This study develops a time-series 30 m building height dataset for China by integrating multi-source remote sensing data, representing a valuable contribution to large-scale mapping of urban vertical dynamics. The dataset has clear potential for urban studies and related applications. However, several issues related to data quality, robustness, and validation—particularly for high-rise buildings and regions with limited reference data—still need to be carefully addressed to strengthen the reliability and usability of the product. The following comments are provided for the authors' consideration.

We express our deep gratitude for the reviewer's positive feedback on our work and sincerely appreciate the constructive comments on highlighting important areas for improvement regarding data quality, robustness, and validation. In the following sections, we address each of the reviewer's points in detail, organizing our responses in a question (black) - response (blue) - revision (red) format to ensure clarity.

1. In Eq. (6), averaging all 1 m pixels (900 pixels) within a 30 m grid mixes building and non-building pixels. Please clarify the rationale for this choice, rather than computing height only over building-covered pixels.

We appreciate the reviewer's professional comment regarding the calculation of building height in Eq. (6) of the original manuscript. Eq. (6) calculates the height value of each 30 m × 30 m grid by averaging the height values of the 900 underlying 1 m × 1 m pixels. For clarity, this terminology is used in responses to related questions: "grid" refers to a 30 m × 30 m grid, while "pixel" refers to a 1 m × 1 m pixel.

$$H^{30m} = \frac{\sum_i^{900} H_i^{1m}}{900} = \frac{\sum_j^{nbc} H_j^{1m}}{900} \qquad \text{(6) of the original manuscript}$$

In this equation, H_i^{1m} denotes the height value of each pixel within the grid, including both building and non-building pixels. For non-building pixels, H_i^{1m} is assigned a value of 0. Therefore, the numerator is equivalent to the sum of height values over only the building pixels, denoted as H_j^{1m}. Here, nbc represents the number of building-covered pixels among the 900 pixels within the grid. Importantly, the denominator is kept as 900 rather than nbc.

This design means that Eq. (6) does not calculate the mean height of building footprints alone. Instead, it derives an area-weighted grid-level building height that preserves both building height information and building coverage density within each grid. If the average is calculated only over building-covered pixels, grids with similar building heights but markedly different building coverage fractions could yield similar mean height values, despite substantial differences in land development intensity. For example, a grid containing sparse high-rise structures and another grid with extensive, contiguous high-rise coverage could show proximate average heights if only building-covered pixels are considered. In contrast, using all 900 pixels as the denominator incorporates the proportion of non-building pixels within the grid, enabling the derived height to more comprehensively represent the overall vertical development intensity of the built environment at the grid scale.

Although the denominator of Eq. (6) is the total number of pixels within each grid, including both building and non-building pixels, the grids themselves are constrained using the Global Artificial Impervious Area (GAIA) dataset and building-related information. This masking procedure ensures

that the retained grids mainly represent buildings and their surrounding built-up environment such as roads and plazas, rather than unrelated natural land-cover types such as vegetation, water bodies, or bare land. As a result, grids without any building information are excluded from the reference sample preparation. In addition, zero-height reference samples are excluded during model training to ensure that the training samples effectively characterize building-related areas.

Through the above processing, this study computes an area-weighted average building height at the grid level. This metric reflects the overall vertical development intensity per unit grid area across the built-up region, rather than merely the average height within building footprints, and therefore captures the combined effects of building height and building density within each grid.

Furthermore, calculating building height at the grid level by considering both buildings and associated built-up surfaces is consistent with established practices in large-scale urban morphology studies. For instance, Zhou et al. (2022) state in their methodology that

> "…In this study, the urban built-up height (unit: meter) was calculated as the mean height of all areas within a 500 m × 500 m grid in the urban domain, including both buildings and non-buildings such as streets, parking lots, and green space…"

Similarly, Li et al. (2020) note that their building height

> "…should be noted that the building height in this study specifically refers to the mean height within the 500 m grid, including buildings and non-buildings like streets and parking lots…"

Other recent high-impact works, such as Chen et al. (2025), also adopt the average RH96 value of all GEDI samples within a grid to represent building height without explicitly filtering for building-only footprints.

In addition, our grid-level calculation aligns with the Gross Building Height definition adopted in the Global Human Settlement Layer (GHSL) framework, which distinguishes between the Net and Gross metrics (Pesaresi et al., 2024). Net Building Height refers to the average height of buildings only, excluding built-up surfaces without buildings, whereas Gross Building Height averages buildings and associated built-up surfaces within the grid. Providing Net Building Height together with building fraction information can offer a more detailed characterization of urban form. However, simultaneously mapping long-term building height and sub-pixel building fractions across decadal scales remains challenging in terms of computational cost and data consistency.

Therefore, considering that grid-level building height is widely adopted in regional and global building height mapping to represent the overall vertical development intensity of the built environment within each grid, Eq. (6) is used to calculate the area-weighted building height.

Thanks again for the reviewer's professional comments. Considering the importance of building height calculation in this study, further clarification is added in Section 2.6 of the revised manuscript:

> "…
>
> To generate rasterized reference data for 2019, building vector footprints are converted to 1m resolution height raster to preserve height information. The 1m height raster is then aggregated to 30m resolution by averaging the values of all 1m sub-pixels within each 30 m × 30 m grid (Zhou et al., 2022).
>
> $$H^{30m} = \frac{\sum_i^{900} H_i^{1m}}{900} = \frac{\sum_j^{nbc} H_j^{1m}}{900} \quad (6)$$
>
> where H_i^{1m} denotes the height value of each 1m sub-pixel. Non-building sub-pixels are assigned a value of 0, so the numerator is equivalent to the sum of

height values over building-covered sub-pixels only, denoted as H_j^{1m}. Here, nbc represents the number of building-covered 1m sub-pixels among the 900 sub-pixels within the grid. The denominator is kept as 900 rather than nbc, because using nbc would produce a building-footprint-only mean height and could assign comparable values to grids with similar building heights but different building coverage fractions, thereby failing to capture variations in built-up intensity.

Although non-building 1m sub-pixels are included in this calculation, the 30m grids are constrained using the GAIA impervious surface mask and building information. As a result, grids without any building information are excluded during reference sample preparation. In addition, zero-height reference samples are omitted during model training to ensure that the training data effectively represent building-related areas. Accordingly, these 30 m grids serve as the basic units for all subsequent processing, model training, validation, and dataset generation, and correspond to the 30m pixels in the final building height dataset.

With these constraints, the derived height better captures the combined effects of building height and building coverage. This grid-level calculation is consistent with established practices in large-scale urban morphology studies (Chen et al., 2025; Li et al., 2020; Zhou et al., 2022) and aligns with the Gross Building Height concept in the GHSL framework (Pesaresi et al., 2024).

…″

Reference:

[1] Zhou, Y., Li, X., Chen, W., Meng, L., Wu, Q., Gong, P., and Seto, K. C.: Satellite mapping of urban built-up heights reveals extreme infrastructure gaps and inequalities in the Global South, Proceedings of the National Academy of Sciences, 119(46), e2214813119, https://doi.org/10.1073/pnas.2214813119, 2022

[2] Li, X., Zhou, Y., Gong, P., Seto, K. C., and Clinton, N.: Developing a method to estimate building height from Sentinel-1 data, Remote Sensing of Environment, 240, 111705, https://doi.org/10.1016/j.rse.2020.111705, 2020

[3] Chen, P., Huang, H., Qin, P., Liu, X., Wu, Z., Zhao, F., Liu, C., Wang, J., Li, Z., Cheng, X., and Gong, P.: Characterizing dynamics of built-up height in China from 2005 to 2020 based on GEDI, Landsat, and PALSAR data, Remote Sensing of Environment, 325, 114776, https://doi.org/10.1016/j.rse.2025.114776, 2025

[4] Pesaresi, M., Schiavina, M., Politis, P., Freire, S., Krasnodębska, K., Uhl, J. H., Carioli, A., Corbane, C., Dijkstra, L., Florio, P., Friedrich, H. K., Gao, J., Leyk, S., Lu, L., Maffenini, L., Mari-Rivero, I., Melchiorri, M., Syrris, V., Van Den Hoek, J. and Kemper, T.: Advances on the Global Human Settlement Layer by joint assessment of Earth Observation and population survey data. International Journal of Digital Earth, 17(1), 2390454, 2024

2. The model performs well in cities with available building height samples (e.g., Beijing and Tianjin). However, in cities without reference height samples (Fig. 14), the heights of mid-rise and high-rise buildings are significantly underestimated. This indicates limited spatial transferability and cross-city robustness of the model. Please address this limitation, and explicitly discuss the applicability and associated uncertainty in such regions.

We sincerely appreciate the reviewer's constructive feedback regarding the spatial transferability

and cross-city robustness of our model. The underestimation observed in cities lacking reference samples, as illustrated in Fig. 14 in original manuscript, is a critical point that requires further clarification and a more transparent discussion of uncertainty.

The observed underestimation in non-sampled regions is largely attributable to the inherent challenges of large-scale building height mapping in China. High-quality, wide-coverage ground truth data is often restricted to proprietary datasets from commercial platforms like Gaode or Baidu (Wu et al., 2023; Yan et al. 2024), which primarily focus on major cities. This uneven spatial distribution of available training data leads to a certain degree of overfitting to the specific urban morphologies and architectural styles of sampled metropolitan areas, thereby limiting the model's generalization when applied to cities with different developmental profiles.

In response to this issue, the model is first optimized using a ratio-controlled sampling strategy, as elaborated in our reply to Question 3. After optimization, the model is trained with improved representation of mid- and high-rise buildings, thereby reducing prediction inaccuracies for such structures and improving overall performance, as shown in Fig. 4 of this response letter.

To further evaluate the spatial transferability and cross-city robustness of the optimized model, we conduct an independent validation using sparse building height information collected from the Lianjia real-estate platform. After being matched and processed with CMAB building footprints (Zhang, Zhao and Long, 2025), the Lianjia data provide community-level building heights. Since these data are neither pixel-level nor footprint-level ground truth, they are not used for model training. Instead, they provide an independent out-of-distribution benchmark for evaluating model performance in extrapolated cities that are excluded from the reference samples.

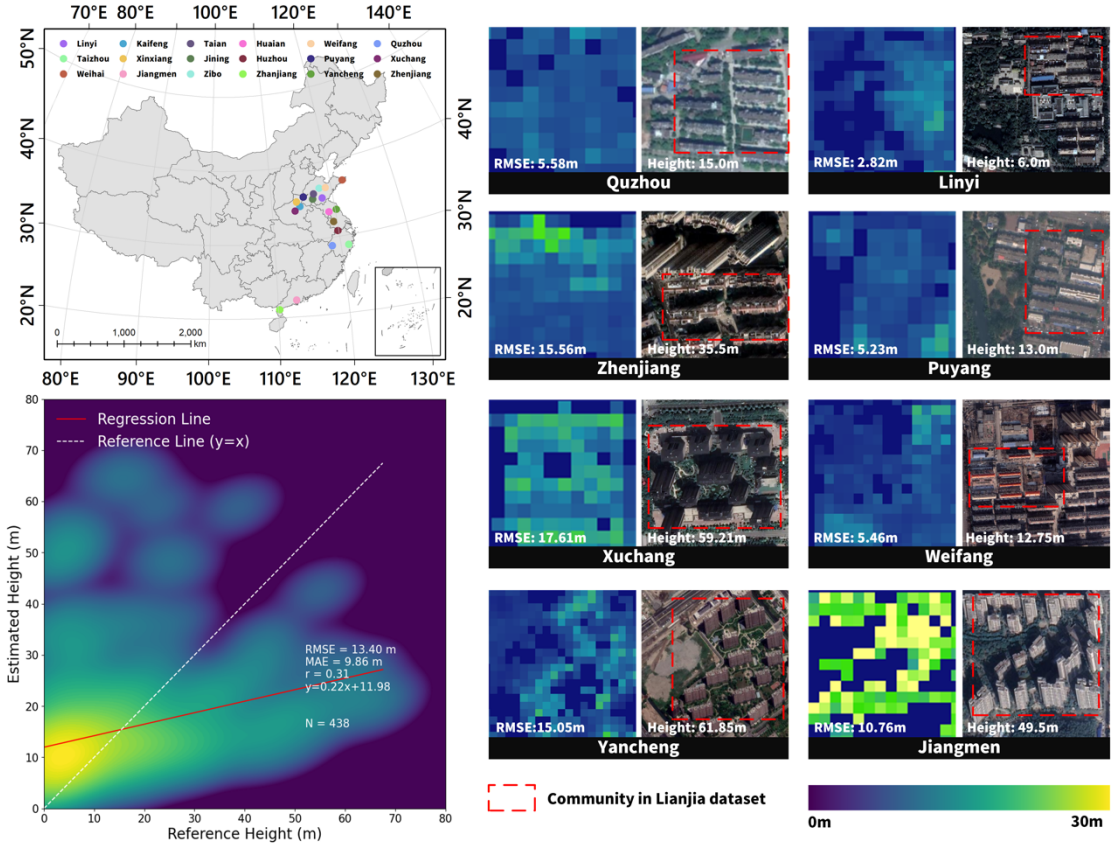

Fig. 1. Accuracy assessment of model performance in cities excluded from the reference samples using Lianjia real-estate data.

The results in Fig. 1 of this response letter indicate that the optimized model maintains reasonable estimation performance in extrapolated cities, with an overall RMSE of 13.40 m. For representative cities, the RMSE values are mostly within 3–6 m in low-rise built-up areas, including Linyi, Puyang, Weifang, and Quzhou. For mid- and high-rise areas, the RMSE values range from 10 to 17 m in Zhenjiang, Jiangmen, Yancheng, and Xuchang.

The relatively larger errors for taller buildings are consistent with patterns reported in existing large-scale building height studies. Che et al. (2024) show that high-rise buildings tend to contribute more strongly to regional errors; for example, in Africa, the RMSE for buildings above 50 m reaches 25.52 m, which is much higher than that for buildings below 20 m. Therefore, although underestimation still occurs in some high-rise areas, the extrapolation results remain comparable to existing findings and demonstrate the optimized model's cross-city applicability and generalization capability.

We thank the reviewer again for the constructive comment. To further evaluate the model's spatial transferability and cross-city robustness, we add a new Section 4.2.3, "Cross-city transferability assessment":

> "…
>
> To further evaluate the spatial transferability and cross-city robustness of the models, an independent validation is conducted using sparse building height information collected from the Lianjia real-estate platform. After being matched and processed with CMAB building footprints (Zhang, Zhao, and Long, 2025), the Lianjia data provide community-level building heights. Since these data are neither pixel-level nor footprint-level ground truth, they are not used for model training. Instead, they provide an independent benchmark for evaluating model performance in extrapolated cities that are excluded from the reference samples.
>
> The extrapolation validation is conducted using the XGB-S1S2 model trained with the 2019 reference samples, because this validation only requires a single-year cross-sectional assessment. The results show that this model maintains reasonable estimation performance in cities excluded from the reference samples, with an overall RMSE of 13.40 m. For representative cities, the RMSE values are mostly within 3–6 m in low-rise built-up areas, including Linyi, Puyang, Weifang, and Quzhou. For mid- and high-rise areas, the RMSE values range from 10 to 17 m in Jiangmen, Zhenjiang, Yancheng, and Xuchang.
>
> The relatively larger errors for taller buildings are consistent with patterns reported in existing large-scale building height studies. Che et al. (2024) show that high-rise buildings tend to contribute more strongly to regional errors; for example, in Africa, the RMSE for buildings above 50 m reaches 25.52 m, which is much higher than that for buildings below 20 m. Therefore, although underestimation still occurs in some high-rise areas, the extrapolation results remain comparable to existing findings and demonstrate this model's cross-city applicability and generalization capability.
>
> …"

We also clarify that while the model remains highly applicable for regional-scale urban analysis, caution should be exercised when interpreting absolute height values in regions far removed from the training data distribution. As for the current manuscript upon evaluating the inherent uncertainties in model predictions, a confidence layer has been developed to quantitatively assess prediction reliability. This uncertainty analysis is presented in a dedicated section, with detailed discussion available in the response to Question 9(2) and in the newly added Section 4.6. This

limitation will be a primary focus of our future work, where we aim to integrate self-supervised learning and multi-source remote sensing data to further enhance cross-city robustness.

Reference:
[1] Wu, W. B., Ma, J., Banzhaf, E., Meadows, M. E., Yu, Z. W., Guo, F. X., Sengupta, D., Cai, X. X., and Zhao, B.: A first Chinese building height estimate at 10 m resolution (CNBH-10 m) using multi-source earth observations and machine learning, Remote Sensing of Environment, 291, 113578, https://doi.org/10.1016/j.rse.2023.113578, 2023
[2] Yan, W., Wu, J., Zhang, C., Chen, X., Ren, J., Xiao, Z., Liao, Z., Lafortezza, R., and Su, Y.: Developing an annual building volume dataset at 1-km resolution from 2001 to 2019 in China, International Journal of Digital Earth, 17(1), 2330690, https://doi.org/10.1080/17538947.2024.2330690, 2024
[3] Che, Y., Li, X., Liu, X., Wang, Y., Liao, W., Zheng, X., Zhang, X., Xu, X., Shi, Q., Zhu, J., Yuan, H., and Dai, Y.: 3D-GloBFP: the first global three-dimensional building footprint dataset, Earth Syst. Sci. Data, 16, 5357–5374, https://doi.org/10.5194/essd-16-5357-2024, 2024
[4] Zhang, Y., Zhao, H., and Long, Y.: CMAB: A Multi-Attribute Building Dataset of China, Scientific Data, 12(1), 430, https://doi.org/10.1038/s41597-025-04730-5, 2025

3. The model shows limited capability in predicting high-rise buildings. Please clarify the maximum building height represented in the training samples. Moreover, this limitation may be related to the uneven distribution of building height samples. In that case, more appropriate sample setting strategies or model optimization schemes could be considered. In addition, the height bins in Fig. 4 (e.g., "<10 m" and ">60 m") should be more clearly defined (e.g., "3–10 m", "60–X m").

We would like to thank the reviewer for these insightful suggestions regarding the training sample distribution and the clarity of our data categorization. We have taken these comments seriously and have implemented a new ratio-controlled strategy to improve the model's performance on high-rise buildings.

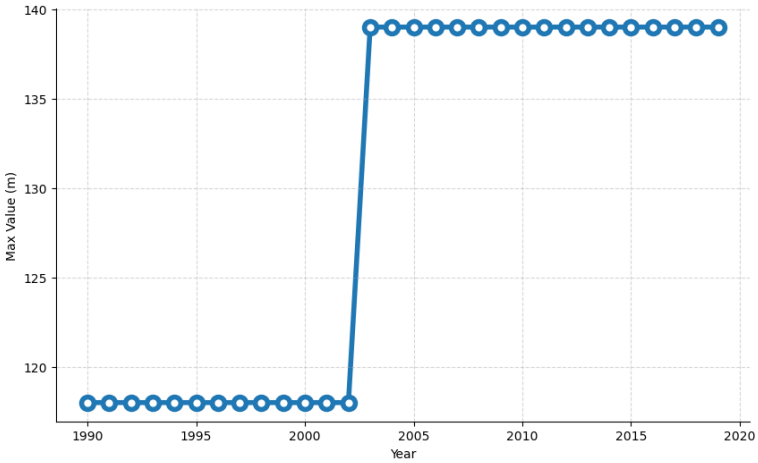

Fig. 2. Maximum height in training samples each year

The maximum building height represented in the training samples is reflected in 0 of this response letter, with 139 m being the highest from 2003 to 2019, 118 m from 1990 to 2002, demonstrating

the existence of high-rise buildings in the samples. Although the Jinmao Tower broke China's building height record after 1999, reaching 420.5 m, the sample preparation procedure averages building height together with surrounding built-up surfaces, as clarified in response to Question 1. Therefore, the resulting value represents an area-weighted grid-level height rather than the height of the individual building itself, which lowers the maximum height value represented in the dataset. Based on the maximum building height represented in the reference samples, building heights are divided into four intervals: (0 m, 10 m], (10 m, 30 m], (30 m, 60 m], and (60 m, 140 m]. Section 4.1 of the manuscript is updated to report the maximum heights:

> "…The highest sample is at 139 m from 2003 to 2019, and 118 m from 1990 to 2002, demonstrating the existence of high-rise buildings in the samples.…"

Regarding the sample distribution, a significant imbalance indeed exists in our original dataset, where buildings under 10 m accounted for over 70% of the total samples. The initial attempt is to balance all height intervals equally through resampling; however, this strategy produces suboptimal results. This is likely because low-rise buildings constitute the dominant building type in the study area, and maintaining a higher proportion of low-rise samples is essential for preserving the model's prediction accuracy for these prevalent building types (Moraes, Campagnolo, and Cartano, 2024).

Consequently, a more nuanced ratio-controlled resampling scheme is adopted to balance minority-class enhancement and majority-class preservation. This approach is consistent with established practices for handling imbalanced samples in remote sensing and urban studies (Chawla et al., 2002; Naboureh et al., 2020). In the updated strategy, the sample size of each height interval above 10 m is capped at 1/10 of that of the (0 m, 10 m] category. This design largely preserves the dominance of low-rise buildings while increasing the relative proportions of the (30 m, 60 m] and (60 m, 140 m] high-rise groups. Specifically, down-sampling is applied to over-represented height bins, whereas oversampling is applied to under-represented bins. As the number of newly added pixels through oversampling exceeds the number of pixels removed through down-sampling, the total sample size increases, as shown in Fig. 3 of this response letter. Although the proportion of the (10 m, 30 m] interval decreases, the experiment shows no negative effect on model accuracy. This strategy maintains the prediction accuracy for predominant low-rise buildings and simultaneously enhances the representation of high-rise samples, thereby improving the model's ability to estimate taller buildings. Section 4.1 of the manuscript is updated accordingly:

> "…The original height distribution of reference samples is highly imbalanced, with low-rise buildings accounting for the majority. This imbalance may limit the model's ability to accurately estimate relatively high buildings. Although equal-proportion resampling can balance different height intervals, it reduces the representation of the original majority class and leads to a decline in prediction accuracy for the most prevalent building types (Moraes, Campagnolo, and Cartano, 2024). To improve the model's performance for high-rise buildings while preserving its accuracy for dominant low-rise structures, a ratio-controlled resampling strategy is adopted.
>
> Specifically, building heights are divided into four intervals based on reference samples: (0 m, 10 m], (10 m, 30 m], (30 m, 60 m], and (60 m, 140 m]. The sample size of each height interval above 10 m is adjusted through down-sampling or oversampling and capped at one-tenth of that of the (0 m, 10 m]

category (Chawla et al., 2002; Naboureh et al., 2020). As the number of newly added pixels through oversampling exceeds the number of pixels removed through down-sampling, the total sample size increases. Although the proportion of the (10 m, 30 m] interval decreases, the experiment shows no negative effect on model accuracy. This strategy maintains the prediction accuracy for predominant low-rise buildings and simultaneously enhances the representation of high-rise samples, thereby improving the model's ability to estimate taller buildings.…″

Based on the optimized sample distribution, the models are retrained and a new dataset is generated. The updated results show substantial improvements in accuracy. The updated results and regenerated dataset are further presented in our response to Question 4, where the accuracy assessment figures (i.e., Fig. 4 of this response letter) are also revised accordingly. Specifically, the RMSE values for 2019, 2017, and 2014 decrease from 6.37 m, 6.60 m, and 6.69 m to 3.90 m, 4.14 m, and 4.11 m, respectively; the MAE values decrease from 3.66 m, 3.79 m, and 3.80 m to 2.70 m, 2.86 m, and 2.83 m, respectively; and the R^2 values increase from 0.46, 0.43, and 0.42 to 0.80, 0.78, and 0.78, respectively. These results demonstrate the effectiveness of the ratio-controlled sampling strategy in improving model accuracy.

It should be clarified that the ratio-controlled resampling strategy is applied only during model training to mitigate sample imbalance. The accuracy assessment is conducted using the original reference samples, rather than the resampled training samples. This ensures that the reported metrics reflect the model's predictive performance on the actual reference data distribution and are not affected by the artificial sample distribution introduced during resampling (Elreedy et al., 2019).

Notably, these optimizations directly address the reviewer's concern regarding the model's capability to predict mid- and high-rise structures and further enhance the model's cross-city robustness and spatial transferability, as shown in the response to Question 2.

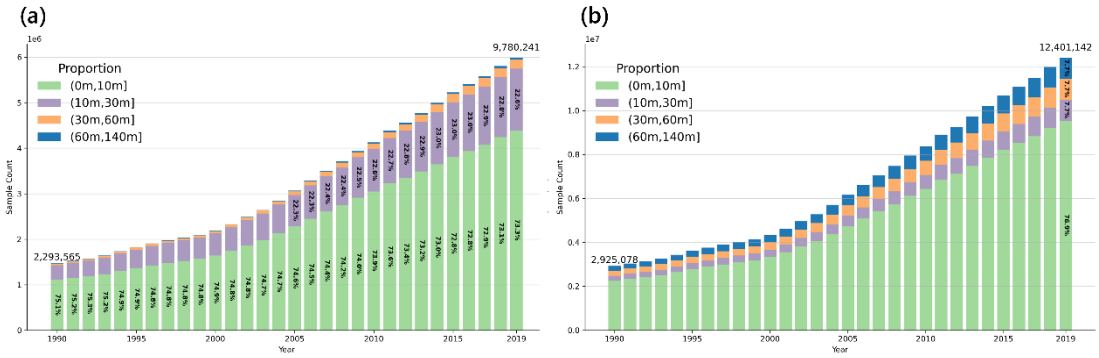

Fig. 3 (the revised Fig. 4 in the original manuscript). Number of reference samples and height proportion each year. (a) Number and proportion before resample; (b) Number and proportion after resample.

Regarding the height-bin notations, we sincerely appreciate the reviewer's suggestion to define them more clearly. The height-bin notations in Figs. 4 and 7 of the original manuscript are revised to (0 m, 10 m], (10 m, 30 m], (30 m, 60 m], and (60 m, 140 m]. This updated notation is more precise, as it explicitly defines the mathematical range of each height interval.

As explained in our response to Question 1 and 4, the area-weighted grid-level height is used as the reference height, which may result in height values between 0 m and 3 m, lower than the height

of a single floor. Since pixels without any building information are excluded during model training, the first height interval is denoted as (0 m, 10 m]. For the last interval, (60 m, 140 m], the upper bound of 140 m is selected because it is higher than any building height represented in the reference samples, as shown in Fig. 2 of this response letter.

References:

[1] Moraes, D., Campagnolo, M. L. and Caetano, M.: Training data in satellite image classification for land cover mapping: a review. European Journal of Remote Sensing, 57(1), 2341414, 2024

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357, 2002

[3] Naboureh, A., Li, A., Bian, J., Lei, G. and Amani, M.: A hybrid data balancing method for classification of imbalanced training data within google earth engine: Case studies from mountainous regions. Remote Sensing, 12(20), 3301, 2020

[4] Elreedy, D. and Atiya, A. F.: A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Information sciences, 505, 32-64, 2019

4. In Fig. 7, the reference building heights are mainly concentrated within 0–8 m. Pixels with height values below 3 m are likely non-building pixels and could be excluded from validation. Moreover, reporting relative error metrics (e.g., rRMSE) would provide a more informative evaluation.

We appreciate the reviewer's valuable suggestions regarding the height thresholds and additional error metrics for the Fig. 7 of the original manuscript.

The concentration of samples around 0–8 m in the density scatter plot shown in Fig. 4 of this response letter, corresponding to the revised Fig. 7 in the original manuscript, is mainly caused by the following three factors:

First, the 30 m grid height in this study represents an area-weighted height, rather than the absolute height of individual buildings. In other words, the grid-level height is jointly affected by both building height and building coverage fraction. When buildings occupy only a small proportion of a grid, the area-weighted height can still be relatively low even if real buildings are present within the grid. Therefore, the concentration around 0–8 m also reflects the prevalence of low-rise buildings and areas with low building coverage in the reference samples.

Second, the reference samples themselves show a clear concentration in the low-height range. As shown in the sample distribution analysis in our response to Question 3, 30 m grids with heights in the interval of (0 m, 10 m] account for the vast majority of the reference samples. This indicates that low-rise buildings, sparsely built-up areas, and areas with low vertical development intensity are the dominant building-height types in the study area. Therefore, the concentration of samples in the low-height range in the accuracy assessment figure is primarily determined by the height distribution of the reference samples.

Third, the visualization mechanism of the density scatter plot further highlights areas with high sample frequency. The color intensity in the density scatter plot reflects the local concentration of sample points. Since the number of samples in the (0 m, 10 m] interval is much larger than that in other height intervals, this range shows a higher point density in the density plot, making the samples visually more concentrated around 0–8 m.

Overall, the concentration of samples in the low-height range in the density plot is a reasonable

phenomenon jointly caused by the area-weighted height definition, the reference sample distribution, and the visualization mechanism. It does not indicate that the model can only predict low-height buildings, nor does it mean that mid- or high-rise building samples are excluded from the validation.

To avoid this ambiguity, we add the following clarification in Section 4.2.1 of the revised manuscript:

> "…The concentration of samples in the low-height range in the density plot is reasonable. It is mainly caused by three factors: the area-weighted definition of grid-level height; the reference sample distribution, which reflects the dominance of low-rise buildings, sparsely built-up areas, and areas with low vertical development intensity in China; and the visualization mechanism of the density scatter plot. This phenomenon does not indicate that the model can only predict low-height buildings, nor does it mean that mid- or high-rise building samples are excluded from the validation.…"

In our dataset, grids with height values greater than 0m but lower than 3m generally represent low area-weighted grid-level building height, where the summed height of building pixels is averaged over all pixels within the grid. As mentioned in our response to Question 1, a grid containing sparse high-rise structures and another grid with intensive, contiguous high-rise coverage could show similar mean heights if only building-covered pixels are considered, despite substantial differences in land development intensity. Thus, grids with height values greater than 0m but lower than 3m usually indicate that buildings occupy only a limited proportion of the grid area, rather than the absence of buildings. They may correspond to low-rise structures, transitional urban fringes, or sparsely built-up areas that are physically present in real urban environments. Excluding these values would lead to an incomplete assessment of the model's performance in complex, heterogeneous urban settings. By retaining these 30m pixels/grids, the validation provides a more transparent and comprehensive evaluation of the dataset's ability to represent the full spectrum of urban vertical forms. Therefore, after careful consideration, 30m pixels with height values greater than 0m but lower than 3m are retained in the validation process. However, we agree that using 0m in the legend may cause misunderstanding. To avoid this ambiguity, we add the following clarification in Section 4.2.1 of the revised manuscript:

> "…Since each 30m grid represents the area-weighted height of buildings and surrounding built-up surfaces, height values greater than 0m but lower than 3m may occur when buildings occupy only a limited proportion of the grid area, rather than indicating the absence of buildings. Accordingly, for visualization purposes only, the starting points of both the X- and Y-axes are set to 0m.…"

Following the reviewer's suggestion, relative Root Mean Square Error (rRMSE) is incorporated into the accuracy assessment, as shown in Fig. 4 of this response letter. The rRMSE normalizes RMSE by the mean reference height, making it useful for comparing model performance across different urban densities and height ranges. The rRMSE values for 2019, 2017, and 2014 are 0.61, 0.63, and 0.63, respectively. We also update Fig. S1 in the Supplementary Material by adding the rRMSE values for each mapped year.

The rRMSE metric is incorporated, and its formula is added to Section 3.3 of the revised manuscript:

"…Model performance is quantified by four metrics: Root Mean Square Error (RMSE, Eq. (7)), Mean Absolute Error (MAE, Eq. (8)), and coefficient of determination (R^2, Eq. (9)) and relative Root Mean Square Error (rRMSE, Eq. (10))….

$$rRMSE = \frac{RMSE}{\frac{\sum_{i=1}^{n} H_{i,ref}}{n}} \quad (10)$$

…"

All accuracy metrics reported in the manuscript are updated accordingly. For example, Section 4.2.1 is revised as follows:

"…with RMSE values of 3.90 m, 4.14 m, and 4.11 m; MAE values of 2.70 m, 2.86 m, and 2.83 m; R^2 values of 0. 80, 0. 78, and 0. 78; and rRMSE values of 0.61, 0.63 and 0.63, respectively. Linear regression slopes between reference and estimated heights are 0.77, 0.75, and 0.75, with intercepts of 0.64 m, 0.71 m, and 0.71 m.…"

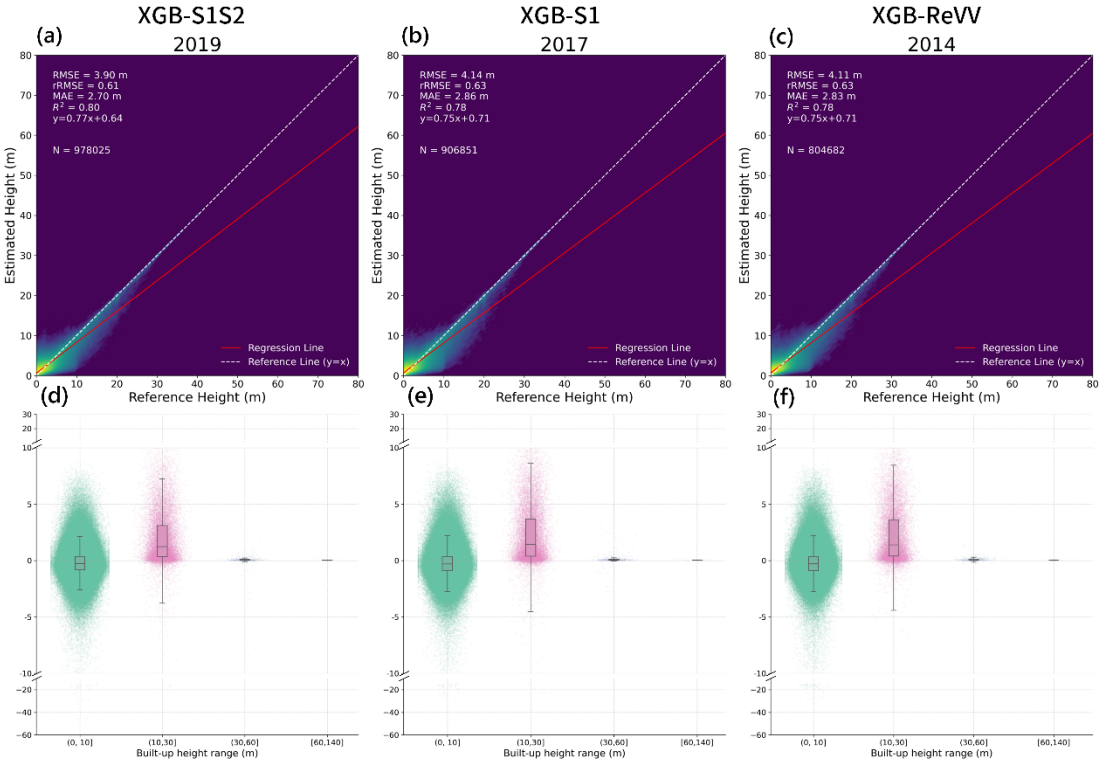

Fig. 4 (the revised Fig. 7 in the original manuscript). Accuracy of XGBoost-based models in their initial years after resampling.

5. In building change areas, the accuracy of estimated height changes should be quantitatively assessed to evaluate whether the model can reliably capture building height dynamics. In stable areas, the temporal consistency of predicted building heights should be quantitatively examined to assess prediction stability.

We appreciate the reviewer's suggestion to quantitatively assess the dataset performance in both building change areas and stable areas. This is indeed crucial for evaluating whether the generated

long-term building height dataset can reliably characterize spatiotemporal urban dynamics.

To validate the reliability of the long-term building height dataset in characterizing building change areas, both quantitative comparisons and historical remote sensing images are used. Fig. 13 and Fig. 14 in the original manuscript are updated as Fig. 5 and Fig. 6 in this response letter, respectively. These two figures are designed to assess three typical urban change contexts as described in Section 3.1.1 of the manuscript, namely persistent built-up areas, complete demolition areas, and new construction areas, in both sampled and non-sampled cities.

Specifically, Fig. 5 in this response letter (Fig. 13 in the original manuscript) focuses on representative areas in sampled cities, including Tianjin, Beijing, and Wuhan. For these cities, predicted heights, reference heights, and historical remote sensing images are jointly compared, allowing a quantitative assessment of the mapped building height dynamics. Fig. 6 in this response letter (Fig. 14 in the original manuscript) focuses on representative areas in non-sampled cities, including Handan, Jining, and Anyang. Since reference height data are unavailable in these non-sampled cities, the quality of the extrapolated building height results is mainly examined using historical remote sensing images as visual evidence of construction, demolition, and redevelopment processes.

It should be further clarified that, due to the lack of long-term records of historical building heights, the annual reference building heights used in sampled cities are not direct observations for all historical building pixels. As described in Section 3.1.1 of the manuscript, they are derived from persistent built-up areas identified based on the 2019 reference data and CCDC breakpoint information. Specifically, for a given year, only pixels with no breakpoint between that year and 2019 are kept as reference data for that year. Therefore, the annual reference data mainly represent areas without clear land-use conversion or building change between the target year and 2019, and do not include pixels that experience complete demolition or substantial change during this period. This explains the local differences between predicted and reference heights in demolition areas such as Jiaohuachang (Fig. 5b in this response letter).

For sampled cities, the predicted heights generally show trends that are consistent with the reference heights and agree well with historical remote sensing images. In persistent built-up and new construction areas, such as Drum Tower in Tianjin (Fig. 5a in this response letter) and Yijiangyuan in Wuhan (Fig. 5c in this response letter), the dataset effectively captures building height dynamics. Taking Yijiangyuan as an example, the reference heights in 2000, 2005, 2013, and 2019 are 5.19 m, 5.26 m, 5.69 m, and 6.00 m, respectively, while the corresponding predicted heights are 5.15 m, 5.30 m, 5.72 m, and 6.11 m, with errors all below 0.11 m. This indicates good consistency between the predicted height changes and the reference data for sampled cities.

For the complete demolition case in the sampled cities, Jiaohuachang in Beijing is selected (Fig. 5b in this response letter). The coke plant is clearly visible in the 2005 historical remote sensing image but experiences demolition between 2005 and 2019. Therefore, this area does not satisfy the "no breakpoint between the target year and 2019" condition and, in principle, is not retained as a 2005 reference sample. In other words, the 2005 reference height cannot fully represent the true building height before demolition. As a result, a difference is observed between the predicted height and reference height in 2005: the predicted height is 5.71 m, whereas the reference height is 3.52 m. This difference mainly results from the applicability limitations of the annual reference data generation method, rather than from model bias. Moreover, historical remote sensing images in 2005, 2010, 2015, and 2018 confirm the demolition process, and the estimated height decreases from 5.71

m to 4.04 m, reflecting the transition from relatively high built-up structures to demolition and redevelopment.

For non-sampled cities, Fig. 6 in this response letter further examines the extrapolated building height results of the dataset under the same three urban change contexts. Although reference height data are unavailable, the mapped height changes are broadly consistent with the construction and demolition processes visible in historical remote sensing images. In Handan Congtai Park (Fig. 6a in this response letter), the area is dominated by bungalows in 2006 and is gradually transformed into clusters of multi-story apartments, which become prominent in the 2013 and 2019 imagry. Correspondingly, the estimated height increases from 5.90 m to 7.40 m, then to 8.04 m, and 8.68 m, indicating that the dataset captures the redevelopment process from low-rise buildings to higher residential structures. In Jining Shiliying Village (Fig. 6b in this response letter), the settlement progressively sinks into a subsidence lake, leading to the abandonment and demolition of houses and the near disappearance of the village by 2019. This process is reflected by a gradual decrease in estimated height from 4.05 m to 3.91 m, then to 3.74 m, and 3.53 m. In Anyang Yiyuan (Fig. 6c in this response letter), new developments are built on previously undeveloped land and expand continuously between 2005 and 2019. The estimated height correspondingly increases from 4.91 m to 5.60 m, 7.22 m, and 8.63 m, showing that the dataset reflects the progressive construction process in this non-sampled city.

Overall, the supplementary analysis demonstrates that the generated dataset can reasonably characterize long-term building height dynamics across persistent built-up, complete demolition, and new construction areas. In sampled cities, the predicted height changes are supported by both reference data and historical imagery. The local differences between predicted and reference heights in demolition areas mainly arise from the applicability limitations of the annual reference data generation method. In non-sampled cities, although reference height data are unavailable, the extrapolated temporal trends are consistent with visually identifiable urban change processes in historical remote sensing images.

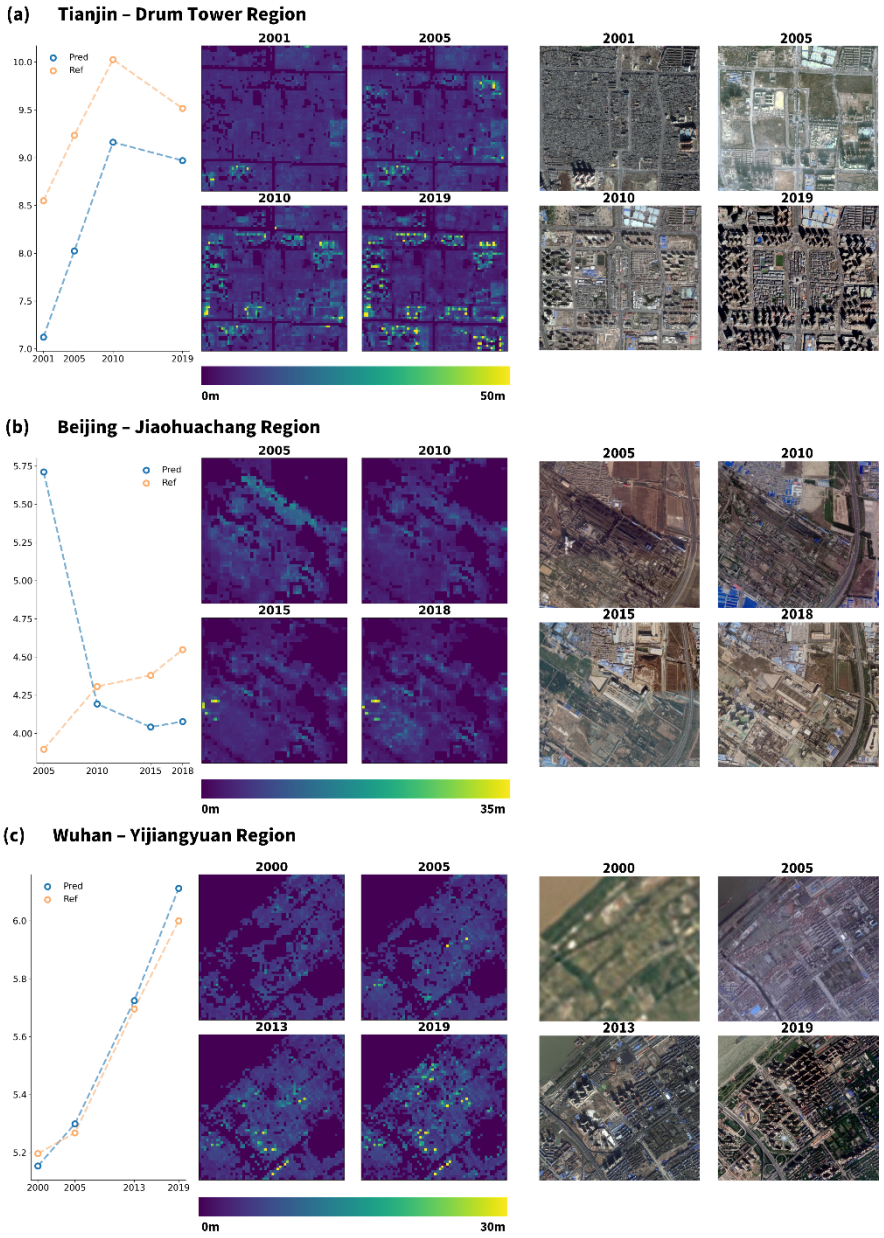

Fig. 5 (the revised Fig. 13 in the original manuscript). Quantitative comparison of urban areas in sampled cities. (a) Drum Tower 435 region, Tianjin; (b) Jiaohuachang region, Beijing; (c) Yijiangyuan region, Wuhan. Remote sensing images are from © Google.

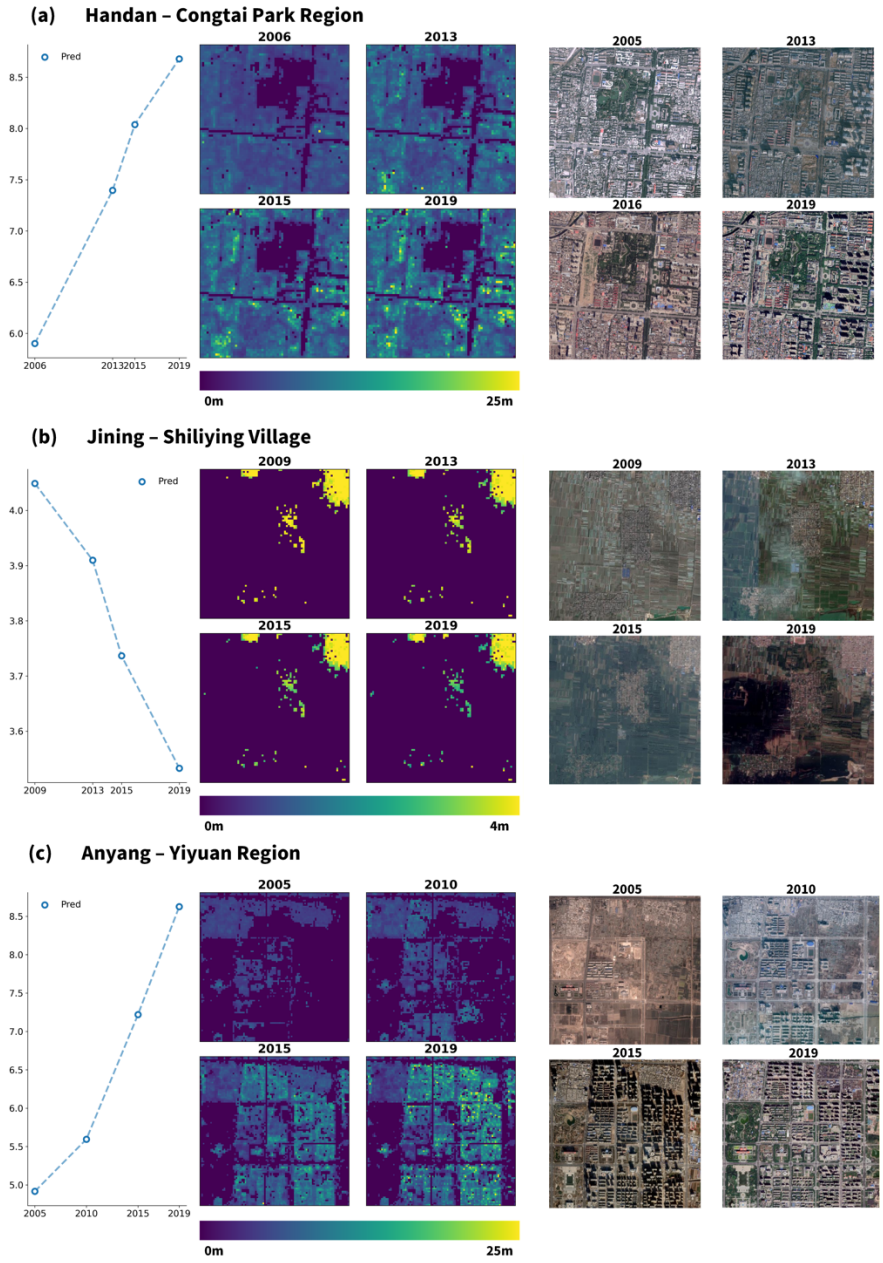

Fig. 6 (the revised Fig. 14 in the original manuscript). Comparison between mapped height changes and historical remote sensing images in non-sampled cities. (a) Congtai Park region, Handan; (b) Shiliying region, Jining; (c) Yiyuan region, Anyang. Remote sensing images are from © Google.

To evaluate prediction stability in unchanged areas, we identify a set of stable areas where building footprints and heights remain unchanged throughout the study period. These areas are defined as pixels with no change year detected by CCDC from 1990 to 2019. As shown in Fig. 7 of this response letter, the estimated mean building heights for these stable areas remain highly consistent from 1990 to 2019. The highest predicted mean height is 7.13 m in 2011, while the lowest is 6.95 m in 1992, corresponding to a fluctuation of less than 0.2 m. The red dotted line represents the mean reference height of the stable areas in 1990, which is 5.58 m and is close to the predicted heights. The orange bars, representing the difference between the predicted height in each year and that in 1990, remain small, with the largest deviation being 0.17 m in 2011. Overall, these results

demonstrate that the dataset does not introduce artificial temporal noise and maintains high temporal stability in areas without physical urban change.

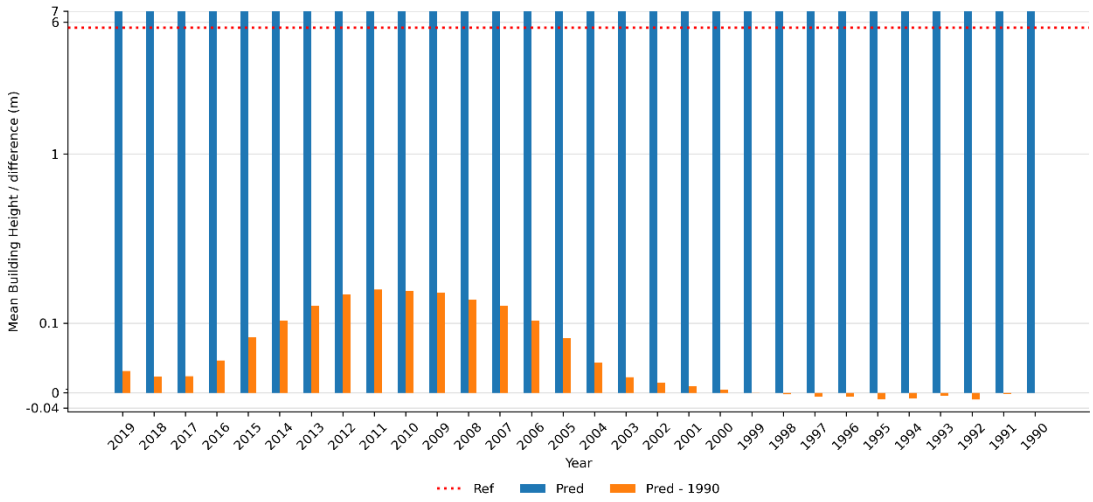

Fig. 7. Predicted height and reference height in stable areas.

Following the reviewer's suggestion, Section 4.4.3 "Comparison with historical remote sensing images" is expanded, reorganized, and retitled as "Comparison with stable areas and historical imagery for temporal reliability validation". This section not only retains the comparison with historical remote sensing images, but also adds a temporal stability assessment in unchanged areas and validation results for building change areas in both sampled and non-sampled cities. Considering that the figure numbering in the manuscript may be adjusted in subsequent revisions, the figure numbers used in the revised manuscript text below are referenced according to the numbering in this response letter.

"…

Section 4.4.3 Comparison with stable areas and historical imagery for temporal reliability validation

To evaluate the temporal stability of the generated building height dataset, stable areas are identified where building footprints and heights remain unchanged throughout the study period. These areas are defined as pixels with no change year detected by CCDC from 1990 to 2019. As shown in Fig. 7 of this response letter, the estimated mean building heights in these stable areas remain highly consistent over time. The highest mean height is 7.13 m in 2011, while the lowest is 6.95 m in 1992, corresponding to a fluctuation of less than 0.2 m. The mean reference height in 1990 is 5.58 m, shown by the red dotted line, and is close to the estimated heights. The orange bars show annual differences from the 1990 estimated height, with a maximum deviation of 0.17 m in 2011. These results indicate that the dataset does not introduce artificial temporal noise and maintains high temporal stability in areas without physical urban change.

To validate the reliability of the long-term building height dataset in building change areas, Fig. 5 and Fig. 6 in this response letter examine three typical urban change contexts described in Section 3.1.1, including persistent built-up, complete demolition, and new construction areas. For sampled cities, predicted heights, reference heights, and historical remote sensing images are jointly compared; for non-sampled cities, where reference height data are

unavailable, historical remote sensing images are used as visual evidence for the extrapolated results.

In sampled cities, the predicted heights generally agree well with both reference heights and historical imagery. For persistent built-up areas and new construction areas, such as Drum Tower in Tianjin (Fig. 5a) and Yijiangyuan in Wuhan (Fig. 5c), the dataset effectively captures building height dynamics. Taking Yijiangyuan as an example, the reference heights in 2000, 2005, 2013, and 2019 are 5.19 m, 5.26 m, 5.69 m, and 6.00 m, respectively, while the corresponding predicted heights are 5.15 m, 5.30 m, 5.72 m, and 6.11 m, with errors all below 0.11 m.

For the complete demolition case, Jiaohuachang in Beijing (Fig. 5b) is selected. The estimated height decreases from 5.71 m to 4.04 m, consistent with the demolition and redevelopment process visible in historical imagery. As described in Section 3.1.1, annual reference heights are derived from persistent built-up areas identified using the 2019 reference data and CCDC breakpoint information, rather than direct observations of all historical building pixels. Therefore, the 2005 difference between the predicted and reference heights, 5.71 m versus 3.52 m, mainly arises from the applicability limitation of the annual reference data generation method in demolition areas. This difference reflects the limited representativeness of the reference height for demolished buildings, rather than model bias or a data error in the generated dataset.

In non-sampled cities, mapped temporal trends are broadly consistent with visually identifiable urban changes. The estimated height at Handan Congtai Park (Fig. 6a) increases from 5.90 m to 7.40 m, 8.04 m, and 8.68 m, reflecting redevelopment from bungalows to multi-story apartments. At Jining Shiliying Village (Fig. 6b), the estimated height decreases from 4.05 m to 3.91 m, 3.74 m, and 3.53 m, consistent with settlement abandonment and demolition caused by subsidence. For Anyang Yiyuan (Fig. 6c), the estimated height increases from 4.91 m to 5.60 m, 7.22 m, and 8.63 m, reflecting progressive new construction on previously undeveloped land. These results indicate that the generated dataset can reasonably characterize long-term building height dynamics across persistent built-up, demolition, and new construction areas.

…″

6. For dataset comparison, it would be more appropriate to focus on comparable 30 m resolution multi-temporal or time-series building height datasets, such as He et al. (2023) and Chen et al. (2025).

We thank the reviewer for suggesting a comparison with other high-quality time-series building height datasets. Following this suggestion, two representative 30m resolution multi-temporal/long-term building height datasets are incorporated into comparative analysis: the 1985–2010 dataset of He et al. (2023) and the 2005\2010\2015\2020 dataset by Chen et al. (2025). Our dataset is compared with these multi-temporal/long-term datasets at both the nationwide scale and the city scale.

Since different datasets vary substantially in reference data sources and building height inference methods, making strict pixel-level quantitative cross-validation not fully appropriate. For example, our study uses vector-based building heights as the source of reference data; He et al. (2023) adopt a sample-free method based on DSM neighborhood analysis; and Chen et al. (2025) mainly extract building heights from GEDI spaceborne lidar observations. Therefore, we focus on comparing temporal height trends and spatial pattern consistency to more reasonably evaluate differences among these datasets.

First, at the nationwide scale, the four datasets all show relatively stable temporal trends in

building height. During 1990–2019, the nationwide mean height of our dataset is close to that of He et al. (2023), with values of approximately 5.17 m and 5.18 m, respectively, whereas Chen et al. (2025) reports a relatively higher mean height of approximately 10.44 m. The CMAB dataset reports even higher values, exceeding 14 m. These differences in building height values are mainly related to differences in sample coverage and reference data sources among the datasets. For example, CMAB focuses more on urban center built-up areas, whereas the other long-term datasets have broader coverage, including rural areas, townships, and low-density built-up areas, resulting in relatively lower overall mean heights. Despite systematic height differences among the datasets, they all show relatively stable temporal trends at the nationwide scale, indicating that our dataset reasonably reflects the stable characteristics of long-term building height changes in China. This stability also suggests that China's urban development maintains a basic balance between horizontal expansion and vertical growth.

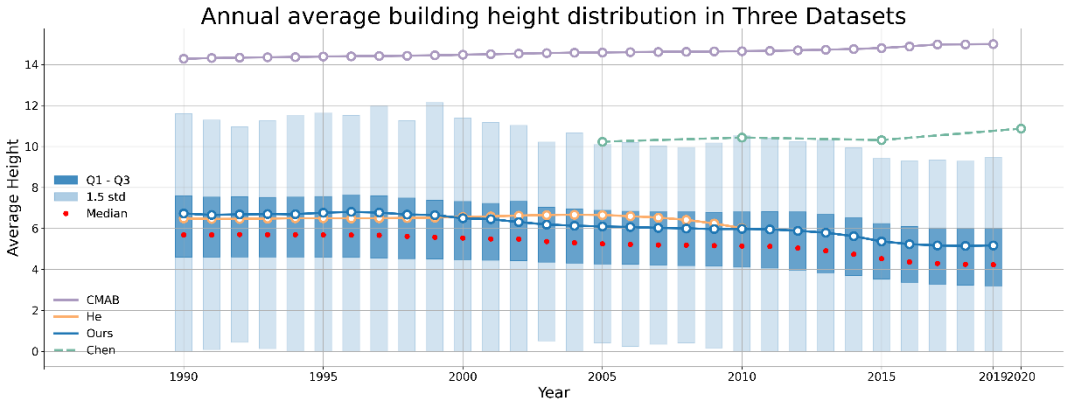

Fig. 8 (the revised Fig. 12 in the original manuscript). Temporal change in annual average height in three long-term building height datasets.

Second, at the urban local scale, Beijing and Guangzhou are selected as two representative cities for detailed comparison. To further examine the ability of different datasets to characterize local building height changes, we select Zhongguancun in Beijing and Liwan in Guangzhou as case study areas with the aid of historical remote sensing images (Fig. 9 in this response letter).

(1) As shown by the historical remote sensing images in Fig.9a in this response letter, Zhongguancun in Beijing experiences obvious urban renewal and demolition–reconstruction processes around 2005, making it suitable for evaluating the ability of datasets to capture dynamic changes in building height. The results show that the dataset of He et al. (2023) fails to sufficiently reflect the demolition and reconstruction process in this area, whereas the dataset of Chen et al. (2025) shows a certain temporal drifting phenomenon, namely unreasonable fluctuations in the heights of some stable buildings across different time steps. In contrast, our dataset better reflects the building height growth and renewal process in the local area.

(2) Liwan District in Guangzhou (Fig. 9b in this response letter) has remained generally stable since 2005, making it suitable for testing the temporal consistency of datasets in stable built-up areas. The comparison shows that the dataset of Chen et al. (2025) also exhibits a certain temporal drifting phenomenon in this area, whereas our dataset maintains better height continuity and temporal consistency in stable built-up areas.

Overall, the supplementary comparison results show that, compared with other long-term datasets,

our dataset maintains reasonable long-term trends at the nationwide scale, better reflects height changes in local renewal areas at the urban scale, and preserves good temporal consistency in stable built-up areas.

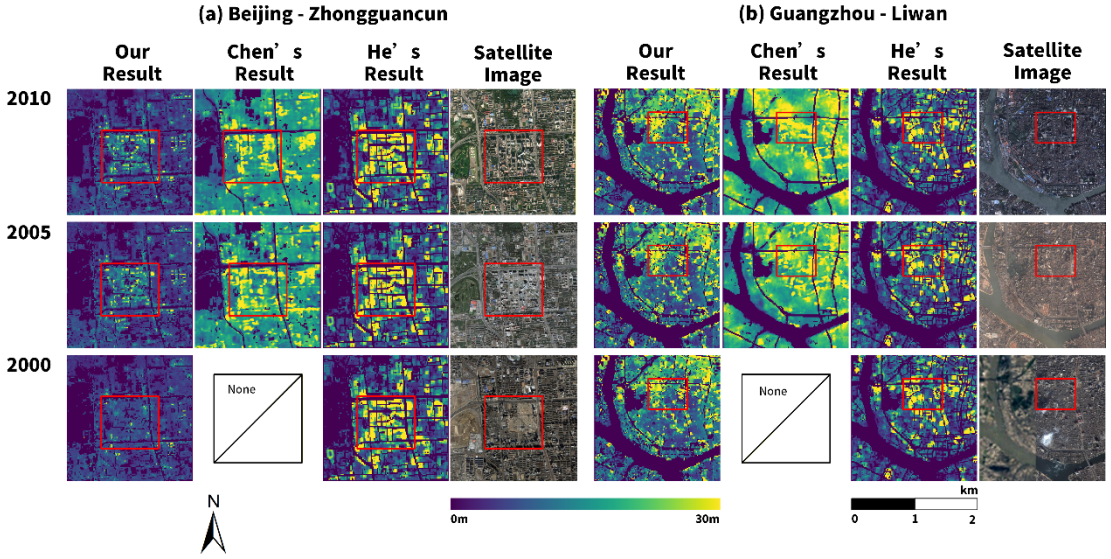

Fig. 9. Comparison of long-term building height datasets in different cities. (a) Beijing; (b) Cuangzhou.

Following the reviewer's suggestion, Section 4.4.2 "Comparison with long-term building height products" is revised as follows. Considering that the figure numbering in the manuscript may be adjusted in subsequent revisions, the figure numbers used in the revised manuscript text below are referenced according to the numbering in this response letter.

"…

Section 4.4.2 Comparison with long-term building height products

To provide a comprehensive comparison with existing long-term building height products, three representative datasets are selected: the 1985–2010 dataset of He et al. (2023), the 2005/2010/2015/2020 dataset of Chen et al. (2025), and the China Multi-Attribute Building (CMAB) dataset (Zhang, Zhao, and Long, 2025). Since He's dataset and CMAB record building construction completion years, the average building height for each reference year is calculated using buildings constructed before that year. For our dataset and Chen et al. (2025), direct annual height averaging is applied.

At the nationwide scale, Fig. 8 in this response letter compares annual mean height trends among these datasets. During 1990–2019, our dataset and He et al. (2023) show similar nationwide mean heights of approximately 5.17 m and 5.18 m, respectively, whereas Chen et al. (2025) reports a higher mean height of approximately 10.44 m, and CMAB exceeds 14 m. These differences are mainly related to sample coverage and reference data sources. For example, CMAB focuses more on urban center built-up areas, while the other long-term datasets cover broader regions, including rural areas, townships, and low-density built-up areas. Despite these systematic height differences, all datasets show relatively stable temporal trends at the nationwide scale, suggesting that our dataset reasonably reflects the long-term stability of building height changes in China.

At the urban local scale, Beijing and Guangzhou are selected for detailed comparison (Fig. 9 in this response letter). In Zhongguancun, Beijing, where

obvious urban renewal and demolition–reconstruction occur around 2005. He et al. (2023) does not sufficiently reflect the local redevelopment process, while Chen et al. (2025) shows temporal drifting with unreasonable height fluctuations for some stable buildings. Our dataset better reflects local building height growth and renewal. In Liwan District, Guangzhou, which has remained generally stable since 2005, Chen et al. (2025) also shows temporal drifting, whereas our dataset maintains better height continuity and temporal consistency.

…″

References:

[1] He, T., Wang, K., Xiao, W., Xu, S., Li, M., Yang, R., and Yue, W.: Global 30 meters spatiotemporal 3D urban expansion dataset from 1990 to 2010, Scientific data, 10(1), 321, https://doi.org/10.1038/s41597-023-02240-w, 2023

[2] Chen, P., Huang, H., Qin, P., Liu, X., Wu, Z., Zhao, F., Liu, C., Wang, J., Li, Z., Cheng, X., and Gong, P.: Characterizing dynamics of built-up height in China from 2005 to 2020 based on GEDI, Landsat, and PALSAR data, Remote Sensing of Environment, 325, 114776, https://doi.org/10.1016/j.rse.2025.114776, 2025

7. Building height results are visualized using both classified and continuous color schemes, with inconsistent value ranges across figures (Figure 11, Figure 13, Figure 14, Figure 15). Adopting a unified classified visualization scheme would improve clarity and consistency.

We thank the reviewer for the suggestion regarding the unification of building height legends and color schemes. We agree that clear and consistent legends are helpful for improving the readability and comparability of the figures.

It should be noted that different classified or continuous color schemes are used in the original manuscript mainly because the figures differ in spatial scale, study object, and visualization purpose. Some figures are designed for the nationwide scale or large municipal administrative regions, which include many low-rise buildings, townships, and rural built-up areas, where building heights are generally low. Other figures are focused on community-scale areas or typical building change areas, where local height differences and temporal change patterns are emphasized. If exactly the same color range and classification intervals are adopted for all figures, subtle height differences in low-rise areas may be compressed, or high-rise areas may become color-saturated, thereby weakening the representation of local details.

Following the reviewer's suggestion and considering the updated dataset, the legends and color schemes of the relevant figures are optimized. For figures requiring cross-region or cross-temporal comparison, more consistent classified color schemes and height intervals are adopted. For figures where local details need to be highlighted, the height ranges and classification standards are explicitly marked in the legends to avoid ambiguity.

Specifically, Fig. 11 in the original manuscript, corresponding to Fig. 10 in this response letter, covers large municipal regions and rural built-up areas with generally low building heights. Therefore, the legend range is set to 0–30 m to better highlight differences in low-rise buildings and urban–rural transitional areas. Fig. 13 (Fig. 5 in this response letter) and Fig. 14 (Fig. 6 in this response letter) mainly show community-scale or local building change areas; therefore, classification schemes suitable for local interpretation are retained, with the corresponding height

ranges clearly labelled. For dense urban core areas with more high-rise structures, the legend range of Fig. 15c–e in the original manuscript (Fig. 11c–e in this response letter) is set to 0–150 m to avoid truncation or color saturation. In addition, a truncated classification scheme is adopted in Fig. 15a–b in the original manuscript (Fig. 11a–b in this response letter) to enhance nationwide visual contrast and avoid excessive homogeneity caused by the dominance of low-rise buildings.

After revision, the building height visualization schemes in Fig. 11, Fig. 13, Fig. 14, and Fig. 15 of the original manuscript (Fig. 10, Fig. 5, Fig. 6, and Fig. 11 in this response letter, respectively) are more standardized, with clearer legends and better balance between cross-figure comparability and local detail representation.

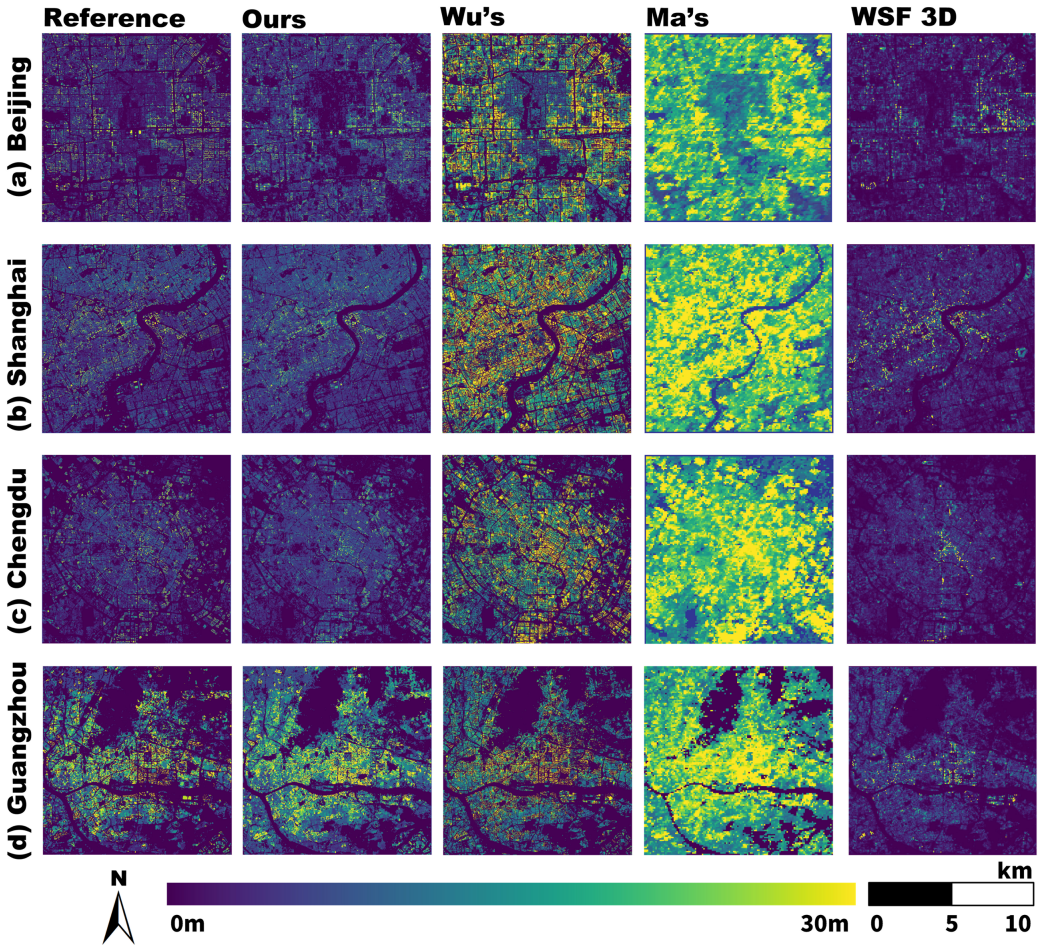

Fig. 10 (the revised Fig. 11 in the original manuscript). Comparison of building height datasets in different cities in 2019. (a) Beijing; (b) Shanghai; (c) Chengdu; (d) Guangzhou.

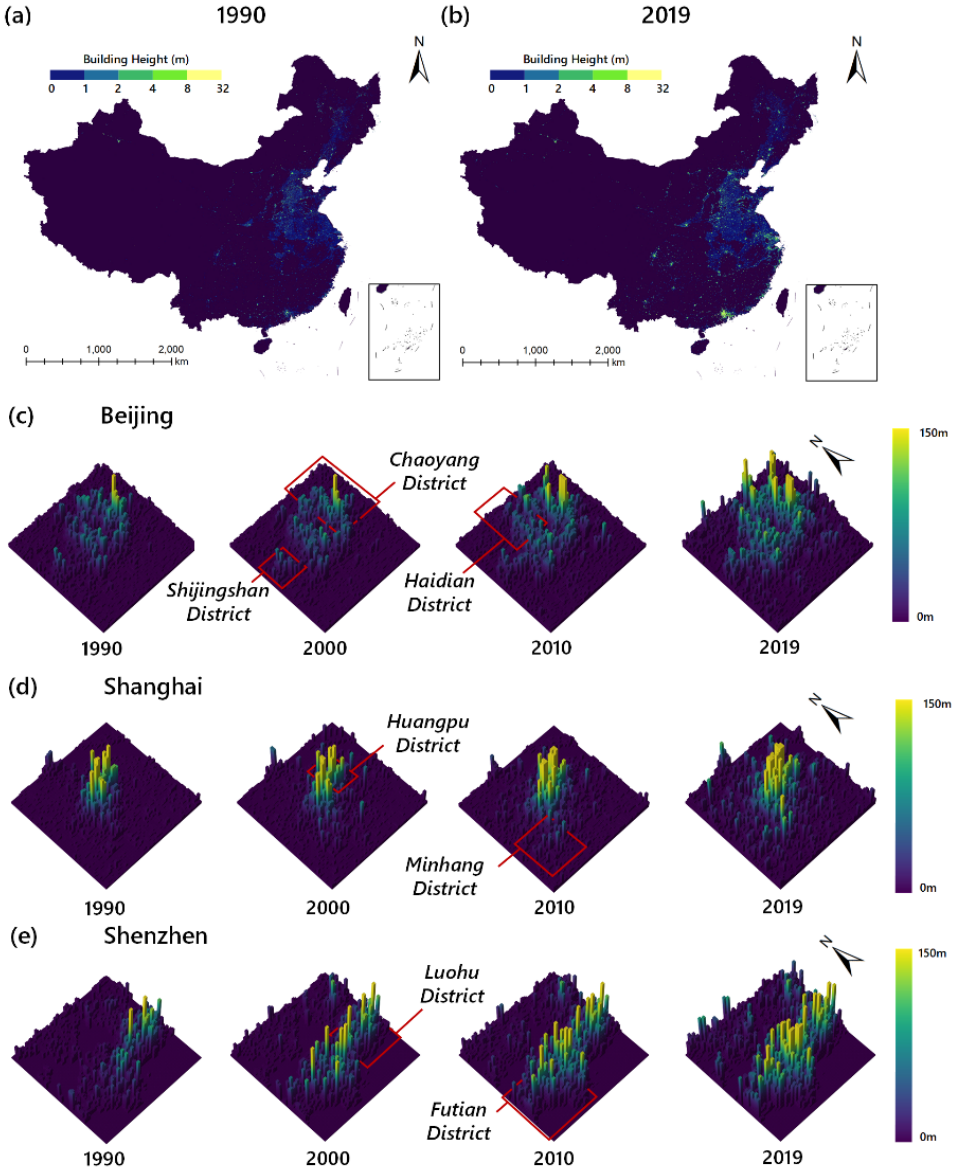

Fig. 11 (the revised Fig. 15 in the original manuscript). Spatial distribution and temporal progression of building heights. (a, b) Spatial distribution of building heights in 1990 and 2019; (c-e) Temporal progression of building height in Beijing, Shanghai and Shenzhen.

8. The comparison presented in Fig. 10 raises several concerns.

We would like to express our gratitude to the reviewer for the meticulous examination of Fig. 10 in the original manuscript, corresponding to Fig. 12 in this response letter. The concerns raised are highly valuable and have guided us in performing a more rigorous re-validation of our comparative analysis.

(1) Different building height products adopt different height definitions (e.g., average height, area-weighted average height), and these definitional differences should be explicitly considered. In addition, it is worth noting that some gridded building height products may include only building information, which appears to differ from the

definition adopted in this study. As a result, direct resampling and comparison across products may not be strictly comparable.

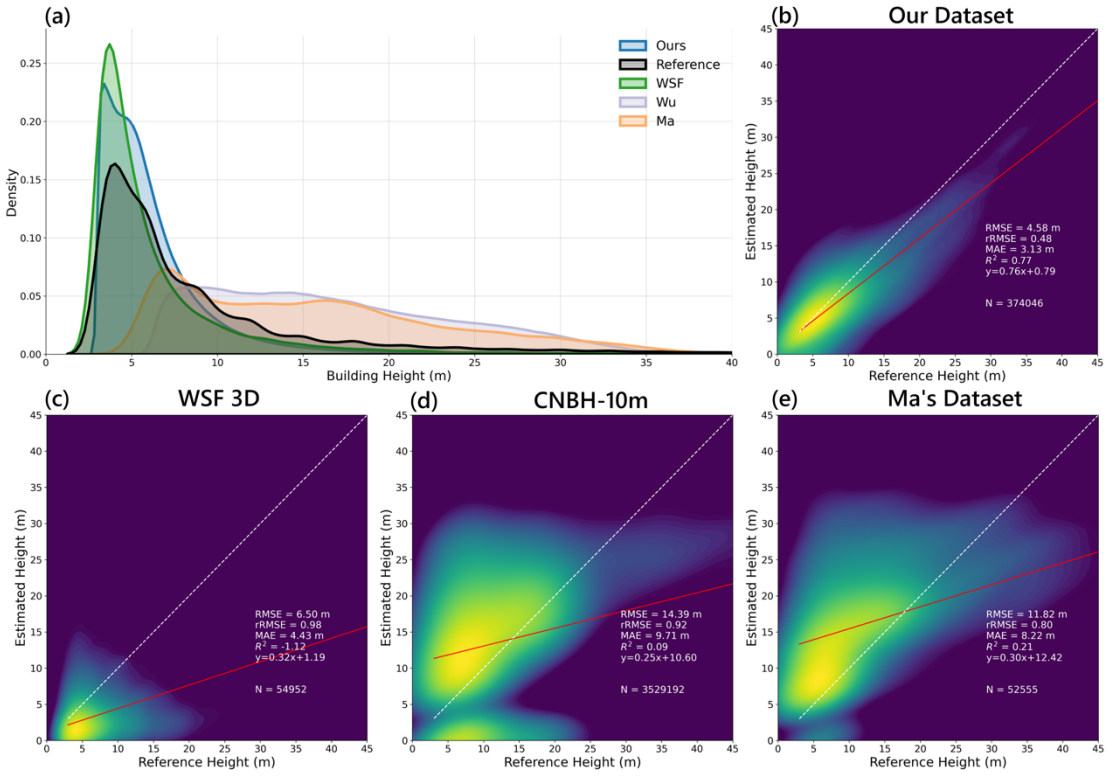

Fig. 12 (the revised Fig. 10 in the original manuscript). Comparison of reference height, our dataset and other datasets in Beijing in 2019. (a) Height distribution of different datasets; (b) Scatter plot of our dataset and reference heights; (c) Scatter plot of WSF 3D (Esch et al., 2022) and reference heights; (d) Scatter plot of CNBH-10m (Wu et al., 2023) and reference heights; (e) Scatter plot of Ma's dataset (Ma et al., 2024) and reference heights.

We thank the reviewer for pointing out the differences in height definitions among different building height products. We fully agree with this comment. Different products may adopt different height definitions, such as average height calculated only over building-covered areas, area-weighted average height that incorporates building coverage fraction, or different treatments of non-building built-up surfaces. If these definitional differences are ignored, direct resampling and comparison across products may lead to limited comparability and affect the fairness and interpretability of the accuracy assessment.

To address this issue, no existing building height product is modified, and the comparison is not based on simply resampling all products to the same spatial scale and directly comparing their absolute height values. Instead, for each evaluated product, corresponding reference data are generated to match its spatial resolution, spatial reference, and height definition. Specifically, the same set of Baidu building height vectors is first converted into a 1m resolution height raster to ensure consistency in the original reference data source. Then, according to the spatial resolution, spatial reference, and height definition of each product, the corresponding aggregation method is applied to generate matched reference data. In this way, consistency in scale, spatial location, and height definition between each evaluated product and its corresponding reference data is ensured as

far as possible.

Specifically, as shown in Table 1 in this response letter, for the datasets of Wu et al. (2023) and Ma et al. (2024), their products mainly represent the height of building-covered areas and exclude non-building areas. Therefore, when the corresponding reference data are generated, only building-covered pixels are averaged and then aggregated to the spatial resolutions of the respective products, namely 10m and 150m. This ensures consistency between the reference height and their "building-only" definition. In contrast, WSF 3D (Esch et al., 2022) derives building height from volume data, and its definition is closer to an area-weighted height that integrates buildings and built-up surfaces, which is more consistent with the grid-level height concept adopted in this study. Therefore, when WSF 3D is evaluated, the corresponding reference data are generated by averaging all pixels within each 90m grid, including non-building built-up surfaces surrounding buildings, to ensure consistency with its volume-based derivation logic.

Therefore, the comparison is not intended to directly compare absolute height differences among different building height products. Instead, while the original products are kept unchanged, each product is evaluated against its definition-matched reference data using accuracy metrics such as RMSE, MAE, rRMSE, and R^2. Through this treatment, comparison bias caused by differences in height definition, spatial resolution, and spatial reference is reduced as much as possible within a more consistent and fair evaluation framework.

Based on the revised comparison results, our dataset achieves the lowest RMSE (4.58 m), rRMSE (0.48), and MAE (3.13 m), as well as a relatively high R^2 (0.77) in Beiijng in 2019. This indicates that, after differences in height definitions among products are explicitly considered and matched reference data are used for evaluation, our dataset still shows good accuracy performance.

Table 1. Resolution and height calculation method for each dataset compared in Fig. 12 in this response letter, BA_i stands for building area with an im resolution pixel.

Dataset	**CNBH-10m (Wu et al., 2023)**	**Ma et al., 2024**	**WSF 3D (Esch et al., 2022)**	**Our dataset**
Resolution	10m	150m	90m	30m
Height definitions	Building only	Building only	Building and built-up surface	Building and built-up surface
Height calculation	$\frac{\sum_i^{BA_{10}} H_i^{1m}}{BA_{10}}$	$\frac{\sum_i^{BA_{150}} H_i^{1m}}{BA_{150}}$	$\frac{\sum_i^{8100} H_i^{1m}}{8100}$	$\frac{\sum_i^{900} H_i^{1m}}{900}$

We thank the reviewer again for the constructive comment. Following the reviewer's suggestion, Section 4.4.1 is revised as follows:

"…

For comparison with single-year building height datasets, Beijing in 2019 is selected as the study area because of its complex urban morphology and heterogeneous building height distribution.

To ensure a fair assessment across products with different height definitions, the original building height products are kept unchanged, and definition-matched reference data are generated from the same Baidu building height vectors. The vectors are first converted into a 1m height raster and then aggregated according to each product's resolution, spatial reference, and height definition. For Wu et al. (2023) and Ma et al. (2024), only building-covered pixels are averaged and aggregated to 10m and 150m, respectively; for WSF 3D

(Esch et al., 2022), all pixels within each 90m grid are averaged to match its area-weighted, volume-derived height definition.

Therefore, each product is evaluated against its matched reference data using consistent accuracy metrics, rather than by directly comparing absolute height values across products with different definitions. In Beijing in 2019, our dataset achieves the best overall accuracy, with the lowest RMSE (4.58 m), rRMSE (0.48), and MAE (3.13 m), as well as a relatively high R^2 (0.77).

…."

(2) The reported RMSE of 748.79 m is physically implausible and requires clarification.

We sincerely apologize for the physically implausible RMSE value (748.79 m) reported in the initial submission. This error occurred during the coding phase, where NaN values in the reference dataset are erroneously included in the numerical calculation. We have conducted a thorough review of our validation scripts and corrected this oversight. The updated RMSE for Ma's dataset is now 11.82 m, which is consistent with the validation results reported in recent literature (e.g., Fig. 15 in Chen et al., 2025). We deeply regret this technical error and have implemented stricter data-cleaning protocols for all reported metrics.

(3) Please note that it is unclear whether the same reference building height data were used consistently across Fig. 10b–e. Reference heights above 20 m appear in some panels (b–d) but not in others (e), and zero-height pixels are shown for Ma's dataset, although such values do not appear in Ma's original released data.

We thank the reviewer for pointing out this issue. As clarified in our response to Question 8(1), the original reference data source is consistent across Fig. 10b–e in the original manuscript, corresponding to Fig. 12b-e in this response letter. The same set of Baidu building height vectors is first converted into a 1m resolution height raster, and then aggregated according to the spatial resolution, spatial reference, and height definition of each evaluated product to generate matched reference data. Therefore, the reference data in Fig. 10b–e (Fig. 12b-e in this response letter) are derived from the same original source, and the apparent differences among panels are mainly caused by the previous data process, rather than by inconsistent reference data sources. We have carefully checked and updated the corresponding comparison results. As shown in the revised Fig. 10e (Fig. 12e in this response letter), the reference height distribution for Ma's dataset now correctly includes values above 20 m, and the estimated heights are consistent with the characteristics of Ma's original released data.

(4) It is unclear whether this part of the validation is limited to buildings with heights below 45 m.

We sincerely appreciate the reviewer's comment. We clarify that this part of the validation is not limited to buildings below 45 m. All accuracy metrics are calculated using the full validation dataset, including buildings higher than 45 m. The axes are capped at 45 m only for visualization purposes in the figure. This is because the majority of data points are concentrated between 0 and 40 m; displaying the full height range would compress most points into a small corner of the plot and make the main data distribution difficult to inspect. As shown in the density plot in the revised Fig. 10 (Fig. 12 in this response letter), the 45 m axis limit provides a better visualization balance and clearer

presentation of the dominant data distribution, but it does not affect the validation samples or the reported accuracy metrics. The following clarification is added in Section 4.4.1 of the revised manuscript:

"…Notably, all accuracy metrics are calculated using the full validation dataset; the 45 m axis limit is used only for visualization to better show the dominant data distribution.…"

9. Overall, Section 4.4 requires careful re-examination.

(1) More detailed descriptions of the accuracy assessment and dataset comparison procedures are required, for instance, whether the validation is conducted at the pixel level or using object-based approaches based on building footprints (with height).

We thank the reviewer for suggesting a more detailed description of the accuracy assessment and dataset comparison procedures. We fully agree that clarifying whether validation is conducted at the pixel level or through an object-based approach based on building footprints is important for ensuring the transparency, reproducibility, and interpretability of the dataset evaluation.

Following the reviewer's suggestion, more detailed descriptions of the accuracy assessment and dataset comparison procedures are added to the revised manuscript, especially regarding data processing, scale matching, and statistical methods used in different validation scenarios.

Specifically, in Section 4.4.1 "Comparison with single-year building height datasets", following our response to Question 8, the pixel-level validation procedure for comparison with single-year building height datasets is further clarified. This section, based on Fig. 12 in this response letter, clarifies that the original products are kept unchanged, while the corresponding reference data are matched and aggregated according to each product's spatial resolution, spatial reference, and height definition. Each product is then evaluated at the pixel level against its definition-matched reference data, rather than through a direct comparison of absolute height values among products with different definitions.

In Section 4.4.2 "Comparison with long-term building height products", following our response to Question 6, the processing procedure for comparing long-term building height products is further supplemented. Based on Fig. 8 and Fig. 9 in this response letter, this section clarifies that validation is mainly based on pixel-level or grid-level statistics. It is also clarified that the building footprints in CMAB are used only to calculate the nationwide average building height in Fig. 8 in this response letter.

In Section 4.4.3 "Comparison with stable areas and historical imagery for temporal reliability validation", following our response to Question 5, the temporal reliability validation procedure for the dataset is further clarified. This section, based on Fig. 5, Fig. 6, and Fig. 7 in this response letter, explains that the validation of stable areas, building change areas in sampled cities, and extrapolated change areas in non-sampled cities is also mainly based on pixel-level or grid-level results. Stable areas are identified using pixels with no change year detected by CCDC, and their multi-year estimated height stability is then assessed. Building change areas are evaluated by comparing pixel-level height change trends with reference data or historical remote sensing images.

Overall, the revised manuscript more clearly distinguishes the data processing methods and statistical scales used in different validation scenarios. The validation and comparison in this study are mainly based on pixel-level/grid-level data processing and statistical assessment. Building

footprint data are used only in specific comparisons as spatial statistical boundaries or reference data sources, and do not constitute object-based validation. We thank the reviewer again, as these clarifications help better explain the validation scale and processing workflow and improve the transparency and reproducibility of the assessment results.

(2) If possible, providing spatial uncertainty or confidence layers would substantially enhance the usability of the dataset.

We fully agree with the reviewer that providing a spatial uncertainty or confidence layer is essential for enhancing the usability and transparency of the dataset. In response to this suggestion, a per-pixel confidence layer is generated for the building height products based on the Mahalanobis distance (MD) in the feature space.

The MD-based confidence layer is used to describe the relative reliability of building height estimates by measuring the distance between each predicted pixel and the multivariate distribution of the training samples. Specifically, the Mahalanobis distance is calculated as:

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \quad , \tag{1}$$

where $D_M(x)$ denotes the Mahalanobis distance from a given pixel x to the distribution, μ is the mean vector of the training samples, and Σ^{-1} is the inverse covariance matrix of the training samples.

A lower MD value indicates that the pixel is closer to the feature distribution of the training samples and therefore has higher confidence in the building height estimate. Conversely, a higher MD value suggests that the pixel is farther from the training distribution, indicating higher uncertainty or lower confidence (Lee et al., 2018). To make the uncertainty information easier to interpret, D_M^2 is further converted into a 0–1 confidence score using the survival function of the chi-squared distribution (Etherington, 2019; Frost, 2020):

$$D_M^2 \sim \chi^2(k) \tag{2}$$

$$confidence\ score = S(D_M^2) \tag{3}$$

where k denotes the degrees of freedom of the chi-squared distribution, corresponding to the number of feature variables used in each year, and $S(\cdot)$ is the survival function of the chi-squared distribution. The confidence score indicates how consistent a pixel is with the training sample distribution, with larger values representing pixels closer to the distribution center. To further generate a binary usability layer U, a practical threshold α is adopted, commonly set to 0.05 or 0.01 (Etherington, 2019). Pixels with confidence scores lower than α are more likely to fall outside the sample ellipsoid and are therefore regarded as potential outliers. In this study, α = 0.01 is used to generate the binary confidence layer:

$$U = \begin{cases} 1 & confidence\ score \geq \alpha \\ 0 & confidence\ score < \alpha \end{cases} \tag{4}$$

where $U = 1$ indicates pixels within the acceptable range of the training feature distribution, and $U = 0$ indicates pixels that may be outliers or extrapolated samples and should be used with greater caution.

To evaluate the usability of the dataset, binary confidence layers are generated for representative

large cities, including Beijing, Shanghai, and Chengdu. As shown in the binary confidence layers from 1990 to 2019 (Fig. 13 in this response letter), most building areas in these cities fall within the 99% confidence range, indicating that their feature distributions are close to the training samples and that the corresponding building height estimates are generally reliable.

Beijing (Fig. 13a in this response letter) and Shanghai (Fig. 13b in this response letter) show good spatial continuity in 1990, 1995, 2000, 2005, 2010, 2015, and 2019, with only a small number of sparsely distributed low-confidence patches. Chengdu (Fig. 13c in this response letter) also shows a high proportion of building areas within the 99% confidence range in most years, although a relatively concentrated low-confidence area appears in 1995. This pattern may be related to the smaller early built-up extent, differences in remote sensing feature quality, or local urban morphology that differs from the training sample distribution. Overall, these results demonstrate the reliability of the dataset across representative Chinese cities.

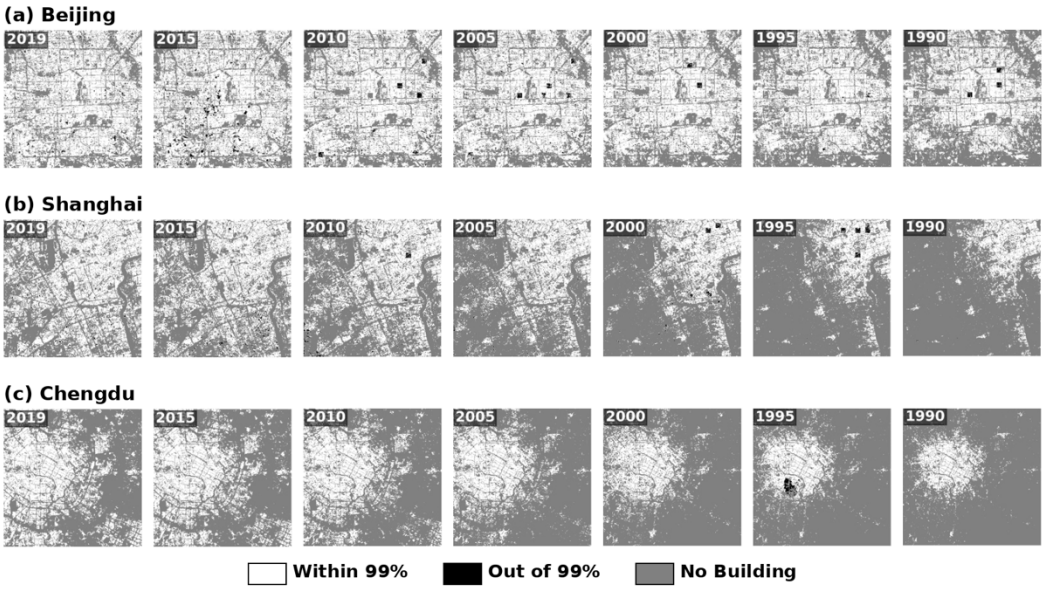

Fig. 13. Binary confidence layers for Beijing, Shanghai, and Chengdu from 1990 to 2019.

Following the reviewer's suggestion, a new Section 4.6, "Limitations and future work", is added to the revised manuscript to further discuss the reliability, uncertainty, and usability of the dataset. Accordingly, the title of the original Section 6 is revised as "Conclusions".

Three main aspects are addressed in the new Section 4.6. First, the spatial uncertainty of building height estimates is discussed based on the confidence layers. Second, the potential underestimation in high-rise building areas is clarified and identified as one of the directions for future improvement. Third, the spatial transferability and cross-city robustness of the model are further discussed using the validation results in non-sampled cities, as supplemented in our response to Question 2.

Through these revisions, the reliability, potential uncertainty, and usability of the dataset are more clearly described in the revised manuscript. This helps users apply the dataset more appropriately across different urban contexts and building height types.

"…

Section 4.6 Limitations and future work

Although the generated long-term building height dataset shows good overall accuracy and temporal consistency, several limitations and uncertainty sources should be further discussed to support appropriate data use.

First, spatial uncertainty varies across regions because the reliability of building height estimates depends on the similarity between predicted pixels and the training samples in the feature space. Areas whose feature distributions deviate from the training samples may have higher uncertainty. To provide additional reliability information, a per-pixel confidence layer is generated using the Mahalanobis distance (MD), which measures the distance between each predicted pixel and the multivariate distribution of the training samples (Lee et al., 2018):

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \quad (11)$$

where $D_M(x)$ denotes the Mahalanobis distance of pixel x, μ is the mean vector of the training samples, and Σ^{-1} is the inverse matrix. A lower MD value indicates that the pixel is closer to the training sample distribution and has higher confidence, whereas a higher MD value indicates greater uncertainty.

For easier interpretation, D_M^2 is converted into a 0–1 confidence score using the survival function of the chi-squared distribution. A threshold of α=0.01 is then used to generate a binary confidence layer, where pixels within the 99% confidence range are regarded as reliable, while pixels outside this range are considered potential outliers or extrapolated samples and should be used with caution (Etherington, 2019; Frost, 2020).

The binary confidence layer is further examined in representative cities, including Beijing, Shanghai, and Chengdu (Fig. 13 in this response letter). From 1990 to 2019, most building areas in these cities fall within the 99% confidence range, indicating that their feature distributions are generally close to the training samples and that the corresponding building height estimates are reliable. Beijing and Shanghai show good spatial continuity across all selected years, with only sparsely distributed low-confidence patches. Chengdu also shows a high proportion of reliable building areas in most years, although a relatively concentrated low-confidence area appears in 1995, possibly related to early built-up extent, remote sensing feature quality, or local morphology differences. Overall, these results demonstrate the reliability of the dataset across representative Chinese cities, while further work is still needed to refine the confidence-layer framework and improve uncertainty characterization across different urban contexts.

Second, underestimation remains in some high-rise areas, mainly due to limited high-rise reference samples. Future improvements will require more high-quality high-rise samples and finer-resolution remote sensing features to better represent vertical urban structures. Third, spatial transferability remains challenging in cities excluded from the reference samples, especially where building morphology differs from the training data, as described in Section 4.2. Further work is needed to develop deep learning models with stronger spatial transferability and cross-city robustness.

…."

References:

[1] Lee, K., Lee, K., Lee, H., and Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks, Advances in neural information processing systems,

31, 2018.
[2] Etherington, T. R.: Mahalanobis distances and ecological niche modelling: correcting a chi-squared probability error, PeerJ, 7, e6678, https://doi.org/10.7717/peerj.6678, 2019
[3] Frost, H. R.: Variance-adjusted Mahalanobis (VAM): a fast and accurate method for cell-specific gene set scoring, Nucleic acids research, 48, e94, https://doi.org/10.1093/nar/gkaa582, 2020

(3) Please carefully check the reference formatting. The citation of Che et al. (2024) should refer to the final published version, rather than the Earth System Science Data Discussions version.

We appreciate the reviewer's careful attention to the reference formatting. The manuscript is thoroughly checked, and the two citations, Che et al. and Zhu et al., are updated as requested:

> "…Che, Y., Li, X., Liu, X., Wang, Y., Liao, W., Zheng, X., Zhang, X., Xu, X., Shi, Q., Zhu, J., Yuan, H., and Dai, Y.: 3D-GloBFP: the first global three-dimensional building footprint dataset, Earth Syst. Sci. Data, 16, 5357–5374, https://doi.org/10.5194/essd-16-5357-2024, 2024. …"
>
> "…Zhu, X. X., Chen, S., Zhang, F., Shi, Y., and Wang, Y.: GlobalBuildingAtlas: an open global and complete dataset of building polygons, heights and LoD1 3D models, Earth Syst. Sci. Data, 17, 6647–6668, https://doi.org/10.5194/essd-17-6647-2025, 2025. …"

A comprehensive check is also performed of all other references to ensure they align with the journal's formatting requirements and reflect the most recent published versions available.