample) even when the instrument is measured at local dawn. Within Africa, the sparsity is also highly uneven. Eight countries (Kenya, Tanzania, Burundi, Zambia, Ethiopia, Guinea, Uganda and Rwanda) record no pressure-level inversion day at all, and five more are at or above 98 percent zeros, so thirteen of the 26 contribute essentially nothing to identifying family A. At the other extreme, Namibia's median cluster-month records 26 inversion days. Three countries, Nigeria, Mozambique and Benin, supply 68 percent of all non-zero observations. In other words, a nominally 26-country design draws, for one of its two instrument families, two-thirds of its identifying variation from three countries. Appendix D.4 shows that no single country drives our results. However, the concentration itself is a feature of the method that published applications do not report.

## 4. Empirical Framework

### 4.1 Six instruments, two families

From the daily dawn temperature profiles we construct six monthly instruments, chosen to span the implementations documented in Table 2 (construction details in Table 4). Family A measures stability against a fixed-altitude reference. Two of its members are built directly from the pressure-level difference $T_{925} - T_{1000}$: inv_days, the count of days with a positive difference (the count implementations), and inv_str, the monthly mean of the positive part (the strength implementations). The third, stab_an, starts from the surface-layer difference $T_{925} - t_{2m}$ but adds an elevation correction at the standard lapse rate, which removes the dependence of that difference on how far the cluster sits below the 925 hPa surface. What remains is referenced to a fixed altitude rather than to the ground, which is why we group it here. The assignment is corroborated rather than assumed. Section 5.3 shows that stab_an is mutually consistent with inv_days in an over-identification test ($p = 0.73$), while its pairing with the surface-layer sfc_str is rejected by the Sargan statistic ($p = 0.044$), though not decisively under cluster-robust weighting ($p = 0.12$). Nothing in Section 5 depends on it, since inv_days and inv_str alone reproduce the family split. Family B operates on the surface-layer difference $T_{925} - t_{2m}$: sfc_days, sfc_str, and sfc_diff, the count, positive-part, and signed continuous versions respectively. The correspondence to the literature is direct. Family A is the seed definition. Jans et al. (2018) instrument with the difference between the 1000 and 925 hPa levels, in both count and strength form and restricted to night, which is what inv_days and inv_str are, and the India studies follow them. Family B is the modal implementation. The China-standard 110/320 m layer comparison measures the same surface-layer stability, and Sager's (2019) surface-versus-925 measure is literally sfc_diff.

Therefore, a researcher's choice of lineage is implicitly a choice of family, and as Section 2.4 shows, seventeen of the 26 coded studies have chosen, presumably without noticing, the family that neither seed paper validated.

**4.2 Specification**

For child $i$ in cluster $c$, survey $s$, region $r$, interviewed in calendar month $m$, we estimate by 2SLS:

$$\mathrm{ARI}_{icsrm} = \beta \mathrm{PM2.5}_{cm} + X'_{cm}\gamma + \delta_s + \delta_r + \delta_m + \varepsilon_{icsrm}$$

instrumenting PM2.5 with one (or, for over-identification tests, two or three) of the six instruments. $X$ contains monthly mean temperature and wind speed, the weather controls standard in this literature, whose role we probe in Section 6. Survey fixed effects absorb country-by-round differences in questionnaire and season of fieldwork, region fixed effects absorb subnational geography, and month fixed effects absorb common seasonality. Standard errors are clustered by DHS cluster.

The exposure, the instrument, and the weather controls are all assigned to the child's actual interview calendar month, taken from the century-month code v008. This is the interview-anchored contemporaneous convention of the literature: Arceo et al. (2016) link weekly mortality to same-week inversions, Jans et al. (2018) link daily health care visits to same-day inversions, and the survey-based applications anchor exposure windows at the interview date (e.g., Deschênes et al. 2020). Because the outcome recalls the two weeks before the interview, children interviewed early in a month have a recall window that partly falls in the previous month. Appendix D.4 shows that the family split of Section 5 is unchanged when the sample is restricted to children interviewed in the second half of the month, whose recall window lies within the exposure month. Appendix D reports region-level clustering, Conley spatial HAC at 25–200 km cutoffs, and wild-cluster bootstrap inference, which matter for the magnitude of standard errors but not for any qualitative conclusion in this paper.

**4.3 Diagnostics**

Three diagnostics run throughout. (i) Placebo outcome: the reduced form of each instrument on diarrhea, which a valid instrument should leave untouched. (ii) Over-identification: Sargan and cluster-robust Hansen $J$ tests for instrument pairs drawn from within a family and across families. Under the null that both instruments are valid, they identify the same parameter. (iii) Physical-units reporting: every first stage is reported as $\mu g/m^3$ of PM2.5 per unit of instrument, alongside the conventional $F$ statistic. In addition, we screen every instrument-sample cell for degeneracy, defined as more than 95 percent zeros or fewer than ten distinct values. The screen is motivated in Section 6.4, where we show what passes through without it.

## 5. Results I: Six Instruments, Contradictory Answers

This section presents the central finding of the paper. We estimate the same two-stage least squares specification, with the same outcome, the same exposure, the same sample, and the same fixed effects and controls, six times, changing only the definition of the thermal-inversion instrument. Each definition is drawn from, or closely follows, a published implementation (Table 4). If these definitions captured a common physical mechanism, they would identify the same local average treatment effect up to sampling noise. However, we find that the estimates contradict one another.

### 5.1 First stages: uniformly strong, uniformly "correct"

Panel A of Table 5 reports the first stage for each of the six instruments on the Sub-Saharan African sample (120,820 children in 8,456 urban clusters across 26 countries and 80 DHS surveys). Judged by the criteria conventionally applied in this literature, every instrument performs well. All six first-stage coefficients carry the theoretically expected positive sign: months with more (or stronger) inversions are months with higher PM2.5. All six are strong by any conventional threshold, with Kleibergen–Paap $F$ statistics ranging from 57 (stab_an) to 356 (sfc_diff), far above the rule-of-thumb value of 10 and comfortably above the Stock–Yogo critical values applied in published work.

A referee reviewing any one of these first stages in isolation would find nothing objectionable, and this is precisely how the literature proceeds. Of the 26 studies we code in Section 2, eighteen report results for a single inversion definition, and not one has placed a pressure-level definition beside a surface-layer definition. A strong first stage is routinely treated as the primary, and often the only, quantitative evidence for the instrument's suitability.

However, the $F$ statistics conceal how narrow the identifying variation is. As Section 3.4 noted, 85 percent of observations record no pressure-level inversion at all, and three countries supply 68 percent of the non-zero observations. In other words, the family-A first stage of $F = 132$ is estimated from a thin tail of a highly skewed distribution across 120,820 observations. Nothing in the conventional reporting of a first stage reveals this feature.

### 5.2 Second stages: the sign depends on the definition

The remaining columns of Panel A show why this practice can be dangerous. The six two-stage estimates of the effect of PM2.5 on child respiratory illness differ not only in magnitude but also in sign, and both signs are "significant" by conventional standards in at least one specification. Figure 2 plots the twelve estimates, six for each region, with their confidence intervals.

The three pressure-level instruments (family A), namely the count of days on which temperature at 925 hPa exceeds temperature at 1000 hPa, the mean strength of that difference, and a lapse-rate anomaly measure, produce positive point estimates of +0.34 to +1.02 percentage points of ARI prevalence per 10 μg/m³ of PM2.5, none of which is statistically distinguishable from zero ($t = 0.26$ to $1.65$).

The three surface-layer instruments (family B), the same three functional forms applied to the difference between 925 hPa temperature and 2-meter air temperature, produce negative estimates of −0.94 to −1.91 percentage points per 10 μg/m³, all nominally significant ($t = -2.16$ to $-3.05$). Taken at face value, family B implies that air pollution protects children against respiratory illness.

Both families correspond directly to published practice. As Table 2, Panel C documents, family B corresponds to the modal implementation in the literature. The MERRA-2 110 m/320 m layer comparison used in the China-based studies that constitute 77 percent of published applications measures stability in the lowest few hundred meters of the atmosphere, as our surface-layer family does, and the surface-versus-925 hPa measure of Sager (2019) is literally our sfc_diff. Family A corresponds to the 1000−925 hPa implementation of Jans et al. (2018) and the India-based studies that follow it. A researcher who adopted the China-standard definition on our data would report a significant negative effect of pollution on child health. A researcher who adopted the seed papers' pressure-level definition would report a positive but insignificant one. Both would report first-stage $F$ statistics in the hundreds, and neither would have any internal indication that anything was wrong.

### 5.3 Diagnostics: what standard tests do and do not detect

We next apply the two validity diagnostics available in this setting: a placebo outcome test and over-identification tests.

Placebo outcome. Diarrhea in the past two weeks is measured in the same survey module as ARI, has similar prevalence (12 percent), and shares its reporting structure, but ambient PM2.5 has no plausible short-run causal pathway to diarrheal disease. A valid instrument should therefore produce a null reduced form on diarrhea. Family A passes cleanly (reduced-form $t = -0.62$ to $0.92$). However, family B does not. The China-standard count measure, sfc_days, moves diarrhea with $t = -2.71$, and the continuous measure sfc_diff is borderline ($t = -1.81$). The surface-layer instruments are correlated with something, plausibly the near-ground meteorology of cold, stagnant mornings, that affects child health through channels other than PM2.5.

Over-identification. Because the two families are constructed from the same reanalysis fields, they can be combined in a single over-identified model. Panel B of Table 5 reports the tests in two forms, the classical Sargan statistic and a cluster-robust Hansen $J$ consistent with the paper's standard errors, because the distinction matters at the margin. The count and continuous cross-family pairings are rejected under both: inv_days with sfc_days gives $J = 17.0$ by Sargan ($p < 10^{-4}$) and 13.9 cluster-robust ($p =$

0.0002); inv_days with sfc_diff gives 11.2 ($p = 0.0008$) and 9.3 ($p = 0.002$); all three together give 17.0 ($p = 0.0002$) and 14.2 ($p = 0.0008$). The strength pairing, stab_an with sfc_str, is rejected by the Sargan statistic ($J = 4.07, p = 0.044$) but not under clustering ($J = 2.41, p = 0.12$). Within family A, by contrast, the instruments are mutually consistent under both statistics: inv_days with inv_str gives $p = 0.72$ and 0.71, and inv_days with stab_an gives $p = 0.73$ and 0.78.

Within family B, the Sargan statistic rejects the pairing of sfc_days with sfc_diff ($J = 3.99, p = 0.046$) and is borderline for sfc_str with sfc_diff ($J = 3.17, p = 0.075$), though neither rejection survives cluster-robust weighting ($p = 0.12$ and 0.19). Therefore, the decisive evidence that the surface-layer family does not agree with itself comes from India (Section 5.4), where a count and a strength measure built from the identical temperature difference have opposite-signed first and second stages, rather than from these marginal statistics. Within the surface layer, the choice of functional form is itself a choice of instrument. The overall pattern is asymmetric in an informative way: the pressure-level family is coherent but detects nothing, whereas the surface-layer family produces significant estimates that fail a placebo test and, in South Asia, contradict one another outright.

Two features of these diagnostics deserve emphasis, because they delimit what applied researchers can and cannot learn from them. First, the placebo test flags only family B. A researcher using family A alone would pass every test available and still obtain an estimate that the cross-family evidence shows cannot be interpreted causally. Second, the over-identification test requires estimating both families, which no study in our coding exercise does. The problem is detectable, but only through a comparison the literature does not perform.

**5.4 Replication in South Asia**

A natural objection is that these results reflect something peculiar about Sub-Saharan Africa, such as its dust, its meteorology, or its comparatively moderate urban pollution levels. South Asia provides the opposite environment. The Indo-Gangetic Plain combines the world's most severe wintertime inversions with its highest urban PM2.5 concentrations, and Bangladesh adds a low-lying delta (median cluster elevation 13 m) where the pressure-level instrument's layers are unambiguously above ground. If the thermal-inversion design works anywhere in the developing world, it should work here. We rebuild the entire pipeline for India, Bangladesh, and Nepal (14 DHS surveys; 128,919 urban children), constructing all six instruments at 00 UTC, which is local dawn and the time at which surface inversions peak, rather than mechanically importing the 06 UTC convention appropriate for Africa (Section 6.4 examines timing directly).

India (Table 6, Panel A). The family split replicates on 104,676 children. Family A detects nothing robust. The instrument inv_days has a first stage of +0.88 μg/m³ per inversion day ($F = 182$) and a 2SLS estimate of −0.21 pp ($t = -0.62$), and the family's largest $t$ statistic, stab_an's $-1.99$, sits exactly at the five percent boundary while carrying a diarrhea placebo $t$ of 1.76. Family B again produces significant estimates with failing diagnostics. The strength measure sfc_str yields +3.52 pp per 10 μg/m³ ($t = 5.02$) while failing the diarrhea placebo ($t = -2.36$).

India also provides the sharpest piece of evidence in this paper. The count measure sfc_days and the strength measure sfc_str are constructed from the identical physical quantity, the difference between 925 hPa and 2-meter temperature, and differ only in functional form. One counts the days on which the difference is positive, and the other averages its positive part. Their first stages have opposite signs (+0.176 versus −1.467 μg/m³ per unit), and their second stages point in opposite directions (−2.50 pp, $t = -2.57$, versus +3.52 pp, $t = +5.02$). No stable structural relationship between atmospheric stability and pollution can generate this pattern. Instead, it can arise only if each functional form loads on a different bundle of correlated meteorology.

Bangladesh (Panel B). In the setting that looks ideal on paper, with sea-level terrain, extreme pollution, and strong winter inversions, all six instruments have significantly negative first stages. The mechanism is seasonal misalignment. PM2.5 peaks in December–February (96–121 μg/m³), driven by monsoon dynamics and transboundary transport, whereas inversion days peak in March (9.1 days per month), by which point PM2.5 has already fallen to 77 μg/m³. The monsoon months of June–September record essentially zero inversions and PM2.5 of 24–30 μg/m³. In other words, local atmospheric stability is not the binding constraint on ambient concentrations in a delta where pollution is imported and removed by rain. Section 6.3 develops this point.

Nepal (Panel C). The highest first-stage $F$ across the six definitions is 8.0, and five of the six fall below 2. Eighty-one percent of observations record zero inversion days, and the sample's bimodal elevation distribution (Terai plains near 100 m; Kathmandu valley near 1,300 m) places much of it where the pressure-level layers are below ground.

Summary. Across two continents, 250,000 children, and every published definition family, the thermal-inversion design produces one of three outcomes: (i) strong first stages with contradictory, definition-dependent second stages, at least one branch of which fails placebo tests (Africa, India); (ii) wrong-signed first stages in the theoretically most favorable setting (Bangladesh); or (iii) no identifying variation at all (Nepal). Section 6 examines why.

## 6. Results II: Why the Instrument Fails

The pattern documented in Section 5, strong first stages together with contradictory second stages and selectively failing placebos, is unlikely to be a statistical accident. This section traces it to four features of the setting, each observable ex ante and none specific to our data, and then to a fifth finding, about the exposure data themselves, that underlies the other four (Section 6.6).

### 6.1 The identifying variation is tiny, and large *F* statistics conceal it

The first-stage coefficient, not the $F$ statistic, measures how much pollution the instrument actually moves. In our Sub-Saharan African sample, one additional inversion day in the interview month shifts measured monthly PM2.5 by 1.26 μg/m³ (SE 0.11), five and a half percent of the sample mean of 23 μg/m³, and the surface-layer count moves it by 0.57 μg/m³. In India the corresponding figure is 0.88 μg/m³ at local dawn. The $F$ statistics attached to these coefficients, 132, 224, and 182 respectively, reflect the precision afforded by hundreds of thousands of observations rather than the size of the experiment the instrument runs.

The instructive comparison is with the papers that validated the design. Jans et al. (2018) report, in their abstract, that inversions raise PM10 by 25 percent. Sager (2019) reports 1.83 μg/m³ per °C of inversion strength, 13.7 percent of his sample mean. Arceo et al. (2016) report that one additional weekly inversion raises PM10 by 3 μg/m³, 3.4 percent of the mean, rising to 5.7 percent in their preferred specification. An instrument that moves exposure by a quarter of its mean can tolerate a good deal of direct meteorological influence on health before the pollution channel is swamped.

The applications that inherited the design report first stages one to three orders of magnitude smaller, and report them as successes. The most-cited China application reports 0.036 μg/m³ per annual inversion day, 0.07 percent of its mean, alongside a Kleibergen–Paap $F$ of 8,249. Zhang et al. (2022) report 0.0066 μg/m³ per inversion day, which they themselves describe as 0.01 percent of the mean, with an $F$ of 668. Our own instruments sit between the two groups, at 5.5 percent of the mean in Africa and 1.8 percent in India. Table 7, Panel A assembles these magnitudes. (Arceo et al. and Jans et al. instrument PM10, while the remaining rows are PM2.5, so the comparison is of relative perturbation rather than of a common quantity.)

Two implications follow. First, the exclusion restriction is not a fixed hurdle. Rather, it must hold to a tolerance set by how much the instrument moves exposure, and that tolerance tightens by orders of magnitude as one moves from the seed papers to their descendants. Any direct meteorological effect on health of comparable size, such as a cold morning's effect on respiratory infection or a stagnant day's effect on indoor crowding, overwhelms a first stage of 0.01 percent while leaving one of 25 percent intact. Second, magnitude is necessary but not sufficient, and this is where our own results enter. Our African first stage is in the seed papers' range and the design still fails, because the perturbation is entangled with season in a way that Sections 6.2 and 6.3 show cannot be undone in this setting. In other words, a weak-instrument problem in the economic sense can hide behind arbitrarily strong instruments in the statistical sense, and a first stage of respectable size can still identify the wrong experiment. We suggest that reporting the first stage in physical units per unit of instrument, relative to the exposure mean, should be as routine as reporting $F$. Both seed papers did so, as do four of the 26 citing studies.

Section 6.6 adds a further qualification. Because the satellite product understates within-city variation, the measured first stage need not equal the true one. Attenuation of the endogenous variable does not bias two-stage least squares, but it means that the size of the experiment cannot be verified at the scale where it runs. In other words, the perturbation is small as measured, and its true size cannot be verified with the exposure data used in this literature.

### 6.2 Relevance and validity trade off against each other

If the instrument's variation is small, one remedy is to seek settings where inversions have stronger effects. However, our data suggest that this remedy fails for a structural reason: the conditions that strengthen the first stage are the same conditions that break the exclusion restriction.

We split the African sample into terciles of the CAMS mineral-dust share of PM2.5 mass (Table 7, Panel B). This is the speciation measure introduced in Section 3.3, distinct from the aerosol-optical-depth share used for the 0.30 sample screen, which is why the top tercile's mean of 0.34 is not in tension with the screen. In the combustion-dominated tercile (dust mass fraction 0.01), the setting where theory says trapped surface emissions should make the instrument cleanest, the first stage is too weak to use, with $F = 5.6$. In the dust-dominated tercile (mass fraction 0.34), the first stage is comfortably strong ($F = 65$), but the diarrhea placebo now fails ($t = 2.97$). In the Sahel, the meteorology that accompanies inversions moves child health through channels that have nothing to do with locally trapped combustion aerosol. The middle tercile sits in between on both dimensions ($F = 28$, placebo $t = -0.21$). Therefore, we find no tercile in which the instrument is simultaneously strong and clean, and the full-sample estimates of Section 5 are averages over this trade-off. Figure 3 shows both halves of this pattern: panel (a) sets the first stage of our instruments against the published record in physical units, and panel (b) plots first-stage strength against the placebo statistic across the dust terciles.

### 6.3 Seasonality is the common confounder

Both failures appear to have a common root. Inversions are not randomly timed weather shocks but a marker of season, and in most of the developing world seasonality affects health through many channels at once.

Bangladesh illustrates this point most clearly. PM2.5 peaks in December–February at 96–121 μg/m³, driven by regional transport and the absence of monsoon washout, while inversion days peak in March (9.1 per month), when concentrations have already fallen by a third. The instrument is therefore negatively correlated with pollution, significantly so for all six definitions, not because inversions disperse pollutants but because both variables follow a seasonal cycle whose phases do not align. Any health outcome with its own seasonal phase will load on the instrument accordingly, with a sign and magnitude that depend on the local calendar rather than on any causal parameter.

The African sample provides a subtler version of the same problem. An antimalarial-use placebo, an outcome that ambient PM2.5 cannot plausibly affect but that the dry season strongly does, loads negatively on the instrument (−2.56 pp per 10 μg/m³, $t = -1.95$; 13,760 children across 44 surveys fielding the relevant module). The loading is essentially unchanged by month fixed effects (−2.62 pp, $t = -1.19$) and vanishes only under country-by-month fixed effects (+4.41 pp, $t = 0.19$), indicating that the instrument's seasonal content is country-specific and cannot be absorbed by generic seasonal controls. Because the module is fielded in only a subset of surveys, this test is under-powered, and we treat it as suggestive rather than decisive. On this point, the seed papers' setting and ours differ fundamentally. Arceo et al. (2016) absorb seasonality with bimonthly-by-municipality fixed effects, comparing weeks within the same municipality within the same two-month block, and enough inversion variation survives that comparison in a single basin city to identify their parameter. In contrast, we cannot make the analogous comparison. The outcome of interest shows the seasonal signature plainly on the full sample. The ARI coefficient rises monotonically as seasonal controls tighten (+0.61 pp, $t = 1.89$; +0.71 pp, $t = 1.65$; +0.99 pp, $t = 0.37$). We read this as evidence that seasonal confounding biases the headline estimate toward zero, with a caveat we state plainly: country-by-month fixed effects absorb most of the instrument's identifying variation, so the +0.99 is estimated too imprecisely to stand on its own. Therefore, the monotonicity, rather than the final coefficient, carries the information. The fact that the control the seed paper could afford is the control that empties our design is itself informative: the seasonal entanglement is not a nuisance left unmodeled but the source of most of the identifying variation in these settings. The direction of that bias, like its size, is an accident of local epidemiology rather than a property of the design.

### 6.4 Implementation details can manufacture results on their own

Two mechanical features of the instrument deserve separate mention because they operate before any economic analysis begins.

Acquisition timing. Reanalysis products report snapshots at fixed UTC hours, and surface inversions exist at dawn. The 06 UTC convention, appropriate for Africa (06:00–09:00 local), corresponds to local noon in South Asia. At that hour the instrument degenerates: 95.2 percent of Indian, 97.7 percent of Bangladeshi, and 99.7 percent of Nepali observations record zero inversion days. However, this degeneracy is not detectable from the usual first-stage statistics. In Nepal, the 26 non-zero observations, all from a single cluster in a single survey, generated a first-stage $F$ of 177 and a "significant" 2SLS estimate before we screened them out. In India, the noon-time first-stage coefficient is three and a half times its dawn value (3.08 against 0.88 μg/m³ per inversion day), so the degenerate instrument looks stronger rather than weaker. Therefore, a researcher who imported the African convention into South Asia, or the reverse, would obtain publishable numbers from an instrument that is nearly a constant.

Functional form and zero inflation. Even at the correct local hour, count-type definitions are heavily zero-inflated. Between 55 percent (Bangladesh) and 85 percent (Africa) of observations record no inversion at all, so identification rests on a thin, highly skewed tail. Continuous definitions eliminate the zeros but, as Section 5.4 showed, need not even share the count measure's sign. Therefore, the choice among functional forms, which the literature makes casually and singly, is in effect a choice among different experiments.

### 6.5 Alternative instruments do not rescue the design

Finally, we verify that the failure is not specific to thermal stability by constructing the two leading alternatives proposed for such settings.

A fire-transport instrument, built as upwind biomass-burning radiative power within 25–500 km, weighted by wind direction, from GFAS data for eleven West and Central African countries (67,477 children), satisfies every criterion in the standard toolkit: strong first stages ($F$ up to 783), correct raw correlations, and a passing diarrhea placebo. However, its 2SLS estimate is significantly negative (−1.33 pp per 10 µg/m³, $t = -2.30$), and the first stage collapses under country-by-month fixed effects ($F = 233$ → 9.9), revealing the same seasonal structure as the inversion instrument. Fires are a dry-season phenomenon, and the dry season is not excludable. A wind-direction × local population instrument in the spirit of Deryugina et al. (2019) shows the opposite pathology: a promising two-survey pilot ($F = 5.6$, correct sign) disappears at scale ($F = 0.1$ on 45 surveys), which suggests that the pilot signal was idiosyncratic rather than transportable. Our exploration of high-altitude cities, where three instrument families fail in three distinct ways, is reported in Appendix B.

The conclusion of this section is not that some particular implementation was flawed. Rather, in these settings the atmospheric variation that the design requires is largely absent. Stability shocks move pollution by too little, and the conditions that make them move it more also open direct channels from weather into health.

### 6.6 The exposure data do not resolve the contrasts the design uses

The failures documented above have a further source. The satellite exposure product does not measure pollution at the scale on which these designs rely. Yamada (2026) validates the product against every usable ground monitor in the public archive and reports a reliability of 0.786 for the temporal contrast and 0.748 for the between-place contrast, but 0.0097 for the within-city contrast, falling to 0.0010 on regulatory-grade instruments alone, where the confidence interval covers zero (Table 9). Only 1.4 percent of the field's spatial variance lies within a 0.10 degree cell, so exchanging the

0.01 and 0.10 degree releases is not a test of resolution. Three implications follow for this paper. First, our estimates remain consistent: classical measurement error in the endogenous variable attenuates the first stage but leaves two-stage least squares consistent, so nothing in Sections 5 and 6 is overturned. Second, the error is not guaranteed to be classical, because the product's sub-grid structure is generated from land-use predictors while our instruments are built from meteorology and geography, and this possibility reinforces rather than softens our conclusion. Third, the contrast our design uses is the temporal one, which is measured. What the validation removes is the presumption that the exposure side of these designs can be taken as given.

## 7. Implications and a Diagnostic Protocol

Our findings do not imply that reanalysis-based instruments should be abandoned. Rather, they imply that the burden of proof currently accepted in this literature, a large *F* and a citation to the seed papers, is too light. Table 8 collects the checks that our results show to be individually necessary. The first eight require no data beyond what every application already possesses. The ninth requires ground observations, and part of what it reveals is that in some regions none are available.

Most of this protocol is not new. As Section 2.4 documents, the two seed papers between them already did six of the first eight things we ask for. Arceo et al. (2016) let the instrument's content vary with municipal altitude (item 1), reported the first stage as both an absolute and a relative change in PM10 (item 5), ran a health-outcome placebo by splitting mortality into internal and externally caused deaths (item 7), and absorbed seasonality with bimonthly-by-municipality fixed effects while naming it as the threat to identification (item 8). Jans et al. (2018) aligned the instrument to local night and justified the restriction (item 2), reported that inversions occur on 25 percent of sample days (item 3), and reported the first stage in the abstract as a 25 percent increase in PM10 (item 5). Six of the first eight items therefore ask applied researchers to return to the seed papers' own standard rather than to adopt a new one. Items 4 and 6, the cross-definition and cross-family comparisons, are genuine additions, and they became necessary only after a single design fragmented into the family of mutually contradictory instruments that Table 2 documents. The ninth item is likewise new, and it became necessary once our validation exercise showed that the exposure side of these designs cannot be presumed adequate (Section 6.6).

1. Layer existence. Verify that the reference layers or pressure surfaces lie above ground across the study area's elevation distribution. (Forty-nine percent of our African clusters sit above the 1000 hPa surface; the most-cited China application changes layers with elevation in a footnote.)

2. Local-time alignment. Convert the reanalysis snapshot hours to local time and confirm the instrument is measured at dawn. (At local noon the instrument is 95–99.7 percent zeros in South Asia, and it still produced $F = 177$ from a single cluster.)

3. Distribution reporting. Report the instrument's zero share and support. (55–85 percent of observations are zero even at dawn; no coded study reports this.)

4. Cross-definition agreement. Estimate with at least one pressure-level and one surface-layer definition; report both. (Ours disagree in sign.)

5. First stage in physical units. Report μg/m³ per unit of instrument relative to the exposure mean, not only $F$. (0.036 μg/m³ per inversion day can coexist with $F = 8{,}249$.)

6. Cross-family over-identification. Where two families are available, run the $J$ test across them, in cluster-robust form. (Our count and continuous cross-family pairings reject at $p \leq 0.002$ cluster-robust and $p < 10^{-4}$ by Sargan; no coded study runs this test.)

7. Health-outcome placebo. Test the reduced form on a same-module outcome without a pollution pathway. (Arceo et al. (2016) do this by cause of death; among the 26 citing studies only Liu & Ao (2021) does, and none runs it separately by definition. It detects our family B.)

8. Seasonal-channel test. Verify robustness to country-by-month fixed effects and test a season-sensitive placebo. (Our antimalarial placebo loads on the instrument until country-month effects absorb it; Bangladesh's sign reversal is purely seasonal.)

9. Exposure validation at the design's scale. Compare the exposure product against ground observations at the spatial and temporal scale of the identifying contrasts, not only in levels. (Against African ground monitors, the satellite product attains a reliability of 0.748 for between-place contrasts and 0.786 for the temporal contrast, but only 0.0097 within cities, and 0.0010 on regulatory-grade instruments alone; Yamada 2026. Applications built on monitor readings, as most Chinese studies are, meet this item where monitors are dense; applications built on satellite products in monitor-sparse regions do not.)

Applied to the 26 coded studies on the first eight items, the protocol yields a clear verdict. None satisfies more than three items, and the modal study satisfies one (a strong $F$). We do not score the ninth item retrospectively, because it arises from a validation exercise that, to our knowledge, this paper is the first to run in this literature. We do not conclude that their estimates are wrong. However, on the evidence their own designs provide, readers cannot tell, and for the regions where the method is now heading, we argue that the presumption of validity should be reversed.

## 8. Conclusion

This paper set out to measure the effect of urban air pollution on child health in the fastest-urbanizing region on earth, using the profession's standard instrument for the task. It ends, however, as a study of the instrument itself. Across 26 African countries, and again across South Asia, six definitions of the thermal-inversion instrument, five of them drawn from published practice and all of them strong and conventionally validated, deliver mutually contradictory causal estimates, fail placebo and over-identification tests in patterned ways, and in the theoretically most favorable setting reverse sign entirely. These failures appear to be structural. Inversions in these regions move pollution by only a small amount, and the inversion measures are strongly correlated with season. Therefore, the design cannot isolate a pollution channel from the direct effects of seasonal meteorology.

Three implications follow. For producers of new estimates, the diagnostic protocol of Section 7 is cheap, and our results suggest it will be decisive more often than the literature expects. For consumers of existing estimates, the geographic concentration of the evidence base, together with our finding that definition choice alone can flip signs, counsels caution in transporting elasticities from Chinese basins to other settings, and suggests that some cross-study disagreement now attributed to context is an artifact of implementation. The same caution extends beyond the instrument. Any design that uses within-city exposure contrasts from satellite products in this region, whatever its identification strategy, rests on exposure data that ground monitors show cannot support them (Yamada 2026). For editors and referees, a strong first-stage $F$ should carry no evidentiary weight about validity. In our data, the strongest $F$ statistics attach to the instruments with the clearest exclusion failures.

Therefore, the question we set out to answer remains open, and it is too important to leave unanswered. The region will add close to a billion urban residents by mid-century, and their children will grow up in cities whose air quality is deteriorating measurably. Whether that air is damaging their health is a question that can be answered, but not with this tool. Yamada (2026) pursues a design that requires no atmospheric instrument at all, comparing siblings within mothers. That comparison identifies from the temporal contrast, the one contrast the exposure data do measure, and it finds that none of twenty-two child-health outcomes survives a multiple-testing correction, with power to detect effects several times smaller than the published benchmark. Better identification is therefore only half of what is missing. The other half is exposure data measured at the scale on which cities actually vary, whether from satellite products trained on local monitors, which are now emerging for parts of the continent (Westervelt et al. 2025), or from the ground networks that the ninth item of our protocol presupposes. In that sense, the negative results of this paper are a necessary first step toward a better answer.

## References


Ahmadi, M., B. Khorsandi, and M. Mesbah (2021). "The effect of air pollution on drivers' safety performance." *Environmental Science and Pollution Research* 28(13): 15768–15781. https://doi.org/10.1007/s11356-020-11687-y

Arceo, E., R. Hanna, and P. Oliva (2016). "Does the Effect of Pollution on Infant Mortality Differ Between Developing and Developed Countries? Evidence from Mexico City." *The Economic Journal* 126(591): 257–280. https://doi.org/10.1111/ecoj.12273

Balakrishnan, U., and M. Tsaneva (2021). "Air pollution and academic performance: Evidence from India." *World Development* 146: 105553. https://doi.org/10.1016/j.worlddev.2021.105553

Balakrishnan, U., and M. Tsaneva (2023). "Impact of Air Pollution on Mental Health in India." *The Journal of Development Studies* 59(1): 133–147. https://doi.org/10.1080/00220388.2022.2120804

Bondy, M., S. Roth, and L. Sager (2020). "Crime Is in the Air: The Contemporaneous Relationship between Air Pollution and Crime." *Journal of the Association of Environmental and Resource Economists* 7(3): 555–585. https://doi.org/10.1086/707127

Cao, W., G. Yang, Y. Ma, and Y. Choi (2025). "The Impact of Air Pollution on Subjective Well-Being in China: County-Level Evidence and Regional Heterogeneity." *SAGE Open* 15(4): 1–19. https://doi.org/10.1177/21582440251396312

Chen, S., C. Guo, and X. Huang (2018). "Air Pollution, Student Health, and School Absences: Evidence from China." *Journal of Environmental Economics and Management* 92: 465–497. https://doi.org/10.1016/j.jeem.2018.10.002

Chen, S., P. Oliva, and P. Zhang (2022). "The effect of air pollution on migration: Evidence from China." *Journal of Development Economics* 156: 102833. https://doi.org/10.1016/j.jdeveco.2022.102833

Chen, S., Y. Chen, Z. Lei, and J.-S. Tan-Soo (2020). "Impact of air pollution on short-term movements: evidence from air travels in China." *Journal of Economic Geography* 20(4): 939–968. https://doi.org/10.1093/jeg/lbaa005

Chen, X., L. Chen, W. Xie, N. D. Mueller, and S. J. Davis (2023). "Flight delays due to air pollution in China." *Journal of Environmental Economics and Management* 119: 102810. https://doi.org/10.1016/j.jeem.2023.102810

Colella, F., R. Lalive, S. O. Sakalli, and M. Thoenig (2019). "Inference with Arbitrary Clustering." *IZA Discussion Paper* 12584.

Conley, T. G. (1999). "GMM estimation with cross sectional dependence." *Journal of Econometrics* 92(1): 1–45. https://doi.org/10.1016/S0304-4076(98)00084-0

Cui, J., S. Huang, and C. Wang (2023). "The impact of air quality on innovation activities in China." *Journal of Environmental Economics and Management* 122: 102893. https://doi.org/10.1016/j.jeem.2023.102893

Deryugina, T., G. Heutel, N. H. Miller, D. Molitor, and J. Reif (2019). "The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction." *American Economic Review* 109(12): 4178–4219. https://doi.org/10.1257/aer.20180279

Deschênes, O., H. Wang, S. Wang, and P. Zhang (2020). "The effect of air pollution on body weight and obesity: Evidence from China." *Journal of Development Economics* 145: 102461. https://doi.org/10.1016/j.jdeveco.2020.102461

Fu, S., V. B. Viard, and P. Zhang (2021). "Air Pollution and Manufacturing Firm Productivity: Nationwide Estimates for China." *The Economic Journal* 131(640): 3241–3273. https://doi.org/10.1093/ej/ueab033

Gao, X., R. Song, and C. Timmins (2024). "The Fertility Consequences of Air Pollution in China." *Journal of the Association of Environmental and Resource Economists* 11(3): 657–688. https://doi.org/10.1086/726316

Gelaro, R., et al. (2017). "The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2)." *Journal of Climate* 30(14): 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Hersbach, H., et al. (2020). "The ERA5 global reanalysis." *Quarterly Journal of the Royal Meteorological Society* 146(730): 1999–2049. https://doi.org/10.1002/qj.3803

ICF. *Demographic and Health Surveys* (various years). Funded by USAID. Rockville, MD: ICF.

Imbens, G. W., and J. D. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2): 467–475. https://doi.org/10.2307/2951620

Inness, A., et al. (2019). "The CAMS reanalysis of atmospheric composition." *Atmospheric Chemistry and Physics* 19(6): 3515–3556. https://doi.org/10.5194/acp-19-3515-2019

Jans, J., P. Johansson, and J. P. Nilsson (2018). "Economic status, air quality, and child health: Evidence from inversion episodes." *Journal of Health Economics* 61: 220–232. https://doi.org/10.1016/j.jhealeco.2018.08.002

Kaiser, J. W., et al. (2012). "Biomass burning emissions estimated with a global fire assimilation system based on observed fire radiative power." *Biogeosciences* 9(1): 527–554. https://doi.org/10.5194/bg-9-527-2012 [GFAS]

Khanna, G., W. Liang, A. M. Mobarak, and R. Song (2025). "The Productivity Consequences of Pollution-Induced Migration in China." *American Economic Journal: Applied Economics* 17(2): 184–224. https://doi.org/10.1257/app.20220655

Li, J., and G. Meng (2023). "Pollution exposure and social conflicts: Evidence from China's daily data." *Journal of Environmental Economics and Management* 121: 102870. https://doi.org/10.1016/j.jeem.2023.102870

Liao, L., M. Du, and Z. Chen (2021). "Air pollution, health care use and medical costs: Evidence from China." *Energy Economics* 95: 105132. https://doi.org/10.1016/j.eneco.2021.105132

Liu, Y.-M., and C.-K. Ao (2021). "Effect of air pollution on health care expenditure: Evidence from respiratory diseases." *Health Economics* 30(4): 858–875. https://doi.org/10.1002/hec.4221

OpenAQ (2026). "OpenAQ Data Archive." OpenAQ, Washington, DC. https://openaq.org (public S3 archive, accessed July 2026)

Priem, J., H. Piwowar, and R. Orr (2022). "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts." *arXiv:2205.01833*. https://doi.org/10.48550/arXiv.2205.01833

Roodman, D., J. G. MacKinnon, M. Ø. Nielsen, and M. D. Webb (2019). "Fast and wild: Bootstrap inference in Stata using boottest." *The Stata Journal* 19(1): 4–60. https://doi.org/10.1177/1536867X19830877

Sager, L. (2019). "Estimating the effect of air pollution on road safety using atmospheric temperature inversions." *Journal of Environmental Economics and Management* 98: 102250. https://doi.org/10.1016/j.jeem.2019.102250

Schiavina, M., et al. (2023). *GHS-POP R2023A — GHS population grid multitemporal (1975–2030)*. European Commission, Joint Research Centre. https://doi.org/10.2905/2FF68A52-5B5B-4A22-8F40-C41DA8332CFE

Shao, S., L. Xu, and M. Fan (2022). "The effect of emission trading system on infant health: evidence from China." *Environmental Geochemistry and Health* 44(9): 3021–3033. https://doi.org/10.1007/s10653-021-01173-w

Shen, S., C. Li, A. van Donkelaar, N. Jacobs, C. Wang, and R. V. Martin (2024). "Enhancing Global Estimation of Fine Particulate Matter Concentrations by Including Geophysical a Priori Information in Deep Learning." *ACS ES&T Air* 1(5): 332–345. https://doi.org/10.1021/acsestair.3c00054

Stock, J. H., and M. Yogo (2005). "Testing for Weak Instruments in Linear IV Regression." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. D. W. K. Andrews and J. H. Stock, 80–108. Cambridge: Cambridge University Press.

Wang, L., Y. Dai, and D. Kong (2021). "Air pollution and employee treatment." *Journal of Corporate Finance* 70: 102067. https://doi.org/10.1016/j.jcorpfin.2021.102067

Westervelt, D. M., J. A. Amooli, and A. Anand (2025). "Twenty Years of High Spatiotemporal Resolution Estimates of Daily PM2.5 in West Africa Using Satellite Data, Surface Monitors, and Machine Learning." *ACS ES&T Air* 2(8): 1468–1477. https://doi.org/10.1021/acsestair.4c00366

Xu, L., H. Cai, and Z. Wang (2023). "Can air pollution reduce technology transfer? Evidence from China's prefecture-level cities." *Science of the Total Environment* 857: 159510. https://doi.org/10.1016/j.scitotenv.2022.159510

Yamada, H. (2026). "What Can Be Identified from Satellite PM2.5? Evidence from 2.9 Million African Births." SSRN Working Paper No. 7277842. https://ssrn.com/abstract=7277842

Ye, H.-J., Z. Huang, and S. Chen (2023). "Air pollution and agricultural labor supply: Evidence from China." *China Economic Review* 82: 102075. https://doi.org/10.1016/j.chieco.2023.102075

Zhang, C., M. Du, L. Liao, and W. Li (2022). "The effect of air pollution on migrants' permanent settlement intention: Evidence from China." *Journal of Cleaner Production* 373: 133832. https://doi.org/10.1016/j.jclepro.2022.133832

Zhang, G., Y. Ren, Y. Yu, and L. Zhang (2022). "The impact of air pollution on individual subjective well-being: Evidence from China." *Journal of Cleaner Production* 336: 130413. https://doi.org/10.1016/j.jclepro.2022.130413

Zhang, X., X. Zhang, Y. Liu, X. Zhao, and X. Chen (2023). "The morbidity costs of air pollution through the lens of health spending in China." *Journal of Population Economics* 36(3): 1269–1292. https://doi.org/10.1007/s00148-023-00948-y

Zhou, T., and N. Zhang (2023). "Does air pollution decrease labor share? Evidence from China." *Global Environmental Change* 82: 102706. https://doi.org/10.1016/j.gloenvcha.2023.102706

# Tables

## Table 1. Use of the thermal-inversion instrument

### Panel A: Size and growth of the corpus

| | Count |
|---|---|
| Distinct citing works of the four seed papers (OpenAlex) | 1,032 |
| Unique records after merging Scopus + WoS and deduplicating | 245 |
| **Records retained by screening (Scopus + WoS)** | **82** |
| **— of which verified false positive on reference checking** | 1 |
| **Confirmed studies using thermal inversions as an instrument** | **81** |
| Same, independent OpenAlex full-text query | 85 |
| — of which published in 2020 | 68 |
| — of which published in 2024 | 198 |

*Note for Panel A: Citing works are counted once each even when they cite more than one seed, so 1,032 is below the sum of the four seeds' individual citation counts (1,373 on the search date). The 2020 and 2024 rows count citing works by publication year on the same de-duplicated basis. The five counts drawn from OpenAlex (1,032; 85; 68; 198; and the 1,373 in this note) are recorded from queries run on 29 July 2026 against a live index; they are reported as recorded and, unlike the rows derived from the archived Scopus and Web of Science exports, cannot be regenerated from the replication archive. Counts from a live index also move over time.*

### Panel B: Confirmed studies by year (81)

| 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | 2025 | 2026 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 8 | 11 | 19 | 16 | 14 | 9 |

*Note for Panel B: The 2026 column runs through the search date, 29 July 2026. The row sums to 81; 2025 excludes the one verified false positive (Appendix E.5).*

### Panel C: Studies by study region (confirmed corpus, 81)

| Region | Studies |
|---|---|
| China | 62 (77%) |
| Europe | 5 |
| United States | 5 |
| India | 2 |
| No region identifiable from record | 10 |
| **Sub-Saharan Africa** | **0** |

*Note for Panel C: Regions are coded from titles, abstracts and keywords and are not mutually exclusive (three records carry two flags). The ten records with no identifiable region include the Taiwan-based study coded in full text (Liu & Ao 2021).*

**Panel D: Studies by outlet (confirmed corpus, 81)**

| Outlet | Studies |
| --- | --- |
| J. Environmental Economics and Management | 6 |
| Int. J. Environmental Research and Public Health | 4 |
| J. Development Economics | 3 |
| J. Assoc. Environmental and Resource Economists | 3 |
| Environmental and Resource Economics | 3 |
| J. Cleaner Production | 3 |
| All other outlets | 59 |

*Notes: The corpus is assembled by a systematic forward-citation search from four seed papers (Arceo, Hanna & Oliva 2016; Jans, Johansson & Nilsson 2018; Deschênes, Wang, Wang & Zhang 2020; Fu, Viard & Zhang 2021). Scopus and Web of Science citations are merged, deduplicated, and screened on title, abstract and keywords; an OpenAlex full-text query is run independently. Appendix E details the protocol.*

## Table 2. Implementation heterogeneity in the thermal-inversion-IV literature

*(Full-text coding of 26 studies; systematic search identified 82 records, of which one was verified to be a false positive; see Appendix E)*

### Panel A: Distribution of implementation choices

| Implementation element | Variant | Studies (of 26) |
|---|---|---|
| **Temperature source** | MERRA-2 | **20** |
| | NCEP | 1 |
| | Ground stations (MIDAS) + AIRS | 1 |
| | Reanalysis, product not named | 4 |
| **Layers compared (main specification)** | **110 m vs 320 m (MERRA-2's two lowest model layers, referenced to sea level)** | **17** |
| | 110 m vs 540 m | 1 |
| | 1000 hPa vs 925 hPa | 2 |
| | Surface air temperature vs 925 hPa | 2 |
| | A pressure surface and a geometric layer both cited | 1 |
| | Not stated precisely enough to classify | 3 |
| **Functional form (categories overlap)** | Count of inversion days or episodes | 14 |
| | Continuous strength | 11 |
| | Daily binary indicator | 6 |
| | Not determinable from the text | 1 |
| **Acquisition timing** | Fixed 6-hourly snapshots aggregated as given, no discussion of local time | **19** |
| | Local nighttime or dawn hours deliberately selected | 4 |
| | Timing not stated | 2 |
| | Not applicable (ground-station instrument) | 1 |
| **Robustness across inversion definitions** | Reports results under ≥ 2 distinct inversion definitions | 8 |
| | Single definition only | 18 |
| | **Of the 8: compare a pressure-level against a surface-layer definition** | **0** |
| | Adds a second instrument of a different kind (wind direction; coal plants) | 2 |

| | | |
|---|---|---|
| **Placebo / falsification (type categories overlap: one study appears in two)** | Reports some placebo or falsification exercise | 18 |
| | — random reassignment of treatment or instrument | 8 |
| | — future or lagged pollution substituted for current | 5 |
| | — an alternative outcome chosen to be unresponsive | 4 |
| | — the study's own controls used as outcomes | 1 |
| | — type not specified | 1 |
| | **A health outcome chosen to be unresponsive (the design used here)** | **1** |
| | **Reports a placebo that fails** | **0** |
| | **Applies a placebo separately to two inversion definitions** | **0** |
| **Over-identification** | Reports a Sargan / Hansen test | 4 (none rejects) |
| | **Instrument set spans both families** | **0** |
| **Instrument's zero share** | Reported at the aggregation used in estimation | **0** |
| **First-stage magnitude** | Reported in μg/m³ per unit of the instrument | 4 |

## Panel B: Overlooked technical problems, as they appear in print

| Problem | Appearance in the literature | Consequence documented in this paper |
|---|---|---|
| **Layer existence at altitude** | Fu et al. (2021, fn. 14): the layers are referenced to sea level, so cells above roughly 320 m have no valid first or second layer; they substitute "the two closest non-missing layers," under which the instrument's physical definition varies with elevation | 48.9 % of our African clusters lie above the 1000 hPa surface (≈ 110 m); Indian clusters reach 5,743 m and Nepali clusters 2,881 m |
| **Sea-level vs above-ground** | **4 of 26 describe the layer geometry incorrectly, or in terms that mix pressure and height: "approximately 110 m *above ground level*" (Cao et al. 2025); "inverse temperature gradient between *ground level* and the next higher level of 320 m" (Chen et al. 2020); "42 *barometric pressure* layers … the altitude of** | The confusion is systematic rather than incidental |

| | | |
|---|---|---|
| | **the different layers *from the ground*" (Xu et al. 2023); "42 *sea-level pressure* layers" (Ye et al. 2023). Two state it correctly; the remaining 20 do not describe it at all** | |
| **Acquisition timing** | 19 of 26 aggregate fixed 6-hourly snapshots with no discussion of local time; 2 more do not state a timing at all | At 06 UTC, local noon in South Asia, the instrument degenerates: 95–99.7 % zeros |
| **Zero inflation** | 0 of 26 report the instrument's zero share at the aggregation used in estimation. Several report the mean and range of an *annual* count, at which aggregation zeros essentially cannot arise (Zhang et al. 2022: minimum 4 inversion days per county-year) | Even at local dawn, 55–85 % of our observations are zero; one Nepali cluster alone produced F = 177 |
| **First-stage magnitude** | **Zhang et al. (2022): 0.0066 µg/m³ per inversion day (the authors themselves write "0.01 % of the sample mean") alongside KP F = 668.3 and a reported 2SLS effect of 2.45 standard deviations. Within that one paper KP F ranges from 2.0 to 668.3, and coefficients are interpreted at F = 2.0** | A large F conceals identifying variation far too small to carry the reported effect |
| **The seed papers' own practice** | **Arceo et al. (2016): inversions declared from balloon soundings "upon finding non-monotonic temperature gradients" anywhere in the profile, not from a two-layer difference; weekly counts; the instrument is interacted with municipal altitude (to instrument two pollutants at once), on the stated reasoning that "PM10 concentrations are lower at higher altitudes, presumably because higher areas are likely to be above the thermal inversion layer"; bimonthly × municipality fixed effects; weather enters as a 4th-order polynomial in mean temperature and 2nd-order polynomials in precipitation, cloud cover and humidity; a health-outcome placebo splitting mortality into internal and** | **The design as practiced in the literature is not the design that was validated. Jans et al.'s definition *is* our family A: coherent, placebo-passing, and null. The modal implementation (17 of 26) is the MERRA-2 surface-layer pair at fixed UTC hours, which matches neither seed, and is our family B: significant, placebo-failing, rejecting against family A, and in South Asia contradicting itself outright. On the Table 8 protocol the two seeds satisfy six of eight; no citing study satisfies more than three and the modal study satisfies one** |

| | | |
|---|---|---|
| | **externally caused deaths ("this can be viewed as a placebo test"); and the statement that temperature controls are needed "since inversions have a clear seasonal pattern." Jans et al. (2018): AIRS profiles, "the two pressure levels closest to the ground (1000 hPa and 925 hPa)," the 925 hPa layer described as measuring "conditions at approximately 600 m above sea level"; nighttime episodes only, reported to occur on 25 % of sample days; count and strength forms; first stage reported in the abstract as +25 % of PM10** | |
| **Robustness that changes the answer** | 8 of 26 test a second inversion definition. When the coefficient moves, it is reported as confirmation: "only the regression coefficient was different … our results remain robust and reliable" (Zhang et al. 2022, on switching from the 320 m to the 540 m layer) | The definitions carry opposite-signed information; the literature's own robustness checks have been recording this without recognizing it |

## Panel C: Literature lineages map onto our instrument families

| Lineage | Representative studies | Our family | Result on our SSA sample |
|---|---|---|---|
| **Seed, sounding-based: profile non-monotonicity** | Arceo, Hanna & Oliva 2016 (EJ) | **neither (no two-layer difference)** | not reproducible from reanalysis |
| **Seed, reanalysis-based: 1000−925 hPa, nighttime** | **Jans, Johansson & Nilsson 2018 (JHE); followed by Balakrishnan & Tsaneva 2021 (World Dev), 2023 (JDS)** | **A (pressure level)** | **Positive but insignificant; placebo passes; internally coherent** |
| **Modal implementation (17 of 26): MERRA-2 110/320 m, fixed UTC** | Fu et al. 2021 (EJ); Deschênes et al. 2020, Chen et al. 2022 (JDE); Chen et al. 2018 (JEEM) | **B (surface layer)** | **Significantly negative; diarrhea placebo fails; rejects against family A, and in South Asia contradicts itself outright** |
| UK lineage: surface−925 hPa (2 of 26) | Sager 2019 (JEEM); Bondy, Roth & Sager 2020 (JAERE) | B variant (sfc_diff) | Negative, borderline |

## Table 3. Descriptive statistics by country

### Panel A: Sub-Saharan Africa, selected countries

| Country | N | Surveys | Clusters | ARI (%) | PM2.5 | Elev. (m) | Inv. days | Zero (%) |
|---|---|---|---|---|---|---|---|---|
| Nigeria | 22,781 | 5 | 1,447 | 3.1 | 28.6 | 107 | 0.65 | 80.9 |
| Congo DR | 10,793 | 3 | 399 | 8.5 | 31.0 | 427 | 0.01 | 99.5 |
| Gabon | 7,735 | 2 | 425 | 8.6 | 21.8 | 69 | 0.02 | 98.5 |
| Mozambique | 7,671 | 3 | 598 | 8.0 | 11.8 | 49 | 2.66 | 48.5 |
| Cameroon | 7,413 | 3 | 549 | 10.2 | 27.9 | 612 | 0.01 | 98.5 |
| Benin | 7,394 | 3 | 496 | 6.3 | 43.0 | 38 | 2.30 | 47.4 |
| Angola | 7,375 | 2 | 413 | 4.0 | 17.7 | 70 | 1.30 | 84.7 |
| Ghana | 6,497 | 5 | 845 | 6.6 | 24.1 | 144 | 0.23 | 85.1 |
| Kenya | 4,635 | 4 | 381 | 7.4 | 11.4 | 148 | **0.00** | **100.0** |
| Tanzania | 3,759 | 4 | 312 | 6.8 | 13.3 | 40 | **0.00** | **100.0** |
| Namibia | 1,091 | 3 | 129 | 11.0 | 10.5 | 30 | **18.15** | 28.2 |
| Ethiopia | 1,465 | 5 | 94 | 7.7 | 17.3 | 455 | **0.00** | **100.0** |
| Uganda | 354 | 3 | 20 | 16.9 | 30.5 | 945 | **0.00** | **100.0** |
| **Total / mean (26 countries)** | **120,820** | **80** | **8,456** | **7.3** | **23.0** | — | **0.75** | **85.1** |

*Notes for Panel A: The panel lists the largest country samples together with countries illustrating the range of the inversion-day distribution; the total row covers all 26 countries, and the complete country listing is available in the replication package. Eight countries (Kenya, Tanzania, Burundi, Zambia, Ethiopia, Guinea, Uganda and Rwanda) record no pressure-level inversion day at all, and five more (Malawi, Gabon, Cameroon, Congo DR, Sierra Leone) are at or above 98 percent zeros. Thirteen of the 26 countries therefore supply essentially no identifying variation for family A.*

### Panel B: South Asia

| Country | N | Surveys | Clusters | ARI (%) | PM2.5 | Elev. (m) | Inv. days | Zero (%) |
|---|---|---|---|---|---|---|---|---|
| India | 104,676 | 2 | 15,863 | 4.9 | 49.3 | 226 | 2.05 | 65.1 |
| Bangladesh | 15,953 | 7 | 1,230 | 11.9 | 63.8 | 13 | 1.83 | 54.8 |
| Nepal | 8,290 | 5 | 706 | 8.1 | 46.3 | 206 | 1.28 | 81.4 |
| **Total / mean** | **128,919** | **14** | **17,799** | **6.0** | **50.9** | — | **1.97** | **64.9** |

*Notes: N is the number of children in the analysis sample; ARI and PM2.5 (μg/m³) are sample means; "Elev." is the median cluster elevation in meters. "Inv. days" is the monthly mean of the instrument inv_days, measured at 06 UTC in Africa and 00 UTC in South Asia, local dawn in both regions. "Zero" is the share of child observations recording no inversion day in the interview month; the corresponding cluster-month share for the African sample is 83.7 percent.*

**Table 4. Construction of the six instruments**

| Family | Instrument | Definition | Interpretation | Literature lineage |
|---|---|---|---|---|
| **A. Pressure level** | inv_days | # days with T925 − T1000 > 0 | Count of inversion days aloft | **Jans et al. (2018), count form; followed by Balakrishnan & Tsaneva (2021, 2023)** |
| | inv_str | mean of max(0, T925 − T1000) | Mean inversion strength aloft | **Jans et al. (2018), strength form** |
| | stab_an | mean of [(T925 − t2m) + 6.5·(762 − h)/1000] | Deviation from the standard lapse rate, elevation-corrected | Our elevation-robust variant (no direct precedent) |
| **B. Surface layer** | sfc_days | # days with T925 − t2m > 0 | Count of surface-based inversion days | China standard (MERRA-2 110 m/320 m layers), count form: Fu et al. 2021; Chen et al. 2018; Deschênes et al. 2020 |
| | sfc_str | mean of max(0, T925 − t2m) | Mean surface-based inversion strength | China standard, strength form: Chen, Oliva & Zhang (2022) |
| | sfc_diff | mean of (T925 − t2m), signed | Continuous surface-layer stability | Sager (2019, JEEM); Bondy, Roth & Sager (2020, JAERE) |

*Notes: All quantities are computed from ERA5 daily snapshots at local dawn (06 UTC in Africa; 00 UTC in South Asia) and aggregated over the child's interview month. T1000 and T925 are temperatures at the 1000 and 925 hPa pressure surfaces; t2m is 2-meter air temperature; h is cluster elevation in meters. Family A is the seed definition: Jans, Johansson and Nilsson (2018) use "the two pressure levels closest to the ground (1000 hPa and 925 hPa)" from AIRS, in both count and strength form, restricted to nighttime episodes, which inv_days and inv_str reproduce; the other seed, Arceo et al. (2016), declares inversions from balloon-sounding profiles rather than from any two-layer difference, so no reanalysis instrument reproduces their measurement (Section 2.4). The instrument stab_an is built from the surface-layer difference T925 − t2m but adds an elevation correction at the standard lapse rate of 6.5 °C/km, which removes the dependence of that difference on how far the cluster sits below the 925 hPa surface; what remains references a fixed altitude rather than the ground, which is why it is grouped with family A. The grouping is corroborated rather than imposed: stab_an is mutually consistent with inv_days in the over-identification tests (J = 0.11, p = 0.73) and rejects against the surface-layer sfc_str (J = 4.07, p = 0.044), and the family split in Table 5 is unchanged by its exclusion, since inv_days and inv_str alone reproduce it. The 110 m/320 m MERRA-2 layers used in the China literature are referenced to sea level and describe the lowest few hundred meters of the atmosphere, the same physical quantity as the surface-layer family. Zero shares of the count and strength forms in the African sample are 85.1 percent (family A) and 62.0 percent (family B) of observations; the continuous forms have no zeros by construction.*

**Table 5. Six instruments, one design: Sub-Saharan Africa**

**Panel A: First stage, 2SLS, and placebo by instrument definition**

| Family | Instrument | First stage (μg/m³ per unit) | F | 2SLS: ARI per 10 μg/m³ | t | Diarrhea placebo (reduced-form t) |
|---|---|---|---|---|---|---|
| A | inv_days (count, T925>T1000) | **+1.259 (0.110)** | 132.0 | +0.71 pp | 1.65 | 0.92 |
| A | inv_str (strength, 925−1000) | **+4.569 (0.571)** | 64.0 | +1.02 pp | 1.05 | 0.71 |
| A | stab_an (lapse-rate anomaly) | **+0.419 (0.056)** | 57.1 | +0.34 pp | 0.26 | −0.62 |
| B | sfc_days (count, T925>t2m) | **+0.567 (0.038)** | 223.7 | **−1.64 pp** | **−3.05** | **−2.71** |
| B | sfc_str (strength, 925−t2m) | **+2.874 (0.348)** | 68.1 | −1.91 pp | −2.17 | −0.97 |
| B | sfc_diff (signed continuous) | **+3.543 (0.188)** | 356.1 | −0.94 pp | −2.16 | −1.81 |

*Note for Panel A: First-stage coefficients are μg/m³ of PM2.5 per unit of instrument (per day for the count forms, per °C for the strength and continuous forms), with standard errors clustered by DHS cluster in parentheses. All six first stages are positive and significant. The diarrhea placebo reports the reduced-form t statistic of each instrument on diarrhea; the bold entry for sfc_days rejects the null at the five percent level, a placebo failure.*

**Panel B: Over-identification (Sargan and cluster-robust Hansen *J*)**

**Cross-family pairings**

| Instrument pair | 2SLS | Sargan J | p | Robust J | p |
|---|---|---|---|---|---|
| inv_days + sfc_days (A×B) | −0.52 pp | 17.01 | **$3.7\times10^{-5}$** | 13.88 | **$2.0\times10^{-4}$** |
| inv_days + sfc_diff (A×B) | −0.39 pp | 11.18 | **$8.3\times10^{-4}$** | 9.27 | **0.0023** |
| stab_an + sfc_str (A×B) | −1.38 pp | 4.07 | **0.044** | 2.41 | 0.120 |
| All three (inv_days + sfc_days + sfc_diff) | −0.54 pp | 17.02 | **$2.0\times10^{-4}$** | 14.15 | **$8.5\times10^{-4}$** |

**Within family A: mutually consistent under both statistics**

| Instrument pair | Sargan J | p | Robust J | p |
|---|---|---|---|---|
| inv_days + inv_str | 0.13 | 0.72 | 0.14 | 0.71 |
| inv_days + stab_an | 0.11 | 0.73 | 0.08 | 0.78 |

**Within family B**

| Instrument pair | Sargan J | p | Robust J | p |
|---|---|---|---|---|
| sfc_days + sfc_diff | 3.99 | **0.046** | 2.36 | 0.124 |
| sfc_str + sfc_diff | 3.17 | 0.075 | 1.75 | 0.186 |

*Note for Panel B: The robust J is the two-step GMM statistic with moments clustered at the DHS cluster, matching the paper's standard errors. The count and continuous cross-family pairings reject under both statistics; the strength pairing and the within-B pairings reject only under the Sargan statistic, and the decisive evidence of within-B incoherence is India's sign flip (Table 6), not these marginal statistics. Within family A, both statistics accept. The asymmetry is informative: the pressure-level family is internally coherent but detects nothing, while the surface-layer family yields significant estimates that fail the placebo and, in South Asia, contradict one another outright.*

*Notes: The sample contains 120,820 children in 8,456 urban clusters across 26 countries and 80 DHS survey rounds, restricted to clusters below 1,000 m elevation with an aerosol-optical-depth dust share below 0.30 (Section 3.4). Each estimate is from a two-stage least squares regression of ARI on PM2.5 (in μg/m³; effects are reported in percentage points per 10 μg/m³) with survey, region, and month fixed effects and controls for monthly mean temperature and wind speed. Standard errors are clustered by DHS cluster.*

## Table 6. Replication in South Asia

### Panel A: India

| Family | Instrument | First stage (μg/m³/unit) | F | 2SLS per 10 μg/m³ | t | Placebo t |
|---|---|---|---|---|---|---|
| A | inv_days | +0.878 | 181.7 | −0.21 pp | −0.62 | −0.29 |
| A | inv_str | +9.615 | 93.4 | −0.53 pp | −1.27 | −0.31 |
| A | stab_an | +1.560 | 267.2 | −0.66 pp | −1.99 | 1.76 |
| B | **sfc_days** | **+0.176** | 41.4 | **−2.50 pp** | **−2.57** | −0.17 |
| B | **sfc_str** | **−1.467** | 92.5 | **+3.52 pp** | **+5.02** | **−2.36** |
| B | sfc_diff | −0.251 | 4.8 | +17.57 pp | 2.06 | −1.01 |

*Note for Panel A: sfc_days and sfc_str are built from the same physical quantity (T925 − t2m) and differ only in functional form; the count form has a positive first stage, the strength form a negative one, and their 2SLS estimates have opposite signs, direct evidence that the instrument does not capture a stable physical mechanism. The bold placebo entry for sfc_str rejects at the five percent level, a placebo failure.*

### Panel B: Bangladesh (all six first stages negative)

| Instrument | First stage | F | 2SLS per 10 μg/m³ | t |
|---|---|---|---|---|
| inv_days | −1.555 | 26.3 | +0.44 pp | 0.46 |
| inv_str | −36.492 | 42.4 | −0.18 pp | −0.20 |
| stab_an | −9.267 | 68.2 | −0.01 pp | −0.01 |
| sfc_days | −0.936 | 22.9 | +0.24 pp | 0.21 |
| sfc_str | −10.527 | 18.5 | −0.25 pp | −0.20 |
| sfc_diff | −9.307 | 68.5 | +0.03 pp | 0.04 |

*Note for Panel B: Inversion days peak in March (9.1 per month) after PM2.5 has already fallen from its December–February peak of 96–121 μg/m³; monsoon dynamics and transboundary transport, not local inversions, drive seasonality (Section 6.3). Median cluster elevation is 13 m, the theoretically ideal setting.*

### Panel C: Nepal (no identifying variation)

| Instrument | F |
|---|---|
| inv_days | 8.0 |
| inv_str | 1.5 |
| stab_an | 1.9 |
| sfc_days | 0.4 |
| sfc_str | 0.2 |
| sfc_diff | 0.8 |

*Note for Panel C: 81 percent of observations record zero inversion days, and the elevation distribution is bimodal (Terai plains near 100 m; Kathmandu valley near 1,300 m), placing much of the sample where the pressure-level layers lie below ground. First stages are too weak for the 2SLS estimates to be informative, so only the F statistics are reported.*

*Notes: Urban DHS clusters with no elevation or dust filter; instruments are measured at 00 UTC, local dawn. India N = 104,676; Bangladesh N = 15,953; Nepal N = 8,290. The specification is that of Table 5: 2SLS of ARI on PM2.5 with survey, region, and month fixed effects, temperature and wind controls, and standard errors clustered by DHS cluster. First-stage coefficients are μg/m³ of PM2.5 per unit of instrument.*

## Table 7. Mechanisms

### Panel A: What the instrument moves, in physical units

| Study / sample | Setting | Instrument | First stage | F | Exposure mean | First stage as % of mean |
|---|---|---|---|---|---|---|
| *The papers that validated the design* | | | | | | |
| **Jans et al. (2018)** | Sweden | nighttime inversion episode | reported as a relative effect | — | PM10 | **+25 %** |
| Sager (2019) | United Kingdom | continuous strength | +1.83 µg/m³ per °C | 327 | 13.3 | 13.7 % |
| **Arceo et al. (2016)** | Mexico City | weekly inversion count | **+3 µg/m³ per inversion** | — | PM10 | **3.4 % (5.7 % preferred spec.)** |
| *This paper* | | | | | | |
| This paper | Sub-Saharan Africa | inv_days (monthly) | **+1.26 µg/m³ per day** | 132 | 23.0 µg/m³ | 5.5 % |
| This paper | Sub-Saharan Africa | sfc_days (monthly) | +0.57 µg/m³ per day | 224 | 23.0 | 2.5 % |
| This paper | India | inv_days (monthly, 00 UTC) | +0.88 µg/m³ per day | 182 | 49.3 | 1.8 % |
| *The applications that inherited it* | | | | | | |
| Balakrishnan & Tsaneva (2023) | India | nighttime inversions | +0.10 µg/m³ per event | 71 (eff.) | — | — |
| Fu et al. (2021) | China | annual inversion days | **+0.036 µg/m³ per day** | **8,249** | 53.5 | 0.07 % |
| **Zhang et al. (2022)** | China | annual inversion days | **+0.0066 µg/m³ per day** | **668** | 72.4 | **0.01 %** |

*Note for Panel A: Aggregation windows differ (weekly, monthly, annual) and Arceo et al. and Jans et al. instrument PM10 while the remaining rows are PM2.5, so the final column, the perturbation relative to the exposure mean, is the only comparable quantity. The papers that validated the design move exposure by 3.4–25 percent of its mean; the applications that inherited it report 0.01–0.07 percent and treat a large F as sufficient. Our own instruments fall between the two groups, which is why magnitude alone does not explain our results: Sections 6.2 and 6.3 show that the perturbation is entangled with season in a way that no fixed effect available to us can undo. The Zhang et al. (2022) row is the extreme case: the authors themselves report the first stage as 0.01 percent of the sample mean, pair it with a Kleibergen–Paap F of 668 and a second-stage effect of 2.45 standard deviations, and interpret coefficients elsewhere in the paper at F = 2.0.*

## Panel B: Relevance and validity trade off (African sample, dust terciles)

| Dust tercile | Mean dust mass fraction of PM2.5 | Countries | First-stage F (inv_days) | 2SLS (pp / 10 μg/m³) | Diarrhea placebo t |
|---|---|---|---|---|---|
| Low (combustion-dominated) | 0.01 | 23 | **5.6** | +6.80 (t = 0.79) | −0.77 |
| Middle | 0.11 | 25 | 28.4 | +2.07 (t = 1.56) | −0.21 |
| High (Sahel-type) | 0.34 | 18 | **65.4** | +0.44 (t = 0.72) | **2.97** |

*Note for Panel B: Terciles are cut on the CAMS mineral-dust share of PM2.5 mass (the speciation of Appendix D.3), not on the aerosol-optical-depth dust share used for the 0.30 sample screen, which is why the top tercile's mean of 0.34 can exceed the screen. The instrument is inv_days; the 2SLS column reports the effect of PM2.5 on ARI in percentage points per 10 μg/m³.*

## Panel C: Seasonality (African sample)

| Fixed effects | ARI, 2SLS (pp / 10 μg/m³) | t | Antimalarial placebo (pp / 10 μg/m³) | t |
|---|---|---|---|---|
| Survey + region | +0.61 | 1.89 | −2.56 | −1.95 |
| + month | +0.71 | 1.65 | −2.62 | −1.19 |
| + country × month | +0.99 | 0.37 | +4.41 | 0.19 |

*Note for Panel C: The instrument is inv_days. Antimalarial use is observed for 13,760 children across the 44 surveys fielding the module; the test is under-powered and reported as suggestive. The monotone rise in the ARI coefficient as seasonal controls tighten indicates downward bias from seasonal confounding.*

## Panel D: Bangladesh (seasonal misalignment)

| Month | 1 | 2 | 3 | 4–5 | 6–9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|
| PM2.5 (μg/m³) | 121 | 99 | 77 | 37–38 | 24–30 | 51 | 83 | 96 |
| Inversion days | 2.9 | 3.2 | **9.1** | 1.9–2.7 | 0.0–0.3 | 0.5 | 1.8 | 1.8 |

*Note for Panel D: Cells are monthly means over the Bangladeshi sample. Pollution peaks in December–February; inversions peak in March. All six first stages are negative (Table 6, Panel B).*

## Panel E: Acquisition timing and degeneracy (South Asia)

| Country | Zero share of inv_days, 00 UTC (local dawn) | 06 UTC (local noon) | Distinct values at 06 UTC |
|---|---|---|---|
| India | 65.1 % | **95.2 %** | 14 |
| Bangladesh | 54.8 % | **97.7 %** | 4 |
| Nepal | 81.4 % | **99.7 %** | **2** |

*Note for Panel E: The 06 UTC snapshot is local noon in South Asia, after morning surface inversions have eroded. Before screening, Nepal's 26 non-zero observations (all from one cluster in one survey) produced a first-stage F of 177 and a nominally significant 2SLS estimate.*

**Table 8. A diagnostic protocol**

| # | Check | What to report | Evidence that it binds |
|---|---|---|---|
| 1 | **Layer existence** | Share of the sample above the reference surface, by elevation | 48.9 % of our African clusters lie above 1000 hPa (≈110 m); Fu et al. (2021, fn. 14) substitute layers by elevation |
| 2 | **Local-time alignment** | Reanalysis snapshot hours converted to local time | 06 UTC = local noon in South Asia → 95–99.7 % zeros |
| 3 | **Instrument distribution** | Zero share, distinct values, skewness | 55–85 % zeros even at dawn; no coded study reports this |
| 4 | **Cross-definition agreement** | Estimates from ≥ 1 pressure-level and ≥ 1 surface-layer definition | Ours disagree in sign (Table 5) |
| 5 | **First stage in physical units** | μg/m³ per unit of instrument, relative to exposure mean | 0.036 μg/m³ per day with F = 8,249 (Fu et al. 2021) |
| 6 | **Cross-family over-identification** | Sargan and cluster-robust Hansen J across families | Count and continuous pairings reject at $p \leq 0.002$ robust, $p < 10^{-4}$ Sargan (Table 5B) |
| 7 | **Health-outcome placebo** | Reduced form on a same-module outcome without a pollution pathway | Diarrhea: family B fails (t = −2.71). Arceo et al. (2016) run one (internal vs externally caused deaths); among the 26 citing studies only Liu & Ao (2021) does, and none runs it separately by definition |
| 8 | **Seasonal-channel test** | Robustness to country × month FE; season-sensitive placebo | ARI coefficient rises +0.61 → +0.99 pp as seasonal controls tighten |
| 9 | **Exposure validation at the design's scale** | Exposure product against ground observations, at the spatial and temporal scale of the identifying contrasts | Against African ground monitors the product attains a reliability of 0.748 between places and 0.786 over time, but 0.0097 within cities and 0.0010 on regulatory-grade instruments alone (Table 9; Yamada 2026) |

*Note: Applied retrospectively to the 26 applications coded in full text, on the first eight items: none satisfies more than three; the modal study satisfies one (a large first-stage F). The ninth item arises from this paper's validation exercise (Section 6.6) and is not scored retrospectively.*

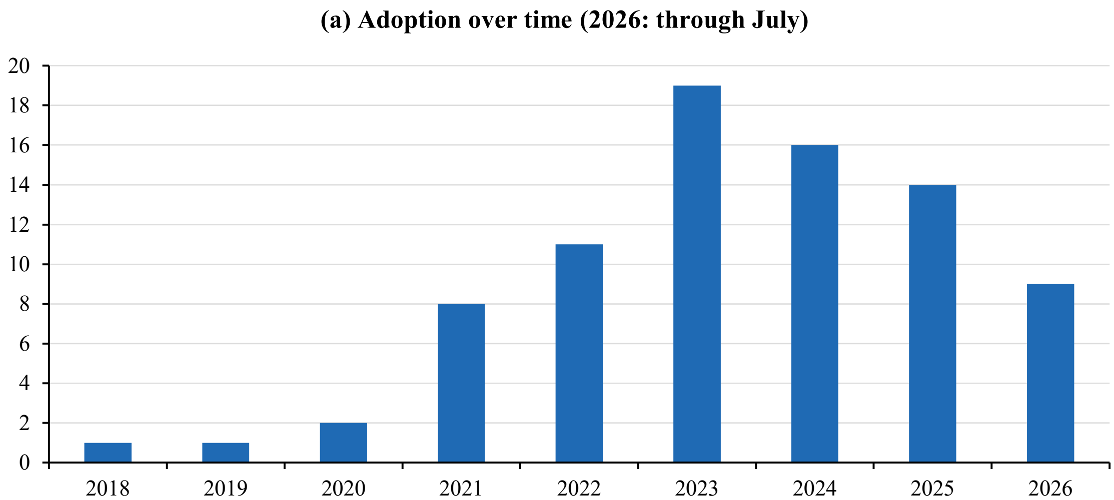
**Table 9. Reliability of the satellite exposure, by identifying contrast**

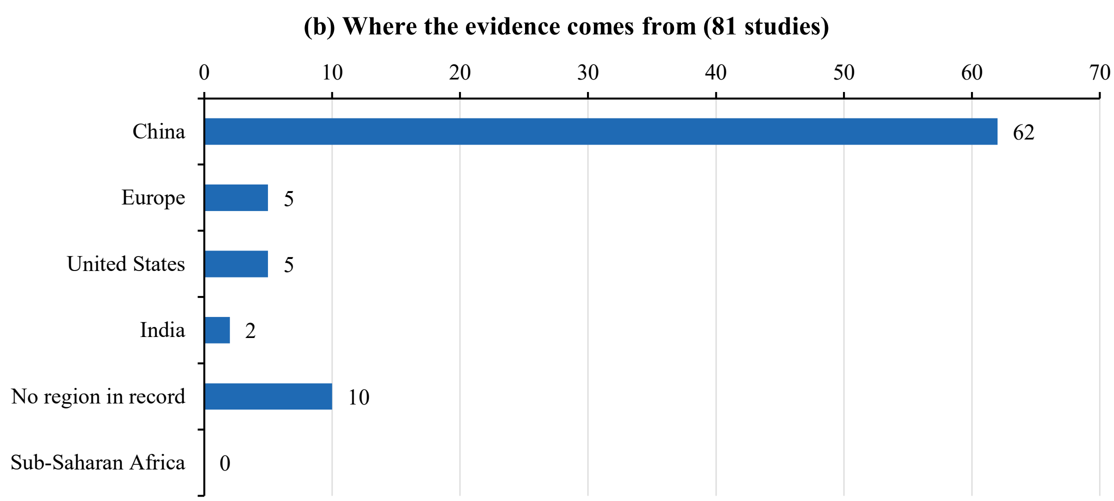
| Contrast | Reliability | Detail |
|---|---|---|
| Temporal (cluster over time, with month fixed effects) | 0.786 | the contrast this paper's instruments move |
| Between places | 0.748 | regulatory-grade monitors, 10 cities; correlation 0.865 |
| Within cities, all sites | 0.0097 | median separation 41.6 km |
| Within cities, regulatory-grade instruments only | 0.0010 | confidence interval covers zero |

*Notes: Reproduced from Yamada (2026), which validates the ACAG V6GL03 product against the ground monitors of the public OpenAQ archive. Reliability is the share of the variance of the ground-measured contrast that the satellite product reproduces. The design of this paper is identified from the temporal contrast (row 1). Only 1.4 percent of the product's spatial variance lies within a 0.10° cell, so the 0.01° release is a smooth interpolation of the 0.10° field rather than an independent measurement at 1.1 km. The within-city rows are the reason item 9 of Table 8 exists; they do not bear directly on the estimates of this paper, which rest on the temporal contrast, but they do bear on any design identified from fine spatial contrasts in this region.*

## Figures

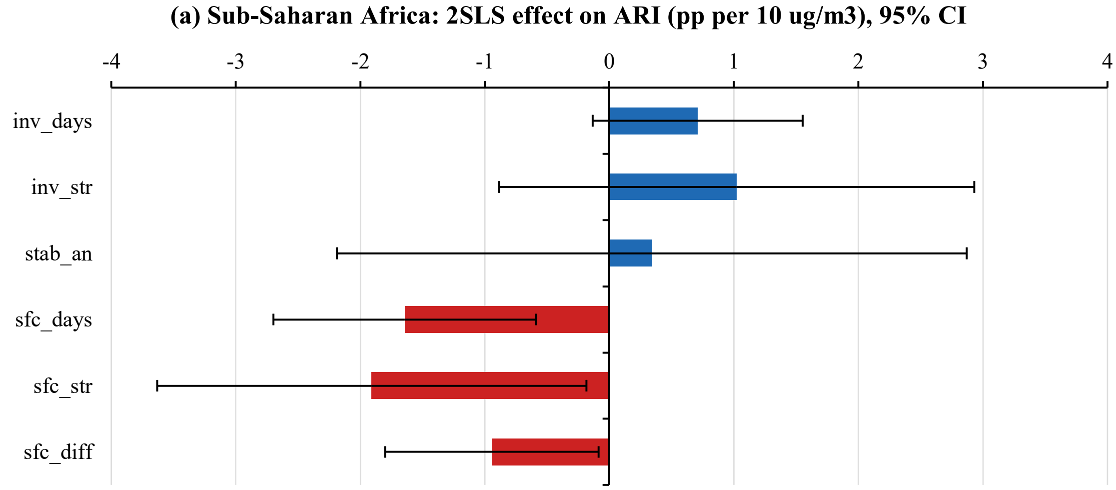
### Figure 1. The thermal-inversion-IV literature: adoption over time and study regions.

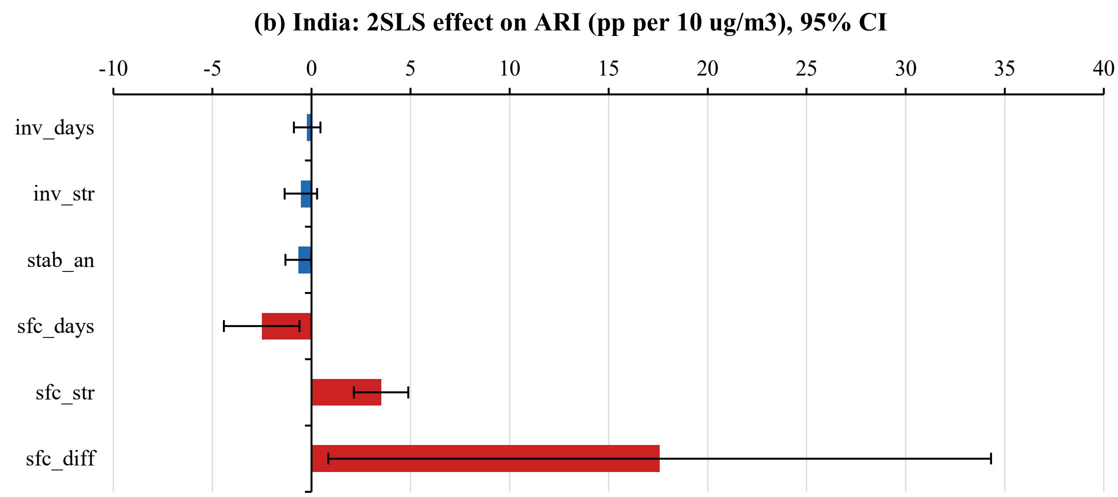



*Notes: The confirmed corpus is the 81 studies retained by the screening of Section 2.2 (Table 1). Panel (a) counts confirmed studies by publication year; the 2026 bar runs through the search date, 29 July 2026. Panel (b) codes study regions from titles, abstracts and keywords; regions are not mutually exclusive. Sub-Saharan Africa appears zero times.*

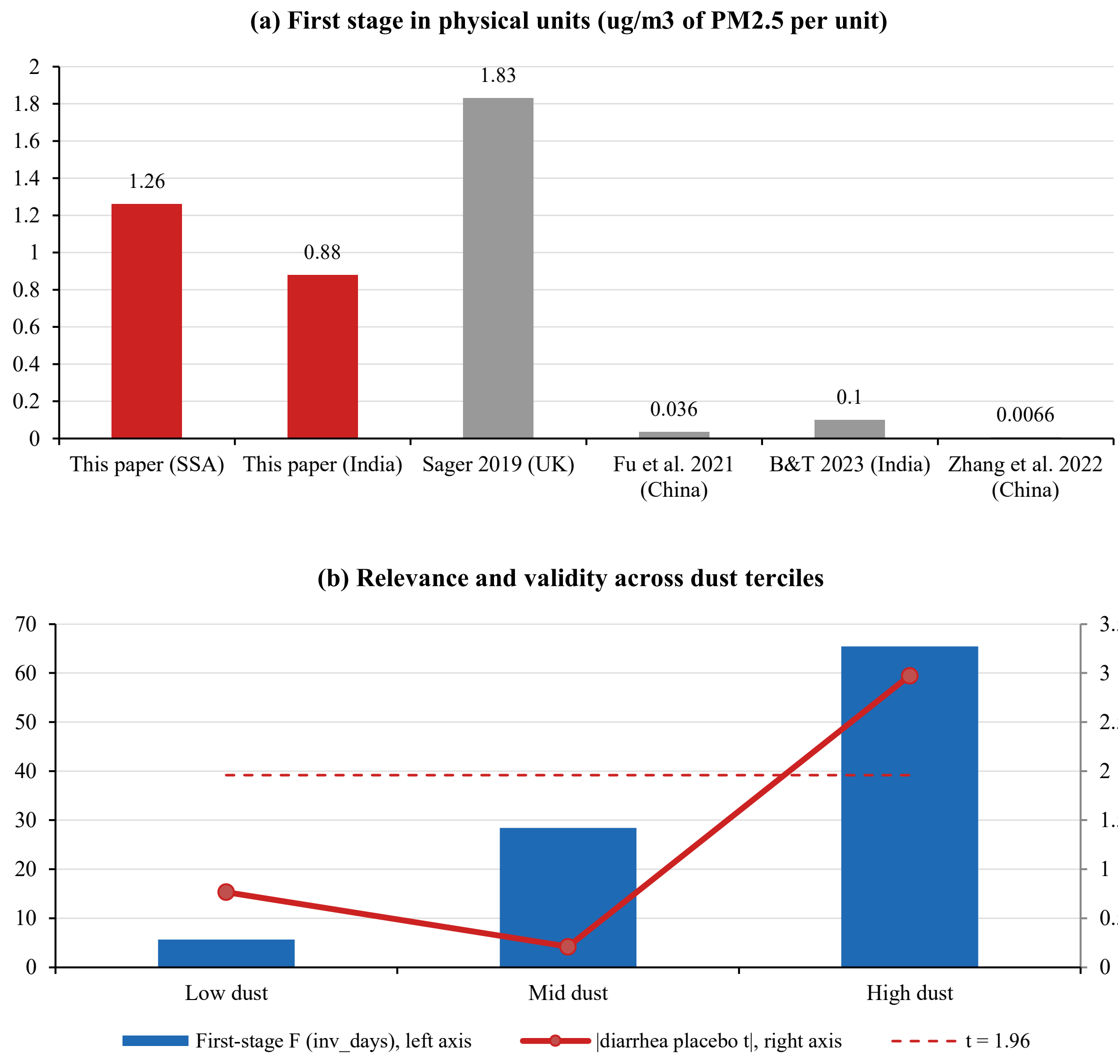
**Figure 2. Six instruments, one specification: 2SLS estimates of PM2.5 on child ARI, Sub-Saharan Africa and India.**




*Notes: Each marker is the 2SLS estimate of the effect of a 10 µg/m³ increase in PM2.5 on ARI prevalence, in percentage points, from the specification of Section 4.2 with the instrument named on the axis; horizontal bars are 95 percent confidence intervals based on standard errors clustered by DHS cluster. Blue markers are pressure-level instruments (family A) and red markers are surface-layer instruments (family B); see Table 4 for the definition of each instrument. The plotted estimates are those of Table 5 (Sub-Saharan Africa) and Table 6, Panel A (India).*

**Figure 3. Mechanisms: the size of the experiment and the relevance-validity trade-off.**




*Notes: Panel (a) plots the first stage in physical units, µg/m³ of PM2.5 per unit of instrument, for this paper's samples and for the published studies assembled in Table 7, Panel A. Panel (b) plots the first-stage F statistic of inv_days (left axis) and the t statistic of the diarrhea placebo (right axis) across terciles of the CAMS mineral-dust share of PM2.5 mass, as in Table 7, Panel B: the tercile where the instrument is strongest is the tercile where the placebo fails.*

# Paper A: Appendices A–E

## Appendix A. Data construction

### A.1 Surveys

The African sample comprises 80 standard DHS surveys across 26 countries (Angola, Benin, Burundi, Cameroon, Congo DR, Côte d'Ivoire, Eswatini, Ethiopia, Gabon, Ghana, Guinea, Kenya, Liberia, Madagascar, Malawi, Mozambique, Namibia, Nigeria, Rwanda, Sierra Leone, South Africa, Tanzania, Togo, Uganda, Zambia, Zimbabwe). The South Asian sample comprises 14 surveys (India NFHS-4 and NFHS-5; Bangladesh 1999–2022, seven rounds; Nepal 2001–2022, five rounds). A complete survey-level listing, with field dates, sample sizes, cluster counts, and the survey-level distribution of the instrument, is provided in the replication package. Two features of that listing bear on the analysis. Within a single country the zero share of the instrument can swing enormously across rounds: Benin records no inversion day at all in its 2001 round but 33.7 percent zeros in 2011–12, so the instrument's strength is not a stable property even of a fixed location. In addition, the South Asian samples span elevations from sea level to 5,743 m in India and 2,881 m in Nepal, far above any pressure surface used to define the instrument.

Membership is defined by data content rather than by file naming. A survey enters if its archive contains both a children's recode (*KR*FL.DTA) and a geographic shapefile (*GE*FL.shp). For India these are distributed as separate archives and are paired on survey year (the recode's median interview year against the shapefile's DHSYEAR, with a tolerance of three years). This content-based rule matters in practice. Six recently released surveys use a naming convention without the string "GPS" and would be silently dropped by filename matching.

### A.2 Non-Gregorian calendars

Nepali surveys record the interview century-month code in the Bikram Sambat calendar, and Ethiopian surveys through the 2019 round record it in the Ethiopian calendar. Uncorrected, these place interviews in 2057–2078 and 1992–2011 respectively. We convert by −681 months (Nepal) and +92 months (Ethiopia), applying the correction per survey and only where the derived median year deviates from the year in the archive path by more than two years. Ethiopia's 2024 round already reports Gregorian dates, so a country-level rule would corrupt it. The conversion fires on ten surveys, five Nepali (2001, 2006, 2011, 2016–17, 2021–22) and five Ethiopian (2000, 2005, 2010–11, 2016, 2019). Nine of the ten survive into the analysis samples. The 2019 Ethiopian interim round does not field the ARI questions (h31/h31b), so it contributes no outcome data, and the 2024 Ethiopian round enters without conversion because it already reports Gregorian dates. All ten reconcile to within one year of their documented field dates after conversion.

**A.3 Exposure**

PM2.5 is ACAG V6GL03, monthly, 0.01°, taken as the mean of finite cells within 2 km of the cluster centroid (implemented as a NaN-aware uniform filter over the grid followed by point indexing). Cluster-month combinations are drawn for the child's actual interview year and month (from v008), not the survey's nominal year. For surveys spanning calendar years this affects a substantial share of observations. Locations outside the product's regional domain are dropped and logged.

**A.4 Variable definitions**

ARI = h31 ∈ {1,2,3} and h31b = 1 (cough with short, rapid breathing in the past two weeks); diarrhea = h11 ∈ {1,2}; antimalarial use = any of h37a–h37h = 1, coded missing where the entire block is inapplicable. Two coding details deserve note because both are easy errors with material consequences. The variables h37a–h37h are antimalarials, while h37i and h37j are antibiotics, and a block of all-missing values must be coded missing rather than zero.

## Appendix B. High-elevation cities: three instrument families, three failures

The elevation restriction in the main sample excludes Africa's fastest-growing highland cities (Nairobi, Addis Ababa, Lusaka, Kampala), so we investigated whether any stability-based instrument can identify effects there. We find that none can. We report the exercise in full because the pattern of failure is informative. Both exercises are rebuilt from their retained inputs and verified in the replication package: B.2 from the highland child files, which carry the cluster-level ERA5 boundary-layer height, and B.3 from scratch, from the retained daily ERA5 model-level temperature fields for the two pilot years.

B.1 Pressure-level inversions. Degenerate by construction. At 1,000–2,000 m both the 1000 hPa surface and much of the 925 hPa surface lie below ground, and ERA5 supplies extrapolated values.

B.2 Boundary-layer height. Using ERA5 BLH for 17 countries, 50 surveys and 63,972 children, the first stage is strong and correctly signed ($t = -7.54$: shallower mixing, higher PM2.5). The 2SLS estimate is nonetheless wrong-signed and insignificant at conventional levels (−2.4 pp per 10 μg/m³, $t = -1.78$). BLH proxies insolation and temperature, which affect respiratory illness directly. In other words, it solves the measurement problem but violates the exclusion restriction.

B.3 Model-level (terrain-following) inversions. ERA5 model levels give genuine above-ground temperature differences in the highlands. Levels 131, 133, 135 and 137 sit at roughly 169, 107, 55 and 10 m above the surface. On the pilot sample (Kenya 2022 and Ethiopia 2016; 5,844 urban children at or above 1,000 m in 668 clusters), at 03 UTC, the hour surface inversions peak locally, the first stage is wrong-signed in all six constructions (count and strength forms at each of the three layer depths, with and without month fixed effects: $t = -2.5$ to $-8.2$), and the 2SLS estimates are unstable and insignificant or marginal ($t$ between −2.0 and +0.3). The one retained 06 UTC pair, a deep layer of roughly 205 m against 10 m, has a weak first stage of the theoretically expected sign ($t = 2.6 - 2.7$) that delivers a wrong-signed health effect (2SLS $t = -1.7$ to $-2.2$). No construction yields an instrument that is simultaneously strong, correctly signed, and clean.

B.4 Interpretation. In highland African cities, the inversion-pollution relationship has the opposite of the sign the design assumes at the hour when surface inversions actually occur. A plausible reason is that sparse local combustion leaves satellite PM2.5 dominated by regional transport, which stable dawn air excludes. No stability instrument recovers a credible effect, and we exclude these cities from the main analysis rather than report estimates we cannot defend.

## Appendix C. Alternative instruments

C.1 Fire transport (West and Central Africa). For eleven countries, 27 surveys and 67,477 children, we construct upwind fire radiative power from GFAS: FRP within a 25–500 km annulus, weighted by $\exp(-d/150$ km) and by upwind alignment $\max(0, -\cos\theta)$ against ERA5 monthly winds. The raw correlation with PM2.5 is positive (+0.28, and +0.36 in logs), the opposite of what we find in the highlands, confirming that fires genuinely drive pollution here. First stages are strong and correctly signed with month fixed effects ($F = 51 - 783$), and the diarrhea placebo passes for every version. The 2SLS estimates are nonetheless significantly negative (log upwind fire: −1.33 pp per 10 μg/m³, $t = -2.30$). Under country-by-month fixed effects the first stage collapses ($F = 233 \rightarrow 9.9$), showing that most fire variation is seasonal. The residual is not excludable, because dry-season timing carries its own health channels (nutrition after harvest, seasonal labor, disease dynamics). Notably, the placebo fails to detect this problem: a respiratory-specific seasonal channel passes a diarrhea test.

C.2 Wind direction × local population. Following Deryugina et al. (2019) in spirit, we weight GHS-POP population within 60 km by upwind alignment with a 20 km distance taper, controlling for the symmetric local population so that identification comes from wind direction alone. A two-survey pilot (Kenya 2022 and Ethiopia 2016; 5,844 children in 668 clusters) was encouraging, with a first-stage $F$ of 5.6 with the correct sign (2.8 with month fixed effects) and a raw correlation of +0.45 between the instrument and

PM2.5. At scale (45 surveys, 17 countries, 58,202 children, 4,562 clusters), the first stage vanishes ($F = 0.1$, or $F = 0.4$ with month fixed effects). This suggests that the pilot signal was idiosyncratic. We report this exercise because it illustrates a general hazard: instrument pilots on small samples are uninformative about transportability.

## Appendix D. Robustness

D.1 Inference. The main text clusters at the DHS cluster. We computed alternatives for inv_days on the pre-SSA-restriction sample. Region-level clustering (322 regions) raises the standard error by roughly 36 percent. Conley spatial HAC gives essentially the region-clustered standard error at every cutoff from 25 to 200 km (0.00075 vs 0.00075). Wild cluster bootstrap and a bootstrap Anderson–Rubin test agree with the analytic region-clustered p-values, and randomization inference permuting the instrument across clusters within survey gives $p = 0.017$. Because this paper's conclusions rest on disagreement across instruments rather than on the significance of any one estimate, the choice of variance estimator does not affect them. We report the comparison for completeness.

D.2 Exposure product and sampling window. Like D.1, this comparison is computed on the pre-Sub-Saharan-restriction sample, so its levels are not directly comparable with Table 5. The four exposure variants in Table D2 are correlated 0.989 to 1.000, and moving across them shifts the first stage only from 1.537 to 1.576 and the 2SLS estimate from +1.30 to +1.33 pp. An earlier draft read this stability as evidence that resolution does not matter. The ground validation of Section 6.6 shows that the correct reading is the opposite one. The variants agree because they carry the same information: the 0.01° product is a smooth interpolation of the 0.10° field, with only 1.4 percent of its spatial variance inside a 0.10° cell (Yamada 2026), so swapping one for the other cannot detect a resolution problem even where one exists. That share is the variance-decomposition term, the size-weighted mean of the within-cell population variance divided by the total. One alternative deserves a direct answer, because it is the one a reader is most likely to reach for. Dividing the same sum of squares by the residual degrees of freedom rather than by the number of observations gives the pooled unbiased estimator of the within-cell variance, which returns 1.77 percent on this sample. It is a sound estimator of a different quantity, namely the within-cell variance of the process that generated the field, rather than the share of the observed variance that lies within cells. The two differ by the factor n/(n − c), and the consequence is visible in the accounting. The between-cell component is 98.6 percent. Added to the decomposition term it gives 100.0 percent, and added to the pooled term it gives 100.4 percent. Only the first apportions the variance, and because this paper states both halves of the split, the first is the one it uses. The decomposition is taken over the cells that hold at least two clusters, which are the only cells in which a within-cell contrast is observable. Table D2 documents the equivalence of the exposure products, not the robustness of the design to resolution. Appendix D.3's corollary, that multiplying a high-resolution total by a coarse composition fraction does not create an independent exposure measure, applies here one level up. In other words, the high-resolution total is itself not measured at high resolution.

D.3 Speciated exposure. Multiplying total PM2.5 by CAMS combustion and dust mass fractions produces exposures whose 2SLS $t$-statistics are identical to three decimal places across all three variants (+2.41, +2.42, +2.31 for inv_days). This is mechanical. In a just-identified model the 2SLS $t$-statistic equals the reduced-form $t$-statistic, which does not involve the exposure, and rescaling the endogenous variable divides the coefficient by the first stage while leaving inference unchanged (1.43 → 2.62, a ratio of 1.83, against an inverse mean combustion share of 1/0.550 = 1.82, with the residual difference due to rounding). These estimates, like D.1 and D.2, are computed on the pre-Sub-Saharan-restriction sample. A corollary worth stating for practitioners is that multiplying a high-resolution total by a coarse reanalysis composition fraction does not create an independent exposure measure. The dust-weighted exposure retains a strong positive first stage ($F = 62$) because the smooth fraction acts as a near-constant multiplier.

D.4 Sample definition and the concentration of identifying variation. Results are unchanged under aerosol-optical-depth dust thresholds of 0.20–0.30, elevation thresholds of 800 or 1,000 m, and the addition of child and household controls.

The interview-month assignment of exposure is also not driving the results. Because the outcome recalls the two weeks before the interview, children interviewed early in a month have a recall window that partly falls in the previous month. Restricting to the 60,714 children interviewed on or after the 16th of the month, whose recall window lies within the exposure month, leaves the family split intact: inv_days gives +0.57 pp ($t = 0.80$, first-stage $F = 50$) and sfc_days gives −2.00 pp ($t = -2.66$, $F = 153$). The complementary early-month half (60,106 children) gives +0.50 pp ($t = 0.89$, $F = 71$) and −1.57 pp ($t = -2.06$, $F = 96$).

Because the pressure-level instrument is zero in eight of the 26 African countries and above 98 percent zeros in five more, we examine whether the estimates are driven by the few countries that supply variation. Non-zero observations are contributed mainly by Nigeria (24.1 percent), Mozambique (21.9 percent), Benin (21.5 percent), Angola (6.2 percent), Ghana (5.4 percent) and Côte d'Ivoire (4.6 percent). Namibia has by far the highest mean (18.2 inversion days per month against 0.75 overall) but, with 1,091 children, accounts for only 4.3 percent of non-zero observations.

Leave-one-country-out estimates (Table D4) show that neither family is a single-country artifact, but they also show that family A's null is less solid than a null usually is. The point estimate stays positive in every exclusion, ranging from +0.32 pp (dropping

Mozambique, $t = 0.78$) to +1.77 pp (dropping Nigeria, $t = 1.93$). In five of the six exclusions it remains short of conventional significance. In the sixth it does not. Dropping Côte d'Ivoire yields +1.03 pp with $t = 2.53$, which a researcher would report as significant at the five percent level. We report this explicitly because it reinforces the paper's argument from the other direction. The pressure-level family's apparent null is itself sensitive to which mid-sized country happens to be in the sample, so neither family delivers a stable answer. Family B is the more stable of the two. Excluding Namibia, Nigeria or Mozambique moves the estimate only between −1.57 and −1.86 pp, with $t$ between −2.62 and −3.38 against a baseline of −1.64 ($t = -3.05$).

Table D4. Leave-one-country-out

| Excluded | Family A (inv_days) F | Effect (pp) | t | Family B (sfc_days) F | Effect (pp) | t |
|---|---|---|---|---|---|---|
| — (baseline) | 132.0 | +0.71 | 1.65 | 223.7 | −1.64 | −3.05 |
| Nigeria | 60.4 | +1.77 | 1.93 | 195.2 | −1.74 | −2.89 |
| Mozambique | 105.6 | +0.32 | 0.78 | 180.4 | −1.57 | −2.62 |
| Benin | 99.5 | +0.72 | 1.36 | — | — | — |
| Angola | 168.6 | +0.61 | 1.38 | — | — | — |
| Ghana | 124.2 | +0.76 | 1.72 | — | — | — |
| Côte d'Ivoire | 142.6 | +1.03 | 2.53 | — | — | — |
| Namibia | — | — | — | 215.7 | −1.86 | −3.38 |

*Verified by direct recomputation from the analysis parquet: all nine leave-one-out cells and the baseline reproduce exactly. That the concentration exists at all is itself a finding. A design presented as spanning 26 countries draws two-thirds of its pressure-level identifying variation from three of them.*

D.5 Region definition. An earlier version of the sample was delimited by the satellite product's African grid, which extends to 38° N and 60° E and therefore included Egypt, Jordan and Morocco. Imposing an explicit Sub-Saharan definition changes the composition materially (150,346 → 120,820 children) and changes one diagnostic. The instrument inv_days fails the diarrhea placebo on the grid-defined sample ($t = 2.33$) but passes on the Sub-Saharan sample ($t = 0.92$). Therefore, a satellite grid boundary should not be used as a region definition.

## Appendix E. Systematic review protocol

E.1 Seed papers. Arceo, Hanna & Oliva (2016, EJ); Jans, Johansson & Nilsson (2018, JHE); Deschênes, Wang, Wang & Zhang (2020, JDE); Fu, Viard & Zhang (2021, EJ). Chosen because applied papers cite them as methodological authority.

E.2 Search. Forward-citation searches executed 2026-07-28 in Scopus and Web of Science (Core Collection) with the within-results filter inversion, and independently through the OpenAlex API (filter=cites:<id>&search=thermal inversion instrument).

Scopus returned 431 records across the four seeds, Web of Science 124, and OpenAlex 1,032 citing works, of which 85 matched the full-text query.

E.3 Screening. Scopus and WoS records were merged and deduplicated on DOI, falling back to a normalized title, yielding 245 unique records. A record was retained if title, abstract or keywords contained both an inversion term and an instrumental-variables term (82 records). Screening on abstracts alone cannot establish implementation detail, so we treat abstract-level counts as descriptive and rely on full-text coding for all claims about practice.

E.4 Full-text coding. Twenty-six retained studies (10 primary, 16 secondary) were obtained in full text and coded on: reanalysis product; layers compared; functional form; acquisition time; reported first-stage magnitude; reporting of the instrument's distribution; study-area elevation; robustness across definitions; placebo tests; over-identification tests. Table 2 reports the distribution. The coding sheet, one row per study and keyed by DOI, is available in the replication package together with the 82-record screening output.

E.5 Limitations and known errors. OpenAlex full-text coverage is incomplete, so the counts (82 screened from Scopus and Web of Science, of which 81 confirmed, and 85 from OpenAlex) are lower bounds. Region coding from abstracts is imprecise and non-exclusive. The 26 coded studies are, with one exception noted below, the most-cited rather than a random sample, which if anything biases the audit toward best practice.

Both seed papers, Arceo et al. (2016) and Jans et al. (2018), are read in full and discussed in Section 2.4, but they are deliberately excluded from the 26. They are the authorities the corpus cites, not members of the corpus that cites them, and including them would blur exactly the comparison Section 2.4 draws. All quotations from both are taken from the published articles. We flag one point on which an earlier reading of the working-paper versions misled us, because it bears on a claim we make elsewhere. Jans et al.'s analysis of childhood injuries is presented in the published paper as a test for avoidance behavior, with the share of injuries occurring indoors as the outcome, not as a placebo on a health outcome. The health-outcome placebo among the seed papers is Arceo et al.'s, which splits infant mortality into internal and externally caused deaths and reports no effect on the latter.

The coded set is 26 rather than 25 because one study entered it by accident. We selected the coding sample by citation count, but the full text we first obtained under the DOI of Zhang et al. (2022), a Journal of Cleaner Production article on subjective well-being in China, proved to be a different article on the same topic, Cao et al. (2025) in SAGE Open, which is also a member of the screened corpus and also a thermal-inversion application. On discovering the substitution we obtained the correct article and coded it as well, retaining Cao et al. (2025) rather than discarding a study we had already read in full. The coded set is therefore the 25 most-cited retained studies plus Cao et al. (2025), a recent article with few citations that would not otherwise have been selected. We note this because it is a departure from the stated selection rule, not because we think it materially affects the distributions in Table 2. Both articles use the 110 m/320 m layer comparison, and both are counted as such in Panel A.

We also identified one false positive during reference checking. A record carried in the Scopus/Web of Science export under the title "The Productivity Consequences of Pollution-Induced Migration in China" with DOI 10.1257/app.20230108 is in fact a different article, on social assistance in Bangladesh. The title had been mis-assigned in the export, as a corrupted character in the title field suggested. The genuine Khanna et al. (2025) article carries DOI 10.1257/app.20220655 and is retained. We therefore report 82 as the screening output and 81 as the confirmed count, and use 81 wherever the text characterizes the literature. Because we verified only the records we examined in full, we cannot rule out further errors of this kind in either direction. The confirmed count is best read as approximate, and as a lower bound for the reason given above.

## Appendix tables

### Table B1. High-elevation cities: three instrument families

*Highland sample: clusters at ≥ 1,000 m, low dust, urban.*

| Family | Instrument | Sample | First stage | 2SLS (pp / 10 μg/m³) | Verdict |
|---|---|---|---|---|---|
| Pressure level | T925 − T1000 | — | degenerate (levels below ground) | — | not estimable |
| Boundary layer | ERA5 BLH | 17 countries, 50 surveys, 63,972 children | t = −7.54 (strong, correct sign) | −2.4, t = −1.78 | exclusion violated (BLH proxies insolation) |
| Model level | 03 UTC, count & strength, 55/107/169 m vs 10 m, ± month FE | Kenya 2022 + Ethiopia 2016; 5,844 children, 668 clusters | **t = −2.5 to −8.2 (wrong-signed in all six constructions)** | t = −2.0 to +0.3 (unstable) | fails |
| Model level | 06 UTC, day count, ~205 m vs 10 m | same | t = +2.6 to +2.7 (expected sign, weak) | t = −1.7 to −2.2 (wrong-signed effect) | fails |

### Table C1. Alternative instruments

### Panel A: Fire transport (West and Central Africa; 11 countries, 27 surveys, 67,477 children)

| Instrument | FE | First stage | F | 2SLS (pp / 10 μg/m³) | t | Diarrhea placebo t |
|---|---|---|---|---|---|---|
| Upwind FRP | survey + region + month | +3.160 | 50.8 | −1.66 | −2.20 | 0.06 |

| | | | | | | |
|---|---|---|---|---|---|---|
| log upwind FRP | survey + region + month | +4.292 | 232.9 | **−1.33** | **−2.30** | 0.62 |
| Symmetric FRP | survey + region + month | +4.601 | 258.8 | −0.79 | −1.30 | 0.40 |
| log symmetric FRP | survey + region + month | +7.050 | 783.3 | −0.17 | −0.40 | 0.75 |
| log upwind FRP | + country × month | +0.829 | **9.9** | −13.66 | −2.33 | — |

*Note for Panel A: The raw correlation of upwind fire with PM2.5 is +0.281 (+0.356 in logs). The 2SLS estimate for log upwind FRP moves only between −1.21 and −1.33 across no-control, wind-only, temperature, and flexible-temperature specifications. Over-identification against the inversion instrument gives J = 3.01, p = 0.083. First-stage coefficients are µg/m³ of PM2.5 per unit of the z-scored instrument; 2SLS effects are percentage points of ARI per 10 µg/m³.*

## Panel B: Wind direction × local population

| Sample | Surveys | Children | Clusters | First stage F | Note |
|---|---|---|---|---|---|
| Pilot (Kenya 2022, Ethiopia 2016) | 2 | 5,844 | 668 | 5.6 (2.8 with month FE; correct sign) | raw correlation +0.45 |
| Full highland sample | 45 | 58,202 | 4,562 | **0.1 (0.4 with month FE)** | signal does not transport |

*Note for Panel B: Both rows are recomputed in the replication package. The pilot's original extract was superseded and not retained, so the pilot row is rebuilt from the raw GHS-POP raster and ERA5 winds, with the construction calibrated against the surviving full-sample instrument (calibration correlations 0.996 and 0.997). The superseded extract had recorded F = 3.5 and a raw correlation of +0.50, a very similar picture. Either way, the lesson of the row is unchanged: a pilot first stage in the 3–6 range collapses to 0.1 at scale.*

## Table D1. Inference

*inv_days, ARI outcome. Computed on the pre-Sub-Saharan-restriction sample (N = 95,224, 6,497 clusters, 322 regions, 19 countries) for comparability with earlier drafts.*

| Method | Clusters | p-value |
|---|---|---|
| Cluster (DHS cluster) | 6,497 | 0.016 |
| Cluster + wild bootstrap | 6,497 | 0.020 |
| Randomization inference (permute instrument within survey) | — | 0.017 |
| Region-clustered | 322 | 0.077 |
| Region + wild bootstrap | 322 | 0.078 |
| Bootstrap Anderson–Rubin | 322 | 0.074 |
| Country-clustered | 19 | 0.093 |
| Conley spatial HAC, 25 km | — | 0.084 |
| Conley spatial HAC, 50 km | — | 0.079 |

| | | |
|---|---|---|
| Conley spatial HAC, 100 km | — | 0.082 |
| Conley spatial HAC, 200 km | — | 0.135 |

*Notes: The Conley standard error matches the region-clustered one at the 25–100 km cutoffs, down to 25 km, the scale of the ERA5 grid itself, and grows only modestly at 200 km. All p-values are normal-approximation values. With 19 country clusters the analytic p is optimistic, which is what the bootstrap rows are for. The analytic rows (cluster, region, country, Conley) are recomputed in the replication package. The wild-bootstrap, bootstrap-AR and randomization-inference rows are Stata outputs (a wild cluster bootstrap and a cluster-level permutation test on the reduced form), regenerated in the replication package.*

## Table D2. Exposure product and sampling window

| Exposure | Correlation with 0.01° 2 km mean | First stage | 2SLS (pp / 10 μg/m³) | t |
|---|---|---|---|---|
| 0.1° coarse product | 0.989 | +1.5367 | +1.33 | 2.41 |
| 0.01°, grid point | 0.999 | +1.5758 | +1.30 | 2.40 |
| **0.01°, 2 km mean (baseline)** | 1.000 | +1.5677 | +1.31 | 2.40 |
| 0.01°, 5 km mean | 0.998 | +1.5482 | +1.32 | 2.41 |

*Notes: Computed on the pre-Sub-Saharan-restriction sample, as in Table D1; levels are therefore not comparable with Table 5, and the comparison across rows is the object of interest. The correlations in column 2 measure the equivalence of the exposure products, not the adequacy of their resolution (Section 6.6). First stages are μg/m³ per inversion day (inv_days); 2SLS effects are percentage points of ARI per 10 μg/m³.*

## Table D3. Speciated exposure (mechanical invariance)

| Exposure | Mean (μg/m³) | 2SLS with inv_days | t | 2SLS with sfc_days | t |
|---|---|---|---|---|---|
| Total PM2.5 | 24.74 | +1.43 | 2.41 | −0.95 | −2.02 |
| Combustion PM2.5 (× 0.550) | 13.61 | +2.62 | **2.42** | −1.23 | **−2.01** |
| Dust PM2.5 | — | +3.26 | **2.31** | −3.30 | **−1.99** |

*Notes: Computed on the pre-Sub-Saharan-restriction sample (hence a total-PM2.5 mean of 24.74 rather than the 23.0 of Table 3). The point of the table is the invariance, not the levels. The t-statistics are invariant by construction: in a just-identified model the 2SLS t equals the reduced-form t, which does not involve the exposure. The coefficient ratio 2.62/1.43 = 1.83 reproduces the inverse mean combustion share, 1/0.550 = 1.82, to within rounding. The dust-weighted exposure retains a strong positive first stage (F = 62), confirming that a coarse composition fraction acts as a near-constant multiplier rather than creating an independent exposure measure.*