Does a proxy measure up? A framework to assess and convey proxy reliability

F. Garrett Boudinot and Joseph Wilson

1Department of Geological Sciences and Institute of Arctic and Alpine Research, University of Colorado Boulder, Boulder, Colorado 80309, USA
2Department of Philosophy and Center for the Study of Origins, University of Colorado Boulder, Boulder, Colorado 80309, USA

Correspondence: F. Garrett Boudinot (garrett.boudinot@colorado.edu)

Received: 2 February 2020 – Discussion started: 2 March 2020
Revised: 11 August 2020 – Accepted: 18 August 2020 – Published: 28 September 2020

Abstract. Earth scientists describe a wide range of observational measurements as “proxy measurements”. By referring to such a vast body of measurements simply as “proxy”, researchers dilute significant differences in the various ways that measurements relate to the phenomena they intend to describe. The limited language around these measurements makes it difficult for the nonspecialist to assess the reliability and uncertainty of data generated from proxy measurements. Producers and reviewers of proxy data need a common framework for conveying proxy measurement methodology, uncertainty, and applicability for a given study.

We develop a functional distinction between different forms of measurement based on the different ways that their outputs (values, interpretations) relate to the phenomena they intend to describe (e.g., temperature). Paleotemperature measurements, which are used to estimate temperatures of systems in Earth’s past, serve as a case study to examine and apply this new functional proxy definition. We explore the historical development and application of two widely used paleotemperature proxies, calcite δ18O and TEX86, to illustrate how different measurements relate to the phenomena they intend to describe. Both proxies are vulnerable to causal factors that interfere with their relationship with temperature but address those “confounding causal factors” in different ways. While the goal of proxy development is to fully identify, quantify, and calibrate to all confounding causal factors, the reality of proxy applications, especially for past systems, engenders unavoidable and potentially significant uncertainties. We propose a framework that allows researchers to be explicit about the limitations of their proxies and identify steps for further development. This paper underscores the ongoing effort and continued need for critical examination of proxies throughout their development and application, particularly in Earth’s history, for reliable proxy interpretation.

1 Introduction

Proxy measurements can provide information about otherwise elusive properties of systems in Earth’s past, present, and worlds beyond. With a growing interest in quantitatively measuring these properties more precisely and in new environments, the diversity of proxies has increased dramatically. While “proxy” is often used to differentiate “indirect” (e.g., geochemical, physical) measurements from more “direct” forms of observational measurement, neither of those terms provide insight into the reliability or applicability of different measurements. Even direct forms of measurement can be considered proxies in this sense; all involve some level of observational “indirectness”. Earth scientists are particularly aware of the nuances of measurement applicability – as researchers look farther back in time, the reliability of a measurement (i.e., our understanding of what that measurement represents) typically becomes less certain. A standardized framework for conveying how proxy measurements relate to different systems and phenomena would be widely useful for describing these complex associations to nonspecialists, students, modelers, and other proxy users.

The goal of this paper is to describe how methods of observational measurements differ in the ways their outputs (val-
ues, data, interpretations) relate to the phenomena they intend to describe. All forms of observational measurement are influenced by factors that are not the property being measured. We provide insight into the assumptions behind the interpretation and development of different forms of measurement, with the goal of more clearly describing those assumptions and uncertainties in the context of data interpretations.

We examine paleotemperature measurements, which are used to estimate temperatures of systems in Earth’s past, as a case study given the growing interest in paleoclimate, the diversity of measurements available, and the field’s relationship to unknown changes in the Earth–climate system through time. We propose a theoretical framework and language that can more accurately distinguish different measurement–property relationships, which we hope will lead to more robust measurement calibrations, more transparent measurement outputs, and stronger interpretations. While paleoclimate is the focus below, the ideas described here apply to observational measurements across many fields of science.

2 Functional distinctions for proxy measurements

The placement of measurements in two overarching groups, proxy and direct, is particularly common in climate sciences (NOAA National Centers for Environmental Information, 2020; Jansen et al., 2007). Recent philosophical work points out the need for clarification behind the definition of proxy measurements as indirect and non-proxy measurements as direct and questioned how proxies can provide reliable measurements in spite of such perceived indirectness. While many have referred to oxygen isotopes in calcite (δ18Ocal) as a proxy for temperature and the mercury thermometer as a direct measurement of temperature (NOAA National Centers for Environmental Information, 2020; Jansen et al., 2007), both scientists and philosophers of science have pointed out that neither measurement technique truly represents direct observation (e.g., Ruddiman, 2008; Wilson and Boudinot, 2020). The mercury thermometer measures temperature via the observable thermal expansion of mercury as a function of temperature, while δ18Ocal measures paleotemperature via observable variation of 18O incorporation into calcite (CaCO3) as a function of temperature resulting from the differences in vibrational energies of different oxygen isotopes (i.e., 16O, 17O, 18O). In other words, neither produces a direct measurement of temperature; both rely on the observation of some effect of temperature in a system.

Each of these measurements is also influenced by other non-temperature causal factors. Mercury expansion is not only a function of temperature, but also of the partial pressure of the atmosphere and expansion dynamics of liquid mercury. Similarly, δ18Ocal is influenced by the δ18O of the surrounding water (δ18OH2O; Urey, 1948), the pH of the surrounding water (Spero et al., 1997), and, if biominal-

ized by calcifying organisms, biological kinetic effects on 18O incorporation (Bemis et al., 1998; Ravelo and Hillaire-Marcel, 2007). Given philosophical arguments attuned to the conceptual and epistemic issues regarding different forms of scientific measurement (e.g., Suppes, 1951; Franklin, 1990; Chang, 2004; Van Fraassen, 2010; Wilson and Boudinot, 2020), we propose that proxies differ from other forms of measurement in how they account for these confounding causal factors (CCFs; see the “Glossary of terms”; Wilson and Boudinot, 2020).

Under this definition, non-proxy measurements are those that have been designed and manufactured to eliminate all of the potential effects of known CCFs on the measurement output. Because these non-proxy measures control which parts of the system contribute to the final measurement outputs, we refer to them as controlled measurements (see the “Glossary of terms”). Mercury thermometers, for example, are manufactured with a glass casing that controls the atmospheric pressure within the thermometer. The glass case eliminates variation in non-temperature CCFs (e.g., changes in atmospheric pressure, potential for fluid exchange) such that the measured signal can only represent the phenomenon in question, temperature. The lines on the thermometer are calibrated to the thermodynamic properties of mercury such that a specific volumetric expansion of mercury is a causal result of the specific local temperature. In this way, the mercury thermometer is used to perform a controlled measurement.

While the process is more sophisticated, the digital thermometers more commonly used today also control all known CCFs within the instrument to provide a single calibrated temperature value. For those digital thermometers that use electrical resistance, for example, the built-in computer immediately converts an electrical resistance reading to temperature and is calibrated to effectively remove the influence of non-temperature effects on such resistance, including the composition, length, and width of the metal probe used in the thermometer. Because digital thermometers account for all CCFs that influence the relationship between electrical resistance and temperature in real time, digital thermometers, too, are used to perform controlled measurements.

Proxy measurements are distinct because their process of measurement does not rule out all CCFs (see the “Glossary of terms”). This means that the original signal from the analytical measurement must be subject to further manipulation, such as incorporation into a calibration. Those calibrations are based on the field’s best understanding of the drivers of that measured property and quantitatively attempt to minimize the influence of CCFs to produce a value that represents the phenomena in question (Fig. 1). For example, δ18Ocal is a proxy measurement because δ18Ocal is measured simply as a ratio of 18O to 16O of a calcite sample compared to an isotopic standard, and alone that analytical measurement does not reflect temperature. To measure temperature using δ18Ocal, researchers must use a calibration that incorporates information about other parts of the system that influ-
ence the inclusion of $\delta^{18}$O into calcite, such as the $\delta^{18}$O$_{H_2O}$ of the surrounding water, and any potential biological effects of calcification. Because most proxy applications do not allow the researcher to produce controlled measurements of each of those CCFs, the output from a proxy is at best an “estimate” (i.e., the $\delta^{18}$O$_{calcite}$ proxy measurement produces paleotemperature estimates).

The term “indicator” is often used synonymously with proxy, or even measurement (e.g., “Application of the Ce anomaly as a paleoredox indicator”, German and Elderfield, 1990; “Using fossil leaves as paleoprecipitation indicators”, Wilf et al., 1998; “Stomatal density and stomatal index as indicators of paleatmospheric CO$_2$ concentrations”, Royer, 2001; “Indicator of relative changes in sea surface temperature”, Hollis et al., 2019; “Palaeoecological proxies... include crustacean Ostracoda... their indicator species... are sensitive to deoxygenation and eutrophication”, Yasuhara et al., 2019).

The use of this term for such a wide range of applications highlights the lack of clarity in the existing literature, which eventually leads to a lack of clarity in the dissemination of resulting information. While all measurements do “indicate” the quality of some property, they do so in different ways and are accompanied by quite different levels of reliability and uncertainty. The proposed distinction between proxy and controlled measurements, and within proxy measurements (see below), is aimed to provide clarity to the discussion of measurements and their outputs – and CCFs provide such clarification.

The importance of CCFs for proxy measurements was recognized in the development of the first quantitative paleotemperature proxy, $\delta^{18}$O$_{calcite}$. Harold Urey first described the thermodynamic relationship between $\delta^{18}$O$_{calcite}$ and calcite formation temperatures through a simple linear calibration that relates $\delta^{18}$O$_{calcite}$ to temperature in degrees Celsius (Urey, 1948). Urey discussed two important CCFs influencing the $\delta^{18}$O$_{calcite}$ relationship with temperature that could have changed significantly through geologic time and space, namely the $\delta^{18}$O$_{H_2O}$ of the (mean) global ocean and $\delta^{18}$O$_{H_2O}$ of local waters surrounding the precipitating carbonate. While the early reports posited that global $\delta^{18}$O$_{H_2O}$ changed on long timescales (millions of years) as a result of rock weathering, later work showed that global $\delta^{18}$O$_{H_2O}$ had changed significantly on much shorter timescales (tens of thousands of years) due to fluctuations of global ice volume (Emiliani, 1955). The uncertainty of mean ocean $\delta^{18}$O$_{H_2O}$ is greater farther back in Earth history due to currently unconstrained conditions such as ancient ocean latitudinal gradient effects (i.e., reduced latitudinal temperature gradient and resultant local $\delta^{18}$O$_{H_2O}$, 100 million years ago) and silicate weathering rates (Urey et al., 1951). Most Earth systems have experienced variability through Earth’s history, contributing to increased uncertainty associated with CCFs moving farther back in geologic time. As such, different temporal applications of a single proxy can dramatically change that proxy estimate’s uncertainty.

The potential for unknown CCFs exists even for well-calibrated proxy systems and control measurements (Wilson and Boudinot, 2020). While the mercury thermometer successfully controls for its relevant CCFs, a hypothetical application that reveals a theretofore unknown CCF would lead us to no longer consider the thermometer a controlled measurement, at least until it were manufactured in a way to also remove the effects of that CCF. The potential for the existence of unknown CCFs necessitates cautious interpretations of all measurements, particularly those in development or under new applications. But how exactly are CCFs incorporated into proxies?

### 3 Assessing a proxy

#### 3.1 Situating proxies on a spectrum

CCFs are incorporated into proxy measurements through a calibration equation (Fig. 1), which provides a quantitative representation of the relative influence of each causal factor that contributes to the measured property. Using the calibration, researchers can account for the influence of CCFs and produce an estimate of the phenomenon in question. However, the extent to which calibrations identify and address CCFs differs greatly between proxies and proxy applications.

We place proxy measurements along a spectrum that can illustrate the diversity of how proxies relate to CCFs (Fig. 2a). Controlled measurements, with all CCFs known and controlled for (e.g., mercury thermometer), occupy one end of the spectrum. On the other end of the spectrum are proxy measures that are not (yet) calibrated to directly account for their CCFs such that only a correlation is proposed (correlation-constrained proxy; see the “Glossary of terms”), carrying uncertainty regarding the nature and precise causal influence of associated CCFs. Between the two ends of the spectrum are proxies that have a calibration that accounts for the CCFs’ influence on the measurement.
ibrations that do not quantitatively account for CCFs. On the right, correlation-constrained proxies have the least direct (quantitative) control of the causal factors, with cal-
tive values for CCFs used in the calibration must be inferred from
CCFs, but the researcher cannot measure the CCFs, so the quantita-
proxies also have a calibration that quantitatively accounts for
left have full control of all causal factors. Observation-constrained
accounts for all causal factors). Controlled measurements on the
of a measurement’s calibrations (i.e., how completely a calibration
target property, as CCFs differ in how they can be accounted
Importantly, many calibrations require a combination of
accuracy and precision in a high-altitude, low-
humidity study site as in a low-altitude, high-humidity site.
Ideally, all proxy measurements would eventually develop
controlled measurements. Unfortunately, and particularly in paleo-applications, the certainty ascribed to the mer-
cury expansion calibration is not easily attainable or vali-
dated. Furthermore, even controlled measurements can be
complicated by work in “extreme” environments, where tem-
peratures may exceed the minimum or maximum range to
which the thermometer is calibrated (e.g., beyond the boiling
point of mercury). Thus, how a measurement’s calibration is
developed and utilized determines the situations and uncertainty
for that measurement’s application.

To illustrate the proxy range of the spectrum, we situ-
ate \( \delta^{18}O_{\text{calcite}} \) as either an observation-constrained proxy or an inference-constrained proxy depending on how CCFs are quantitatively accounted for (Fig. 2a). When the \( \delta^{18}O_{\text{H}_2\text{O}} \) value in the temperature calibration derives from an inde-
pendent measurement (proxy or controlled) of the \( \delta^{18}O_{\text{H}_2\text{O}} \)
of the water from which the calcite precipitated, then the proxy is an observation-constrained proxy; values to account for
the CCFs in the calibration derive from empirical obser-
vations (Fig. 2a). These components of the calibration can be accounted for with information from proxy or controlled
measurements, with the latter contributing less uncertainty given the constraints on CCFs in controlled measurements.

On the other hand, in instances in which \( \delta^{18}O_{\text{H}_2\text{O}} \) cannot be measured, such as in deeper-time applications, the re-
searcher must provide an inference (i.e., reasoned approx-
imation) of local \( \delta^{18}O_{\text{H}_2\text{O}} \). Based on the extrapolation of a
well-known system to a lesser-known system, inference-
constrained proxy measurements inherently present a more
biased estimate due to biases in the researchers’ inference
of that system rather than empirical evidence (Fig. 2b). For
example, some researchers have inferred 100-million-year-
old \( \delta^{18}O_{\text{H}_2\text{O}} \) for the \( \delta^{18}O_{\text{calcite}} \) paleotemperature calib-
ration by applying a first-order estimate of \( \delta^{18}O_{\text{H}_2\text{O}} \) based on
certain characteristics of the system in question, such as a
mean \( \delta^{18}O_{\text{H}_2\text{O}} \) value that applies to any “non-glacial world” (O’Brien et al., 2017). Researchers modified this mean value to
represent the \( \delta^{18}O_{\text{H}_2\text{O}} \) of local waters (where calcite was precipitated) by adjusting the mean \( \delta^{18}O_{\text{H}_2\text{O}} \) based on
modern latitudinal \( \delta^{18}O_{\text{H}_2\text{O}} \) variability (e.g., O’Brien et al.,
2017). This inference is still based on quantitative measure-
ments (e.g., modern \( \delta^{18}O_{\text{H}_2\text{O}} \) latitudinal trends) but requires
several inferences that assume that two systems are similar
(i.e., all ice-free oceans in Earth’s history are isotopically
similar; latitudinal \( \delta^{18}O_{\text{H}_2\text{O}} \) variability is similar between
100 million years ago and the present). Because that inference
is accompanied by uncertainty that is not easily quantifi-
able (e.g., uncertainty associated with assumptions made by
the researcher rather than analytical uncertainty; see below),
the potential uncertainty for inference-constrained proxies is
larger than those that are observation-constrained.

Importantly, many calibrations require a combination of
inference and observation to produce a final estimate of the
target property, as CCFs differ in how they can be accounted
for. In other words, many proxy applications use both obs-
eration and inference constraints to satisfy a calibration.

Figure 2. A spectrum (x axis) of observational measurements as a function of their incorporation of confounding causal factors and re-
related uncertainty. (a) The bottom y axis describes the completeness of a measurement’s calibrations (i.e., how completely a calibration
accounts for all causal factors). Controlled measurements on the left have full control of all causal factors. Observation-constrained
proxies have a calibration that quantitatively accounts for CCFs and allows the researcher to measure those CCFs. Inference-constrained
proxies also have a calibration that quantitatively accounts for CCFs, but the researcher cannot measure the CCFs, so the quantita-
tive values for CCFs used in the calibration must be inferred from
other evidence. On the right, correlation-constrained proxies have the least direct (quantitative) control of the causal factors, with cal-
britures (e.g., uncertainties associated with assumptions made by

https://doi.org/10.5194/cp-16-1807-2020
Moving further away from controlled measurements on our spectrum, we find proxy measurements that are correlated with temperature, but the CCFs are not fully or quantitatively accounted for in a calibration; here, the CCFs are unknown (or roughly understood), though a correlative relationship is identified. It is functionally impossible to accurately assess the uncertainty of estimates produced by these measurements (Fig. 2b), as the causal factors influencing the measurement are not quantitatively represented in a calibration. Not only could the signal from such a correlation-constrained proxy be partially driven by some unknown CCF; it could even be entirely driven by CCFs (e.g., Junium et al., 2018) but would be interpreted as driven by the property in question.

An example of a correlation-constrained proxy is the present incarnation of the TEX$_{86}$ paleotemperature proxy. In 2002, researchers identified a suite of sedimentary hydrocarbons that shared a similar structure but contained a different number of cyclic moieties (Schouten et al., 2002; Fig. 2). Relative abundances of these isoprenoidal glycerol diether glycerol tetraether (isoGDGT) compounds with different cyclic moieties were represented by a ratio (Table 1). When these compounds were recovered from modern sediments and this ratio was calculated, a clear correlation with the surface water temperature at the sample location was identified. In other words, the number of cyclic moieties in the sedimentary isoGDGTs was correlated with the surface water temperatures at the location where they were found. Using statistical (regression) analyses of a suite of modern sediments and sea surface temperature measurements, a calibration was produced, and the authors proposed this molecular ratio as a quantitative paleotemperature proxy (Schouten et al., 2002). A physiological response was posited to explain the relationship—fewer cyclic moieties contributed to a more malleable lipid membrane, which would be advantageous in cooler waters.

In the ensuing years, several questions about the origin and implications of these molecules were raised. They seemed to be produced predominantly by Thaumarchaeota, a type of marine archaea that live well below the sea surface (Schouten et al., 2000) where the temperature correlation was strongest. Additionally, field and culture observations from variable environments produced different calibrations (i.e., different slopes and y intercepts to describe the correlation between the isoGDGT ratio and temperature; Table 1) and even different ratios (e.g., TEX$_{86}$ for low-temperature regions; Table 1). If the ratio of isoGDGT cyclicity directly represented temperature, then why would that ratio be different depending on the study design, location, and time period? And if the calibration accurately accounted for the CCFs contributing to the effect of temperature on isoGDGT cyclicity, why would it be different from place to place?

These questions are driving fundamental research in understanding the mechanistic relationships between TEX$_{86}$ and temperature. Several important advances in this mechanistic understanding have already been produced: culture and field experiments have shown that the cyclic moieties represent a metabolic response to energy demands, growth phase, nutrient availability, and ecosystem composition rather than solely a physiological response to temperature (Elling et al., 2014; Qin et al., 2015; Hurley et al., 2016; Polik et al., 2018). These studies advance TEX$_{86}$ beyond the correlative relationship (i.e., colder temperatures make more cyclic moieties) into a nuanced, yet more accurately representative, understanding of all causal factors and their mechanisms (i.e., relationship between sea surface temperatures and nutrient and oxygen availability, which impacts archaeal metabolic energy demands). However, while work on TEX$_{86}$ drivers suggests that non-temperature factors cause variations in isoGDGT cyclization, TEX$_{86}$ application studies continue to report a specific temperature value. The argument behind continued TEX$_{86}$ applications is the calibration of ammonia oxidation rates and temperature in most modern settings (Hurley et al., 2016). However, many studies have suggested that ammonia or oxygen concentrations in past environments likely varied in a way that did not correlate with temperature (e.g., Liu et al., 2009; Polik et al., 2018). This proxy’s CCFs need full consideration in experimental design and interpretation for it to be truly quantitative and its uncertainty appropriately reported.

### 3.2 Discussing proxy data

A clear distinction should be made between various forms and degrees of uncertainty related to proxy measurements (see the “Glossary of terms”). All proxy measurements are the result of some analysis (e.g., $\delta^{18}$O$_{\text{calcite}}$ as the normalized ratio of $^{18}$O to $^{16}$O in a sample) and incorporation into a calibration (e.g., $\delta^{18}$O$_{\text{calcite}}$ as a function of temperature, $\delta^{18}$O$_{\text{H2O}}$, and biological effects; Fig. 1), from which three forms of uncertainty derive. The first is analytical uncertainty, which is simply the uncertainty associated with the precision and accuracy of the analytical measurement. For oxygen isotopes in calcite, this would include the isotope ratio mass spectrometer’s precision and accuracy when determining the ratio of $^{18}$O to $^{16}$O of a sample normalized to a standard. We argue that analytical uncertainty can always be quantified using standards and is distinct from unquantifiable uncertainties. Unquantifiable uncertainties associated with calibration (including unknown CCFs), as well as sample preparation and analysis, and are grouped into potential uncertainties (Fig. 2b). The distinction between factors that fall into potential versus analytical uncertainty is defined by quantitation. Researchers take many steps to quantify errors and uncertainties associated with sample preparation and analysis. When employed, such efforts reduce the potential uncertainty and more accurately reflect that analytical uncertainty. For example, hydrocarbon standards might be incorporated into a sedimentary sample before hydrocarbon extraction such that the researcher can quantify if any hydro-
calibrations (Table 1), and the lack of codified reporting stan-
practice. For example, the variety of isoGDGT ratios and
for many proxies, the reported uncertainty varies widely in
which should ideally cover (either quantitatively or in discus-
can be quantified (Fig. 2b).
always be higher than an observation-constrained proxy, for
might not be quantifiable (i.e., logical deductions might not
be incorporated to that loss and alteration, more transparently reflecting the
uncertainty in the analysis. However, some potential uncer-
tainties will always exist in a nonquantifiable manner, such as
unknown CCFs or unmeasurable changes in CCFs through
time. Because the error in an inference-constrained proxy
might not be quantifiable (i.e., logical deductions might not
have a quantifiable uncertainty), its potential uncertainty will
always be higher than an observation-constrained proxy, for
which the analytical uncertainty of the CCF measurement
can be quantified (Fig. 2b).
The final type of uncertainty is the reported uncertainty,
which should ideally cover (either quantitatively or in discus-
sion) both analytical and potential uncertainties. However,
for many proxies, the reported uncertainty varies widely in
practice. For example, the variety of isoGDGT ratios and calibrations (Table 1), and the lack of codified reporting stan-
dards used in the expression of TEX$_{86}$-derived paleotemper-
atures, leads to notable variability in the reported uncertainty
associated with TEX$_{86}$. Some TEX$_{86}$-derived paleotemperature estimates are plotted without error bars and are
accompanied by an in-text discussion of the analytical uncer-
tainty from calibration and replicate analyses (e.g., Woelders
et al., 2017), while the analytical uncertainty for others is
not discussed (e.g., Slujis et al., 2006). For some estimates,
the analytical uncertainty derived from only the calibration
is provided (e.g., Hollis et al., 2012; Ho et al., 2014). Ana-
lytical uncertainties from replicate analyses have been com-
bined with the analytical uncertainties of calibration statistics as error windows on plots (e.g., Tierney et al., 2010;
Shevenell et al., 2011), while discussion of potential uncer-
tainties, such as changes in the known (but not calibrated-
to) CCFs, varies greatly between reports (e.g., Tierney et al.,
2010; Shevenell et al., 2011). Because potential uncertainty
is by definition unquantifiable, it might not be incorporated

<table>
<thead>
<tr>
<th>Range ($^\circ$C)</th>
<th>Equation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–30</td>
<td>$T = (\text{TEX}_{86} - 0.27)/0.015$</td>
<td>Schouten et al. (2002)</td>
</tr>
<tr>
<td>22–30</td>
<td>$T = (\text{TEX}_{86} - 0.016)/0.027$</td>
<td>Schouten et al. (2003)</td>
</tr>
<tr>
<td>10–28</td>
<td>$T = (\text{TEX}_{86} - 0.2)/0.016$</td>
<td>Sluijs et al. (2006)</td>
</tr>
<tr>
<td>5–30</td>
<td>$T = -10.78 + 56.2 \times \text{TEX}_{86}$</td>
<td>Kim et al. (2008)</td>
</tr>
<tr>
<td>25–28</td>
<td>$T = (\text{TEX}_{86} + 0.09)/0.035$</td>
<td>Tromer et al. (2009)</td>
</tr>
<tr>
<td>−3–30</td>
<td>$T = 50.475 - 16.332 \times (1/\text{TEX}_{86})$</td>
<td>Liu et al. (2009)</td>
</tr>
<tr>
<td>−3–30</td>
<td>$T = 81.5 \times \text{TEX}_{86} - 26.6$</td>
<td>Kim et al. (2010)</td>
</tr>
<tr>
<td>−3–30</td>
<td>$T = -19.1 \times (1/\text{TEX}_{86}) + 54.5$</td>
<td>Kim et al. (2010)</td>
</tr>
<tr>
<td>−3–30</td>
<td>$T = 49.9 + 67.5 \times (\text{GDGT index-1})$</td>
<td>Kim et al. (2010)</td>
</tr>
<tr>
<td>5–30</td>
<td>$T = 38.6 + 68.4 \times (\text{GDGT index-2})$</td>
<td>Kim et al. (2010)</td>
</tr>
<tr>
<td>10–40</td>
<td>$T = 48.2 \times \text{TEX}_{86} + 1.04$</td>
<td>Kim et al. (2010)</td>
</tr>
<tr>
<td>10–40</td>
<td>$T = -9 \times (1/\text{TEX}_{86}) + 45.2$</td>
<td>Kim et al. (2010)</td>
</tr>
<tr>
<td>10–40</td>
<td>$T = 42.9 \times (\text{GDGT index-1}) + 46.5$</td>
<td>Kim et al. (2010)</td>
</tr>
<tr>
<td>10–40</td>
<td>$T = 52 \times (\text{GDGT index-2}) + 42$</td>
<td>Kim et al. (2010)</td>
</tr>
<tr>
<td>4–30</td>
<td>$T = -14 + 55.2 \times \text{TEX}_{86}$</td>
<td>Powers et al. (2010)</td>
</tr>
<tr>
<td>10–30</td>
<td>$T = 3.5 + 38.9 \times \text{TEX}_{86}$</td>
<td>Tierney et al. (2010)</td>
</tr>
<tr>
<td>−2–30</td>
<td>$T = (\text{TEX}_{86} - 0.3038)/0.0125$</td>
<td>Shevenell et al. (2011)</td>
</tr>
<tr>
<td>14–34</td>
<td>$T = 32.873 \times \ln(\text{GDGT index-1}) + 50.771$</td>
<td>Hollis et al. (2012)</td>
</tr>
<tr>
<td>14–34</td>
<td>$T = 39.036 \times \ln(\text{TEX}_{86}) + 36.455$</td>
<td>Hollis et al. (2012)</td>
</tr>
<tr>
<td>15–35</td>
<td>$T = (\text{TEX}_{86} - 0.21)/0.015$</td>
<td>Qin et al. (2015)</td>
</tr>
<tr>
<td>10–30</td>
<td>$\text{TEX}_{86} = -0.00067T^2 + 0.023T + 0.33$</td>
<td>Qin et al. (2015)</td>
</tr>
<tr>
<td>10–25</td>
<td>$\text{TEX}_{86} = -0.0017T^2 + 0.054T + 0.11$</td>
<td>Qin et al. (2015)</td>
</tr>
<tr>
<td>2–10</td>
<td>$T = 27.898(\text{TEX}_{86}) + 22.723$</td>
<td>Harning et al. (2019)</td>
</tr>
</tbody>
</table>

Table 1. Compilation of TEX$_{86}$ calculations and calibrations as of 2020. Modified from Tierney (2012).
into quantitative data presentation styles, such as Cartesian plots, but can certainly be discussed in light of the existing work on TEX$_{86}$ CCFs.

Importantly, researchers have taken steps to communicate the reliability of proxy data relative to other measurements in reviews, conference sessions, and proxy assessment compilations (e.g., Ravelo and Hillaire-Marcel, 2007; Newman et al., 2016; Hollis et al., 2019; Wilson and Boudinot, 2019). For example, the Paleoclimate Modelling Intercomparison Project (PMIP) appraisal of proxy data for the Intergovernmental Panel on Climate Change (IPCC) reports (Hollis et al., 2019) provides an in-depth description of the paleotemperature proxies used to inform the IPCC reports. The appraisal describes each proxy’s theoretical background, which gives data generators and modelers a better understanding of the biogeochemical processes that relate each proxy to temperature. The assessment then describes strengths and weaknesses of each proxy relative to the other measurements, which can guide users in determining which proxy may be best suited for a given study, as well as providing considerations for the interpretation of the resulting data. Finally, the assessment provides “recommended methodologies”, which includes analytical recommendations, a single recommended calibration, and other best practices for reporting proxy data and interpretations. By providing a consensus presentation of recommended methodologies, the PMIP proxy assessment and similar projects constitute an important means for standardizing data assessment and reporting, as well as guiding proxy users in developing study designs. The framework presented here will improve those methods by providing direct language (e.g., CCFs, types of uncertainty) to more clearly navigate discussions of proxy assessments.

A complete outline of potential uncertainties and the often complex phenomena–measurement relationships is difficult to incorporate into grants, peer-reviewed manuscripts, and educational programs. The lack of extensive discussion of a proxy’s uncertainty can lead to an oversimplification of these relationships (i.e., an under-consideration for CCFs and uncertainties). However, detailing how proxies might relate to some unknown CCFs (as is done here) can make any proxy seem subject to countless unknown CCFs, which may engender an unwarranted dismissal of proxy data interpretations. Because proxy data inform models, manuscripts, and educational lessons, there needs to be a more universally accepted and functional means of discussing and conveying proxy uncertainty that is honest yet robust. Our spectrum of proxy measurements relates measurements to their CCFs, and thus the spectrum and language provide such a means of conveying uncertainty in a universal way.

Many studies, for example, have shown that TEX$_{86}$ trends were driven by changes in nitrogen availability and marine ecology in some paleo-environments (Liu et al., 2009; Hurley et al., 2016; Junium et al., 2018; Polik et al., 2018). How can researchers be sure that TEX$_{86}$ is not driven by these dynamics in other settings, unless those CCFs of nitrogen availability and marine ecology changes are directly assessed? Because uncertainties in estimating these environmental characteristics are often not incorporated (as they are not incorporated in the current litany of quantitative TEX$_{86}$ calibrations; Table 1), we have described the potential uncertainty of TEX$_{86}$ (and other correlation-constrained proxies) as much higher than is often reported (Fig. 2b). By referring to TEX$_{86}$ as a correlation-constrained proxy, modelers, reviewers, and researchers can immediately be aware of this underreporting of uncertainty, which would inform their interpretation of the temperature estimates produced by TEX$_{86}$ in a meaningful yet succinct way.

### 3.3 Development of a proxy

Proxy development is the production and improvement of a calibration that quantitatively accounts for all CCFs that contribute to the measured signal. The controlled characteristic of a mercury thermometer allows the measurement of temperature without needing an external calibration, as the temperature lines are calibrated to the exact expansion of mercury within the glass walls. Prior to the full calibration of the lines on the mercury thermometer, mercury might have served as a proxy: a gram of mercury on a table would expand and contract with fluctuating temperatures, which could be a qualitative, correlation-constrained proxy for temperature (the mercury expanded, so the temperature likely got hotter).

Because proxy measurements do not account for the influence of all known CCFs, quantitative proxy measurements require some external calibration equation to produce reliable estimates. Calibrations express the relative effect of each causal factor (Fig. 1) and provide insight into the applicability of a proxy by addressing the range in which the calibration is useful and the natural variability (uncertainty) associated with that calibration. Proxy applications are limited to the range in which that proxy has been studied and calibrated; applications outside that range do not produce reliable estimates.

Harold Urey’s first description of the thermodynamic relationship between $\delta^{18}$O$_{\text{calcite}}$ and calcite formation temperatures was simply “The calculated slope, 4.4 per mil between 0 and 25 °C” (Urey, 1948). More complex calibrations now exist for the $\delta^{18}$O$_{\text{calcite}}$ paleotemperature proxy, which accounts for its numerous CCFs including $\delta^{18}$O$_{\text{H}_2\text{O}}$ and biological effects (Ravelo and Hillaire-Marcel, 2007; Hollis et al., 2019). While the $\delta^{18}$O$_{\text{calcite}}$ proxy is far from a controlled measurement, its historical development exemplifies the consistent work to make proxies more like controlled measurements, i.e., to eliminate or limit the influence of CCFs. But what does such proxy development look like in practice?

The first step of proxy development is the identification of some corollary relationship between a measurable property (e.g., $\delta^{18}$O of calcite) and a property unable to be measured in a controlled fashion (e.g., temperature of a past environment). At first order, these are usually qualitative and based...
on some hypothesis to describe a system. Mercury expands with increasing temperature due to general fluid dynamics; \(^{18}\text{O}\) is more favorably incorporated into calcite at lower temperatures due to differences in vibrational energies between \(^{18}\text{O}\) and \(^{16}\text{O}\); some organisms alter their cell membranes to maintain homeostasis in variable environments.

Proxies that are based on such a corollary relationship can serve as qualitative proxy measures, which provide useful comparative or relative information. This is the case for some paleotemperature proxies: geological evidence of glacial expansion and retreat in a certain location can indicate relative local temperature change, but variability in numerous (difficult or impossible to constrain) CCFs prohibits a calibration to quantitative temperature changes in degrees Celsius. Such comparative information is appropriate for many paleo-studies, wherein the question is focused on trends and relative changes through time or differences between sites. This corollary relationship can lead researchers into an “optimism phase”, wherein the assumption of a direct cause–effect relationship between a phenomenon and an observation makes users optimistic that a proxy can be used with confidence (Elderfield, 2002).

If researchers aim to use a proxy quantitatively, the relationship between the target property (e.g., temperature), the observable property (e.g., \(^{18}\text{O}_{\text{calcite}}\)), and all CCFs must be accounted for in a calibration (Fig. 1). Quantitative proxies require an (empirically derived) estimation or (logically deduced) inference of the influence of all CCFs represented in a calibration. Calcite precipitation experiments with variable pH, \(^{18}\text{O}_{\text{H}_2\text{O}}\), salinity, and biomineralizing organisms have contributed to calibrations that include those CCFs and represent how they contribute to \(^{18}\text{O}\) incorporation into calcite (Ravelo and Hillaire-Marcel, 2007). Studies using those calibrations must account for those CCFs. For example, calcite-producing organisms live in either bottom waters or surface waters – the temperature from the two will not only have slightly different CCFs, but will also reflect temperature from different parts of the water column. Researchers would identify the type of organism to know where it lived and would address the CCFs specific to that organism (e.g., Bemis et al., 1998). The process of testing CCFs must be extensive to provide confidence in the proxy. Often, this phase of development unearths unforeseen CCFs, such as the role of water-column oxygenation in isoGDGT cyclicity (Qin et al., 2015; Hurley et al., 2016). While some have argued that this can lead to a “pessimism phase”, wherein proxy users might no longer have confidence in that proxy’s utility (Elderfield, 2002), in fact these revelations are essential to proxy development – it is the scientific method at work, and such exhaustive testing of CCFs is a prerequisite for the confident use of a proxy.

The identification and testing of CCFs represent an inherently iterative processes. Urey and others made serious consideration of CCFs before applying the \(^{18}\text{O}_{\text{calcite}}\) paleotemperature proxy. It was proposed that the proxy be used only “if the isotopic composition of the water is known not to differ from the mean of the present seas, or... in the case that it does [differ], if both the isotopic composition of the carbonate and water are determined” (Urey et al., 1951). Urey described local variability in \(^{18}\text{O}_{\text{H}_2\text{O}}\) due to evaporation and salinity as “the greatest difficulty” for accurate temperature measurements but promised that “this problem is being studied from several angles and it is hoped that corrections can be applied in the future” (Urey et al., 1951). Urey’s careful consideration of CCFs, and the subsequent and ongoing investigations into those CCFs, serves as an exemplar for proxy discussion, interpretation, and development.

Sometimes, the development of one proxy can constrain a CCF for another proxy by providing a new means of estimating that CCF. The development of the Mg/Ca paleotemperature proxy, based on the incorporation of magnesium relative to calcium in foraminiferal calcite, provided an independent constraint on temperature at the same time (i.e., mid-1990s) that \(^{18}\text{O}_{\text{calcite}}\) was being developed as a paleotemperature proxy (Hastings et al., 1998). By using Mg/Ca to estimate temperature in the same setting as \(^{18}\text{O}_{\text{calcite}}\), researchers were able to independently constrain temperature and thus use \(^{18}\text{O}_{\text{calcite}}\) to estimate \(^{18}\text{O}_{\text{H}_2\text{O}}\) (Mashiootta et al., 1999). The development of two independent paleothermometers, each with their own CCFs, provided researchers with new opportunities and greater confidence in applying those proxies; \(^{18}\text{O}_{\text{calcite}}\) and Mg/Ca combined helped to identify the degree to which \(^{18}\text{O}_{\text{H}_2\text{O}}\) influenced the \(^{18}\text{O}_{\text{calcite}}\) proxy and resulted in a new means to constrain the CCF of \(^{18}\text{O}_{\text{H}_2\text{O}}\) for future studies. Similarly, multiple studies have compared temperature estimates from TEX\textsubscript{86} with other organic (e.g., alkenones; Huguet et al., 2006; Lee et al., 2008; Li et al., 2013) and inorganic (e.g., Mg/Ca and \(^{18}\text{O}_{\text{calcite}}\); e.g., Hollis et al., 2012; Hetzberg et al., 2016; O’Brien et al., 2017) proxies in the same settings. While those multi-proxy comparative studies are helping to identify CCFs related to TEX\textsubscript{86} and other paleotemperature proxies, the numerous unconstrained CCFs related to TEX\textsubscript{86} make direct testing of CCFs difficult for even those comparative studies. For example, are deviations between \(^{18}\text{O}_{\text{calcite}}\) and TEX\textsubscript{86} due to depth of production in the water column (e.g., Li et al., 2013; Hetzberg et al., 2016), production season (Huguet et al., 2006), or some other CCF like nutrient availability (Hurley et al., 2016)? Some TEX\textsubscript{86} applications have used independent proxies to constrain CCFs related to the environment, such as the use of the BIT index (Hopmans et al., 2004) to estimate changes in the input of isoGDGTs from nonmarine sources (e.g., Weijers et al., 2006; Hollis et al., 2012). Future work integrating the physiological CCFs associated with TEX\textsubscript{86}, such as changes in water-column oxygenation (Qin et al., 2015) and nutrient availability (Hurley et al., 2016), into such multi-proxy comparisons could further constrain the role of different CCFs in TEX\textsubscript{86} paleotemperature estimates.
Alternatively, the use of statistical methods can elucidate CCFs and their impact on proxy measurements. One example is the Bayesian statistical modeling approach, which uses existing data (usually field-produced calibrations) over a wide range of environments to produce a “best-fit” calibration for the range of values measured in a given study. The resulting model allows researchers to identify which environments and/or locations produce a calibration that best fits their data and thus provides a means to investigate environmental conditions and the related CCFs that more fully express the relationship between, for example, TEX$_{86}$ and temperature (Tierney and Tingley, 2014). In fact, the PMIP proxy assessment (Hollis et al., 2019) recommends that TEX$_{86}$ users utilize the Bayesian calibration fit as the best current means to estimate paleotemperatures (Hollis et al., 2019), demonstrating how the field may use these statistical methods to provide best practices for measurement applications. Similarly, stochastic modeling approaches are used in hydrological data interpretations as a means to estimate the partial effects (or confounding effects) of different causal factors contributing to a given signal (Yevjevich, 1987), and such approaches could be utilized by the paleotemperature community.

Additionally, the application of transfer functions, including proxy system models, is used to make inferences about CCFs. Transfer functions provide a theoretical (rather than empirical) constraint on a system’s properties in an attempt to predict the quality of properties rather than observe them (Telford and Birks, 2005). While the reliability of transfer functions is an area of active discussion (e.g., Telford et al., 2004, 2013), transfer functions represent yet another statistical approach used to account for CCFs in lieu of empirical observations and are employed by some to reduce uncertainty for correlation- and inference-constrained proxies. For example, proxy system models use transfer functions to provide an assessment of proxy–phenomenon relationships and the driving mechanisms behind proxy measurement outputs (e.g., Dee et al., 2016, 2018; Okazaki and Yoshimura, 2019). These statistical methods are an important aid in the determination of CCFs on observational signals and can be powerful in the development of proxy calibrations.

Ultimately, a mix of variable-controlled laboratory experiments, statistical analyses, and field validation experiments all contribute to proxy development. The identification and expression of corollary relationships in a statistical regression represent only the first step. Comparisons between laboratory (e.g., culture) experiments and field measurements might produce different calibrations; causes for differences in the regression should be investigated. For TEX$_{86}$, the recognition of significant variability amongst field calibrations led researchers to investigate non-temperature properties, such as physiological effects of Thaumarchaeota, in variable-controlled in-laboratory culture experiments (e.g., Elling et al., 2014; Qin et al., 2015; Hurley et al., 2016). In response, field studies of isoGDGT cyclization were performed in modern and paleo-settings (e.g., Hurley et al., 2016; Ju-
4 Conclusions

The distinction between controlled and proxy measurements, and within proxy measurements, serves a more functional role for interpreting, assessing, and developing proxies than previous distinctions between proxy and “direct” measurements. The language proposed here concerning proxy calibrations (e.g., observation- versus inference-constrained proxy) and uncertainty (e.g., analytical versus potential) succinctly and directly addresses the relationship between measurements and the property they intend to describe and more clearly directs proxy calibration development. Using this language, modelers can more confidently appropriate proxy data outputs into their models, researchers can more efficiently design studies to produce robust measurements, reviewers can more easily assess the reporting of uncertainty and interpretations, and educators can more clearly convey the differences in measurements available for students to learn from, apply, and improve. Readers may find that observational measurements not typically considered proxy measurements in their field may in fact fall on the proxy end of our spectrum. We hope that such realizations might drive researchers to investigate what has been taken for granted in previous interpretations or how future study designs can more accurately assess and account for CCFs. Ultimately, we propose that as much can be learned about a system by developing a proxy as can be learned by applying it.
### Appendix A: Glossary of terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confounding causal factors (CCFs)</td>
<td>Characteristics of an environment that affect the output of a measurement but are not the property being measured</td>
</tr>
<tr>
<td>Controlled measurement</td>
<td>Measurement that has been manufactured or designed to eliminate the potential effects of all known CCFs on the measurement output</td>
</tr>
<tr>
<td>Proxy measurement</td>
<td>Measurement that does not eliminate the influence of all known CCFs on the intended or targeted property</td>
</tr>
<tr>
<td>Observation-constrained proxy</td>
<td>Proxy measurement for which the CCFs are quantitatively incorporated into a calibration and are accounted for with values produced by other proxy measurement estimates or controlled measurements</td>
</tr>
<tr>
<td>Inference-constrained proxy</td>
<td>Proxy measurement for which the CCFs are quantitatively incorporated into a calibration and are qualitatively accounted for using a reasoned approximation (inference) of the value based on comparisons to similar systems, rather than values produced by measurements of the system in question</td>
</tr>
<tr>
<td>Correlation-constrained proxy</td>
<td>Proxy measurement that does not account for known CCFs but is based on a hypothesized relationship between a certain property and a measurement output; uses a calibration that does not quantitatively represent the causal structure of the system</td>
</tr>
<tr>
<td>Analytical uncertainty</td>
<td>The uncertainty associated with the precision and accuracy of the analytical instrument</td>
</tr>
<tr>
<td>Potential uncertainty</td>
<td>The degree to which the measurement or estimated value is affected by something other than the property being measured</td>
</tr>
<tr>
<td>Reported uncertainty</td>
<td>A textual and/or numerical representation of the combined analytical and potential uncertainties associated with a measurement</td>
</tr>
</tbody>
</table>
Data availability. All data described here are presented in previously published literature and are cited as such.

Author contributions. JW and FGB designed the study. FGB wrote the paper, and JW and FGB edited the paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank the Center for the Study of Origins for their support. We thank Thomas Marchitto, Gifford Miller, Benjamin Johnson, Matthew Huber, and F. Douglas Boudinot for their helpful comments that improved the paper and Howard Spero, Julio Sepúlveda, Carol Cleland for discussions that improved the paper. F. Garrett Boudinot acknowledges the Department of Geological Sciences at the University of Colorado Boulder (NSF Division of Earth Sciences Earth-Life Transitions – ELT – program grant no. 1338318) and the American Chemical Society Petroleum Research Fund (ACS-PRF; Doctoral New Investigator Award no. 58815-DN12) for their support. Joseph Wilson acknowledges the Department of Philosophy and the Graduate School at the University of Colorado Boulder for their support.

Review statement. This paper was edited by Alberto Reyes and reviewed by Julie Griffin and one anonymous referee.

References


