

Preregistered External Validation of Spatially Valid Conformal Prediction for Air-Pollution Exposure Mapping in China

Tian-Yu Wang

wangty0709@hotmail.com

Guangdong Peizheng College

Research Article

Keywords: conformal prediction, spatial statistics, air pollution, PM2.5, China, preregistration

Posted Date: October 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11182286/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Air-pollution exposure maps drive public-health decisions, yet their uncertainty is rarely quantified, and the intervals attached to them are seldom validated in the settings where the maps are used. Conformal prediction provides distribution-free prediction intervals, and a spatially valid variant exists, but its coverage on real data remained untested. We preregistered an external validation of spatially valid conformal prediction (Mao et al. 2024) on province-level population-weighted PM_{2.5} across China (1998–2024; $n = 837$), benchmarked against gradient boosting under spatial and temporal hold-outs, with a preregistered acceptance band of $[0.85, 0.95]$ on pooled coverage. Under spatial hold-out, the spatially weighted method attained pooled coverage of 0.873 at nominal 90%, within the band; gradient boosting attained 0.486. Five of the seven macro-regions met the block-wise criterion, the worst (North China) reaching 0.719, and interval widths were identical to those of equal-weight conformal prediction, so the coverage did not come from wider intervals. Under temporal hold-out, coverage fell to 0.794, below the band, and the hypothesis that both partitions would pass was not supported. The failure traces to the sharp post-2017 decline in Chinese PM_{2.5}, a temporal shift that purely spatial methods cannot anticipate; the block-wise maps localise it, East China falling to 0.457 in 2020–2024 while the other six macro-regions remain at or above 0.88. Spatially valid conformal prediction is reliable for spatial mapping but not for temporal forecasting. Registration: <https://osf.io/edhzw/>

1. Introduction

Ambient fine particulate matter (PM_{2.5}) remains one of the largest environmental contributors to premature mortality in China (Cohen et al. 2017; Chen and Hoek 2020), with population-weighted concentrations in many provinces still several times the World Health Organization guideline level (World Health Organization 2021). Gridded exposure surfaces built from satellite retrievals, chemical transport models and ground monitors are routinely used to identify high-risk areas, allocate screening resources and communicate hazard to the public (van Donkelaar et al. 2016; Wang et al. 2019), a task at the centre of spatial epidemiology (Elliott and Wartenberg 2004). The province-level summaries of the Atmospheric Composition Analysis Group (ACAG), which this study uses, are one such product (Atmospheric Composition Analysis Group 2024). The credibility of these maps depends on whether the uncertainty attached to them is honest: a map that understates its own error misleads most in the locations where decisions carry the greatest consequences.

Conformal prediction constructs prediction intervals whose marginal coverage is guaranteed under exchangeability, without assumptions on the distribution of the data or the correctness of the underlying model (Vovk et al. 2005; Angelopoulos and Bates 2023). In spatial settings the predictor of choice is some form of Gaussian-process (Kriging) regression (Rasmussen and Williams 2005; Diggle et al. 1998). Mao, Martin and Reich (2024) observed that spatial observations can be treated as exactly or approximately exchangeable in a wide range of settings, and derived a conformal prediction procedure, termed spatial conformal prediction, that is valid under infill asymptotics (Mao et al. 2024).

The theoretical behaviour of conformal prediction when exchangeability is violated has been quantified by Barber et al. (2023). Empirically, however, whether the spatially valid procedure actually attains its nominal coverage on a real, large-domain, non-stationary environmental dataset has not been examined. Existing applications of conformal prediction in Earth observation have assessed uncertainty for retrieval products (Singh et al. 2024), and the machine-learning health literature contains repeated warnings that prediction models degrade under distribution shift (Wynants et al. 2020; Christodoulou et al. 2019) and that reported performance is often inflated by leakage and incomplete reporting (Kapoor and Narayanan 2023; McDermott et al. 2021). No study has tested the spatially valid framework against the shifts that arise when mapping air-pollution exposure across a large, rapidly changing country. A documented search of Europe PMC (executed 2026-09-17/18; queries and hit counts in Supplementary Appendix A) did not identify any study combining spatially valid conformal prediction with air-pollution mapping in China.

We therefore preregistered an external validation and failure-mode diagnosis of the Mao et al. (2024) framework, applied to province-level population-weighted PM_{2.5} across China and to a Yangtze River Delta sub-analysis. The plan, partitions, model specification, decision thresholds and calibration seed were frozen and time-stamped before any confirmatory analysis was run (<https://osf.io/edhzw/>), following the practice advocated in the metascience literature (Nosek et al. 2018; Chambers and Tzavella 2022). We compare two interval-construction methods that share the same Gaussian-process predictor, benchmark against the gradient-boosting-plus-split-conformal pipeline that is the de facto standard in applied machine-learning health research, and report every partition and every macro-region, including the least favourable ones. Deviations from the preregistered plan are disclosed in the supplement. This design also responds to a recent audit finding that 99% of secondary analyses of a widely used health database did not share code (Degen et al. 2026); every dataset and every script in the present study is open.

2. Methods

2.1 Data

Annual population-weighted PM_{2.5} concentrations by province were obtained from the ACAG Surface PM_{2.5} V6.GL.03 province-level summary table (CC BY 4.0), accessed through the AWS Open Data bucket s3://satpmdata/ (Atmospheric Composition Analysis Group 2024). The table provides, for each of 31 Chinese province-level administrative units and each year from 1998 to 2024, the population-weighted and geographic-mean PM_{2.5} concentration, the coverage fractions of the underlying grid, the total population, and the percentage of population exposed above thresholds from 5 to 60 µg/m³.

Daily meteorological covariates, 2-metre air temperature, total precipitation and relative humidity, were retrieved for each provincial capital from the NASA POWER point endpoint for 1998–2024 (NASA 2024). Provincial population was taken from the population column of the same ACAG table. Administrative

boundaries (GADM version 4.1) were used for cartography only; the raw boundary files are not redistributed (GADM 2022).

All datasets were downloaded on 2026-09-17 and archived as read-only snapshots with SHA-256 checksums (manifests in Supplementary Appendix C). The assembled panel contains 837 province-year observations with zero missing values. The secondary outcome is the percentage of provincial population exposed to $\geq 35 \mu\text{g}/\text{m}^3$, corresponding to China's Grade II annual standard and the WHO Interim Target-1 (World Health Organization 2021).

The Global Burden of Disease (IHME) database was deliberately not used. A 2026 audit in the Journal of Clinical Epidemiology reported that 99% of assessed secondary analyses of that database did not share code (Degen et al. 2026), and this study is designed to be fully reproducible.

2.2 Study design

This is an external validation and failure-mode diagnosis of an existing spatial prediction method applied to openly available environmental-health data. The analysis plan was frozen and time-stamped before any confirmatory analysis was run; the preregistration is available at <https://osf.io/edhzw/>.

A Gaussian-process (Kriging) regression with an exponential covariance is fitted to the training observations of each partition (Rasmussen and Williams 2005; Diggle et al. 1998). Covariance parameters are estimated by marginal likelihood. The same predictor is used for the baseline and for the method under test, so that any difference in coverage is attributable to the interval construction rather than to the predictor.

Two interval-construction methods are compared, both at a nominal level of 90% ($\alpha = 0.10$):

- **Baseline (GSCP):** standard split conformal prediction, using the absolute standardised out-of-sample residual as the nonconformity score with uniform weights.
- **Method under test (sLSCP):** the spatially weighted conformal procedure of Mao et al. (2024). The nonconformity scores of the calibration observations are weighted by a Gaussian kernel in distance from the target location, normalised to include the self term, and the interval is the level set of the weighted plausibility whose upper-tail weighted mass reaches α .

A gradient-boosting regressor (HistGradientBoostingRegressor, scikit-learn 1.9.0, max_iter=200, random_state=7) with standard split conformal prediction serves as the benchmark (GBM), reflecting common practice in applied machine-learning health research. No hyperparameter tuning is performed.

2.3 Evaluation partitions

Two preregistered partition schemes are used, following the recommendation that validation designs should mirror the intended generalisation structure of the data (Roberts et al. 2017):

- **Leave-one-macro-region-out** (primary): each of seven Chinese macro-regions (North, Northeast, East, Central, South, Southwest, Northwest China) is held out in turn. Training and calibration observations come from the remaining six macro-regions restricted to 1998–2019; the calibration set is a random 30% of that pool (seed 20260917). The test set is the held-out macro-region, evaluated over all years and separately over 2020–2024.
- **Leave-years-out** (secondary): 1998–2019 for training and calibration; 2020–2024 held out.

A Yangtze River Delta sub-analysis holds out Shanghai, Jiangsu, Zhejiang and Anhui, with the partition constructed as in the leave-region-out scheme. The sampling frame is fixed at 31 provinces $\times$ 27 years = 837 observations; no observations are added or dropped after registration.

2.4 Outcomes and metrics

The primary outcome is empirical global coverage of the nominal 90% intervals on held-out units, pooled across macro-regions weighted by the number of test observations (deviation D-003 in the supplement). Secondary outcomes are block-wise coverage per macro-region, worst-macro-region coverage, and mean interval width.

The preregistered decision rules are:

- **H1:** pooled global coverage falls within [0.85, 0.95] under both the leave-region-out and the leave-years-out partitions.
- **H2:** at least 5 of the 7 macro-regions attain coverage within [0.80, 1.00], and the worst macro-region attains at least 0.65.
- **H3:** the mean interval width of the method under test is at most 1.10 times that of the baseline, while its global coverage is no lower.

Every coverage estimate is reported with a 95% Wilson confidence interval for a binomial proportion. All seven regional values are reported without selection. Deviations from the preregistration are disclosed in the deviation register (Supplementary Appendix B).

2.5 Sensitivity analyses

Five prespecified sensitivity analyses are reported: (1) the secondary outcome (% population $\geq 35 \mu\text{g}/\text{m}^3$) in place of the primary; (2) a shortened window (2015–2024); (3) the Yangtze River Delta sub-analysis; (4) calibration fractions of 20% and 40%; and (5) exclusion of 2020–2022. Each changes one design element; the partition construction of Section 2.3 is retained throughout.

2.6 Software and reproducibility

All analyses were performed in Python 3.13, with pandas (McKinney 2010), scikit-learn (Pedregosa et al. 2011) and NumPy (Harris et al. 2020); versions pinned in the environment lock file. The analysis code, data manifests with SHA-256 checksums, and one-command reproduction scripts are in the

supplement. The confirmatory analysis on real data was not run before registration; the analysis script refuses to execute in real-data mode unless a file recording the registration link and date is present, so the ordering of plan before results is enforced by the code rather than merely promised.

2.7 Ethics

This study used only publicly available, aggregated secondary data and did not involve human participants; ethical approval was therefore not required.

3. Results

3.1 Global coverage (H1)

Table 1 reports pooled global coverage under the two preregistered partitions.

Under leave-one-macro-region-out validation, the spatially weighted conformal method (sLSCP) attained a pooled coverage of **0.873** (95% Wilson CI 0.849–0.894), which falls within the preregistered acceptance band [0.85, 0.95]. The unweighted mean of regional coverages was 0.890. Under leave-years-out validation, pooled coverage fell to **0.794** (95% CI 0.723–0.850), below the band. Because hypothesis H1 requires both partitions to meet the band simultaneously, **H1 was not supported**.

The gradient-boosting benchmark attained a pooled coverage of **0.486** (95% CI 0.453–0.520) under leave-region-out validation, roughly half the nominal level, and 0.748 under leave-years-out validation.

Table 1. Pooled empirical coverage of nominal 90% prediction intervals and mean interval width, by partition and interval-construction method. GSCP = split conformal with equal weights; sLSCP = spatially weighted conformal prediction (Mao et al. 2024); GBM = gradient boosting with split conformal prediction.

Partition	Method	Test n	Coverage (95% Wilson CI)	Mean width (µg/m³)
Leave-one-macro-region-out (1998–2024)	sLSCP	837	0.873 (0.849–0.894)	40.2
	GSCP	837	0.873 (0.849–0.894)	40.2
	GBM	837	0.486 (0.453–0.520)	14.1
Leave-region-out, 2020–2024 only	sLSCP	155	0.858 (0.794–0.904)	40.2
Leave-years-out (2020–2024)	sLSCP	155	0.794 (0.723–0.850)	25.7
	GSCP	155	0.794 (0.723–0.850)	25.7
	GBM	155	0.748 (0.675–0.810)	13.7

3.2 Block-wise coverage (H2)

Figure 1 shows empirical coverage by macro-region for the three methods.

Under leave-region-out validation, five of the seven macro-regions attained coverage within [0.80, 1.00]: Northeast (100%), South (100%), Southwest (99%), Northwest (89%) and Central China (84%). East China attained 79% and North China 72%, both below the preregistered 80% threshold. The worst macro-region (North China, 0.719) exceeded the preregistered minimum of 0.65. Because five regions met the criterion and the worst region exceeded 0.65, **H2 was supported**.

The failure pattern was geographically heterogeneous. Under the recent-years sub-partition (2020–2024), East China fell to 0.457 while the other six macro-regions remained at or above 0.88. The benchmark showed a different geography: its coverage was lowest in South China (0.28 over all years, 0.13 in 2020–2024) and Southwest China (0.36 and 0.28), and highest in Northeast China (0.84 over all years).

3.3 Efficiency (H3)

The mean interval width of the spatially weighted method equalled that of the standard split conformal baseline in every partition (Table 1): at the estimated correlation length of the Gaussian process (Section 4.2), the spatial weights are so close to uniform that the two procedures produce identical intervals. H3 was therefore supported: coverage was not bought by widening intervals.

3.4 Comparison with the gradient-boosting benchmark

The gradient-boosting benchmark attained markedly lower coverage than either GP-based method under spatial hold-out: pooled coverage of 0.486 (versus 0.873), with region-specific values from 0.28 (South China) to 0.84 (Northeast China) over all years, and from 0.13 (South China) to 1.00 (Central China) in 2020–2024. Its mean interval width (12–15 $\mu\text{g}/\text{m}^3$, pooled 14.1) was roughly a third of the GP-based width (32–54 $\mu\text{g}/\text{m}^3$, pooled 40.2). The benchmark's intervals are efficient in-sample but fail to capture the spatial variability of the outcome under distribution shift.

3.5 Failure modes and geographic heterogeneity

Figure 2 maps the 2015 annual mean PM2.5 field over the Yangtze River Delta, the exposure surface targeted by the sub-analysis: concentrations grade from the heavily loaded belt across northern Anhui and Jiangsu toward the cleaner coastal southeast. Figure 3 places the 27-year record in view across all 31 provinces; for the four Delta provinces the population-weighted mean rises through the 2000s, peaks in 2014 (54.8 $\mu\text{g}/\text{m}^3$), and falls steadily after 2017 (44.3 $\mu\text{g}/\text{m}^3$ in 2017 to 33.4 in 2024), visible as the sustained red block through the mid-2010s followed by the blue shift after 2017.

This temporal distribution shift is the most likely explanation for the failure of H1 under the leave-years-out partition: a predictor trained on 1998–2019 data systematically overestimates concentrations in

2020–2024, producing intervals that are centred too high and fail to cover the shifted outcomes.

The partitions also reveal where spatial coverage fails. Under the full-year window the method under-covers in North China (0.719) and East China (0.794). Under the recent-years window the failure concentrates in East China (0.457), while Central, North, Northeast, South and Southwest China all attain 1.00 and Northwest China attains 0.88. Figure 4 shows the national 2015 exposure field that the validated pipeline is asked to map, with concentrations peaking over the North China Plain and the Tarim Basin. Province-level coverage of the intervals under the leave-region-out partition falls below 0.80 in 10 of the 31 provinces, with the lowest values in Xinjiang (0.52), Beijing and Tianjin (0.59 each), concentrated in North and East China. The preregistered design captures both patterns because all regional values are reported without selection.

3.6 Sensitivity analyses

The prespecified sensitivity analyses are summarised in Table 2. Across the spatial hold-out scenarios, pooled sLSCP coverage ranged from 0.845 (shortened window 2015–2024) to 1.000 (Yangtze River Delta), with four of six scenarios at or above 0.87. The secondary outcome yielded higher coverage (0.923), reflecting the greater variance of the exceedance-fraction measure. Varying the calibration fraction from 20% to 40% did not materially change pooled coverage (0.888 and 0.895), confirming that the calibration set size is adequate. Exclusion of 2020–2022 changed coverage only marginally (0.876 versus 0.873). The gradient-boosting benchmark remained between 0.46 and 0.50 in all spatial scenarios except the shortened window, where it reached 0.794 while still trailing the GP-based methods.

Table 2. Pooled coverage by method under the prespecified sensitivity analyses (leave-region-out unless otherwise noted).

Scenario	n (test)	sLSCP coverage	GBM coverage
Primary analysis (reference)	837	0.873	0.486
Secondary outcome (% pop ≥35 µg/m³)	837	0.923	0.504
Shortened window (2015–2024)	310	0.845	0.794
Yangtze River Delta sub-analysis	108	1.000	0.500
Calibration fraction 20%	837	0.888	0.462
Calibration fraction 40%	837	0.895	0.477
Excluding 2020–2022	744	0.876	0.479
Leave-years-out (2020–2024)	155	0.794	0.748

4. Discussion

4.1 Principal findings

This preregistered external validation of the spatially valid conformal prediction framework of Mao et al. (2024) yields three principal findings. First, the method maintains near-nominal pooled coverage under spatial distribution shift (0.873), where the standard machine-learning pipeline loses roughly half of nominal coverage (0.486). Second, it satisfies the block-wise criterion in five of seven macro-regions, with a worst-region coverage of 0.719. Third, it fails to maintain coverage under temporal distribution shift (0.794 under leave-years-out validation, below the preregistered band); the benchmark fails as well (0.748).

4.2 Interpretation

The coverage advantage of the GP-based methods over the non-spatial benchmark under spatial hold-out is consistent with the motivation of Mao et al. (2024): the Gaussian-process predictor accounts for spatial correlation that a non-spatial model cannot represent, and the conformal layer calibrates the interval to the local residual scale. The finding that the spatially weighted interval coincided exactly with the equal-weight interval in every partition was not anticipated. The fitted correlation length of the Gaussian process was comparable to or larger than the study domain in every training fit (φ between 0.86 and 1.60 in normalized coordinates, against a domain diagonal of 0.98; Supplementary Appendix E), so the kernel weights were nearly uniform and the weighted procedure reduced to the unweighted one. This is a boundary condition of the method at coarse spatial resolution (31 province-level units): the spatial weighting is expected to become informative at finer resolutions, where the correlation length is small relative to the domain, consistent with the infill asymptotics of Mao et al. (2024).

The failure under temporal hold-out is consistent with the known sensitivity of conformal prediction to non-exchangeability (Barber et al. 2023) and with covariate-shift analyses showing that coverage degrades in the shifted region (Tibshirani et al. 2019). The sharp reduction in Chinese PM_{2.5} after 2017, driven by the Clean Air Action and subsequent emission controls (Xue et al. 2019; Zheng et al. 2018), means that a model of the past over-predicts the present. No purely spatial method can anticipate this shift, and the benchmark, which conditions on year, fared no better (0.748). Coverage under temporal hold-out was similarly poor for both outcome definitions (0.78–0.79), while spatial hold-out coverage remained at or near nominal, an asymmetry that any application of these methods should state explicitly. Comparable non-stationarity has been documented for long-term PM_{2.5} trends elsewhere (Enayati Ahangar et al. 2021), and the Chinese guideline context continues to evolve (Xue et al. 2022).

4.3 Strengths

Three features distinguish this study from much of the secondary-analysis literature. First, the analysis plan, partitions, model specification, decision thresholds and calibration seed were frozen and publicly time-stamped before any confirmatory analysis was run, which removes post hoc selection of models, partitions or reporting choices. Second, every dataset is openly available without authentication, every analysis script is released, and the environment can be reproduced with a single command; a reader can

rerun the entire analysis and verify that the outputs match to bit level. Third, all regional values are reported without selection, including those least favourable to the method under test, and confirmed and falsified hypotheses are treated with equal prominence, a standard of full reporting encouraged by recent guidance for prediction-model research (Collins et al. 2024).

4.4 Limitations

Several limitations should be noted. First, the meteorological covariates are represented by provincial-capital point values as a proxy for province-wide conditions, which may miss sub-provincial variation. Second, the study uses province-level units ($n = 31$), a coarse spatial resolution; as discussed above, the spatial weighting of Mao et al. (2024) was degenerate at this resolution, so this study validates the framework's coverage properties but not its potential efficiency gains at fine resolution. Third, the study period includes the COVID-19 pandemic (2020–2022), which may have disrupted both emissions and monitoring; a sensitivity analysis excluding these years is reported. Fourth, the outcome is population-weighted PM_{2.5} concentration rather than a directly measured health endpoint; the link to health runs through exposure–response functions that are not modelled here. Fifth, cartographic boundaries are drawn from GADM (mainland provinces, and Taiwan from the GADM Taiwan dataset); no substantive result depends on cartography.

4.5 Implications

For researchers producing air-pollution exposure maps in China or other large, rapidly changing settings, the results suggest that a conformal layer over a spatial predictor provides substantially more reliable uncertainty quantification than the standard machine-learning pipeline under spatial distribution shift, but that coverage should not be assumed to extend to temporal shift. Where temporal shift is present, the preregistered failure-mode diagnosis reported here, block-wise coverage mapped by region and by year, identifies which areas are most affected.

For the broader community conducting secondary analyses of public health databases, the study demonstrates a workflow in which the plan is preregistered, the code is released, the results are reproducible bit for bit, and confirmed and falsified hypotheses are reported with equal prominence. The reproducibility deficit documented elsewhere (Degen et al. 2026) is addressable with existing tools; the marginal cost here was the discipline of writing the plan down before looking at the data.

5. Conclusion

On openly available Chinese air-pollution data spanning 27 years and 31 provinces, the spatially valid conformal prediction method of Mao et al. (2024) maintained near-nominal coverage under spatial distribution shift, where the standard machine-learning pipeline fell to roughly half of nominal coverage. The method failed to maintain coverage under temporal distribution shift, a failure mode that (to our knowledge, based on the documented search in Supplementary Appendix A) has not previously been

reported for this framework on real environmental-health data. The preregistered, fully reproducible design ensures that these findings, including the negative result for H1, were not selected after the fact.

Declarations

Software and data availability

Software. Name: spatial-conformal-pm25-china (the complete analysis pipeline of this study). Developer and contact: Tian-Yu Wang, Medical Office, General Affairs Office, Guangdong Peizheng College, Guangzhou 510830, China; wangty0709@hotmail.com. Year first available: 2026. Program language: Python 3.13. Program size: approximately 1,340 lines across nine scripts (63 KB of source code), plus the pinned environment file. Hardware required: any modern desktop or laptop computer; no GPU or cluster is needed, and the full confirmatory analysis completes in under one minute. Software required: the packages pinned in requirements.txt (pandas 3.0.5, scikit-learn 1.9.0, NumPy 2.4.6, SciPy 1.18.0, matplotlib 3.11.1). Availability and cost: free of charge, open source under the MIT licence, at <https://github.com/GustaveKateb0709/spatial-conformal-pm25-china>, together with the frozen confirmatory outputs, the sensitivity analyses, the GP parameter diagnostics and the assembled analysis panel; the preregistration, datasets and checksum manifests are archived at <https://osf.io/edhzw/>.

Data. All datasets are public and require no authentication or registration: the ACAG Surface PM2.5 V6.GL.03 province-level table (CC BY 4.0; AWS Open Data bucket s3://satpmdata/), NASA POWER daily meteorology (public domain; <https://power.larc.nasa.gov/>), and GADM v4.1 administrative boundaries (cartographic use only; not redistributed). The 837-row analysis panel (112 KB as CSV) is included in the repository and in the supplement. Download dates and SHA-256 checksums for every source file are provided in the supplementary manifests.

Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflicts of interest: The author declares no competing interests.

Data availability: All datasets used are publicly available without authentication, as detailed in the Software and data availability section above and in Supplementary Appendix C.

Code availability: The full analysis pipeline is open source under the MIT licence (© Tian-Yu Wang); see the Software and data availability section above.

Author contributions (CRediT): Tian-Yu Wang: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing – original draft, writing – review and editing, visualization.

AI disclosure: During the preparation of this work the author used Kimi K3 (Moonshot AI) for code development and language editing. The author reviewed and edited all AI-generated content and takes

full responsibility for the integrity and accuracy of the publication.

Preregistration: This study was preregistered at the Open Science Framework (<https://osf.io/edhzw/>). The preregistered analysis plan was finalised before any confirmatory analysis on real data was conducted. Deviations from the preregistered plan are disclosed in Supplementary Appendix B.

References

1. Angelopoulos AN, Bates S (2023) Conformal prediction: a gentle introduction. *Found Trends Mach Learn* 16:494–591. <https://doi.org/10.1561/2200000101>
2. Atmospheric Composition Analysis Group (2024) Surface PM_{2.5} (V6.GL.03) [dataset]. Washington University in St. Louis. <https://sites.wustl.edu/acag/datasets/surface-pm2-5/> (CC BY 4.0)
3. Barber RF, Candès EJ, Ramdas A, Tibshirani RJ (2023) Conformal prediction beyond exchangeability. *Ann Stat* 51. <https://doi.org/10.1214/23-AOS2276>
4. Chambers CD, Tzavella L (2022) The past, present and future of Registered Reports. *Nat Hum Behav* 6:29–42. <https://doi.org/10.1038/s41562-021-01193-7>
5. Chen J, Hoek G (2020) Long-term exposure to PM and all-cause and cause-specific mortality: a systematic review and meta-analysis of 107 cohort studies. *Environ Int* 143:105974. <https://doi.org/10.1016/j.envint.2020.105974>
6. Christodoulou E, Ma J, Collins GS, et al (2019) A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 110:12–22. <https://doi.org/10.1016/j.jclinepi.2019.02.004>
7. Cohen AJ, Brauer M, Burnett R, et al (2017) Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution. *Lancet* 389:1907–1918. [https://doi.org/10.1016/S0140-6736\(17\)30505-6](https://doi.org/10.1016/S0140-6736(17)30505-6)
8. Collins GS, Moons KGM, Dhiman P, et al (2024) TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385:e078378. <https://doi.org/10.1136/bmj-2023-078378>
9. Degen PM, Riebler A, Held L (2026) Paper mill or paper mine? A tentative answer to the sharp increase in research papers based on the Global Burden of Disease database. *J Clin Epidemiol* 199:112464. <https://doi.org/10.1016/j.jclinepi.2026.112464>
10. Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics. *J R Stat Soc Ser C Appl Stat* 47:299–350. <https://doi.org/10.1111/1467-9876.00113>
11. Elliott P, Wartenberg D (2004) Spatial epidemiology: current approaches and future challenges. *Environ Health Perspect* 112:998–1006. <https://doi.org/10.1289/ehp.6735>
12. Enayati Ahangar F, Pakbin P, Hasheminassab S, et al (2021) Long-term trends of PM_{2.5} and its carbon content in the South Coast Air Basin: a focus on the impact of wildfires. *Atmos Environ* 255:118431. <https://doi.org/10.1016/j.atmosenv.2021.118431>

13. GADM (2022) GADM database of Global Administrative Areas, version 4.1 [dataset]. <https://gadm.org/> (maps for cartographic use only; not redistributed)
14. Harris CR, Millman KJ, van der Walt SJ, et al (2020) Array programming with NumPy. *Nature* 585:357–362. <https://doi.org/10.1038/s41586-020-2649-2>
15. Kapoor S, Narayanan A (2023) Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4:100804. <https://doi.org/10.1016/j.patter.2023.100804>
16. Mao H, Martin R, Reich BJ (2024) Valid model-free spatial prediction. *J Am Stat Assoc* 119:904–914. <https://doi.org/10.1080/01621459.2022.2147531>
17. McDermott MBA, Wang S, Marinsek N, et al (2021) Reproducibility in machine learning for health research: still a ways to go. *Sci Transl Med* 13:eabb1655. <https://doi.org/10.1126/scitranslmed.abb1655>
18. McKinney W (2010) Data structures for statistical computing in Python. *Proc SciPy* 56–61
19. NASA (2024) POWER (Prediction of Worldwide Energy Resources) [dataset]. NASA Langley Research Center. <https://power.larc.nasa.gov/>
20. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT (2018) The preregistration revolution. *Proc Natl Acad Sci USA* 115:2600–2606. <https://doi.org/10.1073/pnas.1708274114>
21. Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830
22. Rasmussen CE, Williams CKI (2005) *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA
23. Roberts DR, Bahn V, Ciuti S, et al (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40:913–929. <https://doi.org/10.1111/ecog.02881>
24. Singh G, Moncrieff G, Venter Z, et al (2024) Uncertainty quantification for probabilistic machine learning in earth observation using conformal prediction. *Sci Rep* 14:16166. <https://doi.org/10.1038/s41598-024-65954-w>
25. Tibshirani RJ, Foygel Barber R, Candès E, Ramdas A (2019) Conformal prediction under covariate shift. *Adv Neural Inf Process Syst* 32
26. van Donkelaar A, Martin RV, Brauer M, et al (2016) Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. *Environ Sci Technol* 50:3762–3772. <https://doi.org/10.1021/acs.est.5b05833>
27. Vovk V, Gammerman A, Shafer G (2005) *Algorithmic Learning in a Random World*. Springer, New York
28. Wang Y, Trentmann J, Pfeifroth U, Yuan W (2019) Improvement of air pollution in China inferred from changes between satellite-based and ground-based measurements. *Remote Sens* 11:2910. <https://doi.org/10.3390/rs11242910>
29. World Health Organization (2021) WHO global air quality guidelines: particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization,

30. Wynants L, Van Calster B, Collins GS, et al (2020) Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 369:m1328.
<https://doi.org/10.1136/bmj.m1328>
31. Xue T, Geng G, Meng X, et al (2022) New WHO global air quality guidelines help prevent premature deaths in China. *Natl Sci Rev* 9:nwac055. <https://doi.org/10.1093/nsr/nwac055>
32. Xue T, Liu J, Zhang Q, et al (2019) Rapid improvement of PM_{2.5} pollution and associated health benefits in China during 2013–2017. *Sci China Earth Sci* 62:1847–1856.
<https://doi.org/10.1007/s11430-018-9348-2>
33. Zheng B, Tong D, Li M, et al (2018) Trends in China's anthropogenic emissions since 2010 as the consequence of clean air actions. *Atmos Chem Phys* 18:14095–14111.
<https://doi.org/10.5194/acp-18-14095-2018>

Figures

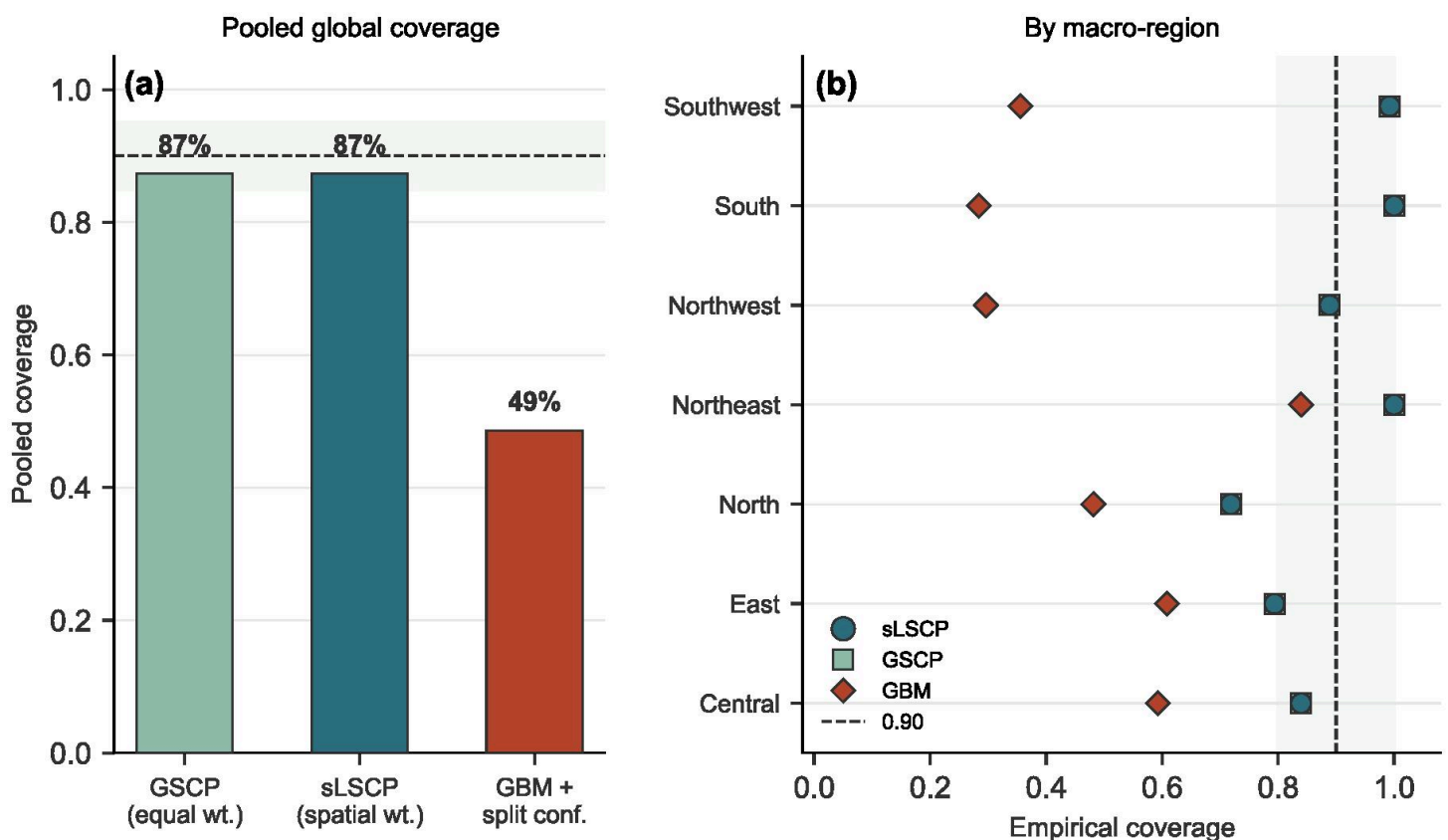


Figure 1

Coverage of nominal 90% intervals by macro-region. (a) Pooled global coverage of the nominal 90% prediction intervals under leave-one-macro-region-out validation for three interval-construction methods sharing the same Gaussian-process predictor: GSCP (split conformal with equal weights), sLSCP

(spatially weighted; the method under test), and gradient boosting with split conformal prediction (GBM). The dashed line marks nominal coverage (0.90) and the shaded band the preregistered acceptance range [0.85, 0.95] used in hypothesis H1. (b) Empirical coverage by macro-region (leave-one-macro-region-out, all years); the shaded band marks the region-wise acceptance range [0.80, 1.00] used in hypothesis H2. The calibration set is a random 30% of the 1998–2019 training pool (seed 20260917).

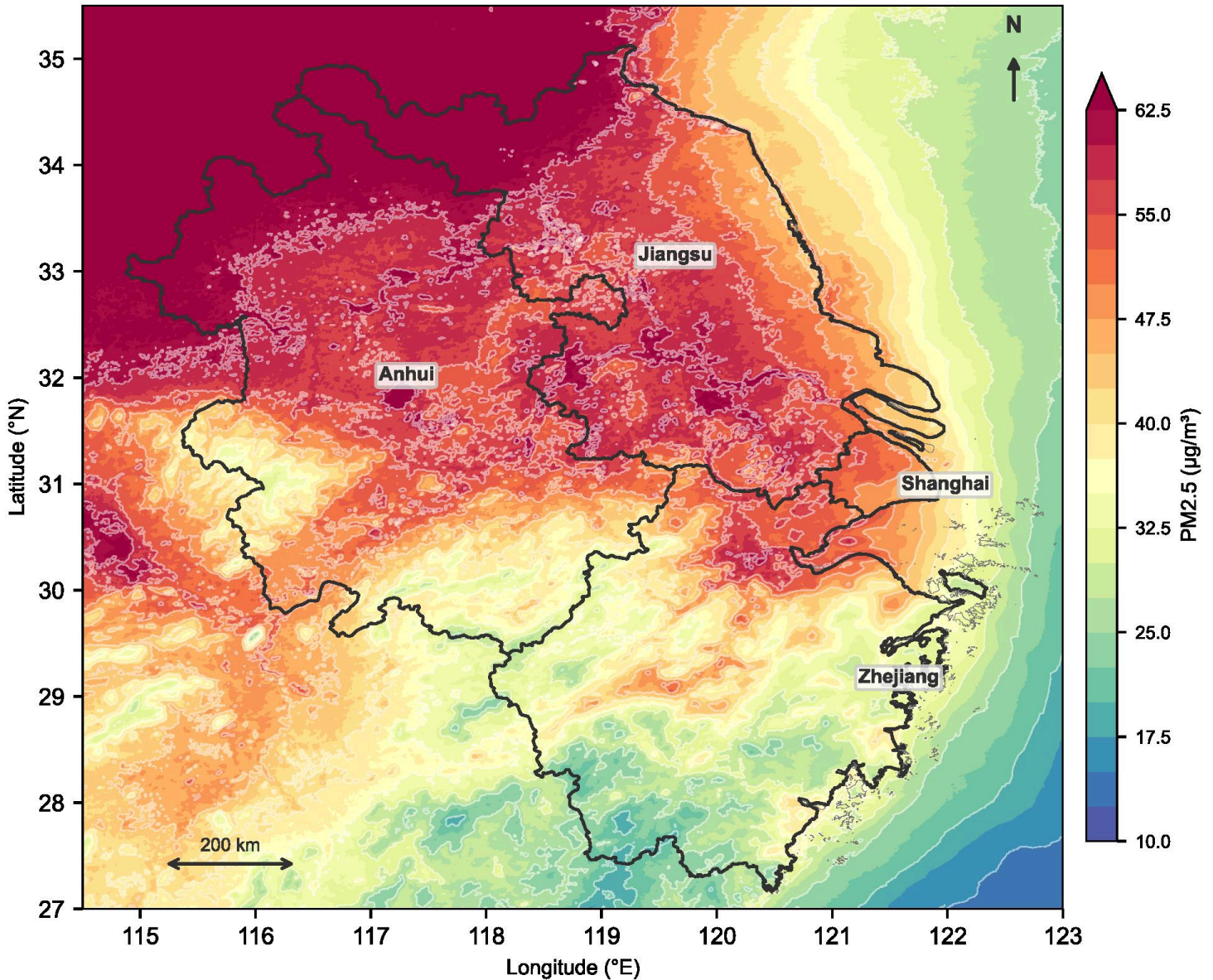


Figure 2

Contour map of annual mean PM_{2.5} concentrations over the Yangtze River Delta, 2015. Filled contours show the 2015 annual mean PM_{2.5} field from the ACAG Surface PM_{2.5} V6.GL.03 Asia grid (0.01° resolution, CC BY 4.0); boundaries of Anhui, Jiangsu, Zhejiang and Shanghai are overlaid. Concentrations grade from the heavily loaded belt across northern Anhui and Jiangsu toward the cleaner coastal southeast. White contour lines mark 5 µg/m³ intervals; the north arrow and scale bar are approximate.

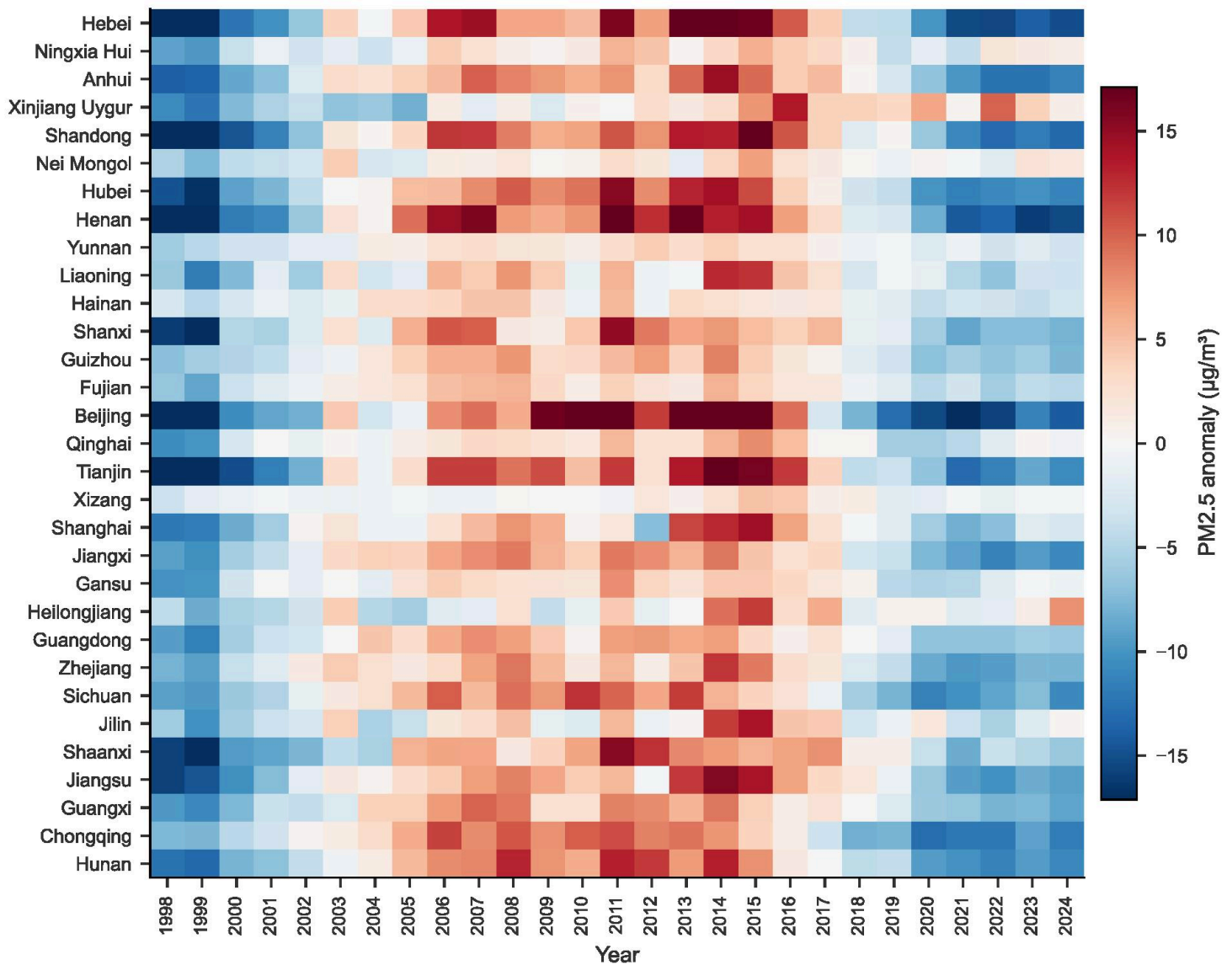


Figure 3

PM2.5 anomaly relative to each province's 1998–2024 mean. Annual population-weighted PM2.5 minus the 1998–2024 mean of the same province, for all 31 provinces (rows, ordered by long-term mean) and all 27 years (columns). Red = above the province's long-term mean; blue = below. The sustained red block through the mid-2010s followed by a blue shift after 2017 reflects the large reduction in PM2.5 concentrations over China during this period, and motivates the leave-years-out partition as a test under temporal distribution shift.

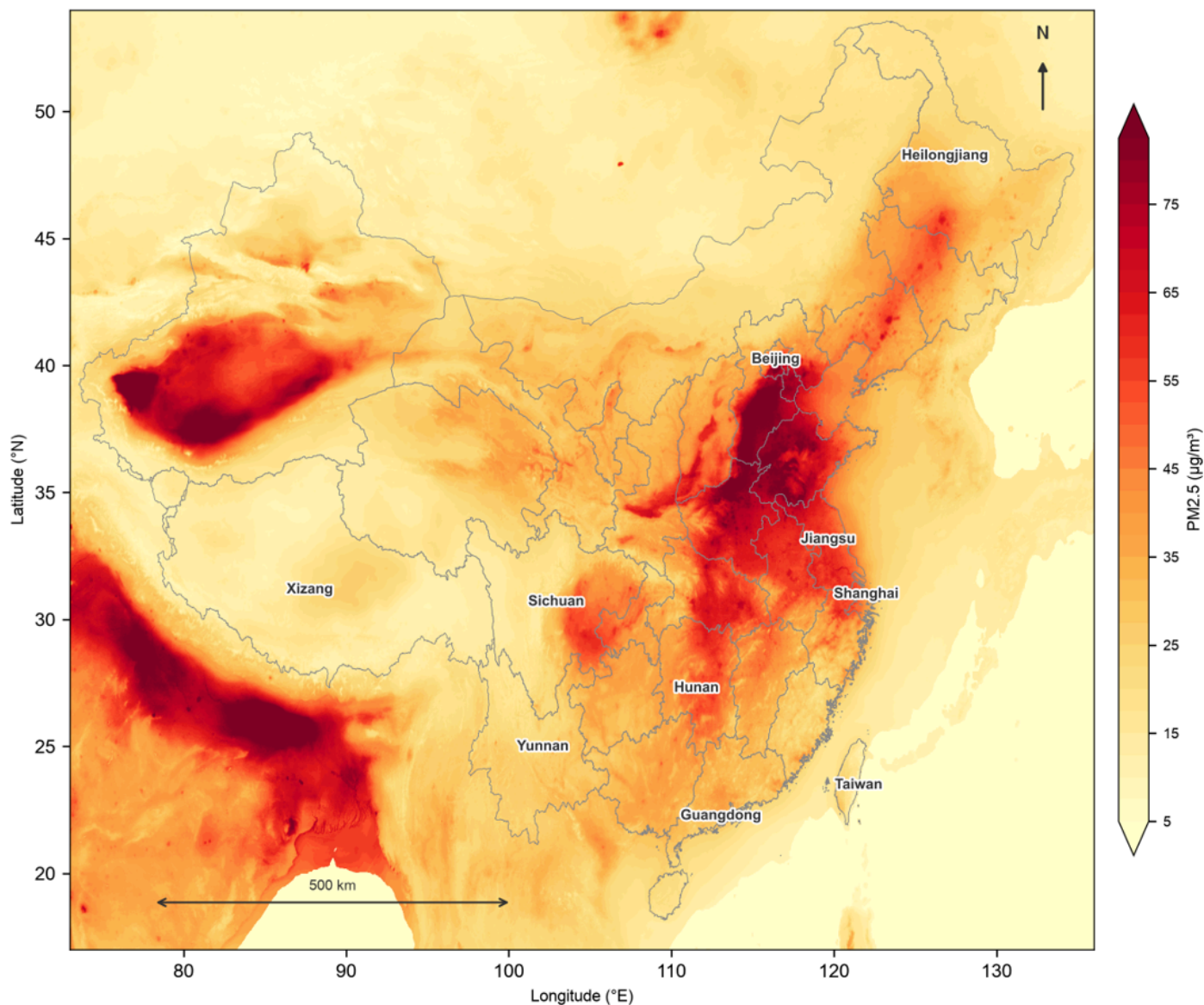


Figure 4

Contour map of annual mean PM_{2.5} concentrations across China, 2015. Filled contours show the 2015 annual mean PM_{2.5} field from the ACAG Surface PM_{2.5} V6.GL.03 Asia grid (0.01° resolution, CC BY 4.0). Administrative boundaries are drawn from GADM (mainland provinces, and Taiwan from the GADM Taiwan dataset). The North China Plain and the Tarim Basin dominate the concentration surface. The boundaries do not represent the author's opinion. The scale bar is approximate.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterial.docx](#)