This paper presents a dataset as a benchmark for developing and evaluating methods to extract PV panels from satellite images. Such a dataset is scarce but timely for the relevant community and will help researchers reduce the time and cost of collecting high-quality samples and will enrich other work related to solar PV energy. Overall, this manuscript is technically sound and well-crafted.

Response: Thank you for your encouraging comments.

I would like to suggest a revision before it can be considered for publication in the Earth System Science Data.

Major concerns:

1. The dataset was collected in Jiangsu Province. Considering that the changes in geographic context may affect the model's performance, how can you guarantee that the model trained with this dataset can be generalized to other regions?

Response: We compared the samples from Gaofen-2 and Beijing-2 images, and found that PV panels exhibit similar characteristic in high-resolution imagery and the main difference comes from the backgrounds. The changes in geographic context will inevitably affect the performance of deep networks. To mitigate such effects, we tried to collect PV samples covering as many backgrounds as possible to enhance the generalization ability of deep networks trained by our dataset. The installed PVs in Jiangsu province are distributed on various land covers, such as, sparse shrubs, low-density grasslands, reservoirs, ponds, saline alkali lands and rooftops, making our dataset representative for most cases. Besides, some skills in deep learning community can be adopted to guarantee the generalization ability in other regions, such as transferring learning, cross-domain feature representation. In completely different areas where direct application may fail, using a small number of samples in the target area to fine-tune the pre-trained model can also achieve satisfactory results. Therefore, although the samples were collected in Jiangsu Province, there is no need to worry too much. Furthermore, our work will continue, and we plan to collect more PV samples in China and around the world through automatic deep learning algorithm and manual discrimination.

2. The dataset is composed of Gaofen-2 and Beijing-2 imagery, which include a near-infrared band in addition to RGB bands. Does this mean that the model trained with this

dataset can only be applied to the satellite imagery with a specific combination of spectrums? Are these two imagery representative of the current common satellite imagery?

Response: Actually, all samples in our dataset are composed of red, green and blue (RGB) bands of used satellite and aerial images. Our experiments (Section 4.1) also demonstrate that RGB bands are enough for deep networks to distinguish PV panels from various backgrounds. Therefore, the near-infrared band of Gaofen-2 and Beijing-2 images is not contained in our samples. We have made it clear in Section 3 as "The shapefile of polygonal annotations was converted to a raster that has the same spatial resolution as satellite or aerial images. The raster and original red, green and blue (RGB) images were then seamlessly cropped into tiles at a fixed size by referring to the sampling grids." The comparison of PV samples from Gaofen-2 and Beijing-2 imagery shows that different PV panels exhibit similar characteristic after histogram equalization. Thus, we believe that the two satellite imagery are representative for common applications.