Response to referee's second review of "The Active Faults of Eurasia Database (AFEAD): Ontology and Design behind the Continental-Scale Dataset" submitted by Egor Zelenin et al.

We are deeply grateful to the reviewer for their careful examination of the manuscript and AFEAD itself. The authors carefully considered the raised issues and regret that the provided justification was considered insufficient. We recognize these reviews as a strong reason for major revisions of the manuscript, and we have made such a revision. Among the main improvements of the manuscript are paragraphs considering other active fault databases, an expanded description of attributes, and a new section on the strategy of database update. In addition to the corrected manuscript and database, we provide the following point-to-point answer (reviewer's comments in *italic, red is the text of the second review,* and blue are answers to it).

Scientific weaknesses

1. *The data collection is based on bibliographical investigations, but most of the bibliographic references are quite outdated. Out of the 657 references (in the Excel file), only 13 are post-2010. Of these 13, three are classified as unpublished information. Of all 657, 55 are classified as unpublished information, most of which are as old as 1996. How reliable can be a piece of information supplied to the authors 25 years ago and never published since then?*

   Indeed, old and unpublished information is the least reliable source. Unfortunately, those cases cannot be considered outdated *sensu stricto* due to the absence of more relevant information. We are grateful that the referee highlighted this topic, but cannot agree that it is a scientific weakness of the database; instead, it displays a bias in active fault studies towards most active or easily accessible fault systems. Referee's concerns on the reliability have already been accounted for in the CONF (level of confidence) parameter.

   Unfortunately, the lowest confidence value "D" in the CONF field of the shapefile mixes up both published and unpublished materials. Also, many items classified as CONF=D are dissolved in regions where updated studies are available. So this reviewer confirms the scientific weakness, and the unclear communication to the users confirms the technical weakness.

   We are constantly working on an update of AFEAD. Since the initial submission, data from 19 recent (2013-2021) studies have been collected. However, the update of the database does not imply replacing all the preceding data within the spatial extent of the update. Collisions between AFEAD and new data are resolved via adjustment of CONF values (see section Update Strategy). The separate metric of reliability is required as the age of publication and even the fact of publication does not directly affect the reliability of data. Objects within the extent of updated studies may remain in AFEAD even if absent in these recent studies, and CONF allows us to manage the likelihood of activity. Of the 55 unpublished studies mentioned by the reviewer, 50 are those compiled for the World Map of Major Active Faults (see sections 3.4 Reference List and 4 Source Data). Their spatial location has been published (Trifonov, 1997, 2004), but with fewer attributes than in AFEAD. The World Map of Major Active Faults was a major compilation of field data and continuous mapping, and indeed, some faults have never been studied since then due to their remoteness or low expected hazard.

2. *In the last decade, several active fault databases have been published containing updated information. Below I list some of them (not necessarily exhaustively) that have significant geographical overlap with AFEAD and contain more up-to-date data than AFEAD.*

Provided data will be included in the forthcoming update of the AFEAD v.2022; a portion of data has been already populated after the AFEAD v.2021 release. However, they are not comparable to AFEAD by extent or detail or both.

The authors seem more concerned to reply to this reviewer than to inform their readers and potential database users about their intentions to update AFEAD or about the existence of these more up-to-date datasets.

We have added a paragraph considering recent active fault databases in the Introduction (lines 44-49). Iterative update of AFEAD was announced in the former section Data Access and Further Development and now is expanded in section 6 Update Strategy.

3. *Apart from those compilations released in the last year, most of these have been around for quite a long time now. In addition to this lack of data, the relationship between the fault representation in AFEAD and the fault representation in the source dataset is not clear. This is of particular concern for the blind faults since only criteria associated with the topographic signature are recalled. On the one hand, not considering the latest fault compilations prevents AFEAD from listing the newly recognized active faults. On the other hand, it also prevents AFEAD from eliminating those faults that were once considered active but are currently considered not active based on new evidence. Unfortunately, the CONF parameter does not consider the recency of the information.*

   A workflow of transferring source data to the AFEAD representation is presented in section 4. Source Data. We have expanded this section to clarify the workflow, especially in the cases of contradiction among data sources. There is no direct relation between the recency of the information and its accuracy, so any join of recent data requires a comparison of the reasoning behind older and recent objects. The result of the comparison affects CONF in either its elevation or decrease and even deletion from the database.

   The added explanations sound more like an excuse not to add references to the most recent and likely more accurate works on active faulting than AFEAD. That there can be a time lag between the appearance of a publication and its ingestion into a database is perfectly understandable even without saying. Some of the data products mentioned by this reviewer are over ten years old already, and not only did the authors not consider those data for inclusion in AFEAD, but they also neglected them in their discussion.

   All the provided references have been thoroughly considered and included in the AFEAD, as well as some recent studies not mentioned by the reviewer. We have provided additional information on the workflow of transferring source data to AFEAD in section 6 Update Strategy.

4. *The compilation of the fault parameters also remains rather obscure in several aspects. For example, of the 47,363 faults, 22,270 (47%) have no parameter assigned (field "Parm" is NULL). Of the 25,093 faults with the field "Parm" not NULL, only 6,849 reports a "Rate=" value; how was then the Rate (rank) parameter assigned to the remaining faults?*

   Objects of null "Parm" are typically those collected from fault maps with no parameterization. Please note that RATE=3 means "no measured rate above 1 mm/yr" (see Table 2), so it addresses all those cases.

   This reply does not clarify the issue. Firstly, there are 542 records with "Parm" = NULL and Rate < 3. Secondly, the definition of Rate=3 does not distinguish between "no measures at all" and "measures below 1 mm/yr but above 0 mm/year."

   Populating of derivative attributes, including RATE, has been described in the expanded section 3.3 Derivative Attributes (lines 166-169) and the new section 6 Update Strategy.

Technical weaknesses

5. *The AFEAD is distributed as a single shapefile. Technically speaking, it is not even a database apart from the implicit relation between geographic features and their attributes. No relational table is provided between AFEAD and any of its linked information. In other words, it should be classified as a geographical flat-file, not a proper database.*

   According to Wikipedia, “A database is an organized collection of data, generally stored and accessed electronically from a computer system.” (https://en.wikipedia.org/wiki/Database), and AFEAD satisfies this definition of a database. However, it may not meet the definition of a relation database. Depending on the editor’s decision, we can identify AFEAD as a "dataset" as it affects neither its inner structure nor representation. However, our experience in hosting and distribution of tectonic data shows that user-friendly shapefile format gets better reception among the researchers. Most AFEAD use cases require basic spatial analysis and text search on the user device without DBMS software.

   The authors retained only the first few words of the definition given by Wikipedia (https://en.wikipedia.org/wiki/Database). AFEAD has some linked information in a separate table which is not properly related to the main table.

   Considering the policy of the Earth System Science Data journal, of 16 geology-related databases, not datasets, published in ESSD since 2017, only two are relation databases published in SQL format. The remaining 14 are flat-files, excel tables, or shapefiles, still named “databases”. Judging by these cases as well as our understanding of the broad “database” term, we consider “database” acceptable. However, we are ready to refer to AFEAD as a "dataset," rather than a "database," based on the editor's decision and to make any necessary changes to the manuscript.

   The fields in the shapefile attribute table are very poorly organized. First of all, none of the fields can be identified as a primary key. The lack of a primary key prevents the user from uniquely identifying any records and establishing their possible relations with external information. Also, the user cannot make an explicit reference to an individual AFEAD record when using it, including this review.

   A primary key has been added (field “FID”).

   The FID field does not appear in the linked shapefiles (https://doi.org/10.13140/RG.2.2.10333.74726 last access on 26/02/2022).

   Unique Fault ID has been added explicitly (https://doi.org/10.13140/RG.2.2.32655.05280).

6. *Both the “Auth” and “Parm” fields contain long text strings that, in the next update, could become even longer and easily exceed the limitations imposed by the shapefile format. Notice that the maximum number of characters in a text field of a shapefile is 254, see Attribute limitations in ESRI documentation at: https://desktop.arcgis.com/en/arcmap/latest/manage-data/shapefiles/geoprocessingconsiderations-for-shapefile-output.htm#GUID-A10ADA3B-0988-4AB1-9EBA-AD704F77B4A2 or https://support.esri.com/en/technical-article/000012081*

   Even accounting for shapefile standard limitations, we consider it the best format to distribute among researchers in the field of active faulting. It requires no proprietary software but supports spatial analysis and data queries. Only few objects are close to the maximum string length in AUTH or PARM and this could easily be resolved by removal of outdated or least relevant sources. In the current AFEAD schema, field limitations do not affect data presentation and usability.

   It is not the choice of the shapefile questioned but its use.

   Reasoning behind the questioned approach has been described in both expanded section 3.2 (lines 116-120) and new section 6 Update Strategy.

7. *These two fields are also very difficult to explore, especially the Parm field that contains very heterogeneous parameters. This poor organization makes it hard for the user to use the database. For example, selecting the faults that have a certain "depth" information would require a very complex query, which would discourage the non-experts in SQL and expose the users to uncertain results. Also, the Parm field takes up more bytes than needed by repeating within the field the word to identify the parameter type, such as "Sense=" or "Rate=" or "Depth=", occasionally also including the reference to the parameter itself.*

   Indeed, PARM is designated for ease of reading, not querying. Below, the reviewer proposes to "separate the "Parm" attributes into different columns, paying attention to storing single numerical values in individual columns." A schema of the spatial database of the World Map of Major Active Faults (DB96) was exactly what the reviewer suggest, and we intentionally changed this approach in AFEAD. The suggested schema leaves no room for different estimates of the same parameter and references for these estimations. A defined domain of values will distort citing of data (e.g. single numerical value is required where only value range or upper estimate is known). Finally, well above 90% of such fields will be empty, which hampers visual interaction with data. However, if any parameter, e.g. depth, becomes credible for a large amount of data, it will be recorded to an individual column (say, DEPTH), like it was done for fault sense (fields SENS1, SENS2) and uplifted side (field SIDE).

   The first statement in this reply contradicts the Database definition the authors proposed to adopt by referring to the Wikipedia definition. As for the ease of reading, do the authors think it is easy to scroll up and down a table with over 47 thousand entries for locating those with some parametric characterization? It's a shame, however, to learn the authors already had such a more appropriate database schema and downgraded their work to this confusing and inefficient design of AFEAD. The schema suggested by this reviewer requires only some data manipulation and reorganization that would improve the AFEAD usability. A properly designed database including one-to-many relational tables would solve the issue regarding the multiple interpretations.

   In addition to the previous answer, the current database design is a trade-off between the ease of computer processing and user accessibility. The former "separate attributes" design of DB96, even seeming more appropriate, proved ineffective due to the excessive number of null values and the intention to record different estimates of the same parameter. As for the previous issue, the reasoning behind the questioned approach has been described in both expanded section 3.2 (lines 116-121, 135-138) and the new section 6 Update Strategy.

8. *The use of the "+" (plus) sign in the "Side" field is unnecessary because all the non-null values are a plus. It could also be troublesome because the plus sign can be automatically converted when importing the data in other systems (try saving the attribute table into the Microsoft Excel format, for example).*

   SIDE is a text field, and any DBMS may handle mathematical symbols in text strings. We were unable to reproduce problems when opening .dbf attribute table in MS Excel. In active faults databases, it is common to label a downthrown side as well, so the plus sign serves as a reminder about an uplifted side.

   Open AFEAD shapefile in QGIS, save the layer as CSV, open AFEAD.csv in MS Excel and see all the values in column "E" (SIDE) showing the Excel message "#NAME?" with the content of the cells reading "=+W" since the "+" sign is interpreted as being part of a formula. Simply renaming column SIDE as UPSIDE and getting rid of the "+" would have solved the problem.

   Converted to UPSIDE, "+" removed.

Other issues (omitted are those solved after the first review)

9. *Table 1: Is the strike-slip with unknown sense contemplated?*

   In AFEAD, strike slip with unknown sense is considered equal to unknown sense (SENS1=U).

   The SENS1 definition remains unclear and insufficient. SENS1=V and SENS1=U are not mutually exclusive.

   SENS1 definition expanded (Table 1 and lines 170-173).

Recommendations

10. *The following few technical fixes are necessary to make AFEAD suitable for using it in a proper DBMS.*

    We consider shapefile to be the most suitable data format for the distribution of AFEAD at the moment. The provided guidelines will be essential for a redesign of AFEAD when demand for relation database managed by DBMS software increases.

    The problem raised by this reviewer has nothing to do with the shapefile format.

    Each technical advice is answered below:

    - *Establish a primary key that uniquely identifies each record (fault) of the shapefile.*

    Unique fault ID has been added.

    - *Separate the "Parm" attributes into different columns, paying attention to storing single numerical values in individual columns.*

    Following the described strategy, some parameters are likely to be separated in the future, as it was done for RATE or SENS1. However, it will not affect "PARM" due to concerns described in section 3 Database Model and the new section 6 Update Strategy.

    - *Establish a primary key for the table of bibliographic references.*

    There is a primary key – a "citation" column.

    - *Create a relational table (many to many) that connects the fault table primary keys with the bibliographic reference table primary keys.*

    The relation table is provided.

    - *Once the relational table is created, the column "Auth" can be deleted from the shapefile.*

    Deletion of AUTH will not improve usability and may impede the most common use case of textual analysis of spatial query to AFEAD by an inexperienced user.

    - *Remove all "+" "-" "=" and similar signs/symbols from all columns. Use the "+" or "-" sign only with numerical values.*

    Only leading symbols may affect representation of data in external software, and those have been removed.

11. *The European plate boundary along the Mid-Atlantic Ridge should be completed to make AFEAD adhere to its name (it could be disappointing for the AFEAD user to find data in the African plate and not the complete European plate).*

    Faults in the Mid-Atlantic Ridge will be included in the forthcoming update of the AFEAD v.2022

    Again, the authors should inform their potential readers, not just the reviewer.

Faults in the Mid-Atlantic Ridge have been included. Potential readers have been informed about the spatial domain of the AFEAD.

12. *More explanations are needed to make the user understand the source of information used to assign the Rate ranks.*

    Explanations have been added to the manuscript and AFEAD web map interface.

    What was added is not a sufficient explanation for the AFEAD user to understand the source of information.

    More explanations have been added to the section 3 Database Model, an example of data processing, including rate assignment, have been added to the section 6 Update Strategy

13. *A justification is needed for not considering all the recent fault data compilations published in the last decade. The authors should also discuss the implications due to the lack of updated information and warn the users about the limitations in using AFEAD instead of more up-to-date regional/local data.*

    Explanations have been added to the manuscript and AFEAD web map interface.

    Repeated from above: The added explanations sound more like an excuse not to add references to most recent, and very likely more accurate works on active faulting than AFEAD. That there can be a time lag between the appearance of a publication and its ingestion into a database is perfectly understandable even without saying. Some of the data products mentioned by this reviewer are over ten years old already, and not only the authors did not consider those data for inclusion in AFEAD but they also neglected them in their discussion.

    All the provided references have been included in the AFEAD, as well as other recent papers. We have added a notification about lacking publications to the Introduction according to the reviewer advice (lines 53-54).