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 are deeply appreciative of the detailed and constructive feedback provided by the reviewers for our manuscript titled "OneDZ: A Global Detrital Zircon Database and Implications for Constructing Giant Geoscience Database." Your insights have been invaluable in guiding us to enhance both the functionality of our database and the clarity of our manuscript. We have meticulously reviewed each comment and have implemented comprehensive revisions to address the concerns raised.

Specifically, we have made substantial progress in improving the OneDZ user interface. The initial issues with the web platform, including the 404 errors and inactive links, have been resolved through a server migration to a more robust infrastructure. The new web platform is now accessible at https://www.onedz.top/, and we have enriched its functionalities, including the development of a contribution data module that is currently under testing. Additionally, we have addressed the certificate issues that were previously blocking HTTP API requests, ensuring smoother data downloads and interactions. We have also taken steps to improve the international accessibility of our database by incorporating automated translation tools and recommending open-source software like DBeaver for database interactions.

In the manuscript, we have made several critical revisions to enhance its accuracy and completeness. We have expanded the introduction to include references to similar databases such as EarthBank and Geochron, acknowledging their contributions and differentiating our work. We have also revised the section on discordance ratio calculation to provide a clearer methodology based on Andersen et al. (2019).

Furthermore, we have added a comprehensive Entity-Relationship (ER) diagram to illustrate the database schema and have detailed the resampling methods used in our study, making the code available on GitHub for transparency and reproducibility.

We are committed to ongoing improvements and are actively working on further enhancements to the OneDZ database. We believe that the revisions we have made address the reviewers' concerns comprehensively and will significantly benefit the scientific community. We are grateful for the opportunity to improve our work and look forward to any additional suggestions that may help us achieve our goal of creating a dynamic and accessible global detrital zircon database.

Discussion #1

User-serving components

1. There seems to be some missing functionality on the OneDZ user interface, e.g. https://dedc.geoscience.cn/onedz/HomePage.html returns a 404 error and the other two menu links are inactive.

Response: Thanks for carefully checking the web platform. In fact, after the publication of this preprint, the web platform made great attention and the page view increased rapidly. The overwhelming traffic congestion on the server was beyond our expectations. Our original server was deployed in the data center of the Chinese Academy of Geological Sciences. Due to access restrictions imposed by the data center, the website frequently encounters 404 errors. Currently, we have completed the migration of the website. The new web platform is https://www.onedz.top/. After preliminary testing, there has been a certain improvement in network quality. Meanwhile, the website's functions have been enriched to some extent. Currently, the contribution data module has been developed and is undergoing testing (expected to be deployed starting from July 25th, with specific progress available at https://www.onedz.top/news.html). Thank you again for your feedback, which has sparked new thinking on how to better utilize web platform services for scientific research. We acknowledge that with our geological background, it is difficult to run the web platform well enough in a short period of time, but currently we have successfully

attracted professional developers to participate in the development of the web platform (see details in https://www.onedz.top/about.html). A platform exclusive email has been established to better promote platform construction and strive to build OneDZ into a dynamic and scalable scientific database.

2. When performing a coordinate search on the OneDZ user interface there seems to be a certificate issue blocking HTTP API requests over the HTTPS domain and preventing data download. This should be addressed and part of regular maintenance if the database frontend is intended as a community resource.

Response: Thank you for your suggestion. In fact, we have received feedback on the interception issue after the preprint was sent out. At present, this work has been carried out by two Huawei engineers. We first rented Huawei Cloud servers again, got rid of the access restrictions on the internal network, and successfully completed the server migration. Subsequently, we applied for a website domain name (www.onedz.top) and SSL certificate, and completed the configuration. Under the extensive tests in Chinese Mainland and hundreds of limited tests in the UK (due to the limitations of partners, we were unable to carry out large-scale tests worldwide), the HTTP security problem has been preliminarily solved. At the same time, we have found through testing that some of the download blocking issues may also be due to logical problems with the JavaScript script code in our front-end and back-end. We have also completed the necessary modifications (see details at https://www.onedz.top/news.html).

3. The table data extraction tool in DeepShovel seems to have better accuracy than most commercially available OCR products.

Response: Thank you for your attention and support to DeepShovel. As both we and DeepShovel are from the DDE project, we are developing a lightweight online inference script for the DeepShovel model.

4. Navicat is a commercial software, if the intention is to "enhance user-friendliness", while providing an interface that is accessible consider using an open source software

(e.g. DBeaver).

Response: Thank you for your suggestion. We do not have a fixed recommendation for a database software to provide a visual interface. But in order to facilitate the use of the database by more users, we have added recommendation of DBeaver and MongoDB in the attachment files.

Manuscript comments

1. There is no reference throughout the paper of other existing databases such as EarthBank (https://ausgeochem.auscope.org.au/map) or Geochron (https://www.geochron.org/geochronsearch.php), which provide a similar product, with more user-friendly interfaces.

Response: Thanks for carefully checking the web platform. We first thoroughly researched databases with similar functions. Subsequently, after experiencing the use of the database, relevant descriptions were added in introduction lines 84-86.

2. The authors stated that "Although almost no previous research summarized the difficulties in collecting data sources". There is an extensive body of literature on this topic, here is just a recent example https://doi.org/10.3390/rs16091484.

Response: Thank you for providing the reference. After careful reading, we have summarized the current difficulties in collecting unstructured data in Earth science by combining literature, including: non-repeatability, uncertainty, multi-dimensionality, computational complexity, and frequent updates. We have made modifications to the content in the introduction, specifically in lines 107-112.

3. The authors mention "To ensure accessibility and inclusivity, Chinese-language papers on detrital zircons have been meticulously translated into English." This is a major effort that is highly welcomed by the international community. To further ensure accessibility and inclusiveness of data access, consider translating the menus and buttons of the user interface as well.

Response: Thank you for your recognition of our international cooperation in the entire

DDE project. In fact, all of these tasks were manually completed by over 20 master's and doctoral students in the entire project team over the past five years. The motivation for this work is to consider that every year Chinese researchers produce a large amount of data on detrital zircons in literature published in Chinese, but due to language limitations, many data are difficult for international researchers to effectively utilize. Therefore, the initial stage of the DDE project was completed through large-scale manual translation. However, relying on manual labor is never a long-term solution. For example, more than 20 students have already graduated, and with the shift in project focus, it is difficult to ensure that a large amount of manual labor will be invested in the translation industry. Therefore, we have turned to large language models. Currently, we have collaborated with Zhijiang Laboratory to build an automated translator based on the open-source big language model GeoGPT. Not only can database information be translated, but also the graphic and textual information of Chinese documents can be translated, further increasing the accessibility of data. At present, the author of this article and more personnel are involved in this larger data annotation and manual reinforcement fine-tuning work. But we must also acknowledge that this is a daunting challenge. The first challenge is that few Chinese journals currently adopt open access publishing methods, so fine-tuning GeoGPT may pose copyright risks. At present, the project team is collaborating and discussing with the publishing house to obtain the opportunity to use Chinese literature on detrital zircons. At the same time, the project team is also prepared to rely entirely on manual annotation to fine tune the GeoGPT model in situations where obtaining usage licenses is difficult. The project team will divide future development work into two parts. In the short term, we will set up the GeoGPT API on the web to enable users to add buttons to display the original text (Chinese or other non English information) while displaying the translated version. The long-term goal is to link to the GeoGPT server of Zhijiang Laboratory after clicking the button and enable complete full-text translation services. Thank you again for your suggestion, which has provided effective guidance for the long-term development of the entire project. In the current updated version of the database, we have made every effort to correct potential language issues by combining manual efforts with large

language models.

4. Regarding the documented "spatial skew" of the OneDZ dataset – a side-by-side comparison of OneDZ sample distribution maps with AusGeochem/EarthBank (a global compilation that began as a nationally focused effort) reveals that curatorial priorities also contribute to regional data availability.

Response: Thank you for highlighting the spatial skew in the OneDZ dataset. We recognize that the sample distribution in OneDZ does show regional biases, which are influenced by curatorial priorities and the historical focus of data collection efforts. While OneDZ is a comprehensive global dataset, its spatial skew is evident when compared to other compilations like AusGeochem/EarthBank. To address this, we are actively expanding our data collection to include more samples from underrepresented regions and implementing quality control measures to standardize data across all regions. The "spatial skew" is an unavoidable issue for current databases. We think only more data be contributed can solve this issue. To expanding data, a contributed module has been developed. Also, in some databases with nationally focused effort like *AusGeochem/EarthBank*, some data were not from public paper. The global and regional databases should be coordinated used for accurate researches.

5. How is the discordance ratio defined in the database and was it calculated for all papers in the same way? See https://doi.org/10.1016/j.earscirev.2019.102899 for discussion.

Response: Thank you for your question on Discordance. In fact, we first report based on the results of the original text. If the original text records discord, then this result will be recorded in the database. The reason for doing so is that a considerable number of articles do not provide complete isotopic ratio information for $^{206}Pb/^{238}U$ and $^{207}Pb/^{235}U$ in the provided data. Therefore, we cannot verify the information of discord calculation for this type of data again. Meanwhile, if there are samples with information of $^{06}Pb/^{238}U$ and $^{207}Pb/^{235}U$ in the original text, the calculation of discord refers to the calculation method of Andersen et al. (2019). Meanwhile, in the website contribution

data module, the calculation of Discordance also adopts the calculation method proposed by Andersen et al. (2019). Although not mandatory, we strongly recommend that users add complete isotope ratio information for $^{206}Pb/^{238}U$ and $^{207}Pb/^{235}U$ when contributing data.

6. Since this contribution is focused on an SQL database, the most useful figure would be a database schema with tables and keys and relationships noted

Response: We are grateful for your insightful comments and suggestions regarding the representation of our database schema. In response to your recommendation, we have incorporated a comprehensive Entity-Relationship (ER) diagram of the OneDZ database as Figure 2 in our manuscript. This diagram delineates the interconnections among the core tables, including main, age, geography and U-Pb/Lu-Hf geochemistry. All tables are linked via the primary key zircon_id. We believe this visual representation will significantly enhance the clarity and understanding of our database structure. Furthermore, we have expanded our description of the database schema in lines 160-164 to provide a detailed account of each table's purpose, field names, data types, and their relational integrity. This additional exposition is intended to offer a more thorough comprehension of how the data is organized and can be effectively queried and analyzed. We appreciate your guidance and are confident that these enhancements will greatly benefit our readers and the broader scientific community interested in detrital zircon data.

7. "Class-2 and Class-3 types provide a more nuanced classification based on grain size" - Class-2 seems to provide a classification based on lithology (conglomerate, sandstone, mudstone, etc.).

Response: Thank you for your insightful comment regarding the classification system used in the OneDZ database. We appreciate your attention to detail. Firstly, we followed Puetz et al. (2024) three classes method. Then we have carefully reviewed the classification system and have made the necessary clarifications in the manuscript. In

the revised manuscript, we have updated the description of the classification system to better reflect its structure and purpose. Specifically, we have clarified that Class-2 provides a more detailed classification based on lithology, including conglomerate, sandstone, mudstone, and other rock types. This classification serves as a supplement to the broader categories of Class-1, which includes clastic, meta-clastic, and pyroclastic rocks. Class-3, on the other hand, provides an even more detailed classification based on grain size, following the particle size classification scheme proposed by Udden (1918), Wentworth (1922), and Krumbein (1938). These clarifications have been made in lines 230-236 of the manuscript, where we now explicitly state that Class-2 and Class-3 types provide a more nuanced classification based on both lithology and grain size, respectively. This ensures that users of the OneDZ database can better understand the classification system and its application in sedimentary provenance studies.

8. Please make the code you used for the two resampling methods and SMOTE available in the Github, supplements, and mentioned around rows 255 and 395 respectively in the preprint.

Response: Thank you for your careful inspection. We have updated the resampling method code about time and space in the GitHub repository. There are a total of four scripts, namely temporal_mc_resampler.py, temporal_bootstrap_mc.py, spatial_smote_mc.py, spatial_grid_smote_mc.py. Four scripts are refactored based on the parallel computing characteristics of supercomputing platforms. We are currently applying to expand graphics card resources on the local server in order to deploy fast parallel computing on the web side.

9. The term "Paleo globality" is not frequently used in Earth Sciences. Consider rewording to paleo reconstruction of spatial distribution (or equivalent) to avoid reader confusion.

Response: Thank you for your advice on word choice professionalism. In order to facilitate readers' understanding of the content we want to express, we have replaced

"Paleo globalization" with "Paleo spatial reconstruction". We want to emphasize the distribution of data at the paleogeographic scale. All changes have been highlighted.

10. "Therefore, the evaluation results based on OneDZ, the world's largest detrital zircon database, indicate that the global scope of zircon big data research needs further assessment." It would be useful to postulate what types of assessment you are implying e.g. which current day areas require more sampling. Comparisons with other databases seem useful as well.

Response: We thank the reviewer for this constructive suggestion. Following the recommendation, we have added a new supplementary figure (Fig. S10) that compares the spatial coverage of OneDZ with four widely used databases, including Wu et al. (2024), Puetz et al. (2024), GeoRoc and EarthBank (at 1° × 1° resolution). The kernel-density maps confirm systematic under-sampling of continental interiors (Amazon, Congo and Siberian cratons) and high-latitude regions (>60° N and S) across all compilations. In the revised manuscript we have expanded the relevant sentence (lines 343-359) to articulate these specific gaps and to propose targeted infill sampling and legacy-data integration as the next steps for global detrital-zircon big-data research.

11. "The impact of data sparsity is controlled by the 2 σ error" While the errors might help with outlier identification, they do not control data sparsity. Consider rewording this sentence.

Response: Thank you for your insightful comments on our manuscript. We appreciate your suggestions and have carefully revised the relevant section to address your concerns. We have reworded the sentence to more accurately reflect the role of the 2σ error in our analysis. Specifically, we have revised the sentence as follows in lines 366-368:

"Firstly, we selected the best age data from zircon U-Pb data for time resampling experiments. In addition to comparing Bootstrap and Monte Carlo resampling methods, we assessed the impact of data sparsity using the 2σ error to identify potential outliers

and quantify the uncertainty."

This revision clarifies that the 2σ error is used to identify potential outliers and quantify the uncertainty associated with our data, rather than controlling data sparsity. We believe this change better aligns with the scientific intent of our study and addresses the potential confusion that you highlighted. We appreciate your careful review and feedback. We are confident that these revisions will improve the clarity and accuracy of our manuscript.

Technical Corrections

1. The organization of the Zenodo archival dataset is confusing. The first version of the dataset contains SQL files without any description. The SQL files are then referenced as strongly recommended for use in the description of version v2 but are not present in the file list. To improve findability of key files SQL files should be added to v2, or at least a note clarifying that the SQL files should be downloaded from v1. The warnings in notes 1-3 while pertinent, are not very specific to this dataset. Since there are known and systematic errors, they should be specifically documented (e.g. which Chinese, Latin and Arabic characters have not been converted correctly) and/or fixed, either with excel macros or AI cleaning. Documenting the cleaning process of the transformed dataset would result in an important contribution for the community at large and improving LLMs that also struggle with these types of data transformations.

Response: Thank you for your valuable feedback. I have incorporated the necessary changes to address your concerns regarding the encoding issues within our dataset. We have leveraged GeoGPT to automatically rectify the dataset, focusing on common encoding artefacts such as Chinese, Latin, and Arabic characters that were mis-rendered during prior UTF-8/ANSI conversions. Due to resource constraints and the departure of several graduate students who originally compiled the database, we were unable to manually catalogue every single encoding artefact. However, we have ensured that the dataset is now provided in English, and all records have been processed to minimize the occurrence of garbled text. We understand that due to varying local encoding settings across different PCs, there may still be instances of text rendering issues. To

mitigate this, both the v1 SQL dump and all CSV/XLSX files in v2 are encoded in UTF-8. We strongly recommend that users ensure their client is configured to UTF-8 before importing the files. In addition, our team is actively developing and testing GeoGPT-based repair and translation modules, and we are confident that a future web release will provide a more definitive solution to these encoding challenges. We hope these updates meet your expectations and enhance the usability of our dataset.

2. Some of the Github python scripts contain the same header block which states "This module is mainly designed to remove duplicate samples", even for modules that have over functions e.g. latitude and longitude estimation. Accurate code documentation is essential for reusability.

Response: Thank you for bringing this to my attention. I completely agree that accurate code documentation is crucial for reusability. I have recently gone through all the Python scripts on our GitHub repository and updated them. The header blocks have been revised to accurately reflect the specific functions of each module, whether it's removing duplicate samples, estimating latitude and longitude, or any other functionality. This should make it much clearer for anyone who wants to reuse the code.

Discussion #2

User-serving components

1. Navigating to the https://www.onedz.top/ downloads tab, I am unable to get the search and download buttons to function. It is unclear to me how to enter the longitude (with an east or west or should this be done somehow with a negative and positive number?). Nothing appears to happen when I press the 'search' button and then I receive an "下载失败：Failed to fetch" error message when I press the 'download' button.

I unfortunately do not have the skill set to check the SQL files.

Looking through some of the .csv files there appears to be an issue with the age and uncertainty columns (e.g., Published 206Pb/238U age (Ma) Published 206Pb/238U 1œÉ uncert. Published 206Pb/238U 2œÉ uncert.). The listed age is the same as the 1 sigma uncertainty and the 2 sigma uncertainty is the number from the preceding two cells doubled (e.g.,1769 1769 3538). The 'best age' and 'best age uncert.' Appear to be correct.

The corrections made already to the Zenodo files following reviewer one's comments improve the understanding and thus accessibility of the data that can be downloaded there, thank you.

Response: Thank you for your positive feedback. We did not anticipate that our modest attempts would attract such significant attention. However, we must acknowledge that, as sedimentologists without a background in computer science or databases, the challenges we encountered during our learning and development process far exceeded our initial expectations. This has led us to continually push the boundaries of our knowledge and explore how earth science researchers can rapidly construct large-scale databases in the context of the rapid development of computer science, particularly artificial intelligence. Nevertheless, we recognize that there are still many areas for improvement and that our database is far from perfect. We hope to have the opportunity to continuously demonstrate our process of creating new tools and enhancing the quality of our database, providing new references for future research.

We appreciate your careful review. Our current testing suggests that there may be two potential causes: (1) The query range is too broad, resulting in low retrieval efficiency. (2) Under the constraints of network bandwidth, the number of repeated requests is too low. In response to these potential issues, we are optimizing the retrieval methods of MySQL to improve retrieval efficiency. We are also optimizing the backend to accommodate network conditions by increasing the number of requests. Updates on our progress will be posted in the News section.

Regarding the issue of abnormal data, after our inspection, we found that this type of data was caused by meta data. As a data integration database, OneDZ must record the original data information as much as possible. Therefore, we did not process the abnormal data. We hope to expand the database as much as possible. The judgment of academic rigor needs to be determined by the downstream scientific research users themselves.