Response to Reviewers

Dear Editors and Reviewers:

Thank you for your letter and for the reviewers' comments concerning our manuscript entitled "***A global base temperature dataset for residential building energy demand modeling***" (**ID: essd-2025-709**). We have studied the comments carefully and have made corrections which we hope will meet with your approval. The main corrections in the paper and the response to the reviewers' comments are as follows (For clarity, our point-by-point responses to the reviewers are shown in **blue**, and the corresponding changes in the manuscript are highlighted in **red**).

Reviewer #1: Overall, this paper presents an important dataset to show the base temperature globally. Without doubt, their effort will speed up the method to benchmark/set the base temperature, thereby regulating the adoption of air-conditioning systems or central heating systems for heating or cooling. This paper is overall well-written and logical. However, I suggest authors should make some revisions.

1. Authors may reflect residential buildings in the title of this paper.

Response: Thanks very much for the professional comments. Following the suggestions, we have revised the title to specify residential buildings, which better reflects the composition of our energy demand data. The revised title is "A global base temperature dataset for residential building energy demand modeling".

2. The introduction can be rewritten to shorten the information presented before line 100, while the sentence in Line 103-105 and subsequent descriptions in this paragraph are odd. These contents do not really show a solid research gap. I encourage authors should re-visit how existing methods are costly and impractical. The mechanism to benchmark the base temperature in existing studies should also be criticized with their strengths and weaknesses.

Response: Thanks very much for the professional comments. We have (1) condensed the background in the first two paragraphs; (2) rewritten the research gap to make explicit that energy-signature and performance-line methods, despite providing empirically grounded, region-specific base temperatures, require high-resolution building energy data matched with meteorological records, which confines them to data-rich regions and makes global benchmarking costly and impractical; and (3) added an explicit discussion of the strengths and weaknesses of these methods (Introduction, paragraphs 3–4, lines 70-105).

> Accurate base temperature selection is essential for reliable degree-day modeling. The energy signature approach and performance line analysis are commonly used to determine optimal base temperatures at regional scales by integrating building energy consumption data and local meteorological data (Bhatnagar et al., 2018; Staffell et al., 2023; Anjomshoaa & Salmanzadeh, 2017; Woods & Fuller, 2014). In the energy signature method, energy consumption is correlated with outdoor temperature to identify a non-linear relationship, and the inflection point of the resulting curve is taken as the base temperature (Anjomshoaa & Salmanzadeh, 2017; Meng et al., 2018). The performance line method estimates the base temperature by plotting the best-fit straight line of energy

consumption relative to the HDD or CDD (Bhatnagar et al., 2018). A major strength of both approaches is that they use observed energy use that integrates local weather conditions and occupancy behaviour, providing empirically grounded base temperatures for specific regions. For example, Bhatnagar et al. (2018) applied both the energy signature method and the performance line method to determine base temperatures across 60 Indian cities, reporting median base temperatures of 18.3°C for cooling and 17.4°C for heating. In Kerman, Iran, the L-curve relationship between average daily temperature and natural gas consumption is derived using the energy signature method, revealing base temperatures of 15.42 and 21.18°C for the region (Anjomshoaa & Salmanzadeh, 2017). Furthermore, Staffell et al. (2023) evaluated indoor temperature thresholds for heating and cooling across several countries and regions based on the energy signature method. However, these benchmark methods also have important limitations. They typically require building energy consumption data at high temporal resolution together with matched meteorological records, which restricts their application to data-rich regions and makes large scale or global benchmarking costly and impractical. Consequently, although these methods are valuable for local and regional calibration, they cannot be readily extended to produce consistent base temperature estimates across data scarce regions worldwide.

When empirical base temperature calibration is unavailable, regional and global assessments of temperature-dependent building energy demand often rely on harmonized base temperatures rather than locally derived values (e.g., Deroubaix et al., 2021; Scoccimarro et al., 2023; Falchetta et al., 2024). This reliance may obscure spatial heterogeneity in building climate responses and reduce the reliability of degree-day-based energy modeling at broader scales. These challenges highlight the critical need for a practical methodology that can predict base temperatures at a global scale using limited data—a prerequisite for accurate global energy demand modeling. To address this gap, this research develops a global regional-scale base temperature dataset by combining energy demand derived labels from segmented linear regression with a BiLSTM framework incorporating an attention mechanism. Using limited building energy observations together with regional weather and socioeconomic data, cooling base temperature (*Tcool*) and heating base temperature (*Theat*) are mapped at provincial resolution and upscaled to continuous global maps. An integrated validation framework is further established to assess the reliability of the predicted base temperatures for building energy demand modeling. The resulting dataset provides spatially explicit base temperature information to support regional energy demand modeling and low-carbon energy system planning.

3. Another comment on the introduction is that authors have well-charted the factors which affect energy use and the methods to characterize and benchmark the base temperature. However, authors have not presented the factors the terrain or topography. The terrain has also shaped energy use, occupant behaviours, and methods to calculate them. Moreover, in many countries, the determination of base temperature is associated with the culture and behaviours. There are also many studies linking the base temperature with indoor thermal comfort which is diverse among different populations.

Response: We thank the reviewer for this valuable point. We have expanded the discussion of factors influencing base temperature in the Introduction to include (1) terrain and topography, which affect local microclimate, energy use, and occupant responses, and (2) the link between base temperature and indoor thermal comfort, which differs across populations and leads to region-specific thresholds for initiating heating or cooling. We further note that the wide range of such influencing factors is precisely why uniform base temperatures are inadequate, motivating our data-driven approach, which implicitly reflects these combined effects through observed energy demand (Introduction, paragraphs 1–2, lines 30-60).

Reductions in energy demand and carbon emissions in the building sector play a crucial role in achieving the low-carbon energy system transition and the Paris Agreement climate goals (Camarasa et al., 2022). Buildings accounted for 36% of global end-use energy and contributed approximately 39% of energy and process-related CO_2 emissions in 2018, with space heating and cooling representing a major component of total building energy use (UN, 2019). However, building energy demand varies substantially across regions because it is strongly shaped by local climate, ambient temperature, occupant behaviour, and socioeconomic conditions (Deroubaix et al., 2021; Wenz et al., 2017; Staffell et al., 2023). An increase in ambient temperature will reduce heating demand but increase cooling demand across different regions (Abajian et al., 2025; Wang et al., 2023; Filahi et al., 2025; Clarke et al., 2018; Zhou et al., 2013), and identical building configurations exhibit distinct energy demand patterns across different climate zones (Mahdavinejad et al., 2024; Yu et al., 2014). Beyond climate zones, terrain and topography further modulate local microclimate, ventilation, and solar exposure, thereby shaping heating and cooling energy demand and occupant responses (Papada & Kaliampakos, 2016; Zhang et al., 2024a). Occupant responses to temperature changes, such as adjusting clothing, using air-conditioning, or changing activity patterns, further contribute to this spatial heterogeneity (Bae et al., 2025; Gupta et al., 2024; IEA, 2024). Accurate characterization of building energy demand therefore requires approaches that account for regional differences rather than assuming uniform responses across locations.

The degree-day method is one of the most commonly used approaches for modeling heating and cooling energy demand in buildings. Degree-day models exhibit substantial advantages in terms of data availability and geographic generalization when compared to physical and statistical models (Prataviera et al., 2021; Ahmad et al., 2018). The base temperatures define thresholds for heating (below the heating base temperature) or cooling (above the cooling base temperature) in buildings (De Rosa et al., 2015; Martinopoulos et al., 2019). Degree days are calculated as the difference between the average daily temperature and this base temperature, with values below the threshold being the heating degree day (HDD) and those above the threshold being the cooling degree day (CDD). Base temperature values are influenced by local climatic conditions, occupant behaviour, building types, and thermal properties (Ramon et al., 2020; Hao et al., 2022), which suggests significant spatial heterogeneity in base temperature selection. Moreover, base temperature has been linked to indoor thermal comfort, which differs across populations, so the temperature thresholds for initiating heating or cooling vary between regions (Staffell et al., 2023). In the United States, 18.3°C is the most common base temperature, above which cooling demand is considered to increase and below which heating demand increases (Petri & Caldeira, 2015). In the UK, the HDD and CDD are calculated using 15.5°C and 22°C as the base temperatures, respectively (UK Climate Projections, 2014). In China, 18°C and 24°C are widely used as the base temperatures for the degree-day model (Yang et al., 2024; Li et al., 2018). Conventionally, recommended base temperatures have served as key parameters in modeling heating and cooling energy demand across buildings, regions, and countries (Camarasa et al., 2022; Deroubaix et al., 2021). However, the application of uniform base temperatures across broad geographical areas inadequately represents the sensitivity of building energy demand to local climatic variations, raising concerns about the accuracy and uncertainty of degree-day modeling. To illustrate, a 1°C change in the base temperature can result in approximately 20 additional or fewer CDDs per month in San Luis Obispo (Woods & Fuller, 2014). Inaccurate base temperatures also trigger a nonlinear relationship between building energy demand and degree days (Day et al., 2003).

4. Line 210, when describing methods, it is better to show the unit/metric of different variables. Based on this way, we can make sure the adoption of right indicators and basic mechanism.

Response: We thank the reviewer. We have added the units for all variables at their first appearance in the Methods section (Data, paragraphs 1–4, lines 105-155). The units are also summarized in Table 1.

Energy demand data

The energy demand data encompasses multi-timescale time series electricity or natural gas demand data from 172 regions worldwide (see Table A1). These data are utilized to estimate region-specific base temperatures (*Tbase*), which serve as critical input data for subsequent analyses. Two types of energy demand are considered: building-level and region-level. Building-level energy demand data predominantly consists of residential demand across various temporal resolutions (second, minute, and hourly intervals), which are aggregated to a daily scale (kWh d^{-1}) to better reflect seasonal patterns linked to local temperatures and weather conditions. Due to the scarcity of available residential data, a small number of public building energy demand datasets are also included to expand coverage. Region-level energy demand data, available at daily and monthly temporal scales (kWh d^{-1} or MWh $month^{-1}$), can also capture seasonal variations when the time series is sufficiently long, as demonstrated by Wenz et al. (2017), Huang & Gurney (2016), and Staffell et al. (2023). All energy demand datasets are spatially matched to the finest available administrative boundaries using location metadata, thereby enabling integrated analysis with local climatic conditions.

Meteorological variables

Meteorological data are primary drivers of temporal variations in building energy demand. This study utilizes five key variables from the ERA5-Land Reanalysis data (monthly averages from *1950 to present*): 2-meter air temperature (T2m, K), dew point temperature (Td, K), 10-m u- and v-components of wind (u, v, m s^{-1}), and surface downward solar radiation (SSRD, J m^{-2}). The u- and v-components are combined into 10-m wind speed (WS) for analysis.

5. Furthermore, beyond ASHRAE, the IPCC also advocated the UK model to use and its accuracy in Europe is more solid than ASHRAE. Therefore, authors may re-think the adoption of Eq-1 to develop the model.

Response: We thank the reviewer. Eq. (1) estimates region-specific heating/cooling base temperatures from segmented energy–temperature relationships, whereas UKMO specifies how daily min/mean/max temperatures are converted to HDD once a base temperature is given; the two steps are complementary rather than interchangeable. For an independent check in a European setting, we compared *Theat* from Eq. (1) with *Theat* from UKMO HDD plus a performance-line fit for the 20 UK regions with the most complete daily records, using identical energy–temperature cleaning. The methods agree closely (mean absolute difference 0.83 °C, RMSE 0.87 °C; Appendix Text A1, Fig. A1). Eq. (1) remains our labelling rule for global mapping because it requires only mean temperature and is applied consistently across all training units. (Appendix, lines 730-755).

Text A1: Cross-method validation of base temperatures against a UKMO-based approach

To assess our base-temperature estimates against an independent, max/min–based method in a European context, we compared the heating base temperature from Eq. (1) with that obtained from a UK Met Office (UKMO) degree-day formulation combined with a performance-line fit (Deroubaix et al., 2021; CIBSE, 2006; Spinoni et al., 2018). The HDD under the UKMO method is computed from the daily minimum, mean, and maximum temperatures (T_{min}, T_{mean}, T_{max}) relative to a *Theat*:

$$HDD = \begin{cases} 0, & T_{heat} \leq T_{min} \\ \dfrac{T_{heat} - T_{min}}{4}, & T_{min} < T_{heat} < T_{mean} \\ \dfrac{T_{heat} - T_{min}}{2} - \dfrac{T_{max} - T_{heat}}{4}, & T_{mean} < T_{heat} < T_{max} \\ T_{heat} - T_{mean}, & T_{heat} > T_{max} \end{cases}$$

The base temperatures from Eq. (1) agree closely with those from an independent UKMO-based method, indicating that our estimates are robust to the choice of estimation formulation. For each region we scanned candidate base temperatures from 10 to 18 °C in 0.25 °C steps. At each candidate *Theat* we computed the daily HDD series and regressed daily energy demand on HDD; candidates yielding a non-positive slope were discarded. The *Theat* giving the highest R^2 was taken as the UKMO-based heating base temperature (*Theat*, UKMO). To keep the comparison fair, energy demand was cleaned using the same procedure applied to Eq. (1). The comparison used the 20 UK regions with the most complete daily energy records. Across these regions, the two methods agree closely (mean absolute difference 0.83 °C, RMSE 0.87 °C, Pearson r = 0.83), with Eq. (1) giving slightly lower values, consistent with the known tendency of mean-temperature methods to fall below max/min-based methods. As shown in Fig. A1, the R^2 profiles are broadly flat near their maxima, so this offset corresponds to a negligible difference in goodness-of-fit; the two methods are therefore effectively equivalent in how well they explain the observed energy–temperature relationship.

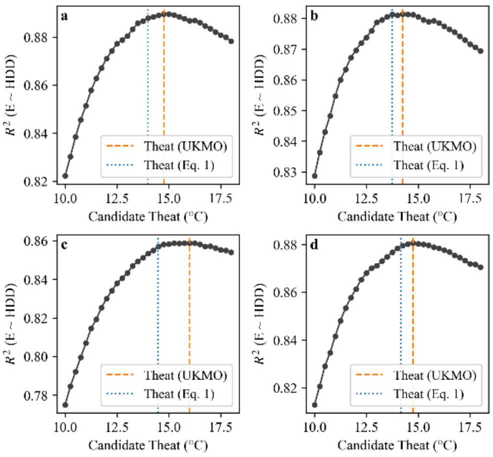

Figure A1: R^2(E ~ HDD) as a function of candidate heating base temperature for four representative UK regions, selected by the median (a), minimum (b), maximum (c), and near-mean (d) difference between the two methods. Dashed line: *Theat* from the UKMO performance line (R^2 maximum); dotted line: *Theat* from Eq. (1).

Reference

Deroubaix, A., Labuhn, I., Camredon, M., Gaubert, B., Monerie, P. A., Popp, M., ... & Siour, G. (2021). Large uncertainties in trends of energy demand for heating and cooling under climate change. Nature communications, 12(1), 5197.

CIBSE, T. (2006). Degree-days: theory and application. The Chartered Institution of Building Services Engineers, 222.

Spinoni, J., Vogt, J. V., Barbosa, P., Dosio, A., McCormick, N., Bigano, A., & Füssel, H. M. (2018). Changes of heating and cooling degree-days in Europe from 1981 to 2100. International Journal of Climatology, 38, e191-e208.

6. Figure 2, authors may reflect some areas that do not have any cooling demand or some areas without heating demand.

Response: We thank the reviewer for this point. Figure 2 shows the predicted base temperatures (*Tcool* and *Theat*), which are outdoor temperature thresholds defining the temperature at which cooling or heating would be initiated. Whether a region currently exhibits actual cooling or heating demand depends on whether local temperatures cross these thresholds—a condition that is dynamic under climate change. In particular, many regions that do not require cooling under the present climate are expected to do so as warming continues. Providing base temperatures for all regions, including those without present-day demand, is therefore essential for assessing how cooling and heating needs will emerge and shift under future climate scenarios. Labelling such regions as "no demand" on the map would obscure this forward-looking application. We have instead clarified in the text and caption that Figure 2 presents base-temperature thresholds rather than realized energy demand, and that actual demand depends on local temperature evolution and population distribution.

Reviewer #2: This manuscript presents a global dataset of cooling and heating base temperatures for building energy demand modelling. Authors combined energy-demand-derived base temperature estimates with a BiLSTM framework, and provided a spatially explicit global dataset that addresses a long-standing limitation in HDD/CDD-based analyses. This topic is timely and relevant, and I believe the resulting dataset could be useful for a broad range of applications, including climate impact assessment and integrated assessment modelling. The manuscript is well organized and the validation effort is extensive (than what is often seen in similar studies).

1. My main suggestion concerns the discussion of robustness and generalizability. The core contribution of this study is the global extrapolation from a relatively limited number of regions with energy-demand-derived base temperatures to thousands of administrative units worldwide. While the reported validation results are encouraging, I felt the manuscript could do more to help readers understand the reliability of this extrapolation. In particular, the training labels themselves are derived through a segmented regression procedure and therefore contain their own assumptions and uncertainties. Since these labels ultimately drives the machine learning framework, I would encourage the authors to discuss/explore how sensitive the resulting dataset may be to choices made during the label-generation process. For example, it would be useful to know whether alternative breakpoint selection approaches, different fitting thresholds, or slightly different parameter settings (no need to do all of them, just as examples) would materially affect the final results. I do not necessarily view this as a weakness of the study, but additional discussion/exploration would help

readers better understand the confidence that can be placed in the dataset.

Response: We thank the reviewer for this constructive suggestion. We have evaluated the sensitivity of the label-generation process along two dimensions and added the results to Section 3.5.

(1) Fitting threshold: Varying the goodness-of-fit criterion for retaining labels from $R^2>0.5$ to 0.4 and 0.6 changed the number of retained units from 123 to 149. However, the overall label distribution remained essentially unchanged: the mean *Theat* varied by only 0.2 °C (14.47–14.67 °C), and the mean *Tcool* by 0.83 °C (20.66–21.49 °C). Both variations are smaller than the model uncertainty interval.

(2) Breakpoint-selection approach: Over the 20 UK regions with the most complete energy records, *Theat* values re-estimated using an alternative UKMO-based performance-line method differed from those obtained via segmented regression by only 0.83 °C on average (Appendix Text A1).

Together, these results demonstrate that the final dataset is robust to reasonable variations in the label-generation process. We have incorporated this sensitivity analysis and discussion into Section 3.5 (Limitations and uncertainties of the *Tbase* dataset, lines 440-460, 790-810).

3.5 Limitations and uncertainties of the *Tbase* dataset

To assess how choices in the label-generation process affect the dataset, we tested the sensitivity to the goodness-of-fit threshold of the segmented regression. Relaxing the inclusion criterion from $R^2 > 0.5$ to 0.4 and tightening it to 0.6 changed the number of retained units between 123 and 149, while the overall label distribution remained essentially unchanged: mean *Theat* ranged over 14.47–14.67 °C (range 0.2 °C) and mean *Tcool* over 20.66–21.49 °C (range 0.83 °C), both smaller than the model uncertainty interval. This indicates that the dataset is insensitive to the fitting threshold. In addition, *Theat* estimated with a UKMO-based alternative breakpoint method over 20 UK regions differs from that from Eq. (1) by only 0.83 °C on average (Text A1), further suggesting limited dependence on the breakpoint-selection approach.

2. A related point is about the representativeness of the training sample. Although the manuscript acknowledges uneven geographical coverage, the model is ultimately applied globally, including regions that appear to be sparsely represented in the training data. I would therefore think that the authors might want to provide a little more discussion of how well the training dataset captures the climatic and socioeconomic diversity of the final application domain. Similarly, the uncertainty analysis currently focuses on model uncertainty, whereas uncertainty associated with limited spatial coverage may also be relevant. Even a qualitative discussion of this issue would be helpful.

Response: We thank the reviewer for this helpful suggestion. In Section 3.5, we now (1) characterize how well the 131 training units represent the climatic and socioeconomic diversity of the global application domain, noting that coverage is good for temperate, mid-to-high-latitude regions but sparser for tropical, arid, and Southern-Hemisphere areas; and (2) distinguish two uncertainty sources, model uncertainty (quantified by MC Dropout) and the additional extrapolation uncertainty from limited spatial coverage, which compound in sparsely trained regions and are reflected in the higher standard deviations over Africa, South America, and western Asia. We also direct users to the uncertainty layer for these regions (Limitations and uncertainties of the *Tbase* dataset, lines 475-485).

3.5 Limitations and uncertainties of the *Tbase* dataset

The representativeness of the training sample also affects the reliability of the global extrapolation. The 131 training units span 24 countries and are mainly concentrated in temperate, mid-to-high-latitude regions. This distribution covers the climatic and socioeconomic conditions of most inhabited areas reasonably well, while tropical, arid, and Southern-Hemisphere regions are covered more sparsely. The MC Dropout analysis above quantifies model uncertainty, whereas limited spatial coverage introduces an additional, extrapolation-related uncertainty; the two compound in sparsely trained regions, consistent with the higher standard deviations found over Africa, South America, and western Asia in the MC Dropout results. Therefore, users are encouraged to consult the accompanying uncertainty layer when interpreting results for these regions, and expanding the training sample there in the future would further improve representativeness.

3. I also wondered whether the manuscript could benefit from a clearer demonstration of the added value of the chosen machine-learning framework. The BiLSTM approach appears to perform well, but readers may naturally ask how much improvement it provides relative to simpler alternatives and common benchmarks. Some discussion of why this architecture was selected and what advantages it offers over more conventional approaches would improve the methodological justification.

Response: We thank the reviewer for this valuable suggestion. In Section 3.1, we have added a benchmark comparison of the BiLSTM against ridge regression, Random Forest, and a unidirectional LSTM using three-fold cross-validation (Table 2). The BiLSTM achieved the highest Pearson correlation on the more challenging heating base temperature prediction task and also outperformed the unidirectional LSTM. Beyond accuracy, we selected the BiLSTM for its ability to model meteorological sequences in an end-to-end manner and to jointly predict both base temperatures even in the presence of missing labels—advantages that conventional models lack (Model performance, lines 315-330, 610-615).

3.1 Model performance

To justify the choice of the BiLSTM architecture, we compared it with three baseline models (Ridge regression, Random Forest, and a unidirectional LSTM) using three-fold cross-validation. This approach provides a robust performance estimate by ensuring every sample serves as part of a held-out validation set across folds. Once the architecture was selected, the final BiLSTM model was retrained on the full training dataset following standard machine learning practice. Therefore, the cross-validation results presented in Table 2 are intended for model comparison and are not directly comparable to the final model performance reported above. In this comparison, BiLSTM achieved the highest Pearson correlation on the more challenging heating base temperature ($r = 0.63$, versus $\leqslant 0.52$ for the other benchmarks) and outperformed the unidirectional LSTM, confirming the benefit of the bidirectional architecture. For the cooling base temperature, which all models predicted well ($r \geqslant 0.70$), simpler models performed comparably. The advantage of BiLSTM is thus most pronounced where the temperature–demand relationship is harder to capture. This strength, together with its ability to model temporal sequences end-to-end and to jointly predict both base temperatures under missing labels, supports its selection.

Table 2 Benchmark comparison of model architectures

Model	$RMSE_{Tcool}$ (°C)	$RMSE_{Theat}$ (°C)	r_{Tcool}	r_{Theat}
Ridge regression	1.08	1.12	0.81	0.49
Random forest	1.16	1.15	0.79	0.50
Uni-LSTM	1.33	1.10	0.71	0.52
BiLSTM	1.24	1.00	0.75	0.63

Note: Values are pooled out-of-fold predictions from three-fold cross-validation and are not directly comparable to the final-model performance in Section 3.1. RMSETcool and RMSETheat are the root mean squared errors (°C) of the out-of-fold predictions of the cooling and heating base temperatures, respectively. rTcool and rTheat are Pearson correlations between the predicted and labeled cooling and heating base temperatures, respectively.

4. Beyond these, a few minor comments: Some parts of the discussion occasionally move beyond what can be directly inferred from the dataset itself. For example, several regional patterns are interpreted through housing quality, adaptation behaviour, or policy conditions. These explanations are certainly plausible, but they are not directly evaluated in the present study. I would therefore suggest slightly more cautious wording in these sections or simply drop them. In addition, parts of the introduction and application discussion could potentially be streamlined. The manuscript already makes a strong case for the importance of regional base temperatures, and reducing some of the "too broad" background discussion would allow greater emphasis on methodology itself, which is likely to be of primary interest to ESSD readers.

Response: We thank the reviewer. We have adopted more cautious wording and removed interpretations that are not directly evaluated in this study—in particular, explanations of regional patterns based on housing quality, adaptation behaviour, and policy conditions have been either removed or rephrased as possibilities consistent with the cited literature rather than as conclusions. We have also streamlined the background in the Introduction and the application discussion to place greater emphasis on the methodology, as suggested (Introduction, Model development, Results and discussion, lines 30-45, 200-220, 330-345).

We hope that the revised manuscript and our point-by-point responses have adequately addressed the reviewers' comments. We look forward to your decision.

Sincerely,

Yuyu ZHOU, AAAS Fellow, Professor
Department of Geography, The University of Hong Kong
The University of Hong Kong, Pokfulam, Hong Kong, PR China
E-mail: yuyuzhou@hku.hk

8 July 2026