



UNIVERSITY OF LEEDS

This is a repository copy of *Curation and Analysis of Global Sedimentary Geochemical Data to Inform Earth History*.

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/172530/>

Version: Accepted Version

Article:

Mehra, A, Keller, B, Zhang, T et al. (34 more authors) (2021) Curation and Analysis of Global Sedimentary Geochemical Data to Inform Earth History. GSA Today. ISSN 1052-5173

<https://doi.org/10.1130/GSATG484A.1>

© The Geological Society of America, 2021. This is an author produced version of an article published in GSA Today. Uploaded in accordance with the publisher's self-archiving policy.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

Curation and analysis of global sedimentary geochemical data to inform Earth history

Akshay Mehra, Brenhin Keller, Tianran Zhang, Nicholas J. Tosca,

Scott M. McLennan, Erik Sperling, Una Farrell, Jochen Brocks, Donald Canfield, Devon Cole,

Peter Crockford, Huan Cui, Tais W. Dahl, Keith Dewing, Joe Emmings, Robert R. Gaines, Tim

Gibson, Geoffrey J. Gilleaudeau, Romain Guilbaud, Malcolm Hodgkiss, Amber Jarrett, Pavel

Kabanov, Marcus Kunzmann, Chao Li, David K. Loydell, Xinze Lu, Austin Miller, N. Tanner

Mills, Lucas D. Mouro, Brennan O'Connell, Shanan E. Peters, Simon Poulton, Samantha R.

Ritzer, Emmy Smith, Philip Wilby, Tina Woltz, Justin V. Strauss

1 **Abstract**

2 Large datasets increasingly provide critical insights into crustal and surface processes on Earth.
3 These data come in the form of published and contributed observations, which often include
4 associated metadata. Even in the best-case scenario of a carefully curated dataset, it may be non-
5 trivial to extract meaningful analyses from such compilations, and choices made with respect to
6 filtering, resampling, and averaging can affect the resulting trends and any interpretation(s)
7 thereof. As a result, a thorough understanding is required of how to digest, process, and analyze
8 large data compilations. Here, we present a generalizable workflow developed using the
9 Sedimentary Geochemistry and Paleoenvironments Project database. We demonstrate the effects
10 of filtering and weighted resampling using Al_2O_3 and U, two representative geochemical
11 components of interest in sedimentary geochemistry (one major and one trace element,
12 respectively). Through our analyses, we highlight several methodological challenges in a “bigger
13 data” approach to Earth Science. We suggest that, with slight modifications to our workflow,

14 researchers can confidently use large collections of observations to gain new insights into
15 processes that have shaped Earth's crustal and surface environments.

16 **Introduction**

17 The study of Earth's past relies on a record that is spatially and temporally variable and, by some
18 metrics, woefully undersampled. Through every geochemical analysis, fossil identification, and
19 measured stratigraphic section, Earth scientists continuously add to this historical record.
20 Compilations of such observations can illuminate global trends through time, providing
21 researchers with crucial insights into our planet's geological and biological evolution. These
22 compilations can vary in size and scope, from hundreds of manually curated entries in a
23 spreadsheet to millions of records stored in software databases. The latter form is exemplified by
24 databases such as The Paleobiology Database (PBDB; Peters and McClennen 2016), Macrostrat
25 (Peters et al. 2018), EarthChem (Walker et al. 2005), Georoc (Sarbas 2008), and the Sedimentary
26 Geochemistry and Paleoenvironments Project (SGP, this study).

27 Of course, large amounts of data are not new to the Earth Sciences, and, with respect to volume,
28 many Earth history and geochemistry compilations are small in comparison to the datasets used
29 in other subdisciplines, including seismology (e.g., Nolet 2012), climate science (e.g., Faghmous
30 and Kumar 2014), and hydrology (e.g., Chen and Wang 2018). As a result, many Earth history
31 compilations likely do not meet the criteria to be called "big data", which is a term that describes
32 very large amounts of information that accumulate rapidly and which are heterogeneous and
33 unstructured in form (Gandomi and Haider 2015; or, "if it fits in memory, it is small data"). That
34 said, the tens of thousands to millions of entries present in such datasets do represent a new
35 frontier for those interested in our planet's past. For many Earth historians, however, and

36 especially for geochemists (where most of the field's efforts traditionally have focused on
37 analytical measurements rather than data analysis; see Sperling et al. 2019), this frontier requires
38 new outlooks and toolkits.

39 When using compilations to extract global trends through time, it is important to recognize that
40 large datasets can have several inherent issues. Observations may be unevenly distributed
41 temporally and/or spatially, with large stretches of time (e.g., parts of the Archean Eon) or space
42 (e.g., much of Africa; Fig. S1) lacking data. There may also be errors with entries—mislabeled
43 values, transposition issues, and missing metadata can occur in even the most carefully curated
44 compilations. Even if data are pristine, they may span decades of acquisition with evolving
45 techniques, such that both analytical precision and measurement uncertainty are non-uniform
46 across the dataset (Fig. S2). Careful examination may demonstrate that contemporaneous and co-
47 located observations do not agree. Additionally, data often are not targeted, such that not every
48 entry may be necessary for (or even useful to) answering a particular question.

49 Luckily, these (and other) issues can be addressed through careful processing and analysis, using
50 well-established statistical and computational techniques. Although such techniques have
51 complications of their own (e.g., a high degree of comfort with programming often is required to
52 run code efficiently), they do provide a way to extract meaningful trends from large datasets. No
53 one lab can generate enough data to cover Earth's history densely enough (i.e., in time and
54 space), but by leveraging compilations of accumulated knowledge, and using a well-developed
55 computational pipeline, researchers can begin to ascertain a clearer picture of Earth's past.

56 **A Proposed Workflow**

57 The process of transforming entries in a dataset into meaningful trends requires a series of steps,
58 many with some degree of user decision-making. Our proposed workflow is designed with the
59 express intent of removing unfit data while appropriately propagating uncertainties. First, a
60 compiled dataset is made or sourced (Fig. S3, i.). Next, a researcher chooses between in-database
61 analysis and extracting data into another format, such as a text file (Fig. S3, ii.). This choice does
62 nothing to the underlying data—its sole function is to recast information into a digital format that
63 the researcher is most comfortable with. Then, a decision must be made about whether to remove
64 entries that are not pertinent to the question at hand (Fig. S3, iii.). Using one or more metadata
65 parameters (e.g., in the case of rocks, lithological descriptions), researchers can turn large
66 compilations into targeted datasets, which then can be used to answer specific questions without
67 the influence of irrelevant data. Following this gross filtering, researchers must decide between
68 removing outliers or keeping them in the dataset (Fig. S3, iv.). Outliers have the potential to
69 drastically skew results in misleading ways. Ascertaining which values are outliers is a non-
70 trivial task and all choices about outlier exclusion must be clearly described when presenting
71 results. Finally, samples are drawn from the filtered dataset (i.e., “resampling”), using a
72 weighting scheme that seeks to address the spatial and temporal heterogeneities—as well as
73 analytical uncertainties—of the data (Fig. S3, vi.). To calculate statistics from the data, multiple
74 iterations of resampling are required.

75 **Case Study: The Sedimentary Geochemistry and Paleoenvironments Project Data**

76 The SGP project seeks to compile sedimentary geochemical data, made up of various analytes
77 (i.e., components that have been analyzed), from throughout geologic time. We applied our
78 workflow to the SGP database to extract coherent temporal trends in Al_2O_3 and U from
79 siliciclastic mudstones. Al_2O_3 is relatively immobile and thus useful for constraining both the

80 provenance and chemical weathering history of ancient sedimentary deposits (Young and Nesbitt
81 1998). Conversely, U is highly sensitive to redox processes. In marine mudstones, U serves as
82 both a local proxy for reducing conditions in the overlying water column (i.e., authigenic U
83 enrichments only occur under low-oxygen or anoxic conditions and/or very low sedimentation
84 rates; see Algeo and Li 2020) and a global proxy for the areal extent of reducing conditions (i.e.,
85 the magnitude of authigenic enrichments scales in part with the global redox landscape; see
86 Partin et al. 2013).

87 SGP data are stored in a PostgreSQL relational database that currently comprises a total of
88 82,579 samples (Fig. 1). The SGP database was created by merging sample data and geological
89 context information from three separate sources, each with different foci and methods for
90 obtaining the “best guess” age of a sample (i.e., the interpreted age as well as potential maximum
91 and minimum ages). The first source is direct entry by SGP team members, which focuses
92 primarily on Neoproterozoic-Paleozoic shale samples and has global coverage. Due to the direct
93 involvement of researchers intimately familiar with their sample sets, these data have the most
94 precise (Fig. 1 a)—and likely also most accurate—age constraints. Second, the SGP database has
95 incorporated sedimentary geochemical data from the United States Geological Survey (USGS)
96 National Geochemical Database (NGDB), comprising data from projects completed between the
97 1960s and 1990s. These samples, which cover all lithologies and are almost entirely from
98 Phanerozoic sedimentary deposits of the United States, are associated with the continuous-time
99 age model from Macrostrat (Peters et al. 2018). Finally, the SGP database includes data from the
100 USGS Global Geochemical Database for Critical Metals in Black Shales project (CMIBS;
101 Granitto et al. 2017), culled to remove ore-deposit related samples. The CMIBS samples
102 predominantly are shales, have global coverage, and span the entirety of Earth’s sedimentary

103 record. When possible, the USGS data are associated with Macrostrat continuous-time age
104 models; otherwise, the data are assigned age information by SGP team members (albeit without
105 detailed knowledge of regional geology or geologic units).

106 **Cleaning and Filtering**

107 We exported SGP data into a comma-separated values (.csv) text file, using a custom structured
108 query language (SQL) query. In the case of geochemical analytes, this query included unit
109 conversions from both weight percent (wt%) and parts per billion (ppb) to parts per million
110 (ppm). After export, we parsed the .csv file and screened the data through a series of steps. First,
111 if multiple values were reported for an analyte in a sample, we calculated and stored the mean (or
112 weighted mean, if there were enough values) and standard deviation of the analyte. Then, we
113 redefined empty values—which are the result of abundance being above or below detection—as
114 “not a number” (NaN, a special value defined by Institute of Electrical and Electronics Engineers
115 (IEEE) floating-point number standard that always returns false on comparison; see IEEE 2019).
116 Next, we converted major elements (e.g., those that together comprise >95% of Earth’s crust or
117 individually >1 wt% of a sample) into their corresponding oxides; if an oxide field did not
118 already exist, or if there was no measurement for a given oxide, the converted value was inserted
119 into the data structure. Then, we assigned both age and measurement uncertainties to the parsed
120 data. In the case of the parsed SGP data, 5,935 samples (i.e., 7.1% of the original dataset) lacked
121 an interpreted age and so no uncertainty could be assigned. For the remainder, we calculated an
122 initial absolute age uncertainty by either using the reported maximum and minimum ages:

$$123 \quad \sigma = \frac{|\text{age}_{\text{maximum}} - \text{age}_{\text{minimum}}|}{2},$$

124 or, if there were no maximum and minimum age values available, by defaulting to a two-sigma
125 value of 6% of the interpreted age:

$$126 \quad \sigma = 0.03 * \text{age}_{\text{interpreted}}$$

127 The choice of a 6% default value was based on a conservative estimate of the precision of
128 common *in situ* dating techniques (see, for example, Schoene 2014). Additionally, we enforced a
129 minimum σ of 25 million years:

$$130 \quad \sigma = \max(\sigma, 25)$$

131 Effectively, each datum can be thought of as a Gaussian distribution along the time axis with a σ
132 of at least 25 million years (the minimum value of which may be thought of as a kernel
133 bandwidth, rather than an analytical uncertainty). The selection of this σ value should correspond
134 to an estimate of the processes that are being investigated (e.g., tectonic changes in provenance).
135 We did not impose a minimum relative age uncertainty.

136 With respect to measurement uncertainties, we assigned an absolute uncertainty to every analyte
137 that lacked one by multiplying the reported analyte value by a relative error. In future database
138 projects, there is considerable scope to go beyond this coarse uncertainty quantification strategy.
139 For example, given the detailed metadata associated with each sample in the SGP database, it
140 would be straightforward to develop correction factors or uncertainty estimates for different
141 geochemical methodologies (e.g., ICP-MS versus ICP-OES, benchtop versus handheld XRF,
142 etc.). Correcting data for biases introduced during measurement is common in large Earth
143 Science datasets (Chan et al. 2019). However, such corrections previously have not been
144 attempted in sedimentary geochemistry datasets.

145 Next, we processed the data through a simple lithology filter because, in the general case of rock-
146 based datasets, only lithologies relevant to the question at hand provide meaningful information.
147 The choice of valid lithologies (or, for that matter, any other filterable metadata) are dependent
148 on the researchers' question(s). As highlighted in the Discussion, lithology filtering has
149 significant implications for redox-sensitive and/or mobile/immobile elements. In this case study,
150 our aim was to only sample data generated from siliciclastic mudstones. To decide which values
151 to screen by, we manually examined a list made up of all unique lithologies in the dataset. We
152 excluded samples that did not match our list of chosen lithologies (removing ~63.5% of the data;
153 Table S1; Fig. S4). Our strategy ensured that we only included mudstones *sensu lato* (see Potter
154 et al. 2005 for a general description) where the lithology was coded. Alternative methods—such
155 as choosing samples based on an Al cutoff value (e.g., Reinhard et al. 2017)—likely would result
156 in a set comprising both mudstone and non-mudstone coded lithologies. In the future, improved
157 machine learning algorithms, designed to classify unknown samples based on their elemental
158 composition, may provide a more sophisticated means by which to generate the largest possible
159 dataset of lithology-appropriate samples.

160 We then completed a preliminary screening of the lithology filtered samples by checking if
161 extant analyte values were outside of physically possible bounds (e.g., individual oxides with
162 wt% less than 0 or greater than 100), and, if so, setting them to NaN. Next, to reduce the number
163 of mudstone samples with detrital or authigenic carbonate and phosphatic mineral phases, we
164 excluded samples with greater than 10 wt% Ca and/or more than 1 wt% P₂O₅ (removing ~
165 66.9% of the remaining data; Fig. S4). Additionally, in order to ensure that our mudstone
166 samples were not subject to secondary enrichment processes, such as ore mineralization, we
167 queried the USGS NGDB to extract the recorded characteristics of every sample with an

168 associated USGS NGDB identifier. We examined these characteristics for the presence of
169 selected strings (i.e., “mineralized”, “mineralization present”, “unknown mineralization”, and
170 “radioactive”) and excluded any sample exhibiting one or more strings. Finally, as there were
171 still several apparent outliers in the dataset, we manually examined the log histograms of each
172 element and oxide of interest. On each histogram, we demarcated the 0.5th and 99.5th percentile
173 bounds of the data, then visually studied those histograms to exclude “outlier populations”, or
174 samples located both well outside those percentile bounds and not part of a continuum of values
175 (removing ~5.7% of the remaining data; Fig. S4). Following these filtering steps, we saved the
176 data in a .csv text file.

177 **Data Resampling**

178 We implemented resampling based on inverse distance weighting (after Keller and Schoene
179 2012), in which samples closer together—that is, with respect to a metric such as age or spatial
180 distance—are considered to be more alike than samples that are further apart. The inverse
181 weighting of an individual point, x , is based on the basic form:

$$182 \quad y(x) = \frac{1}{d(x, x_i)^p},$$

183 where d is a distance function, x_i is a second sample, and p , which is greater than 0, is a power
184 parameter. In the case of the SGP data, we used two distance functions, spatial (s) and temporal
185 (t):

$$186 \quad \begin{aligned} s &= \frac{\text{arcdistance}(x, x_i)}{\text{scale}_{\text{spatial}}}, \\ t &= \frac{|\text{age}(x - x_i)|}{\text{scale}_{\text{age}}}, \end{aligned}$$

187 where *arcdistance* refers to the distance between two points on a sphere, *scale_{spatial}* refers to
188 a preselected arc distance value (in degrees; Fig. S5, inset), and *scale_{age}* is a preselected age
189 value (in million years, Ma). In this case study, we chose a *scale_{spatial}* of 0.5 degrees and a
190 *scale_{age}* of 10 Ma (see below for a discussion about parameter values).

191 For *n* samples, the proximity value *w* assigned to each sample *x* is:

$$192 \quad w(x) = \sum_{i=1}^{i=n} \frac{1}{(s^2 + 1)} + \frac{1}{(t^2 + 1)}.$$

193 Essentially, the proximity value is a summation of the reciprocals of the distance measures made
194 for each pair of the sample and a single other datum from the dataset. Accordingly, samples that
195 are closer to other data in both time and space will have larger *w* values than those that are
196 farther away. Note that the additive term of 1 in the denominator defines a maximum value of 1
197 for each reciprocal distance measure.

198 We normalized the generated proximity values (Fig. S6) to produce a probability value *P*. This
199 normalization was done such that the median proximity value corresponded to a *P* of ~0.20 (i.e.,
200 a 1 in 5 chance of being chosen):

$$201 \quad P(x) = \frac{1}{\left(w(x) * \text{median} \left(\frac{0.20}{w} \right) \right) + 1}.$$

202 This normalization results in an “inverse proximity weighting”, such that samples that are closer
203 to other data (which have large *w* values) end up with a smaller *P* value than those that are far
204 away from other samples. Next, we assigned both analytical and temporal uncertainties to each
205 analyte to be resampled. Then, we culled the dataset into an *m* by *n* matrix, where each row

206 corresponded to a sample and each column to an analyte. We resampled this culled dataset
207 10,000 times using a three-step process: (1) we drew samples, using calculated P values, with
208 replacement (i.e., each draw considered all available samples, regardless of whether a sample
209 had already been drawn); (2) we multiplied the assigned uncertainties discussed above by a
210 random draw from a normal distribution ($\mu = 0$; $\sigma = 1$) to produce an error value; and (3) we
211 added these newly calculated errors to the drawn temporal and analytical values. Finally, we
212 binned and plotted the resampled data.

213 Naturally, the reader may ask how we chose the values for $scale_{age}$ and $scale_{temporal}$ and
214 what, if any, impact those choices had on the final results? Nominally, the values of $scale_{age}$
215 and $scale_{temporal}$ are controlled by the size and age, respectively, of the features that are being
216 sampled. So, in the case of sedimentary rocks, those values should reflect the length scale and
217 duration of a typical sedimentary basin, such that many samples from the same “spatiotemporal”
218 basin have lower P values than few samples from distinct basins. Of course, it is debatable what
219 “typical” means in the context of sedimentary basins, as both size and age can vary over orders
220 of magnitude (Woodcock 2004). Given this uncertainty, we subjected the SGP data to a series of
221 sensitivity tests, where we varied both $scale_{age}$ and $scale_{temporal}$, using logarithmically spaced
222 values of each (Fig. S5). While the uncertainty associated with results varied based on the choice
223 of the two parameters, the overall mean values were not appreciably different (Fig. S7).

224 **Results**

225 To study the impact of our methodology, we present results for two geochemical components, U
226 and Al_2O_3 (Fig. 2). Contents-wise, the U and Al_2O_3 data in the SGP database contain extreme
227 outliers. Many of these outliers were removed using the lithology and Ca or P_2O_5 screening (Fig.

228 2 a, c); the final outlier filtering strategy discussed above handled any remaining values of
229 concern. In the case of U, our multi-step filtering reduced the range of concentrations by two
230 orders of magnitude, from 0 to 500,000 ppm to 0 to 500 ppm.

231 **Discussion**

232 The illustrative examples we have presented have implications for understanding Earth history.
233 Al_2O_3 contents of ancient mudstones appear relatively stable over the last ~ 1500 Ma (the time
234 interval for which appreciable data exist in our dataset), suggesting little first-order change in
235 Al_2O_3 delivery to sedimentary basins over time. The U content of mudstones shows a substantial
236 increase between the Proterozoic and Phanerozoic. Although we have not accounted for the
237 redox state of the overlying water column, these results broadly recapitulate the trends seen in a
238 previous much smaller, and non-weighted, dataset (Partin et al. 2013) and generally may indicate
239 oxygenation of the oceans within the Phanerozoic.

240 Moving forward, there is no reason to believe that the compilation and collection of published
241 data, whether in a semi-automated (e.g., SGP) or automated (e.g., GeoDeepDive; Peters et al.
242 2014) manner, will slow and/or stop (Bai et al. 2017). Those interested in Earth's history—as
243 collected in large compilations—should understand how to extract meaningful trends from these
244 ever-evolving datasets. By presenting a workflow that is purposefully general and must be
245 adapted before use, we hope to elucidate the various aspects that must be considered when
246 processing large volumes of data.

247 Foremost to any interpretation of a quantitative dataset is an assessment of uncertainty. In truth, a
248 datum representing a physical quantity is not a single scalar point, but rather, an entire
249 distribution. In many cases, such as in our workflow, this distribution is implicitly assumed to be

250 Gaussian, an assumption which may or may not be accurate (Rock et al. 1987)—although a
251 simplified distribution certainly is better than none. The quantification of uncertainty in Earth
252 Sciences especially is critical when averaging and binning by a selected independent variable,
253 since neglecting the uncertainty of the independent variable will lead to interpretational failures
254 that may not be mitigated by adding more data. As time perhaps is the most common
255 independent variable (and one with a unique relationship to the assessment of causality),
256 incorporating its uncertainty especially is critical for the purposes of Earth history studies (Ogg
257 et al. 2016). An age without an uncertainty is not meaningful data. Indeed, such a value is even
258 worse than an absence of data, for it is actively misleading. Consequently, assessment of age
259 uncertainty is one of the most important, yet underappreciated, components of building accurate
260 temporal trends from large datasets.

261 Of course, age is not the only uncertain aspect of samples in compiled datasets, and researchers
262 should seek to account for as many inherent uncertainties as possible. Here, we propagate
263 uncertainty by using a resampling methodology that incorporates information about space, time,
264 and measurement error. Our chosen methodology—which is by no means the only option
265 available to researchers studying large datasets—has the benefit of preventing one location or
266 time range from dominating the resulting trend. For example, although the Archean records of
267 Al_2O_3 and U especially are sparse (Fig. 2), resampling prevents the appearance of artificial
268 “steps” when transitioning from times with little data to instances of (relatively) robust sampling
269 (e.g., see the resampled record of Al_2O_3 between 4000 and 3000 Ma). Therefore, researchers
270 should examine their selected methodologies to ensure that: 1) uncertainties are accounted for,
271 and 2) that spatiotemporal heterogeneities are addressed appropriately.

272 Even with careful uncertainty propagation, datasets must also be filtered to keep outliers from
273 affecting the results. It is important to note that the act of filtering does not mean that the filtered
274 data are necessarily “bad”, just that they do not meaningfully contribute to the question at hand.
275 For example, while our lithology and outlier filtering methods removed most U data because
276 they were inappropriate for reconstructing trends in mudstone geochemistry through time, that
277 same data would be especially useful for other questions, such as determining the variability of
278 heat production within shales. This sort of filtering is a fixture of scientific research—e.g.,
279 geochemists will consider whether samples are diagenetically altered when measuring them for
280 isotopic data—and, likewise, should be viewed as a necessary step in the analysis of large
281 datasets.

282 As our workflow demonstrates, filtering often requires multiple steps, some automatic (e.g.,
283 cutoffs that exclude vast amounts of data in one fell swoop or algorithms to determine the
284 “outlierness” of data, see Ptáček et al. 2020) and others manual (e.g., examining source literature
285 to determine whether an anomalous value is, in fact, meaningful). Each procedure, along with
286 any assumptions and/or justifications, must be documented clearly (and code included and/or
287 stored in a publicly-accessible repository) by researchers so that others may reproduce their
288 results and/or build upon their conclusions with increasingly larger datasets.

289 Along with documentation of data processing, filtering, and sampling, it is important for
290 researchers also to leverage sensitivity analyses to understand how parameter choices may
291 impact resulting trends. Here, through the analysis of various spatial and temporal parameter
292 values, we demonstrate that, while the spread of data varies based on the prescribed values of
293 $scale_{spatial}$ and $scale_{temporal}$, the averaged resampled trend does not (Fig. S7). At the same
294 time, we see that trends are directly influenced by the use (or lack thereof) of Ca and P₂O₅ and

295 outlier filtering. For example, the record of U in mudstones becomes overprinted by anomalously
296 large values when carbonate samples are not excluded (Fig. S7 b).

297 **Conclusion**

298 Large datasets can provide increasingly valuable insights into the ancient Earth system.
299 However, to extract meaningful trends, these datasets must be cultivated, curated, and processed
300 with an emphasis on data quality, uncertainty propagation, and transparency. Charles Darwin
301 once noted that the “natural geological record [is] a history of the world imperfectly kept”
302 (Darwin 1859), a reality which is the result of both geological and sociological causes. But while
303 the data are biased, they are also tractable. As we have demonstrated here, the challenges of
304 dealing with this imperfect record—and, by extension, the large datasets that document it—
305 certainly are surmountable.

306 **Acknowledgements**

307 We thank everyone who contributed to the SGP database, including T. Frasier (YGS). BGS
308 authors (JE, PW) publish with permission of the Executive Director of the British Geological
309 Survey, UKRI. We would like to thank the editor and one anonymous reviewer for their helpful
310 feedback.

311 **References**

- 312 Algeo, T.J., and Li, C., 2020, Redox classification and calibration of redox thresholds in
313 sedimentary systems: *Geochimica et Cosmochimica Acta*.
- 314 Bai, Y., Jacobs, C.A., Kwan, M, and Waldmann, C, 2017, Geoscience and the technological
315 revolution [perspectives]: *IEEE Geoscience and Remote Sensing Magazine* 5 (3): 72–75.
- 316 Chan, D., Kent E.C., Berry, D.I., and Huybers, P, 2019, Correcting datasets leads to more
317 homogeneous early-twentieth-century sea surface warming: *Nature* 571 (7765): 393–97.

318 Chen, L., and Wang L, 2018, Recent advances in Earth observation big data for hydrology: Big
319 Earth Data 2 (1): 86–107.

320 Darwin, C, 1859, On the origin of species.

321 Faghmous, J.H., and Kumar, V., 2014, A big data guide to understanding climate change: the
322 case for theory-guided data science: Big Data 2 (3): 155–63.

323 Gandomi, A., and Haider, M., 2015, Beyond the hype: big data concepts, methods, and analytics:
324 International Journal of Information Management 35 (2): 137–44.

325 Granitto, M., Giles S.A., and Kelley, K.D, 2017, Global Geochemical Database for Critical
326 Metals in Black Shales: U.S. Geological Survey Data Release,
327 <https://doi.org/https://doi.org/10.5066/F71G0K7X>.

328 IEEE. 2019. IEEE Standard for Floating-Point Arithmetic. IEEE Std 754-2019 (Revision of
329 IEEE 754-2008), 1–84.

330 Keller, C.B., and Schoene, B., 2012, Statistical geochemistry reveals disruption in secular
331 lithospheric evolution about 2.5 gyr ago: Nature 485 (7399): 490–93.

332 Nolet, G., 2012, Seismic tomography: with applications in global seismology and exploration
333 geophysics: Vol. 5. Springer Science & Business Media.

334 Ogg, J.G., Ogg G.M., and Gradstein F.M., 2016, A concise geologic time scale: 2016. Elsevier.

335 Partin, C.A., Bekker, A., Planavsky, N.J., Scott, C.T., Gill, B.C., Li, C., Podkovyrov, V., et al.
336 2013, Large-scale fluctuations in Precambrian atmospheric and oceanic oxygen levels from the
337 record of U in shales: Earth and Planetary Science Letters 369: 284–93.

338 Peters, S.E., Husson, J.M., and Czaplewski J., 2018, Macrostrat: a platform for geological data
339 integration and deep-time Earth crust research: Geochemistry, Geophysics, Geosystems 19 (4):
340 1393–1409.

341 Peters, S.E., and McClennen, M., 2016, The paleobiology database application programming
342 interface: Paleobiology 42 (1): 1–7.

343 Peters, S.E., Zhang, C., Livny, M., and Re, C., 2014, A machine reading system for assembling
344 synthetic paleontological databases: PLoS one, 9(12):e113523.

345 Potter, P.E., Maynard, J.B., and Depetris, P.J., 2005, Mud and mudstones: introduction and
346 overview. Springer Science & Business Media.

347 Ptáček, M.P, Dauphas, N., and Greber, N.D., 2020, Chemical evolution of the continental crust
348 from a data-driven inversion of terrigenous sediment compositions: Earth and Planetary Science
349 Letters 539: 116090.

350 Reinhard, C.T., Planavsky, N.J., Gill, B.C., Ozaki, K., Robbins, L.J., Lyons, T.W., Fischer,
351 W.W., Wang, C., Cole, D.B., and Konhauser, K.O., 2017, Evolution of the global phosphorus
352 cycle: Nature 541 (7637): 386–89.

353 Rock, N.M.S., Webb J.A., McNaughton, N.J., and Bell, G.D., 1987, Nonparametric estimation of
354 averages and errors for small data-sets in isotope geoscience: a proposal: *Chemical Geology:*
355 *Isotope Geoscience Section* 66 (1-2): 163–77.

356 Sarbas, B., 2008, The Georoc database as part of a growing geoinformatics network, *in*
357 *Geoinformatics 2008—Data to Knowledge*, 42–43. USGS.

358 Schoene, B., 2014, 4.10 - U–Th–Pb geochronology, *in* Holland, H. D. and Turekian, K. K., eds,
359 *Treatise on Geochemistry (Second Edition)*, p 341–378. Elsevier, Oxford.

360 Sperling, E.A., Tecklenburg, S., and Duncan, L.E., 2019, Statistical inference and reproducibility
361 in geobiology: *Geobiology* 17 (3): 261–71.

362 Walker, J.D., Lehnert, K.A., Hofmann, A.W., Sarbas, B., and Carlson, R.W, 2005, EarthChem:
363 International collaboration for solid Earth geochemistry in geoinformatics. AGUFM 2005:
364 IN44A–03.

365 Woodcock, N.H, 2004, Life span and fate of basins: *Geology* 32 (8): 685–88.

366 Young, G.M., and Nesbitt, H.W., 1998, Processes controlling the distribution of Ti and Al in
367 weathering profiles, siliciclastic sediments and sedimentary rocks: *Journal of Sedimentary*
368 *Research* 68 (3): 448–55.

369 **Figure Captions**

370 **Figure 1: Visualizations of data in the SGP database. A.** Relative age uncertainty (i.e., the
371 reported age sigma divided by the reported interpreted age) versus Sample ID. The large gap in
372 Sample ID values results from the deletion of entries during the initial database compilation.
373 This gap has no impact on analyses. **B.** Box plot showing the distribution of relative ages with
374 respect to the sources of data.

375 **Figure 2: Filtering and resampling of Al₂O₃ and U. A and C.** Al₂O₃ and U data through time,
376 respectively. Each datum is color coded by the filtering step at which it was separated from the
377 dataset. In blue is the final filtered data, which was used to generate the resampled trends in **B**
378 and **D. B and D.** Plots depicting Al₂O₃ and U filtered data, along with a histogram of resampled
379 data density and the resulting resampled mean and 2 σ error. Note the log-scale y axis in **C.**