Geoscience Frontiers 15 (2024) 101821

Contents lists available at ScienceDirect

Geoscience Frontiers

journal homepage: www.elsevier.com/locate/gsf

Research Paper

The replication crisis and its relevance to Earth Science studies: Case studies and recommendations

Stephen J. Puetz [a,*], Kent C. Condie [b], Kurt Sundell [c], Nick M.W. Roberts [d], Christopher J. Spencer [e], Slah Boulila [f], Qiuming Cheng [g]

[a] 475 Atkinson Dr, Suite 704, Honolulu, HI 96814, USA
[b] New Mexico Institute of Mining and Technology, Socorro, NM 87801, USA
[c] Idaho State University, Pocatello, ID 83209, USA
[d] Geochronology and Tracers Facility, British Geological Survey, Keyworth, Nottingham NG12 5GG, UK
[e] Queen's University, Department of Geological Sciences and Geological Engineering, Kingston, Ontario K7L 3N6, Canada
[f] Sorbonne Université, CNRS, Institut des Sciences de la Terre Paris, Paris, France
[g] State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences, Wuhan 430074, China

ARTICLE INFO

Article history:
Received 31 August 2023
Revised 6 February 2024
Accepted 6 March 2024
Available online 8 March 2024
Handling Editor: R. Damian Nance

Keywords:
Replication crisis
Replicability
Independent data
Global time-series
Globality index
Filtered data

ABSTRACT

Numerous scientific fields are facing a replication crisis, where the results of a study often cannot be replicated when a new study uses independent data. This issue has been particularly emphasized in psychology, health, and medicine, as incorrect results in these fields could have serious consequences, where lives might be at stake. While other fields have also highlighted significant replication problems, the Earth Sciences seem to be an exception. The paucity of Earth Science research aimed at understanding the replication crisis prompted this study. Specifically, this work aims to fill that gap by seeking to replicate geological results involving various types of time-series. We identify and discuss 11 key variables for replicating U-Pb age distributions: independent data, global sampling, proxy data, data quality, disproportionate non-random sampling, stratigraphic bias, potential filtering bias, accuracy and precision, correlating time-series segments, testing assumptions and divergent analytical methods, and analytical transparency. Even while this work primarily focuses on U-Pb age distributions, most of these factors (or variations of them) also apply to other geoscience disciplines. Thus, some of the discussions involve time-series consisting of εHf, δ^{18}O-zircon, ^{14}C, ^{10}Be, marine δ^{13}C, and marine δ^{18}O. We then provide specific recommendations for minimizing adverse effects related to these factors, and in the process enhancing prospects for replicating geological results.

1. Introduction

A "replication crisis" pervades the sciences. Some also call it a "reproducibility crisis". However, the two terms have slightly different meanings. The Committee on Reproducibility and Replicability in Science (2019) defines "reproducibility" as obtaining consistent results using the same input data, which include computational steps, methods, code, and conditions of analysis. The Committee then defines "replicability" as obtaining consistent results across various studies aimed at answering the same scientific question, each of which has unique, independent data using either the same or different methodologies. Here, we primarily focus on problems associated with replicating U-Pb age distributions, while also giving attention to related time-series, such as variation in εHf and δ^{18}O-zircon over time.

Awareness of significant replication problems in medicine, heath, and psychology emerged at least 20 years ago (Redden and Allison, 2003). Shortly thereafter, groundbreaking research showed that most published medical research findings are false (Ioannidis, 2005). Moonesinghe et al. (2007) attributed the lack of replicable results from factors such as publication bias, selection bias, inappropriate population stratification (the mixture of individuals from heterogeneous genetic backgrounds), and questionable claims of statistical significance. Despite these efforts, the early studies had limited impact on solving the replicability crisis, as shown by a survey conducted by *Nature* nearly a decade later (Baker, 2016). Of the 1,576 researchers participating in the survey, more than 70% had tried and failed to reproduce experiments con-

* Corresponding author.
E-mail address: puetz.steve@gmail.com (S.J. Puetz).

https://doi.org/10.1016/j.gsf.2024.101821

ducted by other scientists, and quite amazingly, more than half failed to reproduce their own experiments. More recently, Chang and Li (2022) found that most economic research is not replicable, while Sileshi (2023) attributed replication problems in agricultural research to poor transparency in the analytical method utilized. Wilson (2022) states that the replication crisis has spread throughout all sciences, and wonders if it can be fixed.

In recent years, we have become increasingly aware of replication problems in our geological research. For instance, results from various granitoid and geochemical databases reveal contrasting variation in Earth's secular composition (Keller and Schoene, 2012; Cox et al., 2018; Zerkle, 2018; Hasterok et al., 2019; Johnson et al., 2019; Liu et al., 2019; Tamblyn et al., 2021; Doucet et al., 2022; Condie et al., 2023; Lu et al., 2023). While investigating this further, we were unable to find publications discussing Earth Science replication problems, which prompted the current work.

Replication problems are likely to differ considerably for specific branches of Earth Science. This research focuses on analyzing and discussing problems that we are familiar with. These include global sampling problems, replicating global age distributions of igneous rocks and sediments, and estimating mean values of various time-series compositional properties. Factors that inhibit replicating geological results include various types of sampling biases, data collection biases, inappropriate analytical methods, and insufficient understanding and testing of assumptions linked to the methods. Conversely, we do not address the closely linked problem of hybrid reproducibility-replicability, which occurs when new data are added to an existing database. Nor do we digress into the divergent interpretations from the same or similar results, which reside in the philosophical domain rather than being empirical. Thus, the remainder of this work avoids interpreting results. Instead, we investigate problems linked to empirically replicating geological/geochemical results, specifically to time-series compositional data. This includes problems related to using the same data but failing to replicate results because of different analytical approaches. In these cases, one method might introduce a bias whereas another method minimizes the bias. Thus, appropriate methodology is central to achieving replicability.

2. Factors affecting replication

The overall approach to replicating results begins with meticulously dissecting data from divergent perspectives to maximize insights into the data properties. We identify 11 key factors that influence replication, and then present them in a logical sequence so that requisite factors are discussed first, while factors that depend on the requisites are discussed last. After describing the key factors, then existing methods as recommended for maximizing replication, which primarily focus on global estimates of U-Pb age distributions. Secondary attention is given to other related types of geological time-series, including εHf(t), δ^{18}O-zircon, and geological proxy data. In these instances, tests and illustrations are presented immediately to reinforce the methodological effectiveness, or lack thereof, before moving on to subsequent related factors.

2.1. Independent data

Legitimate replication is based on analyzing and comparing independent data. This is important because investigators often take a previously published global database and add the latest available data to compile an expanded database – working on the premise that the larger database will be more representative of the true global population. Even while the larger database might lead to revisions to existing hypotheses, the new hypothesis cannot be legitimately validated until the hypothesis is retested with independent data. Even though a legitimate test requires independent data, by randomly dividing the database records into two sub-databases, a test can be conducted to determine the degree to which one sub-database replicates the other. In other words, this legitimate replication test is an inappropriate test of the hypothesis because the data are not independent. In addition to developing hypotheses, a primary reason for collecting independent data is to test existing hypotheses, and there is no shortcut for assembling a completely independent database in this endeavor. To legitimately conduct a replication test, not a single record from the first database can be included in the second database (Committee on Reproducibility and Replicability in Science, 2019). A basic tenet of hypothesis testing is that predictions from an empirical or statistical model, generally given in the form of an equation or probability model, can only be validated after repeated successful testing with independent data. That is, the data for conducting any test must not be used in any way to revise parameters in the existing mathematical model, nor the existing model must not be used in any way to revise the data (Aber, 1997; McDowall, 2004; Waters and Craw, 2006; Crisp et al., 2011; Puetz and Condie, 2022). Once a single parameter is revised, a new hypothesis is formed, and then a whole new set of independent data must be re-collected for a new validation test. A successful hypothesis will survive repeated tests with independent data without having to revise parameters. The tenet of independence is often violated in stratigraphic research, where investigators commonly "tune" the ages of sediments from an ocean drill-core to a time-series aligned with Milankovitch cycles, performed prior to conducting a spectral analysis test (Hilgen et al., 2015; Puetz et al., 2016). Such tuning inflates the statistical significance of Milankovitch-related spectral peaks (Vaughan et al., 2011, 2014). Also, when tuning strata to competing Milankovitch models, the divergent time-series generally produce conflicting results regarding temporal resolution and age uncertainties (Laskar et al., 2004; Hinnov, 2013). Although some justification for this approach might be valid, this type of analytical bias, commonly referred to as circular reasoning, is adamantly discouraged (Waters and Craw, 2006) for certain types of hypothesis testing – in this case, testing to see if the "tuned" time-series exhibits Milankovitch cyclicity (Hinnov, 2013). In recognition of these problems, some have proposed solutions for minimizing circular reasoning in Milankovitch-related research (Hinnov, 2013; Meyers, 2015). Yet, circular reasoning persists in many areas of geological research. Importantly, circular reasoning increases the likelihood of replicating results, but doing so as inadmissible evidence by violating the basic tenet of independence between data and the model being tested.

After several years of continually amalgamating U-Pb detrital zircon databases (Puetz, 2018; Puetz et al., 2018, 2021; Puetz and Condie, 2019), the records have now been segregated into three independent global databases (Puetz et al., 2024). As data independence relates to detrital zircon samples, the same rock type, the same geological formation, and the same geographic area are non-factors for obtaining independent data. Thus, if two investigators sample the same geologic site, the results are considered independent. Disproportionate geographic sampling is a separate problem, of which attempts have been undertaken (Stehman and Selkowitz, 2010; Keller and Schoene, 2012; Puetz et al., 2017) and are ongoing to minimize time-series distortions caused by sampling biases. Section 2.5 (Sample size and non-random sampling) discusses non-random sampling in detail.

A violation of the independence criterion only occurs if U-Pb analyses from one database are then included in another database. This might be best explained by considering that each zircon grain from a sedimentary rock is part of a cluster-sample, and each grain has a unique age, being an independent data point. If one research team collects and analyzes 150 grains from the rock and another

team collects and analyzes 200 different grains from the same rock, then all analyses are independent records – being no different than a single team employing cluster-sampling with 350 zircon grains. The independent U-Pb detrital zircon databases are filtered to include only records with the highest concordance (minimal differences among $^{238}U/^{206}Pb$, $^{235}U/^{207}Pb$, and $^{207}Pb/^{206}Pb$ ages). In turn, the accepted records are then segregated by the year the data became publicly available at publication time, which can sometimes be multiple years after the analyses were performed: (a) DB8 contains 296,315 records published from 1992–2017, (b) DB9 contains 310,360 records published from 2018–2020, and (c) DB10 contains 380,532 records published from 2021 to 2023. These three independent global U-Pb detrital zircon databases eliminate the often-tedious task of compiling large databases to conduct replication tests. At the same time, the three databases allow us to investigate possible inherent systemic biases as a function of time. Yet, any hypotheses developed from these databases will still require gathering more independent data for an acceptable post-hypothesis test.

We risk over-emphasizing that new, revised, and undeveloped hypotheses must be continually retested with new, independent data because this critical step is often omitted in scientific research. We already know this from the replication crisis that pervades most scientific disciplines. Replicating results from a new scientific model is incredibly difficult. Much of science is built around the principle of parsimony, also known as *Occam's Razor*. This principle basically states that the simplest theory is the preferred theory. Other things being equal, science prefers (a) models with fewer parameters and (b) models that make more precise estimates, so that they place more restrictions on the range of data structures they will fit (West et al., 2012). In opposition to parsimony, one approach for making data fit a model is to add more parameters to the model. Many modern classification and regression models are highly adaptable, being capable of modelling complex relationships (Kuhn and Johnson, 2013). Each model's adaptability is typically governed by a set of tuning parameters, which can allow each model to pinpoint predictive patterns and structures within the data. However, these tuning parameters often identify predictive patterns that are not replicable. This is known as over-fitting. An over-fit model generally has excellent predictivity for the samples from which they are built, but the same model has poor predictivity for new samples (Kuhn and Johnson, 2013). For these reasons, knowledgeable researchers generally insist on repeated tests with independent data, with multiple independent research teams conducting the successful tests using an identical, unrevised hypothesis – before finally accepting the hypothesis as valid and legitimate.

2.2. Database globality

Investigators often claim that a database is global, but this description is subjective, without a clear definition of its meaning. To explore the concept of globality, we calculate a "globality index" with five levels of spatial resolution from which global sampling densities are obtained from grid systems with surface areas approximately the size of equatorial trapezoids of 18° × 18°, 12° × 12° (Fig. 1, red), 9° × 9°, 6° × 6°, 4° × 4° (Fig. 1, black), illustrated with grid-centers. The "globality index" is calculated from the percentage of all grids having at least one sample. Various versions of the globality index include: 1) only continents and submerged continental shelves, 2) only the oceans, and 3) the entire globe.

The databases of Puetz et al. (2024) include GPS coordinates recorded in decimal degrees. Published coordinates are increasingly reported as decimals. This format has significant advantages over alternatives such as Degrees-Minutes-Seconds (DMS), Universal Transverse Mercator (UTM), and other arcane formats – of which simplified recording and straight-forward calculations are primary advantages. A database's decimal GPS coordinates are the inputs for calculating our five globality indices, which are the percentages (n/N) of all trapezoidal grids (N) having some samples (n). Thus, globality ranges from 0% to 100% (Data, file 1) and depends on the scale of the grid system being used. Each grid system is defined in terms of three surface areas: continents and continental shelves, oceanic areas, and global total. Quantifying the extensiveness of global sampling is critical because geological and geochemical properties often have unique, heterogeneous populations. Our tentative assessment is that the likelihood of results from one database being replicated by results from another independent database increases with global coverage. Based on limited experiments, a preliminary classification system is defined as: low globality (<30%), moderate globality (30% to70%), and high globality (>70%). A more rigorous classification system will require randomly selecting thousands of subsets of the samples in DB8, DB9, and DB10, constructing time-series from the subsets, finding the correlation coefficients between the time-series, and then producing a table of globality versus expected correlations. While being time-consuming, such as project will ultimately make the globality index far more meaningful.

The five grid systems are constructed because assessments of globality are scale dependent. The high-resolution measurement (4° × 4° grid system) is preferable to the low-resolution measurement (18° × 18° grid system). However, a trade-off exists between high globality and high resolution. Finding an appropriate grid system requires using the same database at all five scales. Globality naturally deteriorates as resolution becomes finer. This is because the 4° × 4° grid system has 2602 grids, whereas the 18° × 18° grid system only has 130 grids. Filling 70%+ of the large 18° × 18° grids is far easier than filling 70%+ of the small 4° × 4° grids. For example, a recent $\delta^{18}O$ database (Puetz et al., 2024) has the following globality measurements for the five grid systems: 87.9% (18°×18°), 74.3% (12° × 12°), 65.4% (9° × 9°), 43.0% (6° × 6°), and 29.0% (4° × 4°). Supplementary Data file 1 contains the detailed calculations for populating the five grid systems, and Supplementary Data file 2 contains details for calculating grid surface areas. A high globality index is preferable to a low globality index when the goal is to study global variation. In this instance, the three highest resolution estimates of globality should be rejected (<70% globality), whereas the 12° × 12° grid estimate is acceptable because it meets the minimum criterion of >70% globality, but the 18° × 18° grid estimate is preferable because it exceeds 80% globality. Thus, the five grid systems provide means for finding the highest resolution globality index that exceeds either the 70% minimum or the 80% preferred threshold.

Although more work is needed, we strive toward compiling large global databases with high globality indices. Using the 12° × 12° globality index for continents and continental shelves, the U-Pb detrital zircon databases DB8, DB9, and DB10 have globality indices of 94.7%, 85.7%, and 82.2%, respectively. In other words, for database DB8, 94.7% of the samples populate grids (Fig. 1, red) associated with continents and oceanic shelves, and likewise for DB9 (85.7%) and DB10 (82.2%). Consequently, assuming other unknown factors do not introduce significant biases, time-series from these databases should be highly replicable. For comparison, the ~76,000 record database of the Sedimentary Geochemistry and Paleoenvironments Project (Farrell et al., 2021) has a 12° × 12° globality index of 68.5%.

2.3. Proxy data

When testing hypotheses and attempting to replicate results, using proxy data can sometimes alleviate the need to compile

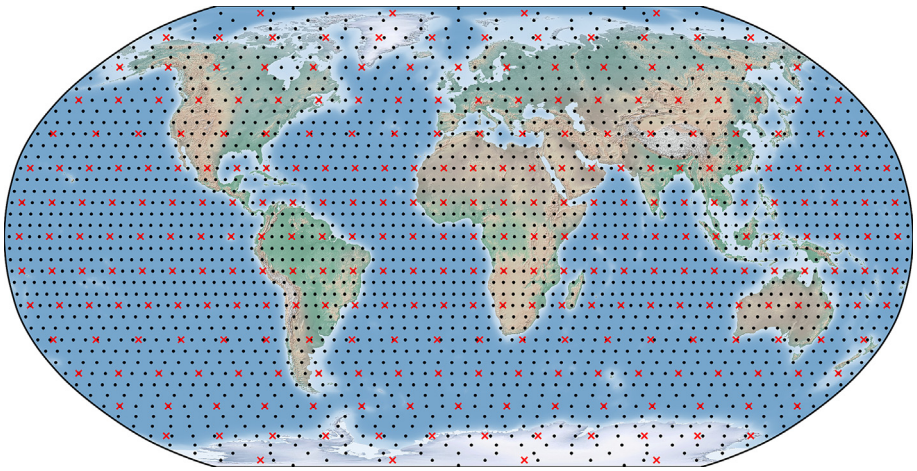

Fig. 1. Robinson projection of 4°×4° grid centers (black dots) and 12°×12° grid centers (red x's). The grids serve two purposes: (1) estimating a globality index from the samples in a database, and (2) weighting database records inversely proportional to sampling densities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

direct data independently. Proxy data are indirect measurements of a process that should serve as a reliable substitute for more direct measurements. The proxy data must be strongly correlated (either positively or negatively) with the primary direct measurements. Numerous types of proxy data are commonly used in various disciplines. In Earth Science, isotope ratios, elemental abundances, and elemental ratios are commonly used as compositional proxies for measurements that cannot easily be obtained from direct measurements, such as thickness of the continental crust, or for measurements in the past that have no means of direct measurement, such as past ocean temperatures. For example, direct measurements of solar activity from 1749 to present are recorded as sunspot numbers from WDC-SILSO, Royal Observatory of Belgium in Brussels. Due to the limited temporal range of reliable sunspot data, researchers sometimes rely on cosmogenic ^{14}C from marine environments (which extends dating by ~50-kyr) and ^{10}Be from polar ice cores (which further extends dating to ~1-Myr) as proxies for variation in solar activity (Usoskin and Kovaltsov, 2012). Similarly, Reimer et al. (2009) use both tree-ring and marine radiocarbon data as proxies for devising a 50-kyr climatic time-series. Time-series of marine carbonate $\delta^{13}C$ and $\delta^{18}O$ are used as carbon-cycle and paleotemperature proxies respectively (Zachos et al., 2001; Jones et al., 2013). Likewise, trace element ratios in detrital zircon are used as proxies for the composition of igneous magmatic activity over time (Barth et al., 2013; Balica et al., 2020). Thus, utilizing proxy data can readily expand upon the possibilities for legitimate independent tests that attempt to replicate results. The tests can involve comparisons between the primary data and proxy data, or alternatively, comparisons between one type of proxy data with another type of proxy data.

2.4. Data quality

Even if independent databases are available with high globality indices, poor data quality can inhibit replication. Poor data quality can occur simply from the omission of key data items, or results might be questionable if one uses data with large uncertainties. Ideally, all key items from published data are available with minimal uncertainties, but that is not always to case. For instance, a database without GPS coordinates eliminates the possibility of using inverse spatial weighting to minimize negative effects from non-random, disproportionate sampling (Stehman and Selkowitz, 2010; Keller and Schoene, 2012; Puetz et al., 2017). For U-Pb data, the method of determining the preferred age (Vermeesch, 2018; Gehrels et al., 2019; Puetz et al., 2024) influences replicability. Specifically, the common practice of choosing between $^{206}Pb/^{238}U$ and $^{207}Pb/^{206}Pb$ ages at a specified cutoff age produces an age distribution with an artificial depression at the cutoff age (Puetz et al., 2021; Puetz and Spencer, 2023). The artificial depression from this common method inhibits replication. For this reason, we advocate the use of Non-Iterative Probability (NIPr) ages as the preferred U-Pb ages (Puetz and Spencer, 2023). Closely associated with data quality, Section 2.8 (Accuracy and precision) expands upon replicating U-Pb age distribution based on reported 2σ precision errors.

2.5. Sample size and non-random sampling

Equally crucial, nearly all geological samples are selected non-randomly in both space and time. The established method is geologically informed sampling, which depends on project objectives. The research team will either seek appropriate archived samples or find a convenient location where specific rocks or sediments are abundant or outcropped. Moreover, sampling tends to be most abundant in wealthier countries and least abundant in developing countries. Understandably, dense tropical forests, deserts, and polar regions tend to be sparsely sampled. For these reasons, the regional densities of samples from large global databases vary considerably (Fig. 1), which differs noticeably from a quasi-uniform distribution expected from large set of randomly generated GPS coordinates.

Additionally, the zircon grains extracted from rocks and sediments are cluster-samples, which serve as subsets of quasi-random samples, but only if the grains are collected via a random process. Moreover, the recommended size of a cluster-sample has steadily increased over time. Dodson et al. (1988) recommended obtaining at least 60 grains from a sample. Later, Vermeesch (2004) recommended 117 grains. More recently, the University of Arizona LaserChron lab (Dobbs et al., 2022; Kushner et al., 2022; Wahbi et al., 2023) and the University of Houston lab (Smith et al., 2023) often obtain 300 + zircon grains in provenance studies. And the University of Calgary CPPATT Lab recently extracted 600 zircon clusters from 9 samples from the Magallanes Basin (Bartelt, 2022). Thus, as preferences for the size of zircon cluster samples continues to increase, regional sampling densities are affected. Regardless of the source of the sampling biases, being either disproportionate convenience sampling or inconsistent cluster-sample sizes, means exist for simulating global random sampling from these non-random samples.

Various methods have been proposed for simulating random sampling by weighting records inversely proportional to sampling densities (Stehman and Selkowitz, 2010; Keller and Schoene, 2012; Puetz et al., 2017). Two common techniques exist: (a) if the time-series is a histogram, frequency plot, or age distribution, then inverse spatial weighting applies (Keller and Schoene, 2012; Puetz et al., 2017); or (b) if the time-series consists of mean values, then inverse spatial–temporal weighting applies (Keller and Schoene, 2012). The steps for calculating inverse spatial weights (Supplementary Data file 1 and file 3) are relatively simple: (i) determine the sampling density by counting the GPS coordinates that fall within each grid, (ii) use the sampling density for each grid with its surface area to find the grid-density relative to the global-density, and (iii) assign the weights from the grid-density inverses ($1/\rho$). Inverse spatial–temporal weighting uses identical methods, with the sole exception being that the process is applied to each bin in the time-series (bandwidth) rather than being applied to the entire time-series.

To test the efficiency of these techniques, we construct four databases from three synthetic samples, each with 50 records from three locations. To simulate disproportionate sampling, several steps are involved. First, the data are restricted to three grids in a hypothetical globe limited to the same three grids. This simplified scenario easily illustrates the concepts (Fig. 2). Then, to test the validity of using inverse spatial weighting for enhancing replicability, four synthetic databases are constructed from three samples by exactly duplicating all 50 records in each sample-set in the following proportions: SynDB1 (blue) 3:1:6 with 100% globality; SynDB2 (red) 1:5:2 with 100% globality; SynDB3 (brown) 5:2:1 with 100% globality; and SynDB4 (green) 1:0:5 with (67% globality). SynDB4 only includes two of the three samples, which simulates estimating the age distribution and mean values from a database with a globality index of 67%. To enhance transparency, the synthetic database details are available in Supplementary Data file 4, which includes the synthetic data, the four databases constructed from the synthetic data, the 12° × 12° grid system tables, all Excel calculations, and the code for Figs. 2 and 3.

The first set of tests illustrate raw age distributions without considering disproportionate sampling. This unweighted approach produces an age-histogram commonly used in geological research – either a probability density plot (PDP) or a Gaussian kernel smoothed version of a PDP, referred to as kernel density estimation (Vermeesch, 2012). With the unweighted approach, the divergent sampling proportions will yield distinctly unique age distributions (Fig. 2a). Of course, this inhibits replicability. The second set of tests utilize inverse spatial weighting to compensate for the disproportionate sampling (Fig. 2b). When this method is applied, despite the samples being added in different proportions, the weighted age distributions with 100% globality (Fig, 2b, blue, red, and brown) give identical results. Thus, disproportionate sampling densities have no effect on the resultant age distributions after applying inverse spatial weighting with 100% global sampling coverage. However, for the database with a globality index of 67% (Fig. 2b, green) the age distribution resembles the three with 100% globality but falls short of perfect replication. This demonstrates that to optimize replicability: (a) inverse spatial weighting should always be applied, and (b) a globality index of ~70% is a minimum, but with the preferred level ~90% or higher – a level which is desirable but perhaps difficult to achieve in practice.

The third set of tests illustrate mean values over time without considering disproportionate sampling (Fig. 3a). Again, this unweighted approach yields distinctly unique mean values, solely

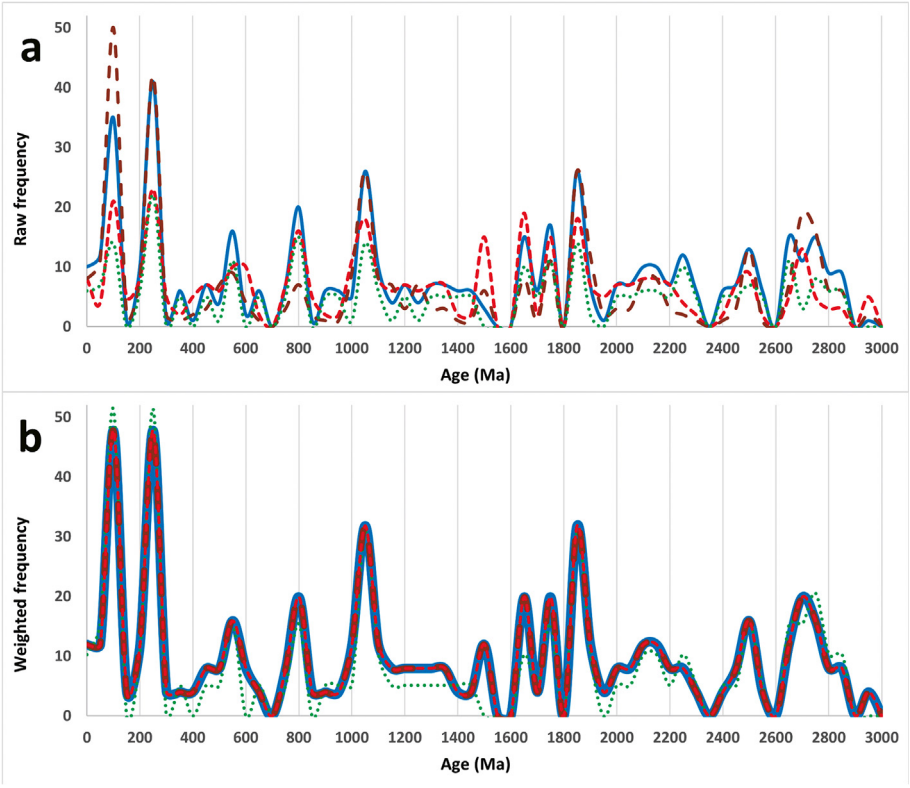

Fig. 2. Raw and weighted age distributions from synthetic databases SynDB1-SynDB4. Color codes: (solid blue) SynDB1 with 3:1:6 sample proportions and 100% globality; (short-dashed red) SynDB2 with 1:5:2 sample proportions and 100% globality; (long-dashed brown) SynDB3 with 5:2:1 sample proportions and 100% globality; and (dotted green) SynDB4 with 1:0:5 sample proportions and 67% globality. Panels: (a) raw age distributions, and (b) age distributions with inverse spatial weighting. Methods are in Supplementary Data file 4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

dependent on variation in sampling densities. The fourth set of tests utilize inverse spatial–temporal weighting (which is a type of weighted average) to compensate for the disproportionate sampling (Fig. 3b). After applying the spatial–temporal weight adjustments, the three mean values with 100% globality (Fig. 3b, blue, red and brown) are again identical. However, for the database with a globality index of 67% (Fig. 3b, green) the mean values fail to replication those with 100% sampling globality.

2.6. Stratigraphic bias

A raw U-Pb age distribution that only includes sedimentary rock samples will naturally contain a stratigraphic bias. The bias exists because the ages for all U-Pb analyses from non-metamorphic detrital zircon grains found to be concordant (where ages match for both the ^{238}U and ^{235}U decay chains) will be older than the stratigraphic age, with no zircon grains younger than the stratigraphic age. If a database contains too many ancient sedimentary rocks and a scarcity of modern sediments and Phanerozoic rocks, then the resulting age distribution might replicate a time-series that only includes rocks with stratigraphic ages > 500 Ma (Fig. 4, blue) – which is atypical of a "global" detrital zircon age distributions because it excludes Phanerozoic ages. Thus, to compensate for stratigraphic bias (Puetz et al., 2021; Puetz and Condie, 2021; Reimink et al., 2021), records are weighted proportionally to the cumulative number of records, binned at fixed intervals, by stratigraphic age. The other curves illustrate the raw detrital zircon age-distribution (Fig. 4, green), the age-distribution from inverse spatial weighting (Fig. 4, red), and the age-distribution from inverse spatial weighting combined with the adjustment for stratigraphic bias (Fig. 4, black). Even while the globality index is poor at 22.9%, the latter time-series (Fig. 4, black) is otherwise optimized for replication. This approach somewhat mimics inverse spatial weighting but weights the time-series based on the cumulative stratigraphic ages – rather than geographic location as done for inverse spatial weighting. Supplementary Data 3 includes a method for adjusting for stratigraphic bias.

The stratigraphic adjustment only applies to sedimentary rock samples and demonstrates that a specific geological methodology might have peculiarities irrelevant to other types of research. Thus, each specialist must seek the relevant biases that are critical to replicating results – which are often biases unrecognized by outsiders.

2.7. Potential filtering biases

If not disclosed, any type of database filtering could introduce a bias that hinders replicability. For example, investigators sometimes intentionally remove records from a database presumed to be of inferior quality or deemed to be outliers. The latter can exist as analytical outliers that may bias the results, or merely natural outliers that exist due to natural heterogeneity in geological processes. Outlier rejection should always be objective, consistent, and openly stated. Which filtering method is chosen depends on the type of study. One would normally expect this type of filtering to enhance replicability. However, the assumption that filtering does not introduce a bias might be incorrect. For instance, in U-Pb geochronology, Gehrels et al. (2019) advise caution when removing discordant U-Pb ages from a time-series – a process that might bias the results. Thus, we tested the possibility that U-Pb discordance filtering introduces a bias by constructing five U-Pb age distributions based on degree of discordance (defined as the smaller of the age differences between $|^{235}U/^{207}Pb-^{207}Pb/^{206}Pb|$

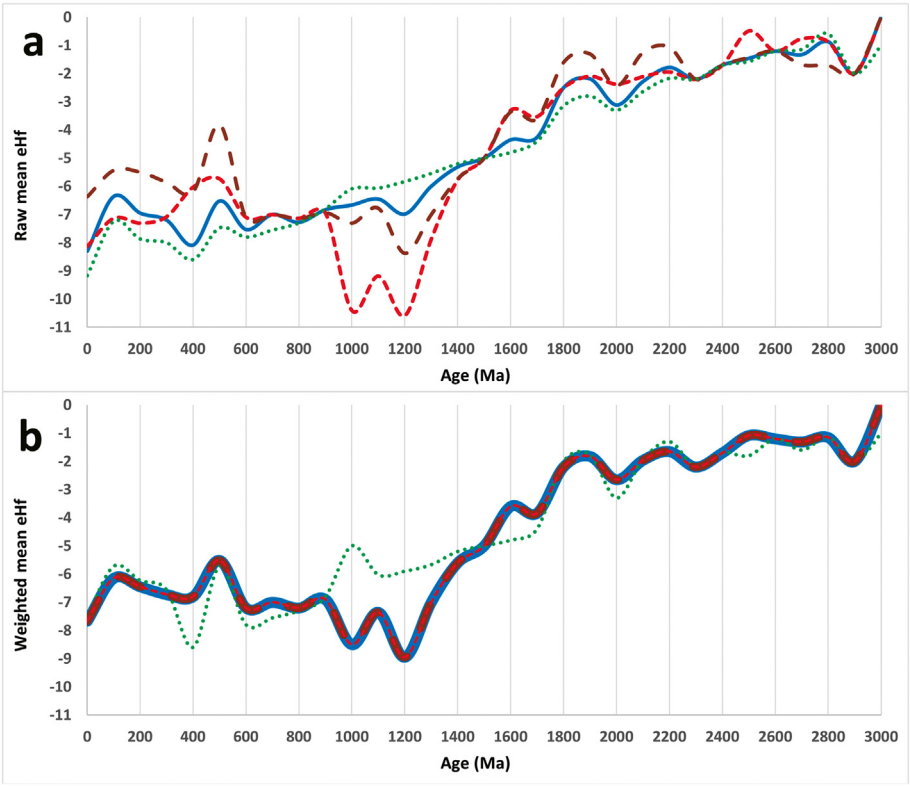

Fig. 3. Raw and weighted mean values from synthetic databases SynDB1-SynDB4. Color codes: (solid blue) SynDB1 with 3:1:6 sample proportions and 100% globality; (short-dashed red) SynDB2 with 1:5:2 sample proportions and 100% globality; (long dashed brown) SynDB3 with 5:2:1 sample proportions and 100% globality; and (dotted green) SynDB4 with 1:0:5 sample proportions and 67% globality. Panels: (a) raw mean values, and (b) mean values with inverse spatial–temporal weighting. Methods are in Supplementary Data file 4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

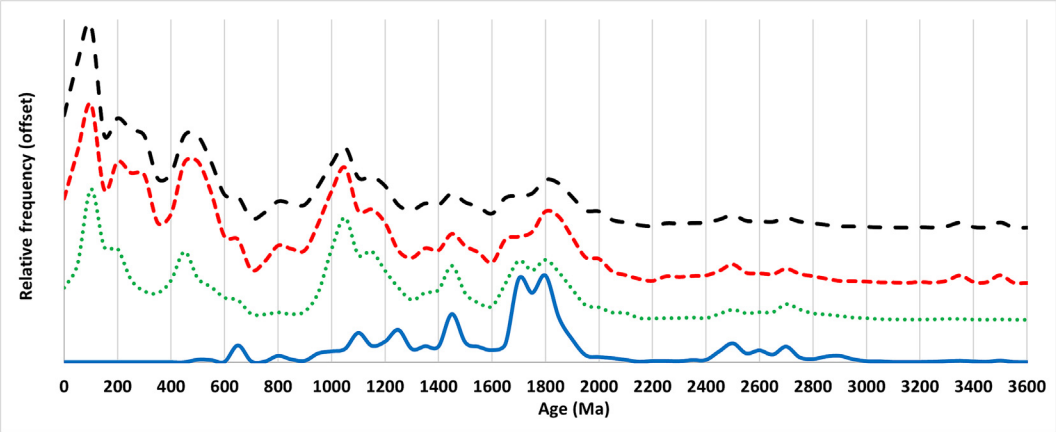

Fig. 4. Four versions of a U-Pb detrital zircon age distribution from a small subset of DB8, with a poor globality index of 22.9%. Color codes: (solid blue) raw time-series for records with stratigraphic ages >500 Ma; (dotted green) raw time-series for all records; (short-dashed red) time-series adjusted by inverse spatial weighting; and (long-dashed black) time-series adjusted by inverse spatial weighting in addition to adjusting for stratigraphic bias. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and $|^{235}U/^{207}Pb - ^{238}U/^{206}Pb|$), with class 1 having the highest concordance and class 5 the lowest (Puetz et al., 2021; Puetz and Spencer, 2023). After constructing the age distributions, each is compared with the other four global U-Pb age distributions to determine if the five produce similar age peaks (Fig. 5). The age peaks become progressively flattened from concordance class 1 (sharp peaks) through class 5 (smoothed peaks), which is due to greater imprecision for classes 4 (Fig. 5, red) and 5 (Fig. 5, dark red). The illustrations also highlight the bias introduced by choosing an arbitrary cutoff-age when the best age is limited to either the $^{238}U/^{206}Pb$ age or the $^{207}Pb/^{206}Pb$ age (Fig. 5a, b). Thus, we use U-Pb preferred ages based on a non-iterative probability method (Fig. 5c) of Puetz and Spencer (2023), which yields age-distributions like the iterative *IsoplotR* approach (Fig. 5d) of Vermeesch (2018). These results indicate that systematic filtering of discordant U-Pb analyses does not significantly bias U-Pb age distributions. Such tests can convert assumptions and guesswork into operational empirical formulations, which in turn enhances replicability. In some instances, methods might not exist for evaluating filtering biases, but when available, such tests should be conducted. Essentially, if the rejected records are of sufficient quantity and quality to compare with the accepted records, then a means might exist for evaluating potential filtering biases.

2.8. Accuracy and precision

When evaluating replicability, one must ponder the question: At what resolution? The answer is highly dependent on the accu-

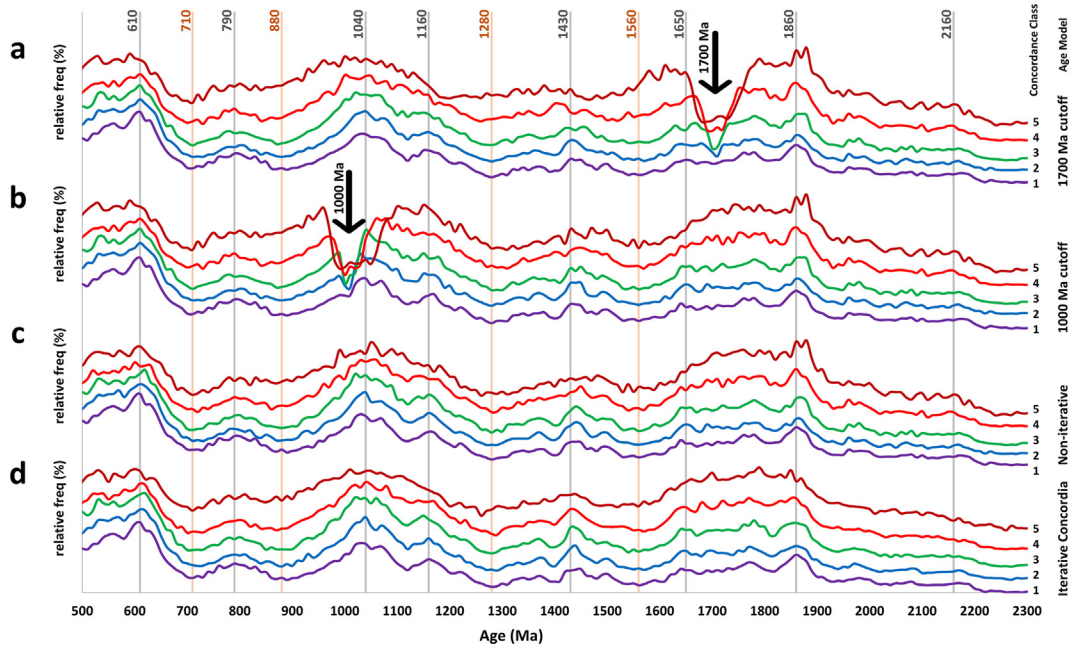

Fig. 5. Stacked U-Pb age distributions (relative frequency probability) with 10-Myr bin-size, from four "best age" models, for concordance classes 1 through 5, as designated along the right column. Global U-Pb detrital zircon data are from Puetz et al. (2021). Color coded concordance classes: (purple) class 1; (blue) class 2; (green) class 3; (red) class 4; and (dark red) class 5. Panels illustrate best age models for: (a) a $^{238}U/^{206}Pb$ versus $^{207}Pb/^{206}Pb$ cut-off age of 1700 Ma; (b) a $^{238}U/^{206}Pb$ versus $^{207}Pb/^{206}Pb$ cut-off age of 1000 Ma; (c) non-iterative probability ages (Puetz and Spencer, 2023); and (d) IsoplotR single grain concordia ages (Vermeesch, 2018). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

racy and precision of the measurements. For instance, a global U-Pb detrital zircon age distribution might be highly replicable when U-Pb ages are binned at 40-Myr intervals, but replicability might be questionable at 1-Myr resolution (Section 2.9). The uncertainties of most geological, geochemical, and isotopic measurements are related to precision rather than accuracy (Schoene et al., 2013), and precision errors are not directly linked to accuracy. As a rule, accuracy remains unknown for geologically related measurements. However, for certain disciplines, accuracy can be determined from independent measurements that are traceable through to scientific units, as is the case for Isotope Dilution Thermal Ionisation Mass Spectrometry (ID-TIMS) U-Pb geochronology (Condon et al., 2015). In addition, it is sometimes possible to devise indirect means for estimating accuracy. For instance, a comparison of age distributions from igneous reference standard materials, using both U-Pb ID-TIMS and U-Pb Laser Ablation Inductively Coupled Mass Spectrometry (LA-ICP-MS) methods, indicates that U-Pb LA-ICP-MS ages are commonly highly accurate, likely being within ±1.5-Myr of the true age (Puetz and Spencer, 2023). If one incorrectly assumes that accuracy is always equivalent to precision, this is far more accurate than one might initially guess, based on the 1%–2% precision errors associated with LA-ICP-MS from previous inter-laboratory tests (Košler et al., 2013) which involved relatively small sample sizes. To partially overcome estimating accuracy from imprecise measurements, we reject analyses for concordance classes >3 (Puetz and Spencer, 2023). The remaining hundreds of thousands of highly concordant U-Pb ages from three independent databases (DB8, DB9, and DB10) are used to construct time-series. Estimating LA-ICP-MS accuracy from age peaks (mode statistics) in these independent time-series requires correlation studies at multiple resolutions.

2.9. Segmented correlations

With the relevant factors identified for developing a replicable time-series, it is possible to take the raw records from the databases and transform them into detrended time-series, with the goal of testing if the time-series are replicable with minimal bias. Or, perhaps more correctly, if multiple, independent, bias-corrected time-series are available, then performing correlation tests among them provides a means for assessing if the time-series are indeed replicable. In addition to correlating the entire independent time-series, a more comprehensive analysis involves correlating segments of the time-series. Segmented correlations are especially important for identifying end-point biases, which occur in many types of nature-based time-series because of difficulties in measuring the beginning and ending points. A paucity of samples and/or measurements with extreme outliers are two common contributors to end-point bias.

A straight correlation between two independent time-series provides one approach for evaluating replicability. After finding the correlation, in which each data point within a time-series is also independent, Student's *t*-test (Owen, 1965) provides a means for assessing statistical significance, as defined by Eq. (1).

$$t = \frac{r/\sqrt{n-2}}{\sqrt{1-r^2}} \tag{1}$$

In Eq. (1), t is the student value (t-statistic) where, r is the sample correlation coefficient required for statistical significance, and n is the sample size. For a time-series with infinite points, the cut-off value t_{95} corresponding to a 95% confidence level, which is 1.645 for a one-sided test and 1.960 for a more conservative two-sided test. For a time-series with finite points, t_{95} values vary dependent on sample size (n), taken from Owen (1965). The differences between the infinite and finite versions are minimal, which is why a t_{95} value of 1.960 is commonly used for a two-sided test involving a finite time-series. Refer to Supplementary Data file 5 for two tables and one figure showing the relationship between sample-size (n) and statistically significant correlation coefficients.

Because the mentioned correlation analysis is based on Pearson's product moment correlation coefficient (r), the time-series must be detrended to form stationary time-series. One way to detrend is to apply a band-pass filter to remove the background trend (i.e., the dominant slope). After detrending, we find correlations among the three independent band passed time-series: DB8, DB9, and DB10, and then compare them with the correlations required for 95% confidence [Eq. (1)]. These threshold correlations provide an initial set of tests for assessing the replicability of U-Pb detrital zircon age distributions. The correlation studies were conducted for the time-series binned at 1, 2, 5, 10, 20, and 40-myr. The extent of each time-series increases with bin-size, which is a consequence of obtaining a sufficiently large number of records per bin to a produce a stationary sequence. Then, each time-series is segregated into three sub-intervals to evaluate replicability as a function of time. Table 1 summarizes the correlation results. The results indicate the extent to which global detrital zircon age distributions are generally replicable (significant for at least 2 of the 3 correlation tests) at the following resolutions: 40-Myr (4240–0 Ma), 20-Myr (2720–0 Ma), 10-Myr (3800–0 Ma), 5-Myr (3400–0 Ma), 2-Myr (1880–0 Ma), 1-Myr (not replicable). Supplementary Data file 6 contains details of Table 1 methods and calculations.

These segmented correlation tests are consistent with the estimated ±1.5-Myr accuracy of U-Pb LA-ICP-MS ages (Puetz and Spencer, 2023). If U-Pb LA-ICP-MS accuracy exceeded ±1.5-Myr, it is inconceivable that the independent DB8, DB9, and DB10 age distributions could be replicated from 1880 to 0 Ma at 2-Myr resolution. Likewise, because the DB8, DB9, and DB10 age distributions fail to produce statistically significant correlations at 1-Myr resolution, U-Pb LA-ICP-MS accuracy is unlikely to be as small as ±1-Myr.

Deciding when to terminate a time-series can be problematic due to a scarcity of samples deeper into time. Likewise, choosing the resolution for binning data can be equally disconcerting because the accuracies of records within a database are often unknown. By using segmented correlation studies with independent time-series, as described in this section with U-Pb detrital zircon time-series, problems with age-related degradation and replicability can be minimized. This understanding should enhance decisions related to the age at which a time-series should be terminated. Supplementary Data file 6 contains the spreadsheet with the correlation calculations for the U-Pb study in this section, which can serve as a template for similar types of time-series studies.

2.10. Testing assumptions and alternative methods

Quite frequently, multiple methods exist for minimizing adverse effects from a known bias. If the methods have equal success attenuating the bias, then selection becomes a personal choice. However, when the success rate is unknown, tests should be conducted to determine the advantages/disadvantages of the competing methods. For example, by employing a grid-based methodology for hypothetical databases with divergent sampling densities, raw age distributions (Fig. 2a) are inconsistent. But when the time-series are adjusted by inverse spatial weighting (Fig. 2b), they are either identical (100% globality) or similar (67% globality). Likewise, considerable inconsistences appear for time-series from raw mean values (Fig. 3a), whereas mean values adjusted by inverse spatial–temporal weighting (Fig. 3b) yield either identical or similar results.

Keller and Schoene (2012) suggest using a similar method for disproportionate sampling densities. While being similar, differences exist. Firstly, Keller and Schoene (2012) do not assess the

Table 1
Correlations among various bandpass filtered detrital zircon time-series.

Variable\Bin-size	40-Myr	20-Myr	10-Myr	5-Myr	2-Myr	1-Myr
Degrees of freedom (*n*-1)	38	68	127	226	470	180
Correlation interval (Ma)	1520-0	1360-0	1270-0	1130-0	940-0	180-0
95% confidence level	0.311	0.235	0.173	0.130	0.090	0.145
DB8 vs DB9 correlation	0.947	0.907	0.799	0.560	0.151	0.098
DB8 vs DB10 correlation	0.954	0.910	0.820	0.616	0.162	0.123
DB9 vs DB10 correlation	0.962	0.927	0.841	0.592	0.117	0.093
Degrees of freedom (*n*-1)	38	68	127	226	470	180
Correlation interval (Ma)	3040-1520	2720-1360	2540-1270	2260-1130	1880-940	360-180
95% confidence level	0.311	0.235	0.173	0.130	0.090	0.145
DB8 vs DB9 correlation	0.859	0.828	0.697	0.440	0.079	-0.106
DB8 vs DB10 correlation	0.924	0.898	0.733	0.463	0.037	0.126
DB9 vs DB10 correlation	0.836	0.803	0.693	0.433	0.176	0.147
Degrees of freedom (*n*-1)	30	68	126	228	460	180
Correlation interval (Ma)	4240-3040	4080-2720	3800-2540	3400-2260	2800-1880	540-360
95% confidence level	0.347	0.235	0.173	0.129	0.091	0.145
DB8 vs DB9 correlation	0.095	-0.015	0.029	0.142	0.096	-0.019
DB8 vs DB10 correlation	0.446	0.456	0.347	0.194	0.069	-0.056
DB9 vs DB10 correlation	0.396	0.143	0.219	0.336	0.033	0.143

Color codes: 95% confidence levels (green) and correlations below the 95% threshold (red).

globality of the samples, as discussed in Section 2.2, but they do provide bivariate kernel density estimates adjusted for sampling densities. Secondly, rather than using grids, they define density from a theoretical perspective, which involves calculations to determine the proximity of a single sample relative to all other samples in the database, as described in the appendix of Keller and Schoene (2012). From these, weights are assigned that are inversely proportional to the density index for each sample. They also provide two versions: (a) inverse spatial weighting for histograms, frequency plots, and age distributions, and (b) inverse spatial–temporal weighting for time-series plotting mean values.

To further assess the appropriateness of these methods, we use the same data as in Section 2.2 (Figs. 2-3) and apply the inverse spatial–temporal approach of Keller and Schoene (2012). Both methods produce nearly identical mean-value time-series (which indicates both methods are acceptable); however, the error bands do occasionally differ noticeably. Here, comparing results from two inverse spatial–temporal methods for weighting time-series serves as an example for assessing replicability.

Every branch of science likely has unique replication problems unrelated to other branches of science. For example, attempts to assign high-resolution ages to otherwise undatable Paleozoic sedimentary layers (Wu et al., 2023) present problems that are far different from assigning low-precision U-Pb ages to detrital zircon. Thus, we cannot prescribe specific tests for all situations, in all disciplines. Yet, the example illustrates how similar comparative studies that use the same data, but with different methods, can provide meaningful insights into the degree to which methodology affects replicability.

2.11. Analytic transparency

Insufficient analytic transparency is a common problem in preventing replication (Horstwood et al., 2016; Sileshi, 2023). To replicate results, it is imperative to know the exact steps that other investigators followed. Obfuscating methods with vague or incomprehensible terms is unacceptable. Methods defined with equations greatly enhance replicability, but only if all terms are clearly stated. Earth Sciences can move forward from its current state by journal editors adopting the minimum transparency standards used in other disciplines. These include (a) providing details of the software used, including exact version numbers, (b) sharing

all the processing data and analytical code via a permanently and openly archived online portal, (c) sharing data packaged together with the processing code and computational environment, so the whole analysis pipeline can be re-run using GitHub releases, Code Ocean, etc., and (d) providing meta-data describing all the variables in all data files and clearly organizing and annotating any analytical code. Supplementary Data file 1 through file 6 include the details and calculations that promote analytic transparency.

3. Discussion

Given full transparency of data and analyses, it is possible to systematically isolate and assess the degree to which various factors might significantly bias results. Minimal biases can often be ignored, whereas a major bias should be addressed to prevent unreliable results. After discovering a major bias, methods often exist for mitigating risks prior to reporting results, which should optimize replication attempts from independent data. Of course, biased data can also be replicated. Thus, replication alone is an insufficient criterion for assessing reliability. Assessing biases and data quality are issues separate from, but generally a requisite to, attempts to replicate reliable results. Some of the methods we have described primarily pertain to U-Pb detrital zircon studies, while other methods likely apply to a broader range of disciplines, including disciplines beyond Earth Sciences.

For our research, multiple requirements and adjustments precede attempts to replicate various global geological and geochemical time-series. The requirements and adjustments include: (a) amalgamating a global database with a bare minimum globality index of 30% to 70% but preferably with a high globality index exceeding 70%; (b) when appropriate, objectively removing questionable data or data of inferior quality; (c) for detrital zircon time-series, adjusting for stratigraphic bias; (d) have some understanding of the precision of the database measurements, which provides guidance for the number of analyses required for assessing accuracy; (e) after adjusting for all known biases, simulating random sampling by producing age distributions with an inverse spatial weighting method; (f) alternatively, simulating random sampling by producing mean-value time-series with an inverse spatial–temporal weighting method; (g) performing segmented correlation analysis for independent time-series to determine if the methods for minimizing biases are sufficient for replicating results for all intervals; and (h) varying the bandwidth for the segmented correlation studies to evaluate data accuracy.

4. Conclusion

This research outlines an approach for developing replicable age distributions from large global databases of U-Pb detrital zircons. Testing replication starts with representative global databases containing independent data. Before developing a time-series, adjustments are essential for known biases such as poor data quality, stratigraphic bias, and geographically disproportionate non-random sampling. Then, segmented correlation analyses from the independent bias-adjusted time-series provide a means for assessing if the adjusts are sufficient for replicating results, the accuracy limits of the data, and the temporal extent to which the time-series are replicable. To augment this approach, Supplementary Data file 3 is an operational spreadsheet containing the documentation and functionality for calculating: five globality indices, inverse spatial weighting, inverse spatial–temporal weighting, and adjustments for stratigraphic bias.

CRediT authorship contribution statement

Stephen J. Puetz: Conceptualization, Methodology, Project administration, Software, Validation, Formal analysis, Data curation, Writing – original draft. **Kent C. Condie:** Writing – review & editing. **Kurt Sundell:** Validation, Formal analysis, Writing – review & editing. **Nick M.W. Roberts:** Writing – review & editing. **Christopher J. Spencer:** Writing – review & editing. **Slah Boulila:** Writing – review & editing. **Qiuming Cheng:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank Peter Voice, Malgorzata Lagisz, and an anonymous reviewer, as well as editor Damian Nance for considerable suggestions and guidance that significantly improved the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gsf.2024.101821.

References

Aber, J.D., 1997. Why don't we believe the models? Bull. Ecol. Soc. Am. 78, 232–233 http://www.jstor.org/stable/20168170.

Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. Nature 533, 452–454. https://doi.org/10.1038/533452a.

Balica, C., Ducea, M.N., Gehrels, G.E., Kirk, J., Roban, R.D., 2020. A zircon petrochronologic view on granitoids and continental evolution. Earth Planet. Sci. Lett. 531,. https://doi.org/10.1016/j.epsl.2019.116005 116005.

Bartelt, C., 2022. Provenance of the Northern Range, Trinidad using Detrital Zircon U-Pb Geochronology: Implications for Northern South American River System Paleogeography. Augustana College, Senior thesis. https://digitalcommons.augustana.edu/geolstudent/8.

Barth, A.P., Wooden, J.L., Jacobson, C.E., Economos, R.C., 2013. Detrital zircon as a proxy for tracking the magmatic arc system: the California arc example. Geology 41, 223–226. https://doi.org/10.1130/G33619.1.

Chang, A.C., Li, P., 2022. Is economics research replicable? Sixty published papers from thirteen journals say "often not". Critical Finance Rev. 11, 185–206. https://doi.org/10.1561/104.00000053.

Committee on Reproducibility and Replicability in Science, 2019. Reproducibility and Replicability in Science. National Academies Press, Washington DC, USA. https://www.ncbi.nlm.nih.gov/books/NBK547546/.

Condie, K.C., Puetz, S.J., Spencer, C.J., Roberts, N.M.W., 2023. Four billion years of secular compositional change in granitoids. Chem. Geol. 644,. https://doi.org/10.1016/j.chemgeo.2023.121868 121868.

Condon, D.J., Schoene, B., McLean, N.M., Bowring, S.A., Parrish, R.R., 2015. Metrology and traceability of U-PB isotope dilution geochronology (EARTHTIME tracer calibration part I). Geochim. Cosmochim Acc. 164, 464–480. https://doi.org/10.1016/j.gca.2015.05.026.

Cox, G.M., Lyons, T.W., Mitchell, R.N., Hasterok, D., Gard, M., 2018. Linking the rise of atmospheric oxygen to growth in the continental phosphorus inventory. Earth Planet. Sci. Lett. 489, 28–36. https://doi.org/10.1016/j.epsl.2018.02.016.

Crisp, M.D., Trewick, S.A., Cook, L.G., 2011. Hypothesis testing in biogeography. Trends Ecol. Evol. 26, 66–72. https://doi.org/10.1016/j.tree.2010.11.005.

Dobbs, S.C., Malkowski, M.A., Schwartz, T.M., Sickmann, Z.T., Graham, S.A., 2022. Depositional controls on detrital zircon provenance: an example from upper cretaceous strata, southern Patagonia. Front. Earth Sci. 2022, 70232516. https://doi.org/10.3389/feart.2022.824930.

Dodson, M.H., Compston, W., Williams, I.S., Wilson, J.F., 1988. A search for ancient detrital zircons in Zimbabwean sediments. Geol Soc. London J. 145, 977–983. https://doi.org/10.1144/gsjgs.145.6.0977.

Doucet, S., Gamaleldien, H., Li, Z.X., 2022. Pitfalls in using the geochronological information from the EarthChem portal for Precambrian time-series analysis. Precambrian Res. 369,. https://doi.org/10.1016/j.precamres.2021.106514 106514.
Farrell, U.C., Samawi, R., Anjanappa, S., Klykov, R., Adeboye, O.O., 2021. The sedimentary geochemistry and paleoenvironments project. Geobiology 19, 545–556. https://doi.org/10.1111/gbi.12462.
Gehrels, G., Sundell, K., George, S., 2019. Short Course modules on U-Pb Geochronology of Detrital Zircons: Best Practices for U-Pb Data Acquisition, Reduction, Analysis, and Archiving. GSA 2019 Meeting, Sept 22-25; Phoenix, Arizona. https://sites.google.com/a/laserchron.org/laserchron/.
Hasterok, D., Gard, M., Cox, G., Hand, M., 2019. A 4 Ga record of granitic heat production: implications for geodynamic evolution and crustal composition of the early earth. Precambrian Res. 331,. https://doi.org/10.1016/j.precamres.2019.105375 105375.
Hilgen, F.J., Hinnov, L.A., Aziz, H.A., Abels, H.A., Batenburg, S., 2015. Stratigraphic continuity and fragmentary sedimentation: the success of cyclostratigraphy as part of integrated stratigraphy. Geol. Soc. London Spec. Pubs. 404, 157–197. https://doi.org/10.1144/SP404.12.
Hinnov, L.A., 2013. Cyclostratigraphy and its revolutionizing applications in the earth and planetary sciences. GSA Bull. 125, 1703–1734. https://doi.org/10.1130/B30934.1.
Horstwood, M.S.A., Košler, J., Gehrels, G., Jackson, S.E., McLean, N.M., Paton, C., 2016. Community-derived standards for LA-ICP-MS U-(th-)pb geochronology – uncertainty propagation, age interpretation and data reporting. Geostand. Geoanal. Res. 40, 311–332. https://doi.org/10.1111/j.1751-908X.2016.00379.x.
Ioannidis, J.P.A., 2005. Why Most published research findings are false. PLoS Med. 2, e124.
Johnson, T.E., Kirkland, C.L., Gardiner, N.J., Brown, M., Smithies, R.H., Santosh, M., 2019. Secular change in TTG compositions: implications for the evolution of Archaean geodynamics. Earth Planet. Sci. Lett. 505, 65–75. https://doi.org/10.1016/j.epsl.2018.10.022.
Jones, T.D., Lunt, D.J., Schmidt, D.N., Ridgwell, A., Sluijs, A., Valdes, P.J., Maslin, M., 2013. Climate model and proxy data constraints on ocean warming across the Paleocene-Eocene Thermal Maximum. Earth-Sci. Rev. 125, 123–145. https://doi.org/10.1016/j.earscirev.2013.07.004.
Keller, C., Schoene, B., 2012. Statistical geochemistry reveals disruption in secular lithospheric evolution about 2.5 Gyr ago. Nature 485, 490–493. https://doi.org/10.1038/nature11024.
Košler, J., Sláma, J., Belousova, E., Corfu, F., Gehrels, G.E., 2013. U-pb detrital zircon analysis – results of an inter-laboratory comparison. Geostand. Geoanal. Res. 37, 243–259. https://doi.org/10.1111/j.1751-908X.2013.00245.x.
Kuhn, M., Johnson, K., 2013. Over-fitting and model tuning. In: Applied Predictive Modeling. Springer, New York. https://doi.org/10.1007/978-1-4614-6849-3_4.
Kushner, B.E., Soreghan, G.S., Soreghan, M.J., 2022. Late Paleozoic cratonal sink: distally sourced sediment filled the Anadarko Basin (USA) from multiple source regions. Geosphere 18, 1831–1850. https://doi.org/10.1130/GES02489.1.
Laskar, J., Robutel, P., Joutel, F., Gastineau, M., Correia, A.C.M., Levard, B., 2004. A long-term numerical solution for the insolation quantities of the Earth. Astron. Astrophys. 428, 261–285. https://doi.org/10.1051/0004-6361:20041335.
Liu, H., Zartman, R.E., Ireland, T.R., Sun, W.D., 2019. Global atmospheric oxygen variations recorded by Th/U systematics of igneous rocks. Proc. Natl. Acad. Sci. 116, 18854–18859. https://doi.org/10.1073/pnas.1902833116.
Lu, G.M., Wang, W., Ernst, R.E., El Bilali, H., Spencer, C.J., Xu, Y.G., Bekker, A., 2023. Evolutionary stasis during the Mesoproterozoic Columbia-Rodinia supercontinent transition. Precambrian Res. 391,. https://doi.org/10.1016/j.precamres.2023.107057 107057.
McDowall, R.M., 2004. What biogeography is: a place for process. J. Biogeogr. 31, 345–351. https://doi.org/10.1046/j.0305-0270.2003.01020.x.
Meyers, S.R., 2015. The evaluation of eccentricity-related amplitude modulation and bundling in paleoclimate data: an inverse approach for Astro-chronologic testing and time scale optimization. Paleoceanogr. Paleocl. 30, 1625–1640. https://doi.org/10.1002/2015PA002850.
Moonesinghe, R., Khoury, M.J., Janssens, A.C.J.W., 2007. Most published research findings are false—But a little replication goes a long way. PLoS Med. 4, e28.
Owen, D.B., 1965. The power of student's t-test. J. Am. Stat. Assoc. 60, 320–333. https://doi.org/10.1080/01621459.1965.10480794.
Puetz, S.J., 2018. A relational database of global U-Pb ages. Geosci. Front. 9, 877–891. https://doi.org/10.1016/j.gsf.2017.12.004.
Puetz, S.J., Prokoph, A., Borchardt, G., 2016. Evaluating alternatives to the Milankovitch theory. J. Stat. Planning Infer. 170, 158–165. https://doi.org/10.1016/j.jspi.2015.10.006.
Puetz, S.J., Condie, K.C., Pisarevsky, S., Davaille, A., Schwarz, C.J., Ganade, C.E., 2017. Quantifying the evolution of the continental and oceanic crust. Earth-Sci. Rev. 164, 63–83. https://doi.org/10.1016/j.earscirev.2016.10.011.
Puetz, S.J., Condie, K.C., 2019. Time series analysis of mantle cycles Part I: Periodicities and correlations among seven global isotopic databases. Geosci. Front. 10, 1305–1326. https://doi.org/10.1016/j.gsf.2019.04.002.
Puetz, S.J., Condie, K.C., 2021. Applying Popperian falsifiability to geodynamic hypotheses: empirical testing of the episodic crustal/zircon production hypothesis and selective preservation hypothesis. Int. Geol. Rev. 63, 1920–1950. https://doi.org/10.1080/00206814.2020.1818143.
Puetz, S.J., Condie, K.C., 2022. A review of methods used to test periodicity of natural processes with a special focus on harmonic periodicities found in global U-Pb detrital zircon age distributions. Earth-Sci. Rev. 224,. https://doi.org/10.1016/j.earscirev.2021.103885 103885.
Puetz, S.J., Ganade, C.E., Zimmermann, U., Borchardt, G., 2018. Statistical analyses of global U-Pb database 2017. Geosci. Front. 9, 121–145. https://doi.org/10.1016/j.gsf.2017.06.001.
Puetz, S.J., Spencer, C.J., Ganade, C.E., 2021. Analyses from a validated global U-Pb detrital zircon database: enhanced methods for filtering discordant U-Pb zircon analyses and optimizing crystallization age estimates. Earth-Sci. Rev. 220,. https://doi.org/10.1016/j.earscirev.2021.103745 103745.
Puetz, S.J., Spencer, C.J., 2023. Evaluating U-Pb accuracy and precision by comparing zircon ages from 12 standards using TIMS and LA-ICP-MS methods. Geosyst. Geoenviron. 2,. https://doi.org/10.1016/j.geogeo.2022.100177 100177.
Puetz, S.J., Spencer, C.J., Condie, K.C., Roberts, N.M.W., 2024. Enhanced U-Pb detrital zircon, Lu-Hf zircon, δ^{18}O zircon, and Sm-Nd whole rock global databases. Sci. Data 11, 56.
Redden, D.T., Allison, D.B., 2003. Nonreplication in genetic association studies of obesity and diabetes. J. Nutr. 133, 3323–3326. https://doi.org/10.1093/jn/133.11.3323.
Reimer, P., Baillie, M., Bard, E., Bayliss, A., Beck, J., Blackwell, P., 2009. IntCal09 and Marine09 radiocarbon age calibration curves, 0–50,000 years cal BP. Radiocarbon 51, 1111–1150. https://doi.org/10.1017/S0033822200034202.
Reimink, J.R., Davies, J.H.F.L., Ielpi, A., 2021. Global zircon analysis records a gradual rise of continental crust throughout the Neoarchean. Earth Planet. Sci. Lett. 554,. https://doi.org/10.1016/j.epsl.2020.116654 116654.
Schoene, B., Condon, D.J., Morgan, L., 2013. Precision and accuracy in geochronology. Elements 9, 19–24. https://doi.org/10.2113/gselements.9.1.19.
Sileshi, G.W., 2023. Analytic transparency is key for reproducibility of agricultural research. CABI Agriculture and Bioscience 4: Article 2. https://doi.org/10.1186/s43170-023-00144-8.
Smith, T.M., Saylor, J.E., Lapen, T.J., Hatfield, K., Sundell, K.E., 2023. Identifying sources of non-unique detrital age distributions through integrated provenance analysis: an example from the paleozoic Central Colorado trough. Geosphere 19, 471–492. https://doi.org/10.1130/GES02541.1.
Stehman, S.V., Selkowitz, D.J., 2010. A spatially stratified, multi-stage cluster sampling design for assessing accuracy of the Alaska (USA) National Land Cover Database (NLCD). Int. J. Remote Sens. 31, 1877–1896. https://doi.org/10.1080/01431160902927945.
Tamblyn, R., Hasterok, D., Hand, M., Gard, M., 2021. Mantle heating at ca. 2 Ga by continental insulation: evidence from granites and eclogites. Geology 50, 91–95. https://doi.org/10.1130/G49288.1.
Usoskin, I.G., Kovaltsov, G.A., 2012. Occurrence of extreme solar particle events: assessment from historical proxy data. Astrophys. J. 757, 92. https://doi.org/10.1088/0004-637X/757/1/92.
Vaughan, S., Bailey, R.J., Smith, D.G., 2011. Detecting cycles in stratigraphic data: spectral analysis in the presence of red noise. Paleoceanography 26, PA2195. https://doi.org/10.1029/2011PA002195.
Vaughan, S., Bailey, R.J., Smith, D.G., 2014. Cyclostratigraphy: data filtering as a source of spurious spectral peaks. Geol. Soc. London Spec. Publ 404, 151–156. https://doi.org/10.1144/SP404.11.
Vermeesch, P., 2004. How many grains are needed for a provenance study? Earth Planet Sci. Lett. 224, 441–451. https://doi.org/10.1016/j.epsl.2004.05.037.
Vermeesch, P., 2012. On the visualisation of detrital age distributions. Chem. Geol. 312–313, 190–194. https://doi.org/10.1016/j.chemgeo.2012.04.021.
Vermeesch, P., 2018. IsoplotR: a free and open toolbox for geochronology. Geosci. Front. 9, 1479–1493. https://doi.org/10.1016/j.gsf.2018.04.001.
Wahbi, A.M., Blum, M.D., Doerger, C.N., 2023. Early cretaceous continental-scale sediment routing, the McMurray formation, Western Canada Sedimentary Basin, Alberta, Canada. GSA Bull. 135, 2088–2106. https://doi.org/10.1130/B36412.1.
Waters, J.M., Craw, D., 2006. Goodbye Gondwana? New Zealand biogeography, geology, and the problem of circularity. Syst. Biol. 55, 351–356. https://doi.org/10.1080/10635150600681659.
West, S.G., Taylor, A.B., Wu, W., 2012. Chapter 13: model fit and model selection in structural equation modeling. In: Hoyle, R.H. (Ed.), Handbook of Structural Equation Modeling. Guilford Press, New York, pp. 209–231.
Wilson, C., 2022. The replication crisis has spread through science – can it be fixed? New Scientist: Humans, April 6, 2022.
Wu, H., Fang, Q., Hinnov, L.A., Zhang, S., Yang, T., Shi, M., Li, H., 2023. Astronomical time scale for the Paleozoic Era. Earth-Sci. Rev. 244,. https://doi.org/10.1016/j.earscirev.2023.104510 104510.
Zachos, J., Pagani, M., Sloan, L., Thomas, E., Billups, K., 2001. Trends, rhythms, and aberrations in global climate 65 Ma to present. Science 292, 686–693. https://doi.org/10.1126/science.1059412.
Zerkle, A.L., 2018. Biogeodynamics: bridging the gap between surface and deep Earth processes. Philos. Tran. Royal Soc. A 376,. https://doi.org/10.1098/rsta.2017.0401 20170401.