Response Letter to Earth System Science Data Submission

Manuscript Number: ESSD-2025-157

Paper title: OneDZ: A Global Detrital Zircon Database and Implications for Constructing Giant Geoscience Database

General response:

We sincerely thank the Editor for their constructive comments and patience. We fully acknowledge that an online database must allow users to find and download data reliably. After careful consideration, we have opted for your alternative (2): we have removed all descriptions of the relational database schema, the MySQL backend, and the OneDZ web portal from the manuscript. The revised manuscript now treats OneDZ exclusively as a compiled, harmonised, and quality-controlled flat-file dataset distributed via Zenodo. This approach ensures immediate data accessibility without relying on any custom web interface.

Point-to-point response:

Comment 1: please add the csv files of all your data points to the Zenodo Data Publication

Reply: Thank you for your careful review of the data repository. In version 5.0, all SQL files have been removed from the primary repository, which now contains only flat CSV files in the root directory, immediately visible upon landing on the Zenodo page (https://doi.org/10.5281/zenodo.19690702, version 5.0). Specifically:

(1) Total U-Pb dataset: 22 sequential CSV parts (zircon_upb_part_01.csv to zircon_upb_part_22.csv), totalling 2,550,738 records with a uniform 64-column schema.

(2) Total Lu-Hf dataset: 3 sequential CSV parts (zircon_luhf_part_01.csv to zircon_luhf_part_03.csv), totalling 297,527 records with a uniform 33-column schema. Note that the Lu-Hf dataset has been checked by experts.

(3) Expert-verified U-Pb subset: 14 sequential CSV parts (expert_upb_part_01.csv to expert_upb_part_14.csv), totalling 1,414,062 records.

All files use UTF-8 encoding with BOM, comma delimiters, LF line endings, and a single header row, enabling direct ingestion into R, Python, Excel, or MATLAB without any database software. A comprehensive README.md file in the repository root details the column definitions, file-splitting rationale (~100,000–130,000 rows per part for download efficiency), and cross-dataset linkage keys (Lead_Author + Year + Published_Sample_ID + Grain). No technical expertise is required to download or open these files.

Comment 2: please remove all description of your relational database and OneDZ data portal from the manuscript and focus on your data compilation and harmonisation

Reply: We have fully implemented this request. The following systematic changes have been made throughout the manuscript and supplementary materials:

(1) Introduction: We removed the sentence “professional database management software such as MySQL was adapted”. The revised text now describes Python scripts developed for data cleaning, format standardisation, and quality control, with all final outputs normalised to a uniform CSV schema (lines 80-94).

(2) Section 3 (formerly “Database”): This section has been entirely rewritten as “Data compilation and harmonization”. All descriptions of the three core tables (main, age, geography), the zircon_id primary-key linkage, and the relational schema have been removed. Figure 2 (schematic of the relational schema) has been deleted as it is no longer relevant. The revised Section 3 now describes the flat CSV structure (64 columns for U-Pb, 33 for Lu-Hf) and emphasises that each row represents a single analysis with no relational infrastructure required.

(3) Section 2.3: The reference to "MySQL scripts" as part of the data distribution workflow has been removed. The text now states that Python scripts perform cleaning and standardisation, with outputs formatted exclusively as flat CSV files.

(4) Section 5.3: The paragraph discussing the development of an “automated processing platform” and a “user-friendly graphical interface” has been deleted entirely. The section now focuses on data-quality control lessons and the role of LLM-driven extraction in large-scale compilation workflows.

(5) Data availability: All references to https://onedz.top/ have been removed. The statement now describes only the Zenodo CSV release and the GitHub repositories for data-cleaning code.

(6) Supplement Section 2: The title has been changed from “Python and MySQL code snippets in data process” to “Python and SQL code snippets for data cleaning”. The text now clarifies that SQL

queries were used solely as an internal utility for temporary deduplication during the cleaning workflow, and that all final published outputs are flat CSV files. The Figure S4 caption has been revised to state that Navicat was used only as an auxiliary inspection tool, not as part of the data distribution system.

Comment 3: work in a functional data portal and don't resubmit the manuscript before having solved your bandwidth and performance issues.

Reply: Thank you for this clear guidance. We have chosen not to pursue the portal-development path. As a PhD student in sedimentology without professional software-engineering support, we lack the resources to bring the web portal to the functional standard required for peer-reviewed data publication within the revision timeframe. Consequently, we have eliminated all references to the OneDZ web platform from the manuscript, the Data availability section, and the Supplement. Data access is now provided exclusively through the Zenodo CSV release. The website (onedz.top) is no longer mentioned anywhere in the paper or its supplementary materials as a data source.

Other response:

We have also updated the dataset contents. During the past year, community contributors have submitted additional PDF files, from which the LLM-driven extraction pipeline has generated new records. These newly extracted records have been incorporated into the Total_UPb_split_parts/ folder. However, as automated extraction can introduce errors, we clearly distinguish between the full compilation (Total_UPb_split_parts/, containing all records including unchecked LLM-extracted data) and the quality-controlled subset (Experts_checked_UPb_split_parts/, containing only records verified by domain experts). The Lu-Hf dataset (Total_LuHf_split_parts/) has been fully expert-checked and is provided as a separate, curated release. This dual-track structure allows users to select the appropriate data product for their research needs.