Dear Editors and Reviewer 1,

We appreciate the opportunity to revise the manuscript, “Methane Emissions Estimation from China's Leading Coal Production Hub: A Hybrid Hyperspectral Satellite Observations and Emission Inventory Framework” (Manuscript ID: EGUSPHERE-2025-4610), which was submitted for publication in Atmospheric Chemistry and Physics.

We have gone through a detailed revision on this manuscript with utmost effort. We have addressed all the points raised, and revised the manuscript accordingly. We hope that the editor and the reviewer can find the revised version has been significantly improved. Please find enclosed a point-by-point response to each comment. The text in black is a copy of the comments. Our responses are in text labeled with blue, and our revisions are in text highlighted in red.

We look forward to hearing from you regarding the outcome of our paper.

Sincerely yours,

Yongguang Zhang on behalf of all the authors

Reviewer #1:

General comments

This manuscript presents an approach to estimate coal mine methane emissions in Shanxi by combining multi-satellite hyperspectral observations with a hierarchical Bayesian inversion framework. The topic is timely and relevant, particularly in the context of increasing interest in facility-scale methane monitoring using emerging satellite observations. The use of multiple hyperspectral datasets and the attempt to address sparse temporal sampling through a hierarchical Bayesian framework are both valuable contributions. The integration of satellite observations with inventory data is also well motivated, and the results provide useful insights into emission factors as well as spatial and temporal patterns of methane emissions.

Overall, the manuscript is clearly written and addresses an important problem. The methodology is promising, and the results are of potential interest to both the remote sensing and atmospheric science communities. While a few aspects could benefit from further clarification and discussion, I believe these can be addressed with relatively minor revisions.

Response: We sincerely thank the reviewer for the constructive comments and positive assessment of our work. We have carefully addressed each comment and revised the manuscript accordingly. These revisions have substantially strengthened the scientific rigor of our study.

Major comment

One point that may deserve a bit more careful discussion concerns the use of satellite observations to characterize the distribution of methane emissions. While the general approach is well motivated and has been explored in previous studies, it is important to consider the detection limits of hyperspectral instruments. For example, the detection threshold for instruments such as EnMAP is on the order of ~500 kg/h under favorable conditions, and this threshold may be higher in regions with complex or heterogeneous surface conditions, such as Shanxi.

In this context, it would be helpful for the authors to elaborate on how such detection limits might influence the inferred emission distribution. In particular, there may be a tendency for satellite observations to preferentially capture stronger emission events, while smaller but more frequent sources remain undetected. This could potentially affect the characterization of the emission distribution and the relative importance of extreme events. In related studies that focus specifically on emission distributions, airborne campaigns are often used, partly because they can achieve substantially lower detection limits and therefore provide a more complete sampling across the full range of emission strengths. A brief comparison or discussion along these lines would help place the current results in context.

Related to this, the choice of Shanxi as the primary study region is understandable given its importance in coal production, but its heterogeneous surface conditions may pose additional challenges for plume detection. A brief discussion on how representative the observed emission distribution is, given these observational constraints, would help strengthen the interpretation of the results.

Response: Thank you for this important comment. We have added a dedicated section (Sect. 2.5) to quantify satellite detection limits under conditions representative of Shanxi. Using controlled plume-injection experiments with LES plumes superimposed on real retrieval backgrounds from all seven sensors, we derived sensor-specific detection limits that explicitly account for surface heterogeneity and wind dependence. We also performed leave-one-sensor-out sensitivity tests to assess the robustness of the inferred emission factor to the composition of the satellite archive. These analyses confirm that weaker sources are preferentially missed and that the observed distribution is therefore weighted toward stronger emitters. We have revised the discussion accordingly and now interpret the inferred distribution within these observational constraints.

Revised:

2.5 Inter-satellite comparability and detection limits

We designed a controlled plume-injection experiment by combining large-eddy simulations with real observations from seven satellite sensors over a methane-free scene in Changzhi, a major coal-producing region of Shanxi Province.

Large-eddy simulations were performed with WRF-v3.9.1 under real meteorology, with ERA5 providing initial and boundary conditions, over two typical coal-producing districts of Changzhi, Shanxi. The triply-nested configuration used horizontal resolutions of 3630, 330 and 30 m over domains of 200, 51 and 9.9 km, with 110 vertical layers of which 40 lie within the first kilometre. Methane was released as an inert tracer from the surface at the centre of the innermost domain (Varon et al., 2018). Simulations were run for January, April, July and October at prescribed ambient wind speeds of 3, 5, 8 and 10 m/s. After model-physics equilibration and the development of three-dimensional turbulence, the final half hour of 10-minute output was used to construct plumes at 30 and 60 m resolution. Methane column enhancements retrieved over an emission-free coal-producing district of Changzhi covered by all seven satellites provide the sensor-specific noise, preserving both its local spatial heterogeneity and its overall distribution. Combining the two yields 6 km × 6 km synthetic observations for seven sensors at emission rates of 100, 300, 500, 1000, 3000, 5000 and 10000 kg/h, 672 scenarios per sensor and 4704 in total (**Figure A5-1 and A5-2**). Every scenario was passed through the plume detection, segmentation and IME chain of Sect. 2.2 by the same analysts, with non-detections explicitly recorded. The noise-free simulation processed through the same chain provides the reference quantification of each scenario.

Comparison is made on the ratio $R = (\mathrm{IME}/L)_{\mathrm{sensor}}/(\mathrm{IME}/L)_{\mathrm{LES}}$, it isolates the combined effect of retrieval noise, mask extraction and IME integration and remains orthogonal to the wind term of **Eq. (8)**. Effective detection limits over Shanxi span a factor of seven across the constellation, from $\mathrm{DL}_{50} = 74$ kg/h for GHGSat to 486 kg/h for EMIT, with DL_{90} ranging from 144 to 912 kg/h (**Table C3, Figure A6a**). Accordingly, weaker sources near or below these thresholds are less likely to be represented in the detected plume archive, particularly for sensors with higher detection limits, resulting in comparatively sparser sampling of the lower-emission range. We find that the commonly quoted detection threshold of ~500 kg/h commonly quoted for PRISMA and EnMAP class instruments under favourable conditions corresponds to our DL_{90} for PRISMA (505 kg/h) rather than to DL_{50}. Wind is a strong secondary control, raising DL_{50} by a factor of 1.4 to 2.4 between ambient winds of 3 m/s and 10 m/s, most steeply for the coarser or noisier instruments (**Figure A6b**). Because the background fields are real retrievals over the heterogeneous, low-albedo surfaces of the Changzhi mining district, these limits embody the surface complexity of the study region rather than that of a bright homogeneous test site.

The seven sensors nevertheless quantify consistently. Over the 2687 valid quantifications with Q⩾ 1000 kg/h, the relative bias $b_s = \mathrm{median}(R) - 1$ spans only 13.7 percentage points and the within-sensor dispersion u_s lies between 6.3 and 8.9 % (**Table C3**). All sensors share a positive offset relative to the noise-free chain $\bar{b} = +31.4\,\%$, which arises from differences between idealized and realistically detectable plume masks and is already embedded in the U_{eff} calibration. Accordingly, only the inter-sensor spread in b_s is propagated in **Eq. (12)**.

$$\sigma_{\mathrm{sat}} = \sqrt{\left[\sum_s f_s \left(b_s - \bar{b}\right)^2\right]} = 5.8\% \tag{13}$$

Sensitivity of the emission factor to the composition of the archive was tested by removing each sensor in turn, re-estimating the hyper-parameters on the remaining plumes, repeating the inversion and recomputing the emission factor over the mines that still carry a detection, against a baseline evaluated on the same mines (**Figure A7**). Removing GHGSat, which contributes 70 % of all observations, changes the emission factor by 6.2 % and removing GF5-02 AHSI by −5.5 %. Every other sensor changes the result by at most 2.0%, with all variations remaining within σ_{EF}. For comparison, switching off the seasonal layer of the inversion changes the emission factor by −3.5%, whereas replacing the inversion with arithmetic means across all 13 mines raises it by 20.0%.

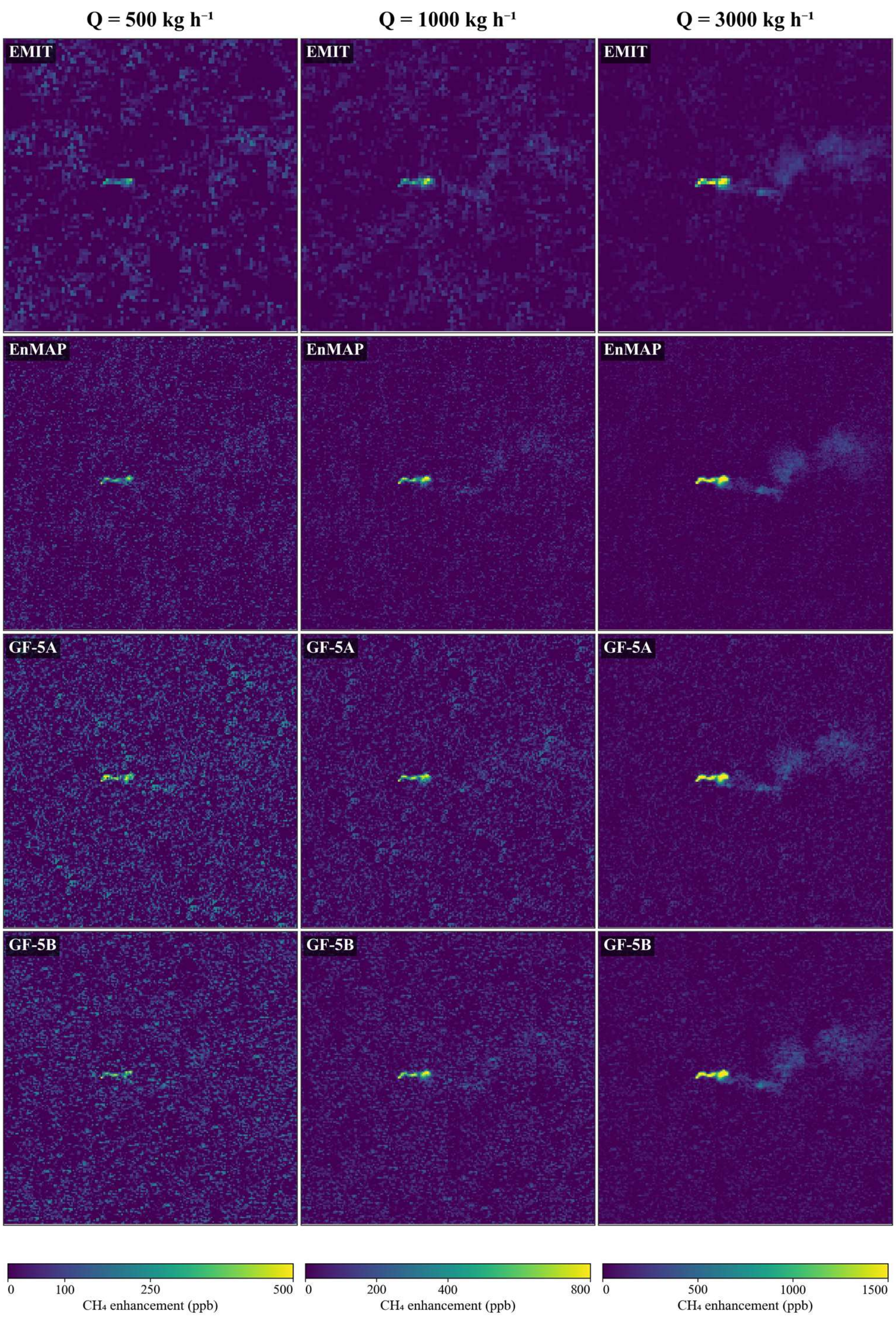

Figure A5-1. Synthetic observations of the plume-injection experiment.

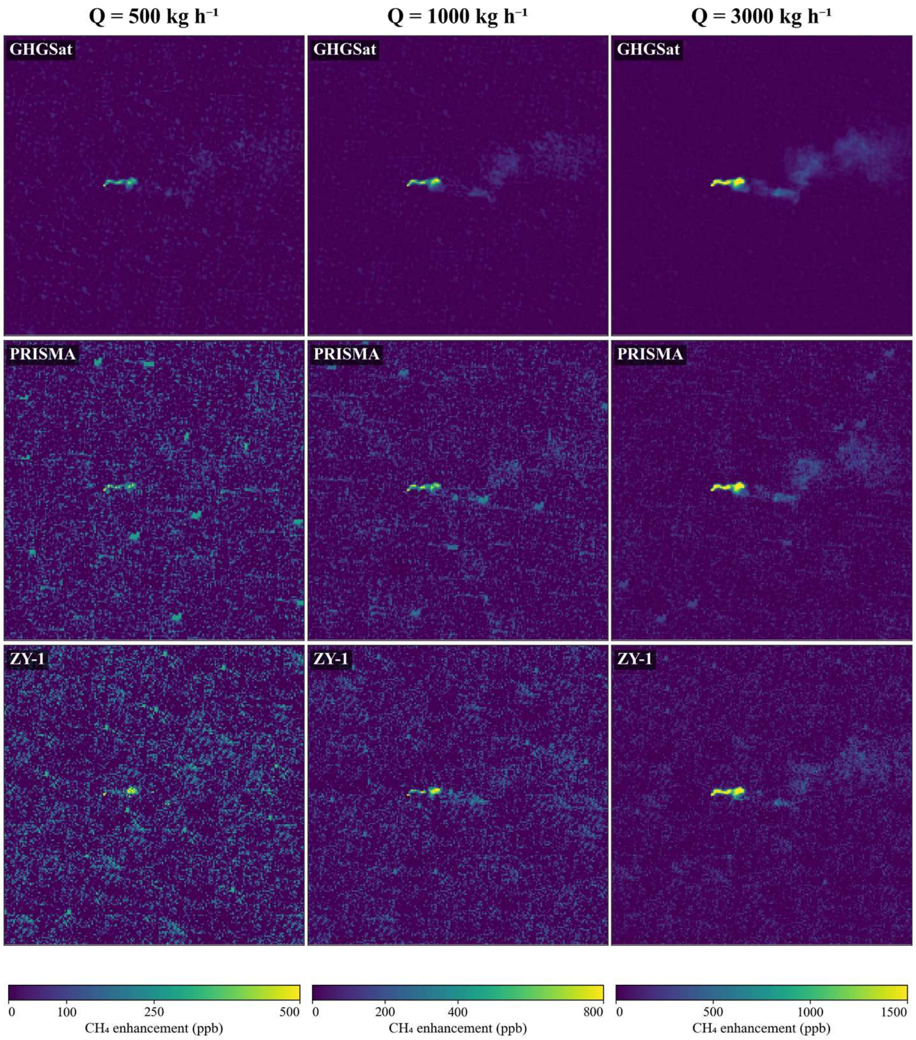

Figure A5-2. Synthetic observations of the plume-injection experiment.

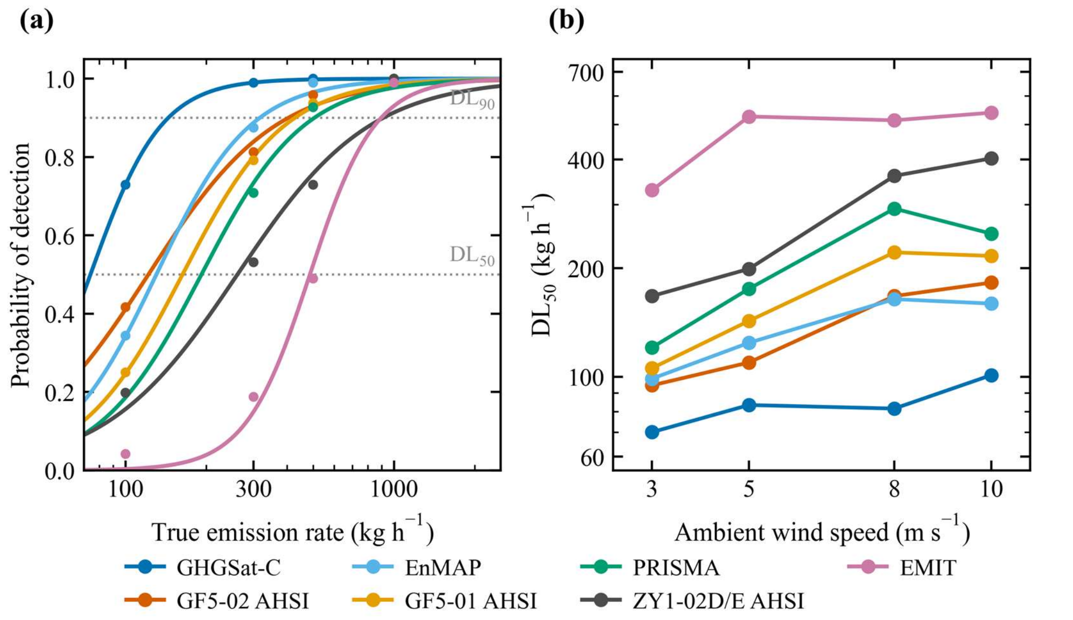

Figure A6. Detectability of CMM plumes in the plume-injection experiment. (a) Probability of detection against true emission rate, curves are logistic fits and symbols the observed detection frequencies over the 672 scenarios per sensor, with dotted lines marking the 50% and 90% levels that define DL_{50} and DL_{90}. (b) DL_{50} as a function of the prescribed ambient wind speed of the large-eddy simulation.

Table C3. Per-sensor results of the plume-injection experiment (672 scenarios per sensor). DL_{50} and DL_{90} are the emission rates at which the probability of detection reaches 50% and 90%. b_s and u_s are the relative bias and the relative dispersion of the quantification ratio R for $Q \geq 1000\ kg/h$.

Sensor	DL_{50} (kg/h)	DL_{90} (kg/h)	POD_{300}	POD_{500}	b_s (%)	u_s (%)
GHGSat	74	144	0.99	1.00	+27.6	7.5
GF5-02 AHSI	122	404	0.81	0.96	+40.4	8.0
GF5-01 AHSI	163	427	0.79	0.94	+40.8	8.3
PRISMA	192	505	0.71	0.93	+40.9	8.6
EMIT	486	896	0.19	0.49	+28.7	6.7
EnMAP	130	316	0.88	0.99	+29.6	6.3
ZY1-02 AHSI	261	912	0.53	0.73	+41.3	8.9
Archive-weighted					$\bar{b} = +31.4$	$\sigma_{\mathrm{sat}} = 5.8$

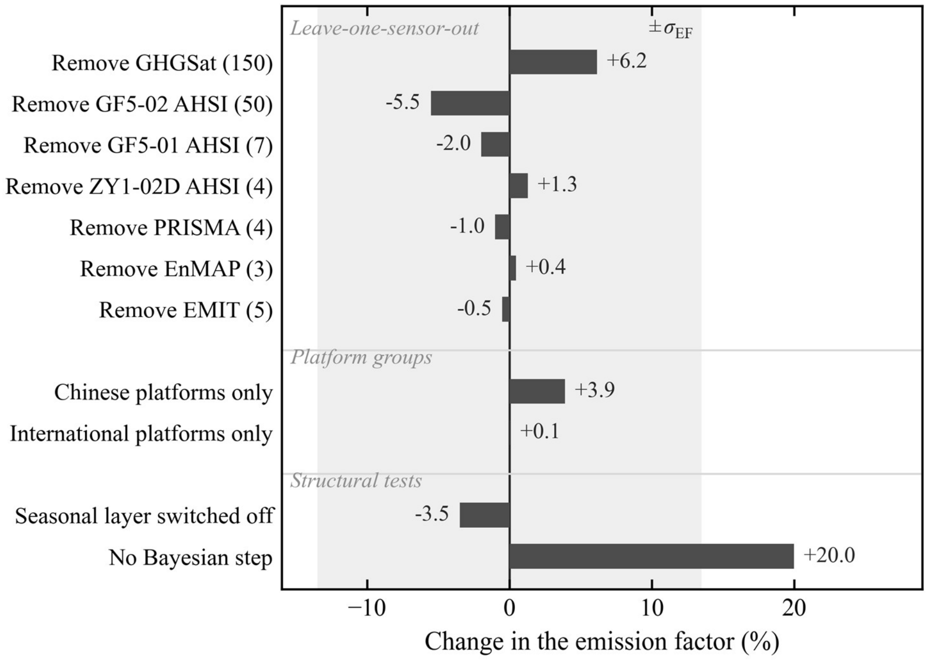

Figure A7. Change in the provincial emission factor when individual sensors, platform groups or structural elements of the inversion are removed, each evaluated against a baseline over the same surviving mines. Numbers in parentheses are the plumes withdrawn and the shaded band is $\pm\sigma_{EF}$.

Minor comments

1. The description of the hierarchical Bayesian framework could be slightly clarified to improve readability, particularly for readers less familiar with this type of method.

Response: Thank you for this suggestion. We have clarified Sect. 2.3 by explicitly defining the hierarchical structure, variance components, shrinkage mechanism, posterior estimation, and cross-validation procedure, with additional figures illustrating the prior-to-posterior update and model performance.

Revised:

2.3 Hierarchical Bayesian Inversion Algorithm

To accurately estimate annual methane fluxes and reduce estimation errors caused by limited hyperspectral satellite revisit frequencies, we developed a novel Hierarchical Bayesian inversion algorithm. The algorithm extracts global and seasonal characteristics from 2019-2023 Shanxi time-series observations, combining these with limited target mine observations to estimate annual methane emission fluxes, as illustrated below.

Let q_{ij} denote the *j*-th plume quantification at mine *i*.

$$\log q_{ij} = \mu_0 + s_{season(j)} + \theta_i + \varepsilon_{ij}, \theta_i \sim N(0, \sigma_b^2), \varepsilon_{ij} \sim N(0, \sigma_w^2), \sum_S S_S = 0 \quad (1)$$

The global level μ_0 is the log-scale emission level of Shanxi coal mines as a population. The seasonal level consists of four fixed offsets S_S constrained to sum to zero. The mine level θ_i is the departure of mine *i* from the provincial level. σ_b measures between-mine heterogeneity and σ_w within-mine dispersion, the latter containing both genuine temporal variability and quantification error. The log-normal specification is confirmed by comparison with gamma, Weibull, exponential, and normal alternatives using the Akaike information criterion and the Kolmogorov–Smirnov test, evaluated both for the full archive and separately by season (**Table C1**).

Hyper-parameter estimation: $(\mu_0, S_S, \sigma_b, \sigma_w)$ are estimated by maximum likelihood from the 215 plumes at the 26 time-series mines. The exchangeable covariance of mine i, $\boldsymbol{V}_i = \sigma_w^2 \boldsymbol{I}^2 + \sigma_b^2 \mathbf{1}\mathbf{1}^T$, is used directly, so that mines with different numbers of observations are weighted correctly rather than being pooled as independent samples.

Posterior and annual emission rate: For a target mine with n observations and de-seasonalised log mean $\hat{y}_i$, the mine effect has a conjugate normal posterior,

$$\hat{\theta}_i = w_n(\hat{y}_i - \mu_0), w_n = \frac{n\sigma_b^2}{n\sigma_b^2 + \sigma_w^2} \tag{2}$$

$$\sigma_\theta{}^2 = \frac{\sigma_b^2 \sigma_w^2}{n\sigma_b^2 + \sigma_w^2} \tag{3}$$

The shrinkage weight w_n is fixed entirely by the variance components and the sample size, interpolating between the provincial mean and the mine's own record, so that no single detection is ever taken at face value. The annual mean emission rate follows as the four season average evaluated at the posterior mine effect,

$$\hat{Q}_i = \frac{\sum_S exp\left(\mu_0 + \hat{\theta}_i + S_S + 0.5\sigma_w^2\right)}{4} \tag{4}$$

$$\sigma_{Q,post} = \hat{Q}_i \sqrt{[exp(\sigma_\theta{}^2) - 1]} \tag{5}$$

where $0.5\sigma_w^2$ is the standard log-normal correction from a log-scale location to an arithmetic mean. Prior, individual overpasses and posterior estimates for the six mines with limited observations, together with the dependence of w_n and the posterior dispersion on the number of observations, are shown in **Figure A3**.

The established Hierarchical Bayesian Inversion is validated on the eight mines with at least ten observations, for which a well constrained annual mean is available. Each mine is held out in turn. For a held-out mine, $m \in \{1,2\}$ observations are drawn at random to emulate a limited-observation mine, the hyper-parameters are re-estimated from the remaining 25 mines, and the subset is inverted. The estimate is compared with the arithmetic mean of the full observation record of that mine, which is the quantity reported for time-series mines in this study. Enumerating all mines and subsets gives 921 scenarios, benchmarked against the arithmetic mean of the drawn observations.

The inversion has a median relative bias of −7.7 % and a relative RMSE of 37.1 %, against −5.4 % and 44.3 % for the arithmetic mean (**Table C2**). When the drawn observations include a plume above 10^4 kg/h, the relative RMSE is 27.9 % for the Hierarchical Bayesian Inversion and 51.5 % for the arithmetic mean. Shrinkage towards the provincial prior thus prevents a fortuitous detection of a rare event from propagating into an annual mean (**Figure A4**). A single observation is already informative, at $m = 1$ the RMSE is 40.6 %, against 62.0 % for the arithmetic mean. Nominal 68.3 % and 95.4 % posterior intervals from **Eq. (5)** contain the reference in 67.2 % and 86.2 % of scenarios. The cross-validated RMSE is adopted as the mine-level uncertainty in Sect. 2.4.

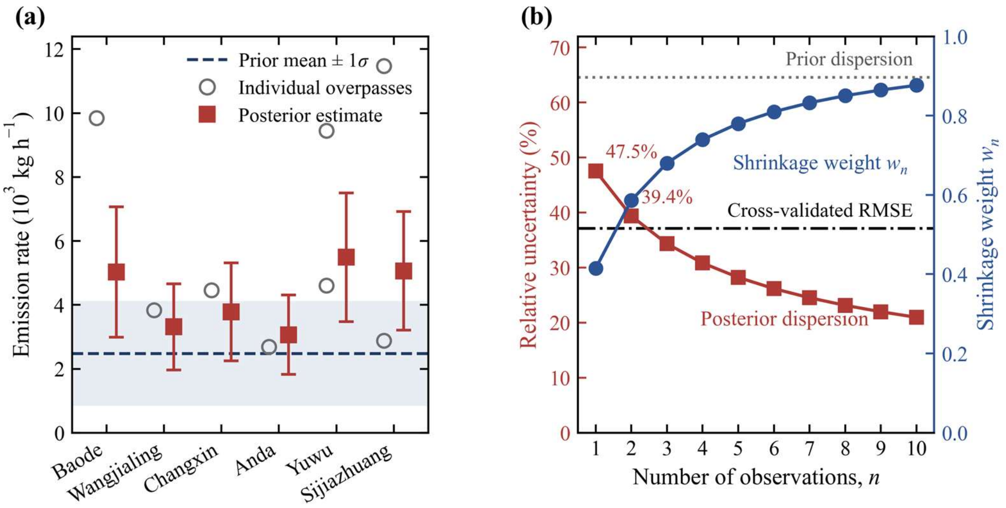

Figure A3. (a) Prior, individual overpasses and posterior estimates for the six mines with limited observations. (b) Posterior dispersion (left axis) and shrinkage weight w_n (right axis) against the number of observations. Dotted line, prior dispersion. Dash-dotted line, cross-validated RMSE.

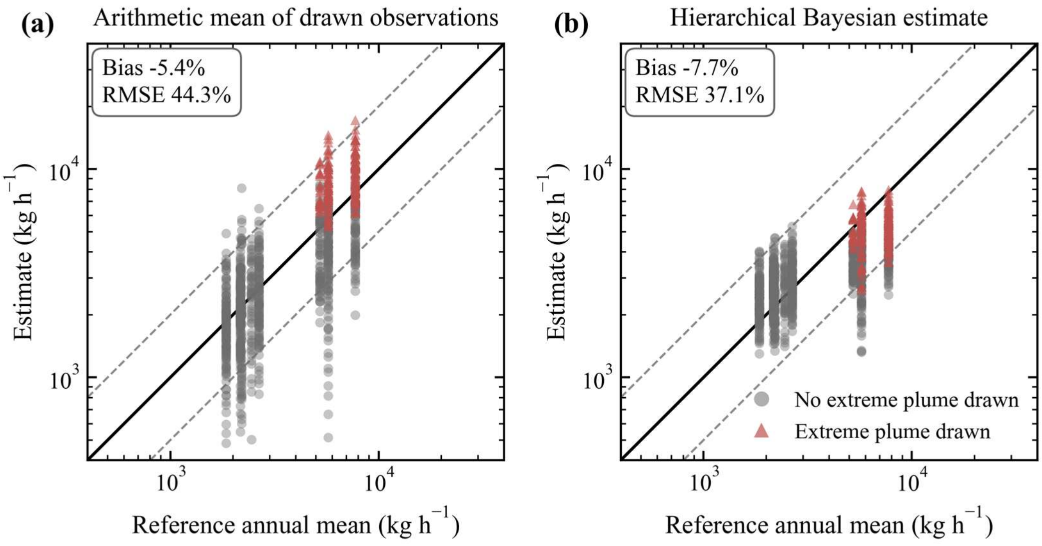

Figure A4. Cross-validation of the hierarchical inversion. (a) Arithmetic mean of the drawn observations and (b) hierarchical Bayesian estimate. Solid line, 1:1. Dashed lines, factor of two.

Table C1. Comparison of distributional support for the hierarchical model using 215 plume emission rates from 26 time-series mines. ΔAIC is relative to the log-normal distribution in each column.

Distribution	ΔAIC (pooled, n=215)	ΔAIC Spring (42)	ΔAIC Summer (18)	ΔAIC Autumn (74)	ΔAIC Winter (81)	KS *p*
Log-normal	0.0	0.0	0.0	0.0	0.0	0.996
Gamma	16.9	4.6	0.4	7.5	9.7	0.171
Weibull	22.1	6.7	1.2	8.8	12.0	0.242
Exponential	25.6	-	-	-	-	0.068

Table C2. Cross-validation performance, given as median relative bias/relative RMSE (%).

Stratum	n	Arithmetic mean	Hierarchical inversion
All scenarios	921	-5.4/44.3	−7.7/37.1
m=1	115	-9.9/62.0	−12.9/40.6
m=2	806	-4.4/41.2	−6.6/36.6
Subset contains an extreme plume	178	33.4/51.5	−20.5/27.9
Subset contains no extreme plume	743	-14.2/42.4	-2.2/39.0

2. It would be helpful to briefly clarify the extent to which plume identification relies on manual interpretation. In addition, in Figure 2 (lower right panel, EMIT), part of the detected plume appears visually correlated with underlying terrain features, which may suggest a potential false positive. A brief clarification on how such cases are handled would improve confidence in the plume identification procedure.

Response: Thank you for this comment. We have clarified that plume identification involves manual interpretation guided by GEOS-FP 10-m winds and facility location. Our primary criterion is the presence of a distinct high-enhancement plume core near an emission facility. If this core exhibits a physically consistent plume orientation and dispersion pattern, downstream enhancement that follows or accumulates along complex terrain can still be considered plausible within a limited area rather than automatically classified as a false positive.

Revised: GEOS-FP 10-m wind data (U_{10}) were used to support manual plume identification; a plume was deemed valid when its elongation direction formed an acute angle with the wind vector (**Figure A2a–c**) (Lauvaux et al., 2022). The criterion was applied to the plume core rather than the full dilated-mask outline, as downstream plume morphology may be progressively modified by wind shear and turbulent meandering during transport. Plume origins were attributed to specific coal mines using Google Earth imagery, and high-value pixels within each plume were assigned to the corresponding emission facility (**Figure A2d**). In this case, all three plumes traced back to ventilation shafts at the same mine.

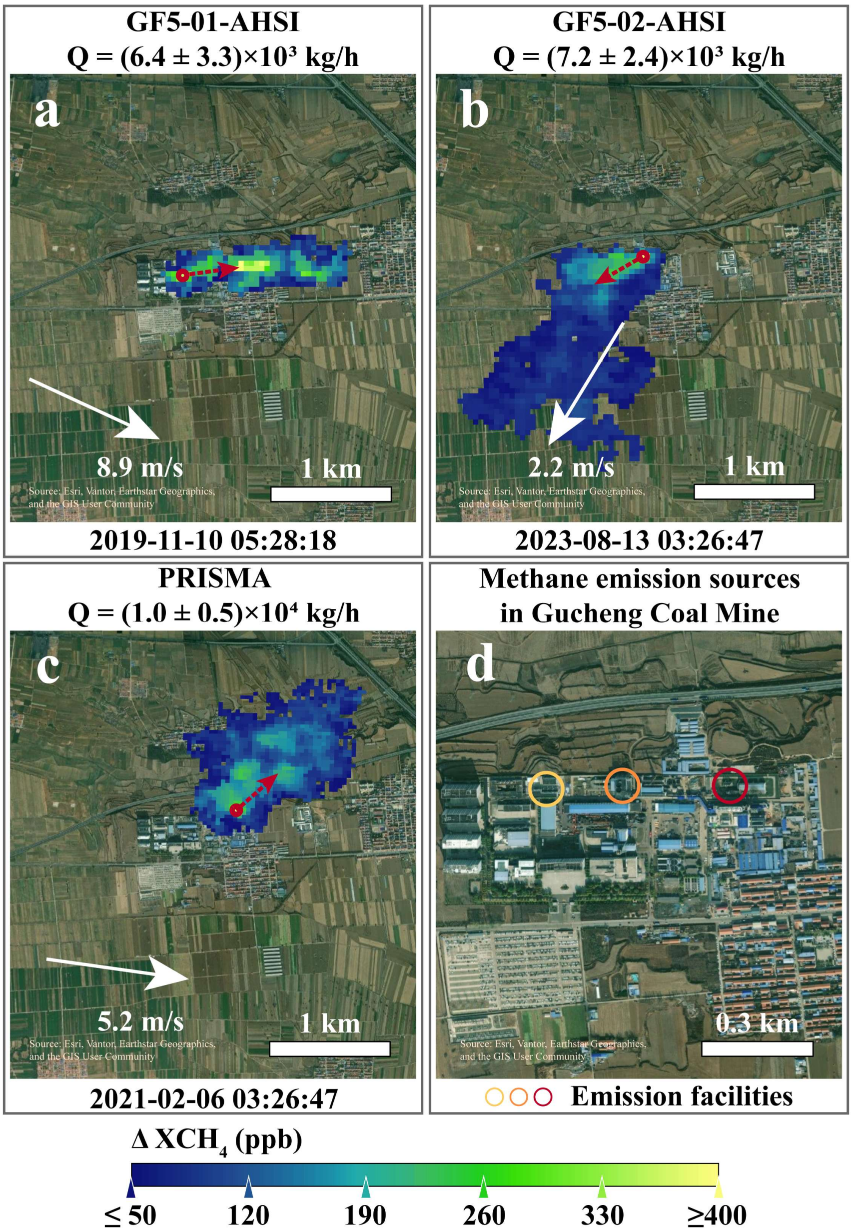

Figure A2. Examples of methane plume identification and source attribution. (a)–(c) Methane plumes observed by GF5-01, GF5-02, and PRISMA, respectively, and attributed to the Gucheng coal mine. (d) Attribution of the three methane plumes to their respective emission sources within the coal mine. The red dashed arrow denotes the methane concentration gradient direction near the plume source. The circled features indicate the attributed facilities. Background RGB imagery in (a)-(d) is Powered by Esri.

3. The discussion on detection limits could be expanded slightly to note potential implications for smaller emission sources.

Response: Thank you for this suggestion. We have expanded the discussion of detection limits to clarify that weaker sources near or below the sensor-specific thresholds are less likely to be represented, leading to sparser sampling of the lower-emission range.

Revised: Effective detection limits over Shanxi span a factor of seven across the constellation, from DL_{50} = 74 kg/h for GHGSat to 486 kg/h for EMIT, with DL_{90} ranging from 144 to 912 kg/h

(**Table C3, Figure A6a**). Accordingly, weaker sources near or below these thresholds are less likely to be represented in the detected plume archive, particularly for sensors with higher detection limits, resulting in comparatively sparser sampling of the lower-emission range.