RC3: This manuscript proposes a lightweight Embedded Seamless Database (ESD) based on global Earth observation images, which features high compression and high reconstructive fidelity. The research in this dataset is of great significance, as it provides a different solution from AEFs for global-scale Earth analysis. Judging from the download volume on the data portal (https://data-starcloud.pcl.ac.cn/iearthdata), the dataset appears to be very popular. Overall, it is a interesting dataset and a well-written manuscript. I would like to offer the following comments for reference.

Response: We sincerely thank the reviewer for the positive assessment of the dataset and the manuscript, and for noting the uptake of the product on the data portal. The comments below have helped us substantially improve the clarity of the core figures and the completeness of the reproducibility information.

3.1 Figures 1 and 2 are the core illustrations that convey the essential methodological logic of the paper, directly shaping readers' understanding of the ESD pipeline and the ESDNet architecture. However, the current version suffers from serious ambiguity and insufficient information delivery. In particular, subfigures (b) and (c) in Figure 2 give the impression of being redundant and lacking substantive meaning due to design deficiencies. In addition, there is some logical overlap between Figures 1 and 2, and the authors may consider improving the way the relationship between these two figures is presented. The relationships among the three subfigures (a), (b), and (c) in Figure 2 also require more detailed explanation and clearer correspondence. Overall, as these are central components of the paper, it would be helpful to include more key information and technical background regarding ESDNet and the FSQ module.

Response: We thank the reviewer for this important observation. We agree that Figures 1 and 2 are central to readers' understanding of both the ESD production pipeline and the ESDNet architecture, and that the original presentation did not explain their relationship clearly enough. In the revised manuscript, we have retained the three subpanels of Figure 2, but substantially improved the caption and the corresponding methodological description in Section 3.2 so that the roles of Figures 1 and 2, as well as the correspondence among panels (a), (b), and (c), are stated explicitly. First, we now clarify the division of labor between the two figures. Figure 1 is used to describe the dataset-level production pipeline from multi-sensor observations to the released ESD product, whereas Figure 2 is used to explain the internal computational workflow of ESDNet. This clarification reduces the previous ambiguity in how the two figures should be interpreted together. Second, we have retained panels (b) and (c) of Figure 2 because they serve as schematic expansions of the Encoder and Decoder blocks shown in panel (a). In the revised caption, panel (a) is explicitly described as the overall workflow, while panels (b) and (c) are described as the internal compositions of the Encoder and Decoder, respectively. This makes the correspondence among the three panels much clearer and avoids the impression that panels (b) and (c) are visually redundant. Third, to address the reviewer's concern about missing technical information, we have expanded the description of ESDNet and the FSQ bottleneck in Section 3.2 and added Table 2 to summarize the released architectural hyperparameters. Together, the revised caption, the expanded methodological text, and Table 2 provide the additional technical background needed to interpret Figure 2 more independently and reproducibly.

Revision in manuscript:

We rewrote the caption of Figure 2 to explain more explicitly the roles of panels (a), (b), and (c), and we added a clarifying sentence in Section 3.2 stating that Figure 1 summarizes the dataset-level production pipeline whereas Figure 2 details the internal computational workflow of ESDNet. We also expanded Section 3.2 and added Table 2 to provide the associated architectural and FSQ details.

3.2 The three core dimensional symbols in Figure 1, 365, H, W, C1; 12, H, W, C2; and 12, H, W are not explained anywhere in the manuscript. Their physical meanings should be clearly clarified either within the figure itself or in the figure caption, so that the figure can be understood independently of the main text.

Response: We thank the reviewer for catching this omission. We agree that the dimensional symbols in Figure 1 should be understandable without requiring the reader to infer their meaning from the main text. We have therefore expanded the caption of Figure 1 to define each symbol explicitly. In the revised caption, 365 denotes the annual daily axis of the SDC30 cube, H and W denote the height and width of a processing tile, C_1 denotes the per-time-step input feature dimensionality after preprocessing, 12 denotes the number of monthly latent steps produced at the bottleneck, and C_2 denotes the latent dimensionality at each step. We also verified that these definitions are now consistent with Figure 2 and Table 2.

Revision in manuscript (Figure 1 caption): "*Figure 1: Schematic of the integrated ESD production pipeline. (a) Heterogeneous Landsat and MODIS time-series observations are harmonized and fused into the unified Global 30-m Seamless Data Cube (SDC30). (b) For each processing tile, the annual SDC30 feature cube of shape [365, H, W, C_1] is preprocessed and then transformed by the Encoder into a compact latent representation, which is then discretized by the Finite Scalar Quantization (FSQ) module into 12 monthly latent steps of shape [12, H, W, C2]. (c) The quantized embeddings are used by the Decoder to reconstruct the original surface-reflectance time series and by the MLP task heads to impose auxiliary thematic supervision during training, yielding the released ESD product. Here, 365 denotes the annual daily axis of the SDC30 cube. H and W denote the height and width of a processing tile. C_1 denotes the per-time-step input feature dimensionality after preprocessing. The internal latent representation contains 12 monthly steps, and C_2 denotes the latent dimensionality at each step. The distributed annual ESD product retains the 12-step temporal axis and stores one quantized token index per step, yielding a tensor of shape [12, H, W]. In the released configuration, $C_2 = 6$.*"

3.3 In Table 3, the global annual SDC30 dataset is reported as 0.8 PB, while ESD is reported as 2.4 TB. This implies a theoretical compression ratio of approximately 333, which is somewhat inconsistent with the ~340 description used in the abstract and the main text.

Response: We thank the reviewer for pointing out this ambiguity. The apparent discrepancy arises from the storage-unit convention used in the manuscript. In our calculation, the reported annual SDC30 volume of 0.8 PB is interpreted using the binary convention for data storage, i.e., 1 PB = 1024 TB. Under this convention, the compression ratio is 1024/2.4≈341, which is consistent with

the approximately ~340 reduction reported in the abstract and main text. We have revised the manuscript to clarify the unit convention explicitly in the data-volume comparison table and related text, so that the basis of the reported compression ratio is unambiguous.

3.4 Figure 3 appears very similar to the ESA global classification product, making it difficult to identify its relationship to the samples or training dataset used in this study.

Response: We thank the reviewer for this important observation. We confirm that Figure 3 is not a wall-to-wall ESA WorldCover map, but a visualization of the stratified random sampling sites used to construct the ESDNet training dataset. In the original figure, each sample location was colored according to its dominant ESA WorldCover 2021 class, which made the figure visually similar to the reference product and may therefore have obscured its connection to the actual training samples. This effect is further strengthened by the fact that the training sites were generated using a spatially uniform global sampling design, so the high density of colored points can visually resemble a continuous land-cover map in some regions. To remove this ambiguity, we have revised the figure caption to state more explicitly that the map shows the sampled training locations and that the displayed colors are derived from the dominant ESA WorldCover 2021 class at each site.

Revision in manuscript (Figure 3 caption): "*Figure 3. Geographic distribution and thematic composition of the global training dataset. The map displays the locations of the 223,622 stratified random sampling sites used to train the ESDNet. These sites were generated using a spatially uniform global sampling design, so their high density may visually resemble a wall-to-wall land-cover product in some regions. Each site is colored according to its dominant land-cover class as derived from the ESA WorldCover 2021 product.*"

3.5 The paper only briefly mentions three key hyperparameters in the ablation experiments, such as 12 temporal steps, 10 residual blocks, and a 65,536-dimensional embedding space. However, the manuscript does not provide a complete set of training and inference hyperparameters for ESDNet, making it difficult for third-party researchers to fully reproduce the reported results.

Response: We thank the reviewer for this important suggestion. We agree that the original manuscript did not provide a sufficiently complete description of the settings needed for independent reproduction. In the revised manuscript, we therefore supplemented the reproducibility information at two levels. First, a new Table 2 in Section 3.2 summarizes the released architectural hyperparameters of ESDNet. Second, a short paragraph in Section 3.4 summarizes the optimization and deployment settings of the released configuration.

Revised paragraph (Section 3.4): "*For reproducibility, the optimization and deployment settings of the released configuration are summarized as follows. Training used the Adam optimizer with a learning rate of* 5×10^{-4}*, without warmup or learning-rate decay. The public training script defaults to a batch size of 1024, uses 8 data-loading workers, evaluates every 2000 steps, saves checkpoints every 10000 steps, and trains for up to 1000 epochs with early stopping after 10 evaluation intervals without improvement. No custom weight-initialization scheme was applied beyond the default PyTorch initialization. For the released loss configuration,* $\alpha = 1.0$ *and*

β = γ =0.1. For inference, annual production tiles are generated tile-wise at 3600×3600 *pixels and partitioned into non-overlapping 600 × 600 subwindows for memory-efficient processing. Each annual sequence is converted into 13 input features and temporally interpolated to length 96 before encoding. No overlap blending is used during tile generation.*"