12-2021

Machine Learning Applied to a Modern-Pleistocene Petrographic Dataset: The Global Prediction of Sand Mineralogy (GloPrSM) Model

Isaac Johnson
University of Arkansas, Fayetteville

Machine Learning Applied to a Modern-Pleistocene Petrographic Dataset: The Global Prediction of Sand Mineralogy (GloPrSM) Model

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Geology

by

Jah Isaac Johnson
Trinity University
Bachelor of Science in Geosciences, 2017

December 2021
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

_____
Glenn R. Sharman, Ph.D.
Thesis Director

_____
Eugene Szymanski, Ph.D.
Committee Member

_____
Xiao Huang, Ph.D.
Committee Member

_____
Matthew Covington, Ph.D.
Committee Member

Abstract

Petrography has long been used as a tool to decipher the sedimentary provenance of sand and sandstone from the relative proportions of framework grain types. Petrographers have also related the proportions of quartz (Q), feldspar (F), and lithic (L) grains to the processes that form and modify sediments within sediment routing systems. This past work has shown that factors including source lithology, climate, transport history, and tectonism work in concert to modify the framework mineralogy of sand. However, there is a lack of a quantitative understanding of the interactions and feedbacks between these factors and how they modify sand mineralogy. This research aims to establish a predictive framework that constrains the relationship between sand framework grain mineralogy and the factors that influence it, including bedrock lithology, topography, and climate. Specifically, this study asks, "to what degree can the final modal composition of sand be predicted if the boundary conditions that generate sediments are known?".

This question is investigated by analyzing a globally extensive modal point count dataset of 3,522 Pleistocene to modern sand samples from 51 published sources. A petrographic data model was created to standardize 287 reported petrographic labels to a final list of 54 labels. An inline series of random forest (RF) machine learning algorithms were trained on a subset of 3,208 fluvial and marine samples whose boundary conditions are known with a high degree of confidence. Data for precipitation, temperature, elevation, slope, basin area, and seven generalized source lithologies were extracted from sample catchments and used to train 100 RF meta-estimators that predict the logarithms F:Q and L:Q ratios as well as eight Q-F-L subcompositions, resulting in R^2 scores of 0.654 ± 0.031 (1-sigma) and 0.706 ± 0.023 (1-sigma)

for ln(F/Q) and ln(L/Q) models, respectively. Mean Q-F-L prediction error within one standard deviation is 2.6% ± 15% for Q, -1.1 ± 9.4% for F, and -1.5% ± 15.4% for L.

The Global Prediction of Sand Mineralogy (GloPrSM) model was generated by applying the 100 RF meta-estimators to a global dataset of fluvial watersheds (mean area of ~1,500 km^2). The resulting Q-F-L prediction includes an estimate of spatial uncertainty based upon variability in the 100 predictions. In general, the GloPrSM model predicts quartz enrichment in low latitudes (35°N to 35°S), feldspar enrichment near plutonic and metamorphic crystalline terranes in middle to high latitudes, and lithic enrichment near active margins and flood basalts. Low model confidence is exhibited in catchments draining large igneous provinces, in sedimentary terranes in middle to high latitudes, and in orogenic settings. Feature importance algorithms reveal that slope, temperature, metamorphic source abundance, and felsic to intermediate plutonic source abundance are the most important predictors of Q-F-L composition. In addition, partial dependence analysis suggests temperatures higher than 15 ºC and large drainage areas favor quartz enrichment, while steeply sloping environments favor lithic enrichment. The GloPrSM model represents the first, global-scale estimate of sand mineralogical proportions, and illustrates that the spatial distribution of Q-F-L at Earth's surface can be predicted from the first-order factors that generate sediments.

Acknowledgements

I will never be able to fully express the gratitude I feel at this time toward the countless number of individuals that have made it possible for me to compose even this section of my master's thesis. Without writing an entirely separate thesis of acknowledgements, I would first like to thank Dr. Glenn Sharman whose brilliant ideas have led to the final product that is the subject of this document. Admittedly, I was doubtful this research topic would make it as far as it has. From my earliest conversations with Glenn, I recall thinking, "who cares about sand composition and why would anyone want to predict it?" At the time, I was a young, immature, lithic sand moving rapidly downslope. Following my matriculation at the University of Arkansas, I've spent significant time residing in alluvial storage under some of the most intense weathering conditions I've encountered to date. Under Glenn's tutelage, I've become rounded, sorted, and enriched with the ever-increasing confidence to embrace challenging new experiences at the confluence of life's rivers on my way to becoming pQ. This project has reignited my passion for the geosciences and has been worth the many long days and nights spent working on it. Thank you, Glenn, for providing me this opportunity, and for constantly pushing me to think critically as we explore the boundaries of our field.

Thank you to my committee members Dr. Eugene Szymanski, Dr. Matthew (Matt) Covington, and Dr. Xiao Huang for the invaluable feedback they have provided on this project. Eugene has always asked the most thought-provoking questions and helped me conceptually approach my project in novel ways. Through Matt's counsel, I have fundamentally become a better programmer and modeler. Xiao's expert guidance has helped me build a foundation from which I've have been able to and will continue to establish myself as a machine learning practitioner.

Thank you to members of the Detrital Geochronology Laboratory (DGL) consortium, including Chevron and California Resources Corporation, for funding DGL research, and to Dr. Bill Heins for his advice on placing the broader applications of this research in the context actualistic provenance studies. Additionally, Dr. Jonathan Sharman provided assistance with Python coding, particularly determining how to assign classification labels based on ternary composition, and Dr. Kevin Befus provided key suggestions that facilitated automated upstream watershed mapping.

A special thank you goes to Dr. Rohit Goswami, P.E., for his patience as I've worked toward my degree, for his trust in my abilities, and for being my biggest advocate in the professional world.

Thank you to my closest friends who have helped me recharge from the cyclical exhaustion I've felt in every walk of life, and for the endlessly great memories that give me the strength to face each new day. This includes the Sharmanites, Goon Squad, the metalheads (the Monkey and the Aussie), a Daring Physicist, a Hydrogeoarcheologist, and those unmentioned that have been at my side through thick and thin.

I would like to thank my parents, Leann and Scott, for their perpetual support. For every endeavor, they have always been there for me with tireless encouragement. I have felt their love at each step of the way.

Lastly, and most importantly, I would like to thank Caitlin Heller, my partner in crime, inspiration, future wife, best friend, and soulmate. Her unending love, restless patience, and undying support of my ambitions has meant more to me than life itself. Nobody else has singlehandedly kept me sane while simultaneously letting me talk to them about sand, science,

and philosophy until they fall asleep. Caitlin, you have always been and will always be the light at the end of the tunnel.

Table of Contents

Chapter 1: Introduction **1**

1.1. Importance of Sand Mineralogical Studies **1**

1.2. Controls on Sand Mineralogy **4**

1.3. Purpose of This Study **9**

Chapter 2: Methods **11**

2.1. Published Data Sources **11**

2.2. Data Model and Methodological Parameters **11**

2.3. Data Compilation – Dependent Variables **13**

2.4. Statistical Treatment of Recalculated Parameters **15**

2.5. Data Compilation – Independent Variables **16**

2.6. Machine Learning **20**

2.6.1. Model Selection – Random Forest Regressor **20**

2.6.2. Model Training and Cross Validation **23**

2.6.3. Permutation Feature Importance **24**

2.6.4. Feature Partial Dependence **25**

2.7. Developing a Global Predictive Model **25**

Chapter 3: Results **26**

3.1. Sand Modal Composition **26**

3.2. Model Performance – Q-F-L **28**

3.3. Global Prediction of Sand Mineralogy (GloPrSM) **30**

3.3.1. Q-F-L Classification **30**

3.3.2. Octonary Classification **36**

Chapter 4: Discussion **40**

4.1. Assumptions and Limitations **40**

4.2. Spatial Uncertainty **45**

4.3. Factors Controlling Sand Mineralogy 48
4.3.1. Elevation and Slope..... 48
4.3.2. Precipitation and Temperature 49
4.3.3. Basin Area..... 51
4.3.4. Lithology 52
4.4. Synthesis..... 54
Conclusions..... 56
References..... 58

Chapter 1: Introduction

Sand petrography has long been used as a tool for deciphering sedimentary provenance and unraveling the geologic history of modern and ancient landscapes (e.g., Krynine, 1942; Potter 1978; Dickinson et al., 1983; Potter, 1994). In many cases, sediments are the only remnants of eroded terranes and thus provide insight to the tectonic and paleogeographic context of Earth's distant past. The idea that grain proportions in sand and sandstone reflect the nature of the source area and transport history suggests that inferences can be made about the boundary conditions that modify sediments, including lithology, climate, physiography, and tectonism (Johnsson, 1993). Furthermore, the mineralogical proportions within sedimentary accumulations have a direct impact on human activities, with implications for civil engineering (Nilsen et al., 1976, Gue and Tan, 2006; Qasim et al., 2013; Tarolli and Sofia, 2016), agriculture (Dalal and Probert, 1997; Pimentel et al., 1987), tourism (Mestanza-Ramón et al., 2020), climate change (Berner et al., 1983; Berner and Berner, 1997; Colbourn et al., 2015), carbon sequestration (Lackner et al., 1995; Krevor and Lackner, 2009), groundwater (Aagaard et al., 1990), energy (Bloch 1994), paleogeographic reconstruction (Dickinson and Suczek, 1979), and more. Thus, it is of paramount importance to understand the relationship between sand mineralogy and the processes that modify it.

1.1. Importance of Sand Mineralogical Studies

The production of siliciclastic sediment via weathering of primary minerals within bedrock and during pedogenesis directly controls the fertility of soil and release of nutrients upon which nearly all terrestrial life depends (White and Buss, 2003; Brantley et al., 2007). Yet the expansion of agricultural practices due to increased global demand for food and biofuels has accelerated nutrient depletion, erosion, and sedimentation rates in developed landscapes (Costa,

1975; Pimentel et al., 1987; Dalal and Probert, 1997; Tan et al., 2005; Montgomery, 2007; Merten and Minella, 2013). At geologic time scales, silicate weathering has been proposed as the principal moderator in controlling fluctuations in global temperature and precipitation (Berner et al., 1983; Brady and Carroll, 1994; Berner and Berner, 1997; Gaillardet et al., 1999; Hilley and Porder, 2008; Colbourn et al., 2015). Through the consumption of atmospheric CO_2, silicate weathering products are eventually stored as carbonates in the world's oceans (West et al., 2005). Thus, mineralogic proportions in sand can record global climate signals through geologic time (Potter, 1978; Hessler et al., 2017).

Perhaps the greatest relevance of sand mineralogy to human prosperity is the exploitation of water- and petroleum-bearing reservoirs in the subsurface. As reservoir quality is a function of the volume (i.e., porosity) a container can hold and the rate at which fluids flow from it (i.e., permeability), the most prolific reservoirs are those which supply large volumes over short time intervals. Numerous studies have established a reciprocal relationship between sediment porosity and permeability (Griffiths, 1958; Chilingar, 1964; Bloch, 1994). The re-equilibration of framework minerals to subsurface conditions during burial diagenesis commonly results in the destruction and dissolution of unstable phases (Nickel, 1973; Siebert et al., 1984; Morton, 1986) and the ensuing development of authigenic phases (Land and Milliken, 1981; Boles, 1982; Helmold and van de Kamp, 1984; Walker, 1984; Morad et al., 1990), cementation (Longstaffe, 1984), and secondary porosity (Bjølykke, 1984; Boles, 1984; Franks and Forester, 1984; Markert and Al-Shaieb, 1984; Moncure et al., 1984; Surdam et al., 1984; Morton, 1986; Nedkvitne and Bjørlykke, 1992), with the extent of diagenetic alteration largely dependent on the starting composition of the material (Walderhaug, 1994; Lander and Walderhaug, 1999; Yang et al., 2017). In general, quartzose sands yield the least rapid reduction in porosity with burial depth

(Sclater and Christie, 1980), followed by feldspathic sands (Zeiglar and Spotts, 1978) and lithic sands (Galloway, 1974), respectively. Grain size also plays a significant role in porosity-permeability characteristics of sandstone (Chilingar, 1964; Masch and Denny, 1966; Russell, 1989) and is in part controlled by mineralogy (Potter, 1978; Blatt, 1982; Ingersoll et al., 1984; Zuffa, 1985; Potter, 1986; Tolosana-Delgado and von Eynatten, 2009; Garzanti et al., 2009).

In addition to the aforementioned applications of sand mineralogical studies, researchers have taken interest in the development and implementation of actualistic models to ascertain sedimentary provenance. The mineralogical characteristics of a sedimentary deposit can be useful in aiding paleogeographic reconstruction. For instance, the identification of unaltered feldspars or solution pits in quartz grains has been used as indicators of arid or tropical climates, respectively (Krynine, 1950; Franzinelli and Potter, 1983). The plate tectonic paradigm that developed during the 1960s and 1970s was followed by the introduction of a revolutionary classification scheme that related the detrital modes of terrigenous sand and sandstone to the tectonic setting of their source terrane (Dickinson and Suczek, 1979; Dickinson, 1985). The "Dickinson model" has seen great success in its application to ancient sedimentary basins (Dickinson et al., 1983; Girty, 1987; Jett and Heller, 1988; Ingersoll, 1988; Ingersoll, 1990) and has equally been met with criticism (Molinaroli et al., 1991; Weltje, 2006). Other actualistic modeling efforts have sought to deduce sand composition from the scale of local drainages to entire basins and continents (Ingersoll et al., 1993; Potter, 1986; Potter, 1994). More recently, there has been advocacy for updating the approach to actualistic provenance with advanced statistical methods (e.g., Weltje, 2002) and the development of dynamic, process-based simulations to address the limitations of these foundational models and account for the complexities observed in natural systems (Weltje, 2012; Garzanti, 2016).

1.2. Controls on Sand Mineralogy

Decades of research have uncovered a complex system of factors that control the composition of clastic sediments (Fig. 1.1; Johnsson, 1993). Clastic detritus undergoes several evolutionary stages prior to burial and lithification, and the interplay and feedbacks between first-order factors such as source lithology, climate, basin physiography, transport distance, and tectonic setting coevally alter the framework proportions in clastic sediments (Suttner et al., 1981; Johnsson, 1993). If the relative influence of these factors on sediment composition can be constrained, so too can their combined influence explain the abundance of framework grains in sand.

Source lithology dictates the petrologic diversity of sediments produced during pedogenesis and supplied to sediment routing systems. Under ideal conditions, sand mineralogy may closely reflect the composition of the sediment source terrane, such as in the absence of chemical weathering (e.g., Potter et al., 2001, Garzanti et al., 2015b) or when sampled from low order drainages (Heins, 1992; Ingersoll et al., 1993). However, most environments at Earth's surface are hostile to silicate minerals that are thermodynamically stable at the depths in which they form, leading to an inevitable divergence in compositional space between source assemblages and their derived sediments with time (Weltje et al., 1998). This process of mineralogic modification within sediment routing systems may produce a sedimentary product that better records the climatic regime than source lithology (Potter, 1978; Franzinelli and Potter, 1983; Johnsson et al., 1988, 1991).

Climate factors include the temperature of the weathering environment and the supply of weathering agents, such as water from precipitation and organic acids from the productivity and decay of vegetation (Jackson and Keller, 1970; Knoll and James, 1987). Bedrock lithology

dictates the starting porosity of unweathered material (Graham et al., 1997; Behrens et al., 2015) which can control turnover rates of weathering fluids (Maher, 2010), the composition of dissolved solids (Aagaard et al., 2010), and the rate at which dissolved solids are introduced to the weathering medium (Lasaga et al., 1994), with the chemical composition and residence time of weathering fluids introduced to the critical zone controlling the reaction rate of unweathered material (Maher, 2010). Fluid residence time is dependent upon the rate of supply of fresh water, where longer residence times correspond to decreasing mineral reaction rates as fluids becomes progressively saturated with dissolved solids (Maher, 2010). Furthermore, reaction rates are temperature-dependent with higher temperatures accelerating silicate weathering (White and Blum, 1995; West et al., 2005). For example, Eq. 31 from White and Buss (2003) predicts that silicate reaction rates should increase by an order of magnitude between 0°C and 25°C when all else is held constant. Additionally, Riebe et al. (2004) measured ^{10}Be cosmogenic nuclides from sampled granitic outcrops in North America and New Zealand to show coupled temperature and precipitation explain 67-84% of variance in the measured chemical depletion fraction across their study area.

Topographic elements such as elevation and slope govern the residence time of sediments and the duration of weathering during pedogenesis. Low slopes favor longer residence times as the aid of gravity-driven transport is diminished, while steep slopes, often associated with high elevation settings, lead to the rapid downslope transport of material and generally correspond to high sedimentation rates in adjacent sedimentary basins (Johnsson, 1993). The weathering regime can be thought of in terms of the relationship between residence time and mineral weathering rates. Where the supply of weathering agents is large relative to the supply of unweathered material and the residence time of the material is long compared to the reaction

rate, minerals may be significantly or wholly altered before removal (Stallard and Edmond, 1983; Johnsson, 1993; West et al., 2005; Maher, 2010, Sklar et al., 2017). This "transport-limited" regime is controlled by the supply of fresh, unweathered solids, where the transport of material is slow compared to chemical weathering rates. Conversely, a "kinetically-limited" regime occurs where silicate weathering is incomplete, as the supply of material is high relative to the rate at which it is altered or if it is removed faster than it is supplied (Stallard and Edmond, 1983; Johnsson, 1993; Sklar et al., 2017). Generally, steep slopes favor high denudation rates, thus limiting the exposure of sediments to weathering in the depositional environment (Johnsson, 1993).

Previously, sediment transport processes were thought to remove softer grains via mechanical breakdown and to promote the relative enrichment of durable constituents (Martens, 1931; Plumley, 1948; Hayes, 1962; Pittman, 1969). However, Garzanti et al. (2015a) showed that physical processes that operate during transport were insufficient to modify sand composition along the 1,750 km Orange littoral cell of the Atlantic coast of Africa. In a tumbler experiment, Osborne et al. (1993) revealed that different grain sizes became statistically homogeneous after significant transport distances and suggested that observed compositional differences are the result of sedimentological processes rather than from mechanical reduction in grain size. Compositional evolution during transport in fluvial systems may be associated with transport-limited weathering processes acting on alluvially stored sediments in fluvial systems (Johnsson, 1993). Despite the relative lack of influence of mechanical processes on the alteration of clastic sediment composition during transport, it is often difficult to separate the effects of mechanical and chemical weathering as the least durable constituents are often chemically labile (Johnsson, 1993). Furthermore, the comminution of grains during transport can accelerate

chemical weathering following a proportional relationship between grain surface-area-to-volume ratio and reaction rates (White and Brantley, 2003; Gabet and Mudd, 2009; Israeli and Emmanuel, 2018).

Tectonic setting has been regarded as the overarching control of clastic sediment composition due to its influence on several of the aforementioned factors (Pettijohn et al., 1972). With rapid uplift, climate in the source area can be modified as temperatures fall and liquid precipitation is reduced. Tectonism determines topographic characteristics of a drainage basin where rapidly uplifted settings favor higher denudation rates and associated kinetically limited weathering regimes. With the development of plate tectonic theory, the global distribution of rock types could be better explained in terms of plate tectonic setting (Dickinson and Suczek, 1979; Valloni and Maynard, 1981; Dickinson et al., 1983, Yerino and Maynard, 1984). This association has led to the development of classification schemes that relate framework modes of a sample to the geodynamic setting of its source terrane (Dickinson and Suczek, 1979; Dickinson, 1985).

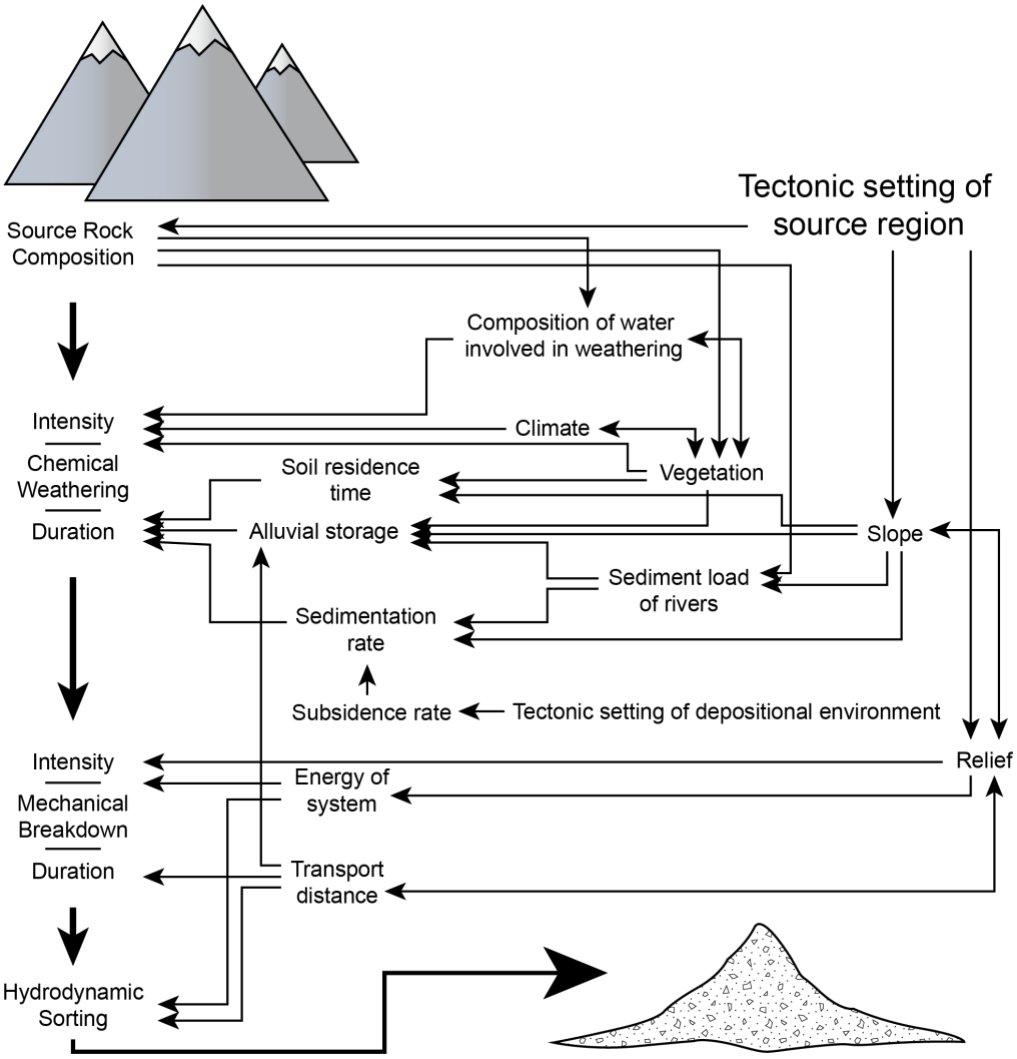

Figure 1.1. Schematic diagram of the system that controls the composition of clastic sediments, modified after Johnsson (1993). Complex interactions and feedbacks at Earth's surface drive progressive changes in mineral abundances prior to final deposition and lithification.

The notion that sediments generated in different geodynamic settings will plot in distinct fields on a ternary diagram has faced criticism since the conception of the Dickinson model (Molinaroli et al., 1991; Weltje, 2006; Garzanti, 2016, 2019). Molinaroli et al. (1991) showed that Dickinson et al. (1983) successfully classified the tectonic setting of no more than 85% of North American Phanerozoic sandstones in their database when using the Dickinson model. Further evaluation by Weltje (2006) suggested that Molinaroli et al. (1991) overestimated the discriminatory power of the Dickinson model, with an updated success ratio ranging between

64% and 78% for the four ternary classifications used by Dickinson et al. (1983) (Q-F-L, Qm-F-Lt, Qm-P-K, Qp-Lv-Ls). Factors such as source lithology, sampling scale (Ingersoll et al., 1993), basin slope (Stallard, 1988), and climate (Nesbitt et al., 1996) can alter sediment composition along sediment routing systems (Fig. 1.1) and are not accounted for in the Dickinson model. Furthermore, not all tectonic settings are represented in this classification, some of which represent significant areal proportions of Earth's surface, such as anorogenic flood basalts, transform margins, and other hybrid settings. In effect, sediments derived from vastly different source terranes can plot in the same provenance fields (e.g., Hunter, 1967; Johnsson et al., 1991; Garzanti et al., 2015b).

1.3. Purpose of This Study

Provenance analysis often applies inverse modeling approaches to infer cause from effect (e.g., tectonic setting from point counts) (Graham et al., 1976; Valloni and Maynard, 1981; Maynard et al., 1982; Bhatia, 1983; Dickinson et al., 1983; Bhatia and Crook, 1986; Roser and Korsch, 1986; Ingersoll, 1990; Marsaglia, 1992; Marsaglia and Ingersoll, 1992; Marsaglia et al., 1995a, 1995b; Weltje et al., 1998; Weltje and Prins, 2003; Verma and Armstrong-Altrin, 2013). However, few studies have forward modeled sediment generation based on fundamental controlling factors (Weltje, 2012). Heins and Kairo (2007) introduced SandGEM, an empirical, Bayesian belief network-based forward model, to predict sand mineralogy and texture as a function of environmental controls. Others have used data-calibrated theoretical approaches to quantify mineral weathering rates, major element geochemistry, mass transfer, and soil formation in response to chemical, mechanical, and geomorphological processes (Gabet, 2007; Yoo and Mudd, 2008; Gabet and Mudd, 2009; Yoo et al., 2009; Ferrier et al., 2010; Hilley et al., 2010; Maher, 2010; Brantley and Lebedeva, 2011; von Eynatten et al., 2012, 2016; Israeli and

Emmanuel, 2018). Despite these significant advances, there is not yet a framework that integrates a suite of environmental factors and processes (e.g., Fig. 1) to predict the spatiotemporal evolution of mineral abundances in clastic, sand-sized sediments at the global scale.

This research aims to better understand how Earth-surface processes are manifested in sediment and the sedimentary archive. Herein we introduce the Global Prediction of Sand Mineralogy (GloPrSM) model that is derived from a series of machine learning models trained on a global petrographic database compiled from 51 published sources. Following the methodology of previous actualistic studies (e.g., Ingersoll, 1990, 1993; Garzanti, 2016, 2019), GloPrSM model is calibrated to Pleistocene-modern samples whose boundary conditions can be inferred with a high degree of confidence. The data-driven, forward modeling approach presented herein draws on first-principle relations between sand composition and its controlling factors, incorporating globally extensive climatic, topographic and lithologic datasets. The GloPrSM model includes predictions for several sand classification schemes, including Q-F-L (Dickinson, 1970; Garzanti, 2019), Qm-F-Lt (Dickinson, 1985), and an internally developed scheme to predict Q-F-L proportions and eight Q-F-L subcompositions. The GloPrSM model provides insights into the fundamental controls on sand mineralogy through stochastic simulations and novel model interrogation techniques (Pedregosa et al., 2011; Molnar, 2020). The result is a high-resolution, first-order estimate of the global distribution of sand mineralogy explained in the context of Earth's present-day source-to-sink systems.

Chapter 2: Methods

2.1. Published Data Sources

This research is based upon a compilation of sand mineralogical (i.e., modal point count) data acquired from published sources (Supplemental Table A). Several criteria must have been met for a data source to be included: 1) published data were reported as raw sedimentary point counts or recalculated percentages of raw point counts, 2) point counts were collected from unconsolidated, sand-sized sediments (Wentworth, 1922), 3) samples are Pleistocene or younger, 4) sample locations were reported, either as coordinates or plotted on study area maps, and, 5) data were reported for individual samples instead of as averages of multiple samples. The sand modal composition database was compiled from 51 published sources, the majority from peer-reviewed journal articles and university theses (Supplemental Table A).

2.2. Data Model and Methodological Parameters

Point count parameters may vary by label and definition depending on the published source. Furthermore, methodological approaches may vary by point counting operator. To address the lack of a consistent format for published point count data, a data model was designed to standardize modal point counting parameters (Supplemental Table B) and to capture methodological metadata (Table 2.1). First, petrographically distinct grain types were identified from source publications. For instance, chert grains may be reported in different source tables as the labels ‘Lh’, ‘Lch’, or ‘Lsch’ (e.g., Marsaglia et al., 1995a, 1995b; Rittner et al., 2016; Garzanti et al., 2019). Moreover, chert may be defined simply as “chert” (e.g., Malkowski and Hampton, 2014), or more specifically as “sedimentary chert or cherty argillite lithic” (e.g., Marsaglia and Ingersoll, 1992; Van Grinsven and Marsaglia, 2019). As each of these labels and definitions describe the same, or similar, petrographic observations, the data model standardizes

chert's label as 'Qch'. Using this approach, 287 unique point counting labels were distilled to a final list of 54 point counting parameters, following these guidelines.

1. The most specific point counting label was used, whenever possible. For example, if an author reported counts of both monocrystalline quartz and polycrystalline quartz, these data would be reported separately and not combined under the umbrella quartz (Q) category. However, some authors reported data in more detail than captured in the data model. In this case, data were generalized into the appropriate point counting parameter. For example, Marsaglia et al. (1995a, 1995b) reported grain types of "quartz-mica tectonite lithic" and "quartz-feldspar-mica aggregate lithic" which are both grouped under 'Lmh' (Supplemental Table B).
2. Point counts are never included twice. For example, if a source reports counts for K-feldspar and plagioclase, these data would not be also included under the "feldspar" (F) point counting parameter.
3. Primary compositions (e.g., Q, F, L) may be counted alongside their corresponding subcompositions (e.g., Qm, K, Lv) if the source did not supply sufficiently detailed definitions of the subcomposition. For example, Chamov and Murdmaa (1995) report P as "clearly identified calcic, intermediate, and sodic plagioclases" and F as "other feldspars with refractive index n < 1.540", yet F does not imply the presence of alkali feldspars (K). Thus, counts of both P and F were added to the database, with counts of P excluded from counts of F.
4. Zero values for a point counting parameter indicate that the parameter was included in the source's counting scheme but not found in the sample. Null values were used to indicate that a point counting parameter was not counted for a given sample.

Additionally, the data model records differences in data acquisition and reporting as methodological parameters (Table 2.1). For instance, the 'Qch_as_Qp' parameter was used to record whether a data source counted chert as polycrystalline quartz and the 'GaDi' parameter was used to indicate whether the Gazzi-Dickinson method (Ingersoll et al., 1984) used during point counting. Methodological parameters were chosen to provide a measure of data quality and consistency.

Table 2.1. List of methodological parameters recorded during data compilation.

Methodological Parameter	Description
Qch_as_Qp	Chert was counted as polycrystalline quartz
GaDi	Gazzi-Dickinson method was used during the point counting procedure
Q_undiff	Total quartz is divided into subcompositions, such as monocrystalline or polycrystalline quartz
Raw_as_pct	Raw point counts are reported as percentages
Ternary_vals	Values are reported in ternary proportions instead of raw counts or percentages
Location_approximate	Georeferencing was used to determine sample locations

2.3. Data Compilation – Dependent Variables

Published datasets were compiled in Microsoft Excel as exact replicas. A modified spreadsheet was created to reassign reported petrographic labels to their new label following the data model (Supplemental Table B). After label reassignment, Q-F-L and Qm-F-Lt ternary values were recalculated following the classification schemes of Dickinson (1970, 1985) and Garzanti (2019) (Supplemental Table F). The Dickinsonian approach seeks to use the framework modes of sediments as a proxy for identifying sedimentary provenance and paleotectonic setting. This classification counts chert toward the Q-pole in a Q-F-L system due to its stability (Dickinson, 1970). Furthermore, counts of extrabasinal carbonate grains are excluded from recalculations of lithic proportions due to their propensity to chemical alteration or removal in certain weathering environments, and their ease of confusion with intrabasinal clasts (Dickinson,

1985). Both polycrystalline quartz and chert are included with the total lithics (Lt) of the Qm-F-Lt scheme of Dickinson (1985). On the other hand, Garzanti (2019) counts chert towards the L-pole in a Q-F-L system on the basis that chert cannot always be distinguished from other lithic grain types such as felsitic volcanic lithics and argillitic grains. Garzanti (2019) also argues that all extrabasinal lithic grain types be included in recalculations of lithic proportions, as the exclusion of carbonates on the notion of lability is inconsistent with the inclusion of even more labile grain types such as shale and gypsum. This exclusion of weathering-susceptible sedimentary grains leaves very few observations of sedimentary grains being recorded (Garzanti, 2019).

Following the principles of Dickinson (1970, 1985) and Garzanti (2019), an internal classification scheme was developed to calculate grain proportions for eight commonly observed groups of parameters (Supplemental Table F). Monocrystalline quartz (Qm), polycrystalline quartz (Qp), and chert (Qch) were subdivided from the Q-fraction of the Q-F-L global prediction, while alkali feldspar (K) and plagioclase feldspar (P) were subdivided from the F-fraction, and volcanic lithics (Lv), sedimentary lithics (Ls), and metamorphic lithics (Lm) were subdivided from the L-fraction. The purpose of this octonary classification is to increase the granularity of model predictions while still being able to regroup these parameters in Q-F-L proportions. In this hybrid model, chert is classified under total quartz in a Q-F-L ternary system while counts of extrabasinal carbonate lithics are counted toward the L- and Ls-poles. Despite arguments against classifying chert as a quartzose grain type (Garzanti, 2019), we have chosen to group Qch under the Q-pole, in part due to its chemical stability, but, more importantly, due to inconsistency in the reporting of chert as polycrystalline quartz (Table 2.1).

Recalculated proportions from each scheme are the dependent variables for the GloPrSM models and were calculated in Python. Recalculated parameters are sum-to-one values and were not calculated if any of the categories lacked counted parameters (i.e., null values). For example, Qm-F-Lt would not be calculated if an operator did not count Qm as separate category.

2.4. Statistical Treatment of Recalculated Parameters

By nature, compositional (i.e., sum-to-one) data are constrained and unable to assume any value greater than 1 and less than 0. For this reason, compositional data cannot be statistically treated by methods designed for use with unconstrained data (Chayes, 1960; Aitchison, 1982; Weltje, 2002). To account for this, an additive log ratio transformation (Aitchison, 1982) was applied to each classification's recalculated proportions:

$$y_i = ln\left(\frac{x_i}{x_k}\right) \; for \; 1 \leq i \leq k-1 \qquad \textbf{(1)}$$

where x_i represents the relative abundance of the *i*th component in a system with *k* constituents (*k* = 3 for ternary systems), and the *k*th component's value, x_k, is specified from the sum of the other *k – 1* components. Ratios alone (e.g., F:Q or Q:F) are asymmetric depending on the value assigned to the denominator, and arbitrarily choosing one value or the other can yield different results of mean or variance (Weltje, 2012). By taking the logarithms of ratios, symmetry is restored and variances are identical regardless of a components place in the log ratio's numerator or denominator (Weltje, 2012).

Prior to log transformation, an imputation was applied to the compositional dataset where zero values were replaced by 0.001, implying 1 occurrence of a component for every 1000 recorded observations (Martin-Fernandez et al., 2003; Weltje, 2012; Vermeesch, 2018). The purpose of the imputation was to, 1) maximize the size of the training dataset as the natural log of zero is indeterminable, and 2) to retain important information contained by end-member

values (e.g., Q = 100%). With the imputed dataset, log ratios can now assume any value between $ln\left(\frac{0.001}{1}\right)$ and $ln\left(\frac{1}{0.001}\right)$. All possible combinations of log ratios were calculated for each classification, resulting in $k^2 - k$ combinations for each system (e.g., 6 possible combinations for ternary systems). An inverse transformation was applied return log-ratio values to their corresponding sum-to-one values (Supplemental Table G):

$$z_i = \begin{cases} \frac{e^{y_i}}{1+\sum_{i=1}^{k} e^{y_i}} \; for \; 1 \leq i \leq k-1 \\ \frac{1}{1+\sum_{i=1}^{k} e^{y_i}} \; for \; i = k \end{cases} \quad \textbf{(2)}$$

where z_i is the inverse transformed composition of the ith log-ratio in a system with k components.

2.5. Data Compilation – Independent Variables

Prior to the widespread use of GPS, most authors did not report sample locations as coordinates. For these sources, sample locations were georeferenced from study area maps in ArcGIS software using the WGS84 datum. Samples with reported coordinates were imported to ArcMap as point features, converted to the WGS84 datum if necessary, and subsequently georeferenced to flow accumulation lines corresponding with the original sample location. Where possible, published data tables containing the names of streams and beaches from which samples were collected were used to inform the georeferencing workflow.

Two flow accumulation datasets were employed for georeferencing tasks, following Sharman and Malkowski (2020). For samples south of 60°N, the HydroSHEDS HydroRIVERS 15 arc-second global river network was utilized (Lehner et al., 2008). Only fluvial and littoral samples were georeferenced to HydroRIVERS, as samples from other depositional environments (e.g., aeolian, marine, lacustrine) do not necessarily correspond to river and coastline locations. Fluvial sample point features were manually moved to intersect flow accumulation polylines

representing the river reaches corresponding to their original sampled locations. This practice ensured that the calculated watershed for that sample correctly captured its upstream drainage area. Littoral sample locations were drawn as polygon features that extended from the reported sample location along the reach of coastline to the source of their respective littoral cells, often intersecting several high flow accumulation polylines along the reach of the coastline. This practice was employed to ensure that littoral samples included all relevant fluvial inputs. Where possible, the identification of littoral cell extent was informed by the source publication. Watershed boundaries were delineated using the HydroSHEDS 15 arc-second flow direction grid. Subsequently, these catchment polygons were resampled to the resolution of the ETOPO1 Global Relief Model (Amante and Eakins, 2009).

ETOPO1 is a 1 arc-minute global digital elevation model (DEM) that integrates land surface elevation and ocean bathymetry (Amante and Eakins, 2009). A flow accumulation polyline feature dataset was extracted from the ETOPO1 DEM and employed to georeference samples located north of 60°N and those collected from offshore depositional environments. Source publications were used to guide the placement of marine sample polygon features and determine if they should overlap flow accumulation polylines representing submarine canyon networks or other local sediment routing networks. Catchment boundaries were then delineated from an input flow direction dataset extracted from the ETOPO1 DEM. Once all catchment boundaries were computed as raster data types, each was converted to a polygon feature to calculate catchment area (Fig. 2.1b) and catchment-averaged elevation (Fig. 2.1c). In total, 3,208 catchment boundaries were delineated in ArcGIS for samples spanning fluvial, littoral, nearshore shelf, and deepwater depositional settings.

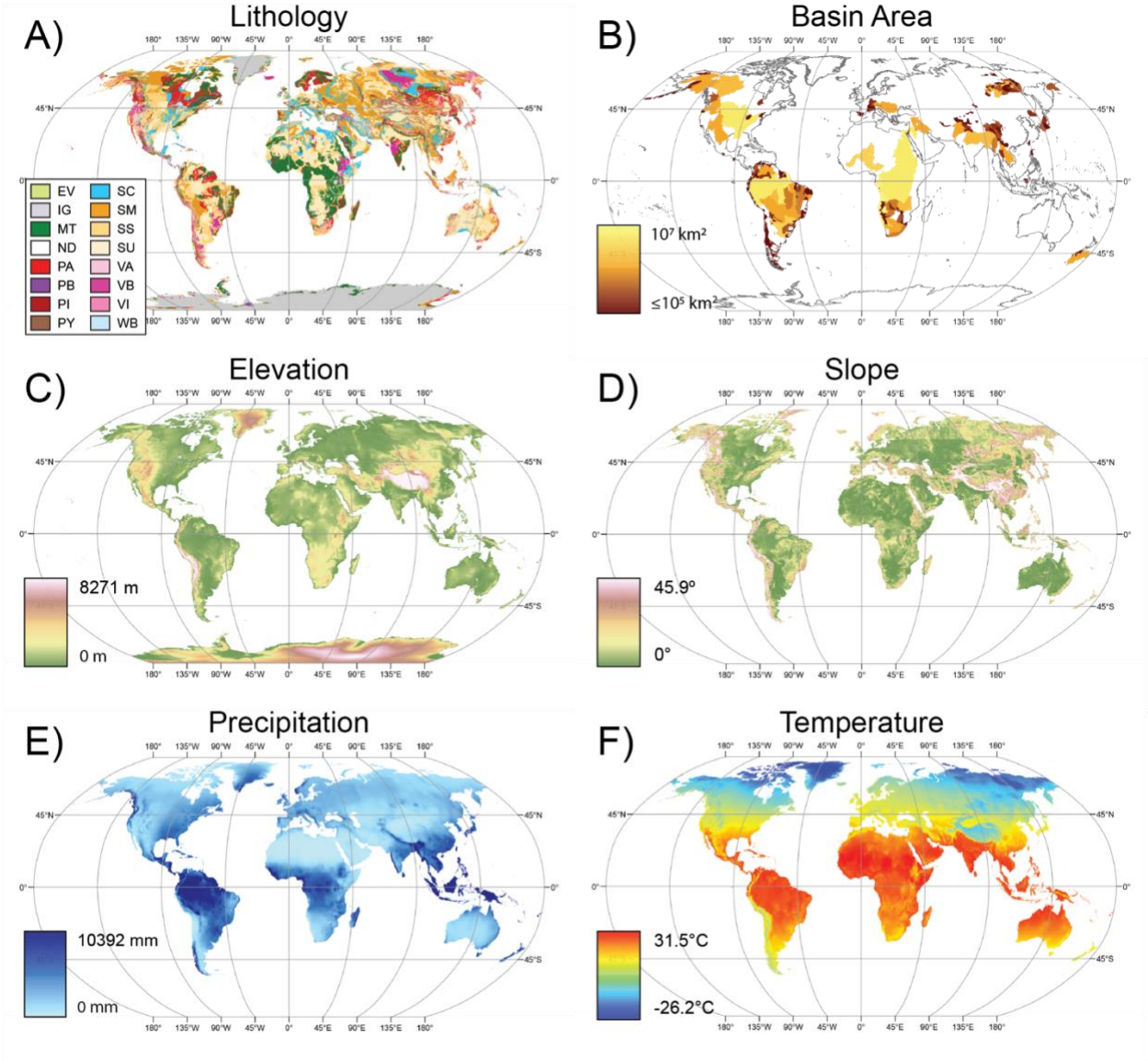

Figure 2.1. Independent variables and their source datasets used in this study: a) bedrock lithology (GLiM, Hartmann and Moosdorf, 2012), b) maximum basin drainage area, c) elevation (ETOPO1, Amante and Eakins, 2009), d) slope (EarthEnv-DEM90, Robinson et al., 2014), e) precipitation (WorldClim v1.4, Hijmans et al., 2005), and f) temperature (WorldClim v1.4, Hijmans et al., 2005).

HydroATLAS is a compilation of two globally extensive, hydro-environmental databases (BasinATLAS and HydroRIVERS; Lehner et al., 2008; Linke et al., 2019). From BasinATLAS, three datasets were extracted at the level 12 catchment scale (Fig. 2.1) to satisfy data requirements for the independent variables, including precipitation and temperature (WorldClim v1.4, Hijmans et al., 2005), terrain slope (EarthEnv-DEM90, Robinson et al., 2014), and bedrock

lithology (GLiM, Hartmann and Moosdorf, 2012). These datasets are stored in the BasinATLAS feature dataset and were downloaded from the HydroSHEDS website (https://hydrosheds.org/) as shapefiles before subsequent being converted to raster format. Independent variable raster datasets were loaded into Python as NumPy arrays using the Rasterio external library (Gillies et al., 2019), and the catchment polygons were loaded using the GeoPandas external library (Jordahl et al., 2020). Catchment polygons were iteratively used to mask the independent variable arrays, allowing basic arithmetic operations to be performed on the intersection. Excluding the GLiM dataset, the average, minimum and maximum values in each catchment polygon were extracted for each independent variable. For GLiM, the relative areal abundance of each "first level" lithologic class (16 total classes, Hartmann and Moosdorf, 2012) present in each catchment polygon were calculated. To reduce the total number of independent variables, each lithologic proportion was assigned to a new category based on shared mineralogic associations between each lithology (Table 2.2); lithologic classes in the 'Unassigned' group were excluded from the GloPrSM model training dataset.

Table 2.2. GLiM database lithologies and their reassigned groupings based on assumed mineralogical associations.

Original Lithologic Class	Label	Reassigned Lithologic Group
Plutonic Acid	PA	Felsic to Intermediate Plutonic
Plutonic Intermediate	PI	
Plutonic Basic	PB	Mafic Plutonic (undivided)
Pyroclastic	PY	Felsic to Intermediate Volcanic + Pyroclastic
Volcanic Acid	VA	
Volcanic Intermediate	VI	
Volcanic Basic	VB	Mafic Volcanic (undivided)
Sedimentary Mixed	SM	Siliciclastic + Mixed + Unconsolidated Sedimentary
Sedimentary Siliciclastic	SS	
Sedimentary Unconsolidated	SU	
Metamorphic	MT	Metamorphic (undivided)
Evaporites	EV	Evaporites + Carbonates
Sedimentary Carbonate	SC	
Ice and Glaciers	IG	Unassigned
No Data	ND	
Waterbodies	WB	

2.6. Machine Learning

2.6.1. Model Selection – Random Forest Regressor

Due to the numerical, multivariate nature of the dependent and independent datasets, the Scikit-Learn Random Forest (RF) Regressor algorithm was selected for machine learning modeling tasks (Pedregosa et al., 2011). RF is a widely employed algorithm that can constrain complex, non-linear relationships between predictor and response variables (Breiman, 2001). In addition, RFs are known for their high accuracy, interpretability, and are capable of handling a large number of predictor variables. RFs are composed of a series of decision or regression trees (Fig. 2.2) and draw upon two elements of randomness in their construction. First, a bootstrapped dataset is built from a subset of the original dataset that is randomly selected with replacement, leaving approximately one-third of the cases unused. Second, a subset of predictor variables are chosen when splitting each node. When applying the RF, each tree in the forest generates an

individual prediction. While individual trees may be weak estimators, their predictions are averaged to produce a final result that is relatively insensitive to noise from new datasets.

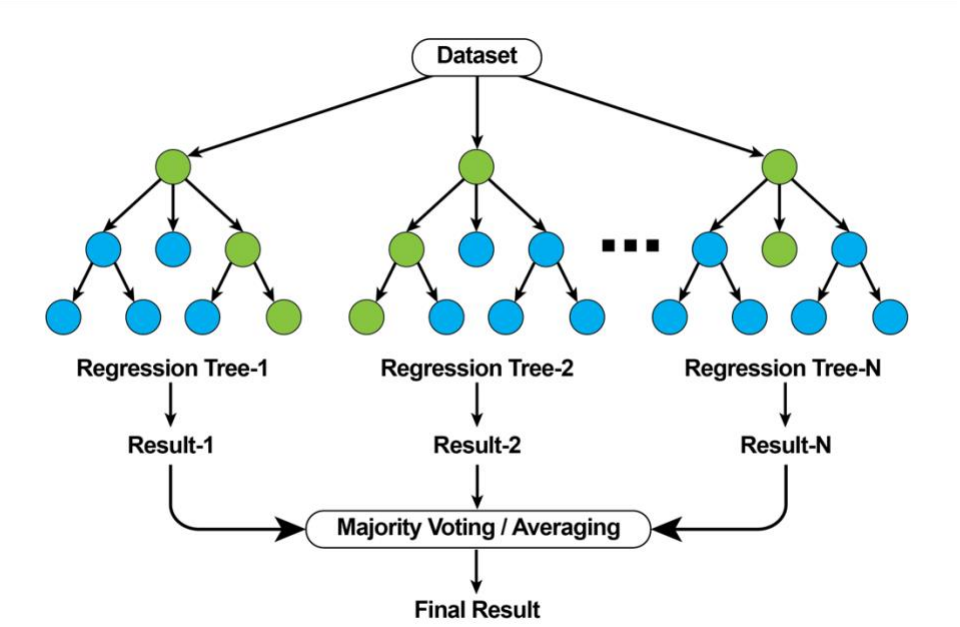

Figure 2.2. Conceptual diagram of a random forest regressor, adapted from TIBCO (https://www.tibco.com/reference-center/what-is-a-random-forest). Random forests comprise a series of N decision or regression trees. When applied to a new dataset, each tree in the forest generates an individual prediction, with all predictions averaged to produce a result that is relatively insensitive to noise.

Regression trees are composed of root nodes, internal nodes, and terminal or "leaf" nodes. In each tree, a random subset of the predictor features is selected without replacement when splitting a node. This subset can contain as few as 2 features or up to as many features as are present in the input dataset; the Scikit-Learn RF Regressor defaults to the latter. At the root node, each feature is evaluated by how well it can predict the target value using the mean square error (MSE) estimated over n samples:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \textbf{(3)}$$

where $\hat{y}_i$ is the predicted value of the *i*th sample and y_i is the corresponding true value. First, each feature's values are compared against the target values (i.e., $\hat{y}$). A split is then made

between each feature value in the sequence, and the mean of the target values is computed on either side of the split (Fig. 2.3). The difference between the true value and the mean is calculated and summed for all individual values on each side of the split. The split value is iterated over each feature value in the sequence, and the MSE computed. The split with the lowest MSE is chosen as a candidate for the node value.

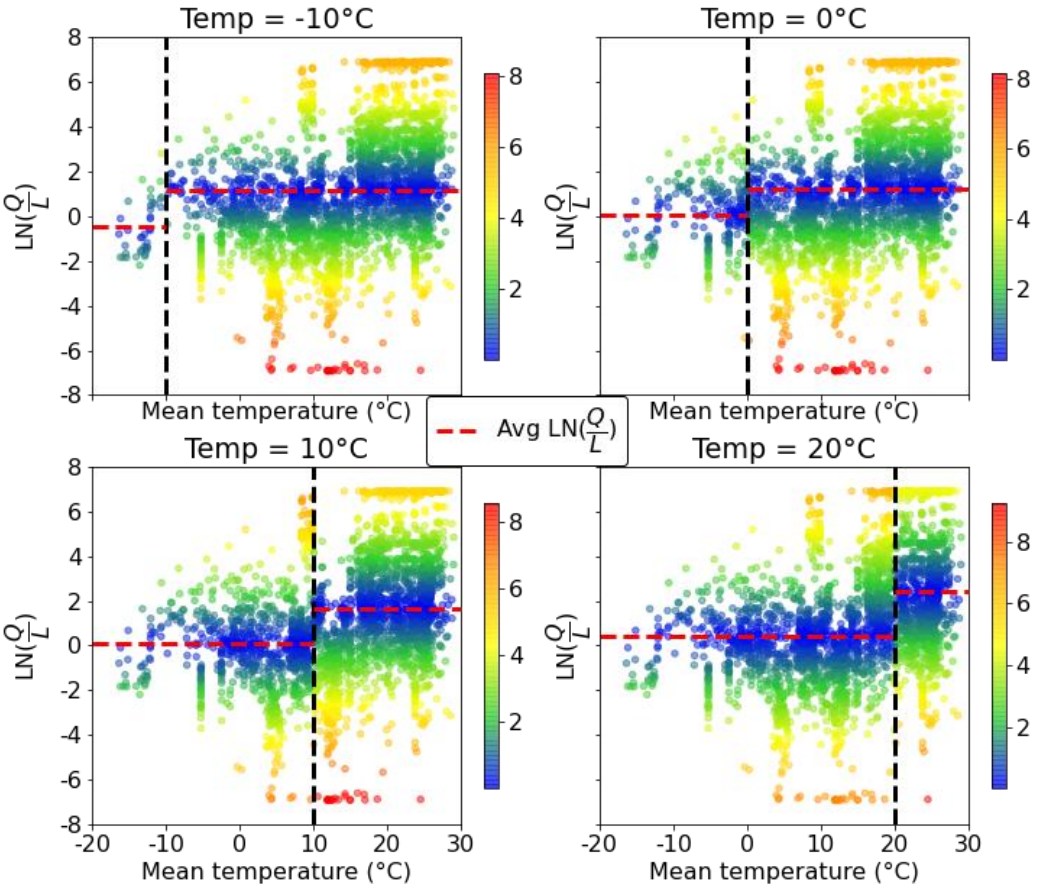

Figure 2.3. The splitting framework employed by a single regression tree, using mean temperature as an example feature with data points colored by their MSE. Each feature's value becomes a split value, with the mean of the target values and each data points MSE calculated on either side of the split.

This process is repeated for all features in the dataset, and the feature with the lowest MSE is selected to be the root node value which is subsequently split into true and false outcomes (Fig. 2.4), and the iterative evaluation of a feature condition by its MSE is performed

for the internal nodes representing these outcomes. Nodes will continue to expand until 1 of 3 scenarios occur, in which case a leaf node is realized: 1) the MSE of an unsplit node is higher than if the node were split, 2) the leaf is pure (MSE equals zero), or 3) the node contains less than the required number of samples needed to split a node (2 samples minimum).

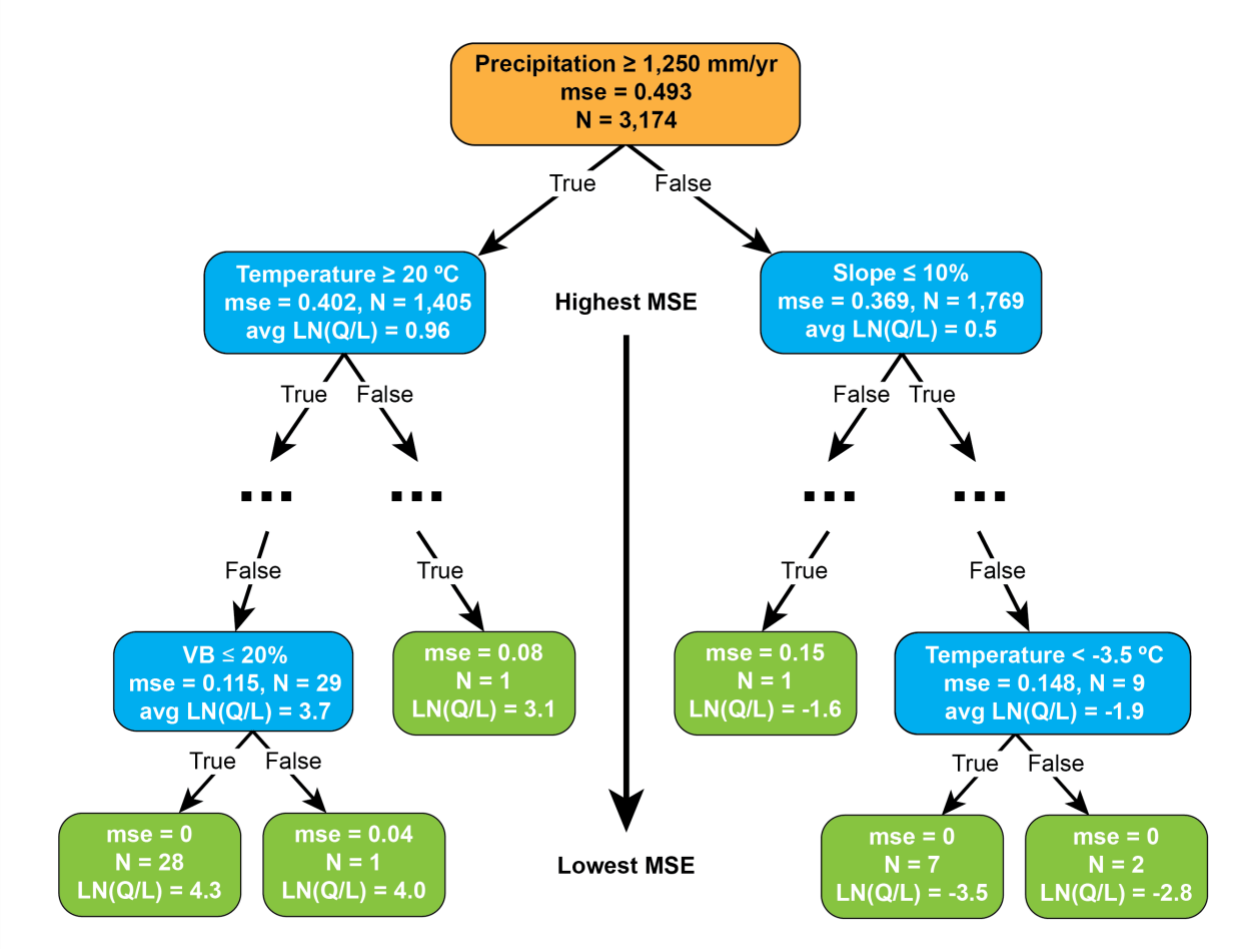

Figure 2.4. Conceptual diagram of a single regression tree within a random forest regressor. Decisions are made in a top-down fashion through root (orange), internal (blue), and terminal or "leaf" nodes (green). MSE is used to evaluate features when assigning a node condition, and, in general, MSE decreases as the terminal nodes are approached.

2.6.2. Model Training and Cross Validation

Cross-validation techniques were employed during model training to prevent overfitting. When training a model, it is common practice to use a held-out test set to validate model performance. Testing the prediction function on the data from which it was derived would repeat

the labels of the samples it has already seen and fail to make a useful prediction from yet unseen data (Pedregosa et al., 2011). As the input dataset is stratified by Source ID and consequently by geographic location, the ShuffleSplit cross validator was chosen. This algorithm randomly selects rows from the input dataset to divide the training and test sets. ShuffleSplit was run with 100 re-shuffling and splitting iterations for each log-ratio combination for each classification scheme. In each shuffle, a random forest regressor was instantiated with the number of estimators (i.e., trees in the forest) set to equal to the length of the dataset, the test size set as 0.2 (train size becomes 0.8 by default), and the random state set to 0 for reproducibility; the remaining hyperparameters such as *max_depth*, *min_samples_split*, and *min_samples_leaf* were kept as their package default values. The RF was then fit to the shuffled training features and target variables (e.g., ln[F/Q]), and a prediction was generated from the test features. The prediction was scored against the test labels using the R^2 coefficient of determination. Model performance was documented with each shuffle, and the log-ratio combination with the best median performance of the 100 shuffles was chosen to generate future predictions (Supplemental Table D).

2.6.3. Permutation Feature Importance

A Scikit-Learn permutation importance algorithm was employed to assess the influence of each predictor variable on model accuracy (Pedregosa et al., 2011). This model inspection technique is applied to a fitted estimator (i.e., RF) by randomly permuting the values of a variable from the test dataset and generating a new prediction using the trained model. The R^2 score is calculated for the new prediction and compared to the original score generated from the validation prediction. This process was repeated 30 times for each predictor variable; variables with the largest residual R^2 score are considered most important as they control the model's

ability to explain variation in the dataset from which they are trained. Due to the extensive processing time and computing power required to implement the permutation importance algorithm, variable importance was only calculated for the highest scoring RFs from 100 ShuffleSplit iterations.

2.6.4. Feature Partial Dependence

The Scikit-Learn partial dependence (PD) algorithm (Pedregosa et al., 2011) was employed to quantify the log-ratio prediction's response to sequential changes in each predictor variable's values (Molnar, 2020). To compute PD of the Q-F-L prediction, a grid is initialized to contain evenly spaced values between the minimum and maximum values of each predictor variable. Iteratively, each variable's original values are replaced by a value from the grid while holding the remaining variable values constant before a new log-ratio prediction is generated and its mean is computed. The result is a series of predicted values representing the sequential change in log-ratio composition as a function of a predictor variables values. Long processing times only allowed for the partial dependence to be calculated for the highest scoring RFs from 100 shuffles.

2.7. Developing a Global Predictive Model

Integrated upstream watershed boundaries were computed for BasinATLAS level 8 catchment polygons and independent variable data (Fig. 2.1) were subsequently collected for each integrated watershed. Catchments in Greenland and Antarctica were excluded as these landmasses have very few watersheds not covered by ice and glaciers. The resulting dataset contains 183,672 watershed polygons with approximate mean and median areas of 27,000 km^2 and 1,500 km^2, respectively. The trained RFs were applied to this dataset to generate 100 log-ratio predictions for each classification system (Supplemental Table D) during the ShuffleSplit

procedure. Statistics were calculated from variations within each classification's 100 predictions and include the prediction mean, median, standard deviation, skewness, 2.5th percentile, 25th percentile, 75th percentile, and 97.5th percentile. These variances were then inverse transformed to ternary values (Eq. 2) and their interquartile range and range of the inner 95% of the predicted distribution were calculated from the percentiles. Results were joined with the BasinATLAS level 8 shapefile to be mapped in ArcGIS software.

Chapter 3: Results

3.1. Sand Modal Composition

The sand modal composition database contains data for 3,512 Pleistocene-modern samples spanning all continents except Antarctica (Supplemental Table C). An average of 364 point counts were conducted per sample, and the dataset thus contains approximately 1.2 million point counts of which ~50% are quartzose grains, 17% are feldspar, and 33% are lithic and other grain types (Fig. 3.1). Samples were categorized into nine depositional environments (Fig. 3.2). Approximately 65.6% of samples are from fluvial environments, 15.2% from marine environments, 15.1% from littoral environments, and 4.1% from aeolian and other environments. Database sample catchments have a median area of approximately 6,400 km^2. With 69.7% of samples deposited in terrestrial environments, data coverage roughly amounts to a terrestrial sample for every 61,000 km^2 of Earth's land area, a beach sample every 1,300 km of coastline, and a sample for every 145,000 km^2 of Earth's total surface area.

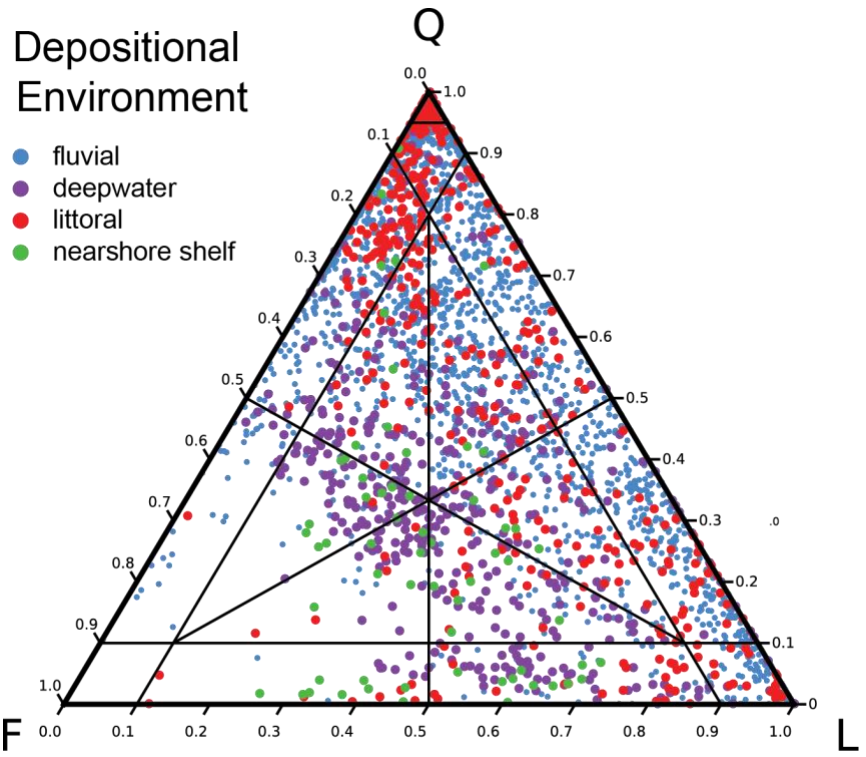

Figure 3.1. Ternary plot of database samples for fluvial, littoral, and marine depositional environments. Feldspathic samples are least abundant and quartzose samples are the most abundant, especially in littoral and fluvial environments.

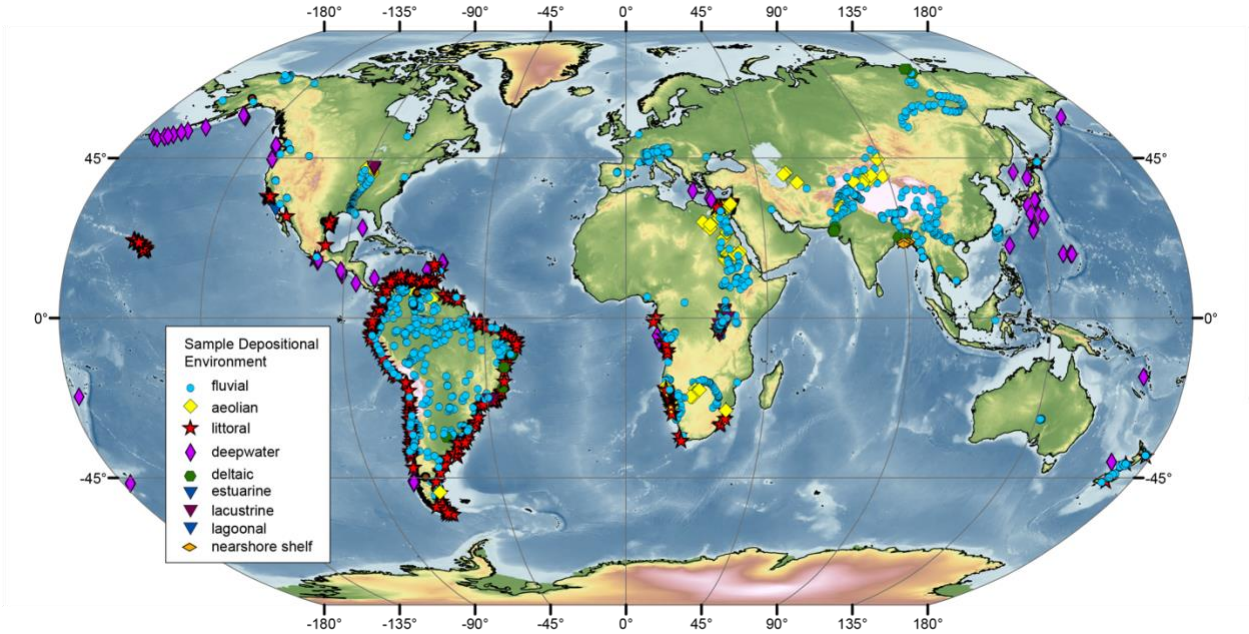

Figure 3.2. Map of 3,512 Pleistocene-modern sand samples and their depositional environments.

3.2. Model Performance – Q-F-L

The mean R^2 score of the 100 ShuffleSplit ln(F/Q) models is 0.654 ± 0.031 (1-sigma) with minimum and maximum scores of 0.560 of 0.730, respectively. For ln(L/Q), the mean R^2 score is 0.706 ± 0.023 (1-sigma) with minimum and maximum scores of 0.654 and 0.770, respectively (Fig. 3.3; see Supplemental Table D for results from all models). The validation dataset used to evaluate model performance during the 100 ShuffleSplit iterations contains 61,100 total samples for each of the ln(F/Q) and ln(L/Q) models. Figure 3.4 shows a scatter of the true log-ratio composition against the validation prediction with 1-, 2- and 3-sigma confidence ellipses. The imputation of the zero values for lithic and quartz data leads to lower prediction confidence for ln(L/Q), as the true imputed values are compared against a wide range of values for the validation prediction (Fig. 3.4).

Log-ratios from the validation prediction were inverse transformed to provide an estimate of prediction error (predicted minus true composition) in Q-F-L ternary space (Fig. 3.5). Error in the prediction of quartz averages 2.6% ± 15% (1-sigma) indicating a slight overprediction of quartz abundance relative to its true values in the validation dataset. Prediction error for feldspar is -1.1% ± 9.4% (1-sigma), and for lithics is -1.5% ± 15.4% (1-sigma), suggesting a slight underprediction of these components relative to their true values.

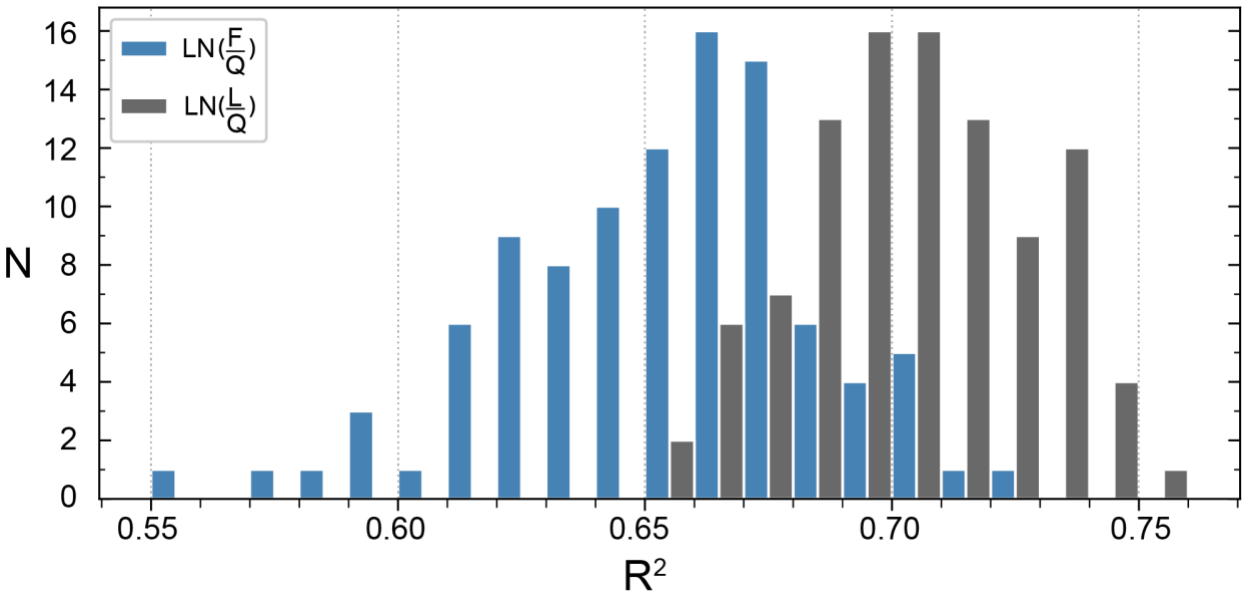

Figure 3.3. Frequency distribution of R^2 scores from 100 ShuffleSplit iterations for ln(F/Q) and ln(L/Q) models. On average, ln(L/Q) models outperform ln(F/Q) models and display a narrower distribution.

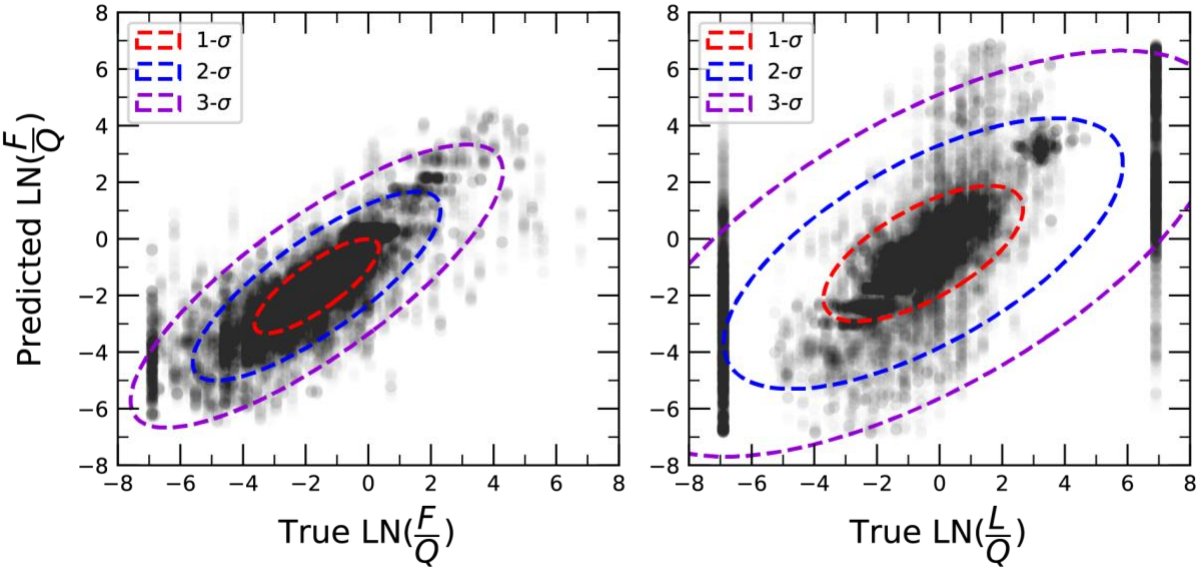

Figure 3.4. Scatter plots of true log-ratio composition against the validation prediction.

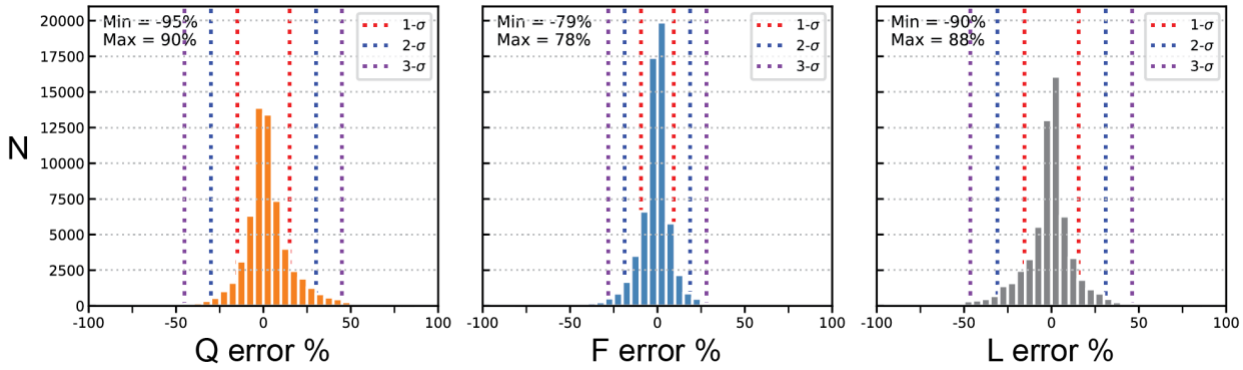

Figure 3.5. Histograms of Q-F-L error percent (prediction minus true) and their standard deviations for 61,100 samples of the validation dataset.

3.3. Global Prediction of Sand Mineralogy (GloPrSM)

3.3.1. Q-F-L Classification

The median of the 100 global ln(F/Q) and ln(L/Q) predictions from this study's internal classification scheme was inverse transformed to ternary Q-F-L values (Eq. 2) and mapped to a cyan-yellow-magenta (CYM) color scale (Fig. 3.6). Total quartz (Q) is predicted to be the most abundant component of sand with an area-weighted abundance of 68.5%. Predicted global Q abundance displays a bimodal distribution with modes at 45-50% and 90-95% (Fig. 3.7a). In general, quartzose sands are most common in low latitudes and within broad, low elevation basins, with relative enrichment predicted to occur in the Gulf Coast Basin of North America, the Amazon foreland basin of South America, and in nearly all of Africa, the Middle East, India, and Australia (Fig. 3.7a).

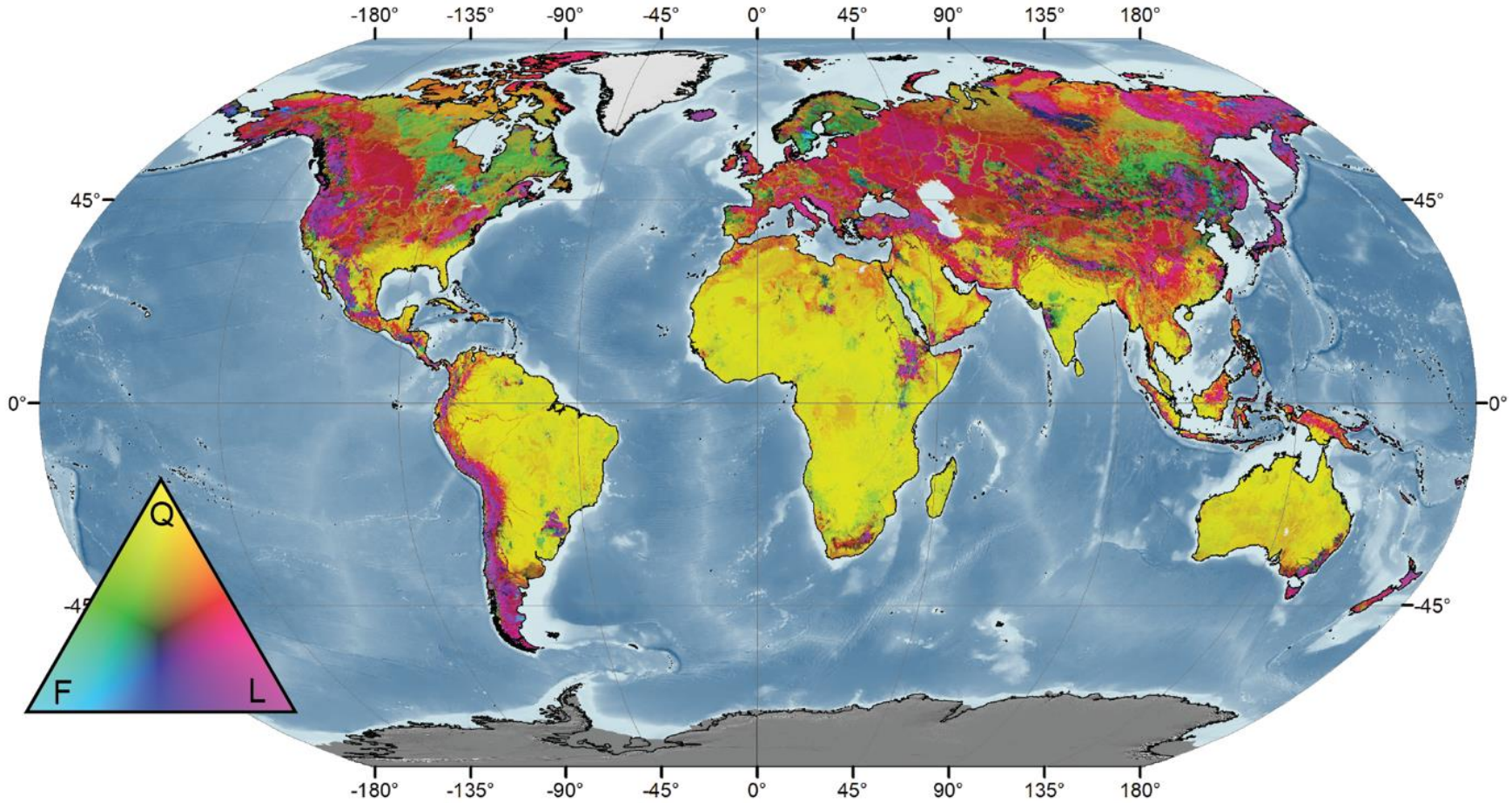

Figure 3.6. Ternary CYM color map of the GloPrSM Q-F-L prediction showing the relative abundance of quartzose (yellow), feldspathic (cyan), and lithic (magenta) grain types, with mixed compositions shown as a gradient between endmember colors.

Lithic grains (L) are predicted to be the second most abundant component of sand, comprising 18% areal abundance globally. Relative enrichment of L is predicted in proximity to convergent plate boundaries, including the circum-Pacific Belt and Alpine-Himalayan belt, ancient collisional mountain belts, intraplate volcanic provinces (e.g., Deccan traps of India), and in catchments that drain these settings (e.g., foreland basins) (Fig. 3.6 and 3.7b). A histogram of globally predicted L shows a mode at 0-5%, reflecting the depletion of this grain type in tropical latitudes. Most catchments are predicted to contain less than ~55% abundance (Fig. 3.7b).

Grains of total feldspar (F) account for 13.5% of the global prediction by area. F is predicted to be most abundant in higher-latitude catchments that are sourced by coarse crystalline bedrock (felsic plutonic and/or metamorphic), including the Canadian Shield, portions of Scandinavia, and the North China Craton (Fig. 3.6 and 3.7c). F is also predicted at modest abundances in association with volcanic rocks, such as those from the East Africa Rift and Deccan Traps, and is generally predicted to be depleted in low-elevation, tropical settings as suggested by the 0-5% mode in the frequency distribution of the prediction (Fig. 3.7c).

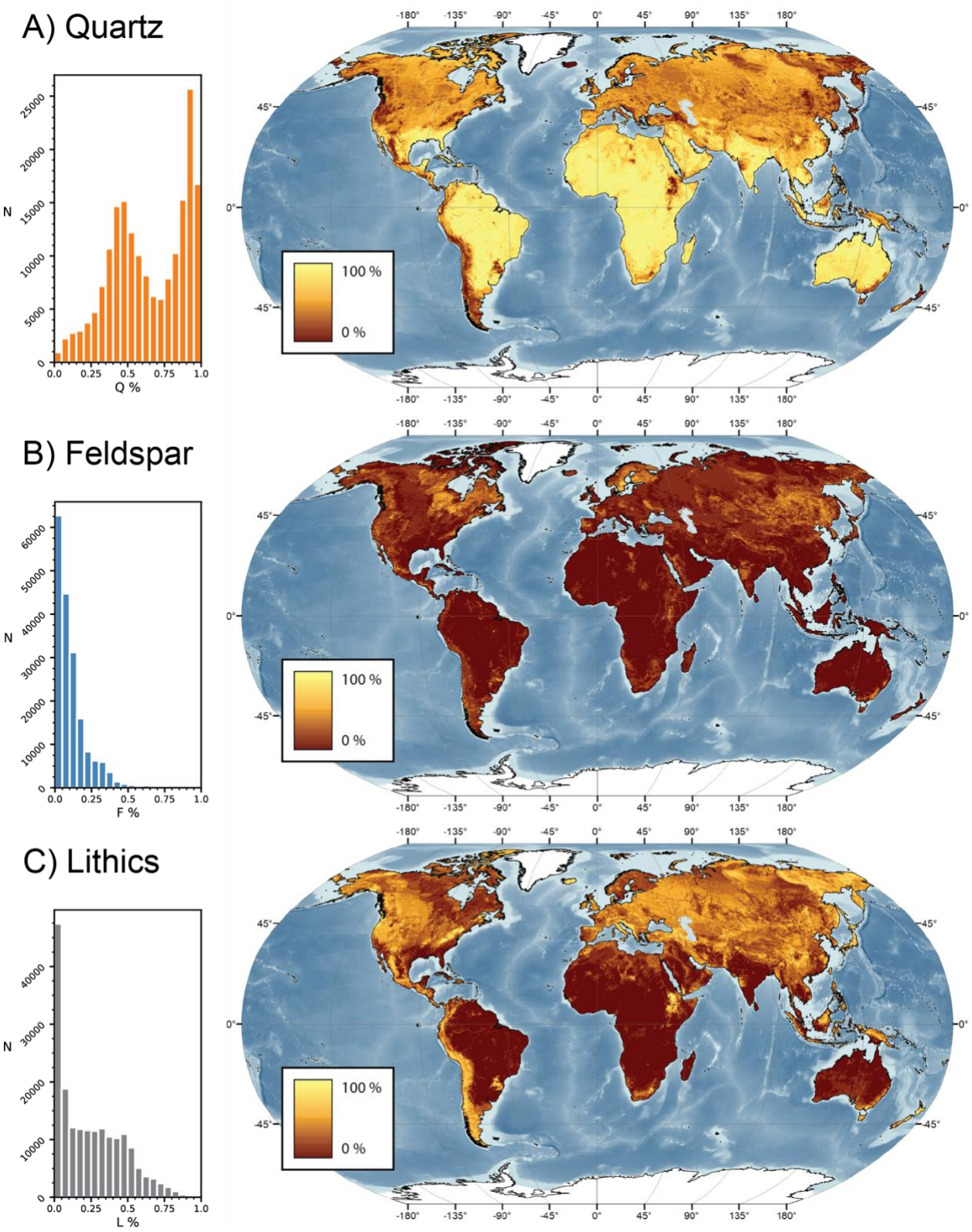

Figure 3.7. Global maps and frequency distribution plots of A) total quartz (Q), B) feldspar (F), and C) lithic (L) abundance from the median of 100 ShuffleSplit iterations.

Figure 3.8 shows the spatial distribution of the inner 95% of the GloPrSM Q-F-L prediction. Globally, inner 95% variance for Q is generally low (< ±5%) where Q is predicted to be most abundant (Fig. 3.8a). These occurrences include vast regions of the Amazon foreland basin, a significant proportion of sub-Saharan Africa, India excepting the Himalaya and the Deccan Traps, northern Australia, and the northern Gulf Coast Basin of the United States. In tropical settings, high variance of Q occurs where volcanic lithic input is significant, particularly in smaller catchments draining the Paraná-Etendeka flood basalts, the East Africa Rift, and the Deccan Traps (Fig. 3.8a). In the northern hemisphere north of the Tropic of Cancer, high variance of Q is observed in east China, northeast Siberia, and the middle reaches of the Mississippi River basin north of the Gulf Coast Plain (Fig. 3.8a).

Predicted F generally displays low variance in many of the catchments where feldspar abundance is low to begin with, particularly in tropical latitudes and settings distant from coarse crystalline bedrock and volcanic input (Fig. 3.8b). In settings where F is locally and regionally predicted at modest abundance, variance is also high suggesting low confidence in the prediction of higher abundances of F. These settings include many of the world's volcanic provinces, such as the Deccan and Siberian Traps (Fig. 3.8b).

For the prediction of L, inner 95% variance is generally lowest in tropical latitudes excepting the aforementioned volcanic provinces (Fig. 3.8c). North of the tropics, variance is low in the vicinity of regions with high predicted L values, including the Siberian Traps and the Columbia River basin of North America (Fig. 3.8c). Variance is notably high locally in western Europe, along the Arctic coast of Siberia, in the northern reaches of the Nunavut territory in Canada, and in the Mississippi River basin (Fig. 3.8c).

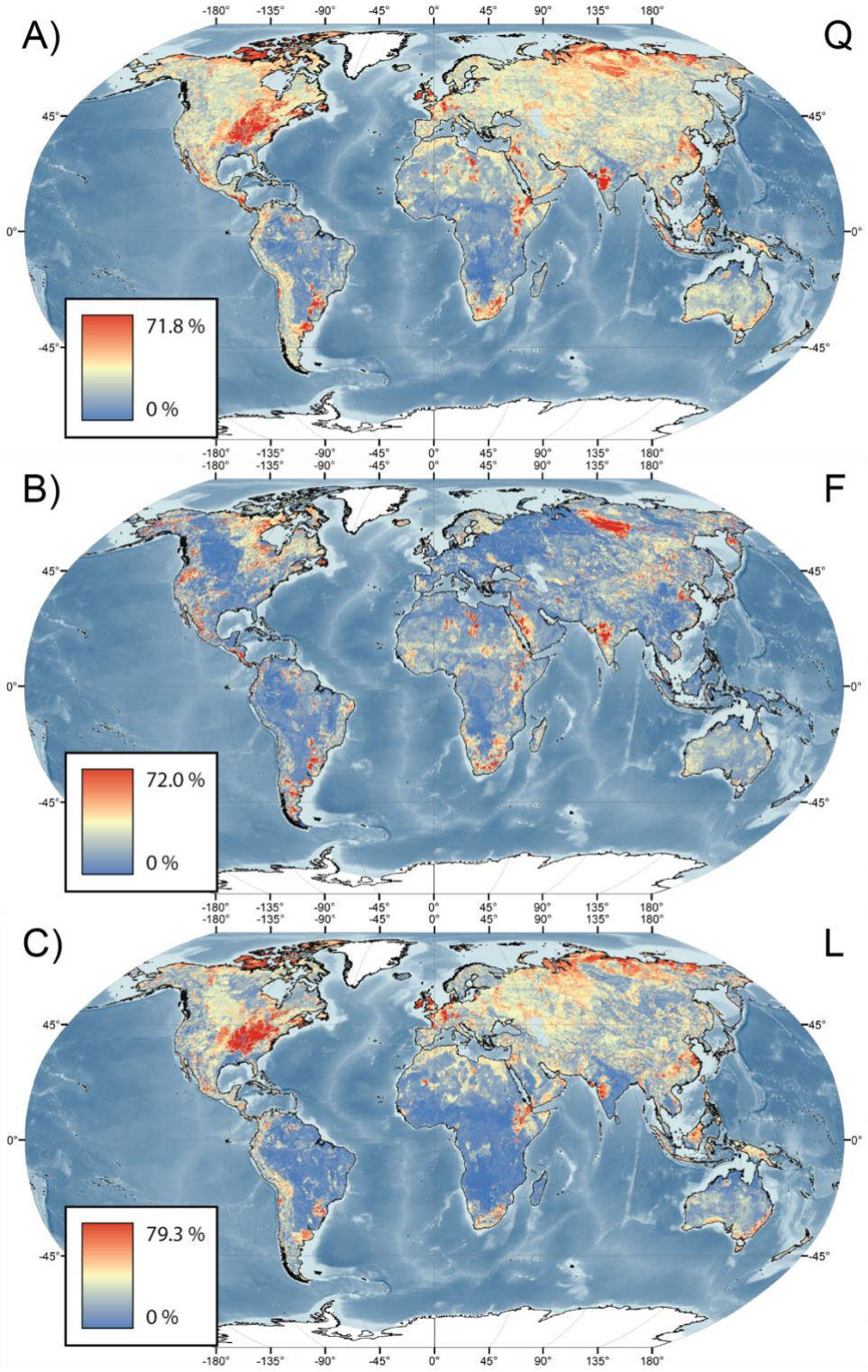

Figure 3.8. Global variance maps (inner 95%) of the Q-F-L prediction with areas of lowest variance in blue and highest variance in red.

An example of the range of Q-F-L predictions from the 100 ShuffleSplit predictions is shown for four BasinATLAS level 8 catchments in Figure 3.9. Some catchment predictions display little variation, with less than ±5% Q, F or L (e.g., green circles). Other catchments display a wider variation with some predicted Q-F-L components experiencing the full range of abundance (e.g., 0–100%; red or black circles). Approximately 1% of watersheds contain bimodal or multimodal distributions of predicted Q-F-L producing some of the highest variances in the GloPrSM model. However, within each mode, variance is typically low.

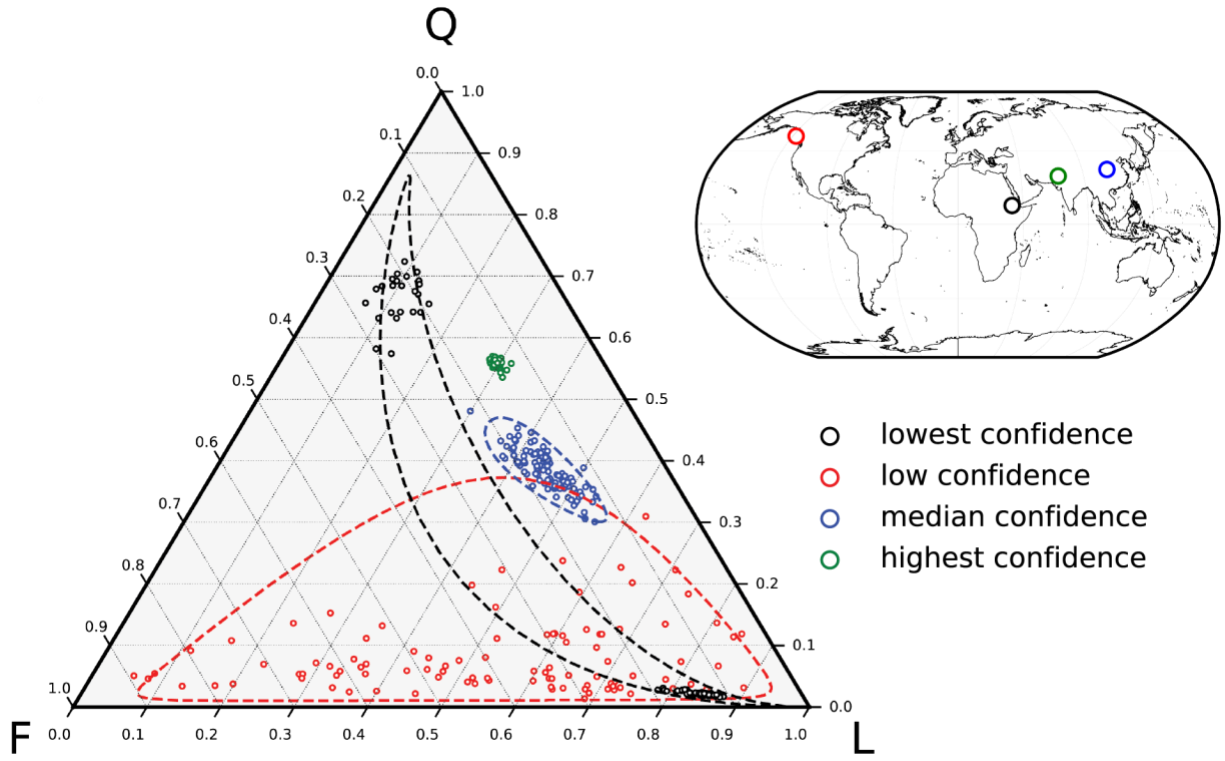

Figure 3.9. Q-F-L ternary diagram showing the 100 ShuffleSplit predictions for four BasinATLAS catchments and their 95% confidence intervals. Locations for each catchment are shown in the inset map.

3.3.2. Octonary Classification

Figure 3.10 shows the predicted global abundances of the eight modeled Q-F-L subcompositions. Monocrystalline quartz (Qm) is the dominant component of the quartzose fraction of Q-F-L and composes 62.3% of sand at Earth's surface, on average (Fig 3.10a). Qm

follows similar spatial trends as the prediction of Q, being most abundant in low latitudes and low-elevation catchments and least abundant in high elevation catchments and in areas dominated by mafic volcanic lithologies. In the northern hemisphere, relative enrichment of Qm is predicted to occur along the course of large rivers draining low elevation basins, such as the Ob and Lena River systems and their major tributaries. Qm values as high as 60-70% are predicted in catchments draining basement exposures such as the Anabar Shield of the Siberian Craton and the Canadian Shield in northeast Nunavut. Polycrystalline quartz (Qp) accounts for approximately 5.8% of the aerially normalized global prediction and never exceeds 53.2% abundance in any drainage. Highest Qp is predicted to exist in broad, low elevation basins and in catchments draining sedimentary bedrock and unconsolidated sediments away from volcanic lithic input (Fig. 3.10b). Examples include western Africa, Saudi Arabia, the Canadian Rockies of North America, northwest Siberia between the Ob and Yenisei Rivers, and northeast Siberia in the Verkhoyansk range. Chert (Qch) is the least abundant subcomposition globally accounting 0.47% of the aerially normalized prediction. Like Qp, Qch is most abundant in catchments containing sedimentary bedrock, never exceeding 12.5% in predicted abundance in any drainage. Notable locations of higher abundance include the Nullarbor Plain of southern Australia, the Paleozoic sedimentary bedrock in the southeastern portion of the Boreal Plains, the Hudson Bay Lowlands of Canada, the Baltica region of western Russia, and the eastern portion (peripheral foreland) of the Siberian craton (Fig. 3.10c).

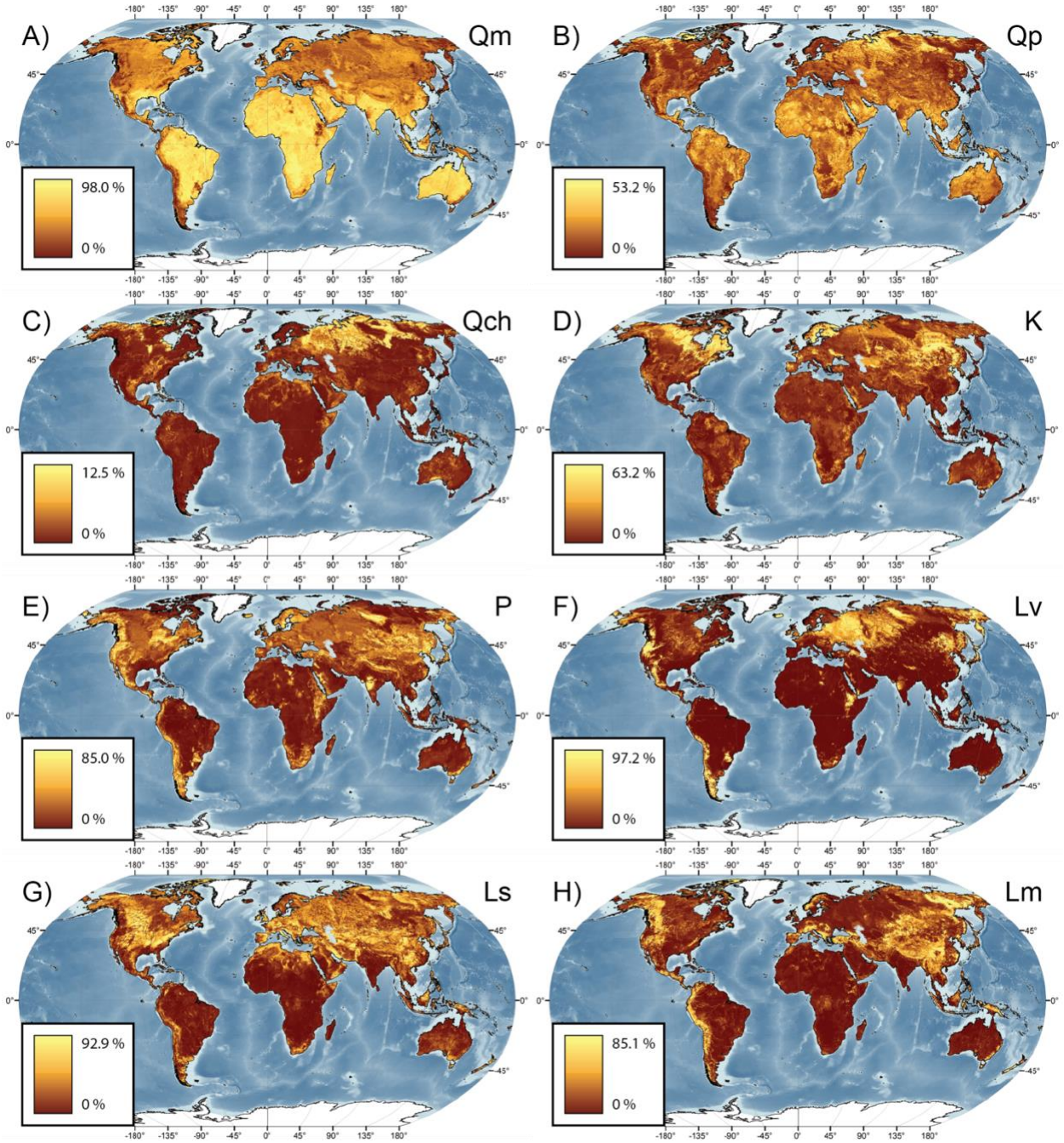

Figure 3.10. Maps of the global prediction for eight Q-F-L subcompositions: a) monocrystalline quartz, b) polycrystalline quartz, c) chert, d) alkali feldspar, e) plagioclase feldspar, f) volcanic lithics, g) sedimentary lithics, and h) metamorphic lithics.

The prediction of alkali feldspar (K) averages 7.2% with low values in low latitudes and highest values in catchments draining the coarse, crystalline basement of ancient cratons (Fig. 3.10d). Regions of high predicted abundance of K include the Canadian Shield of North

America, the Baltic Shield of Scandinavia, the Ukrainian Shield of eastern Europe, and the Aldan Shield of northeast Asia (Fig. 3.10d). Plagioclase feldspar (P) is predicted to be slightly less abundant (6.3%) than K (7.2%) and is predicted to coexist with K in the aforementioned crystalline basement exposures but at lower abundance relative to K (Figs. 3.10e and 3.11). Unlike K, P is more abundant at lower latitudes, particularly near exposures of volcanic bedrock including the Mesozoic Paraná-Etendeka flood basalts of South America and the Late Cretaceous Deccan Traps of west-central India (Fig. 3.10e). P is also predicted at modest abundances within large tropical rivers such as the Nile and Limpopo Rivers of Africa (Fig. 3.10e). At higher latitudes, P is predicted to be common in active margins such as the North and South American Cordillera, and in vast exposures of flood basalts including the Paleozoic-Mesozoic Siberian Traps, and the Cenozoic Columbia River flood basalts of North America.

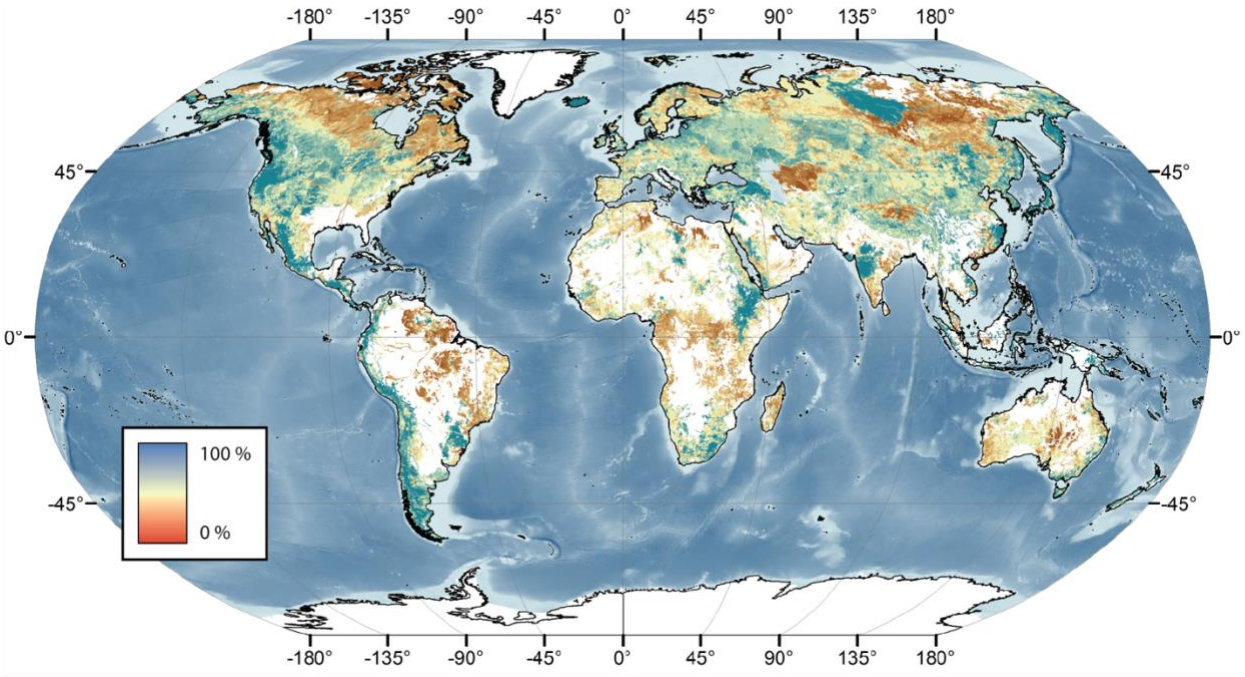

Figure 3.11. Global map of predicted plagioclase abundance as a percentage of total feldspar. Areas with <5% total feldspar are displayed in white.

Volcanic lithics (Lv) are predicted to average 4.8% abundance globally and follow similar spatial trends as P (Fig. 3.10e-f). Elevated Lv predictions are most common in regions

characterized by active volcanism, such as in the East Africa Rift, eastern Iceland, the eastern Kamchatka Peninsula, and central Honshu, Japan. Vast regions of relatively high Lv are also predicted in the broad expanse of the Baltica paleocontinent of eastern Europe (Southern Baltica and Volgo-Uralia) and east of the Ural Mountains of Russia. Sedimentary lithics (Ls) are predicted to comprise an average aerial abundance of 6.1%, with particularly high (>60%) values in catchments draining highlands primarily composed of sedimentary bedrock (Fig. 3.10g) such as the Ozark region of Arkansas, the Appalachian peripheral foreland, the Northern Apennines of Italy, and the eastern Balkan region of southeastern Europe (Fig. 3.10g). Lastly, metamorphic lithics (Lm) are predicted at an average aerial abundance of 7.1%. Lm are predicted to be most abundant in extensive fold-thrust belts and collisional orogens that form many of the world's elevated mountain ranges (Fig. 3.10h), including the North and South American Cordillera, the Alps of Europe, the Pontic Mountains of Turkey, the Tian Shan of central Asia, the Himalaya, the Verkhoyansk and Sikhote-Alin Ranges of east Siberia, the New Guinea Highlands, and Southern Alps of New Zealand (Fig. 3.10h).

Chapter 4: Discussion

4.1. Assumptions and Limitations

1) Point count error is a source of uncertainty in the dataset. The number of counts per sample (typically 200-400) dictates the counting error associated with any given result (Van der Plas and Tobi, 1965). Operator error in misclassifying point counts is another source of uncertainty that is difficult to quantify (Weltje, 2002). Finally, methodological differences in point counting introduce an additional source of uncertainty (Table 2.1). Approximately 61.2% of samples were compiled from sources who report using the Gazzi-Dickinson method (Ingersoll et al., 1984), 31.3% of samples are from sources that

use a different method, and the remaining 7.5% of samples are from source publications that do not report their method. Samples collected by methods other than the Gazzi-Dickinson were included to maximize the size of the training dataset when generating the global prediction (e.g., Potter, 1993). Point count methodology primarily affects the treatment of encountered rock fragments and polymineralic grains, with implications for grain comminution and a corresponding mineralogical dependence on grain size (Ingersoll et al., 1984). Operators who do not use the Gazzi-Dickinson method are likely to count higher values of lithic fragments than operators who use the Gazzi-Dickinson method. Spatial uncertainty in the global prediction resulting from a mixed-method dataset may manifest as the conflation of polymineralic grains (e.g., Lv and Ls) and their monocrystalline derivatives (e.g., P and Qm).

2) The data model (Supplemental Table B) is a standardization of the grain types and descriptions encountered in source publications and was designed to reduce database complexity and allow comparison between samples containing similar petrographic observations. Information may have been lost and uncertainty introduced in the process of distilling 287 uniquely reported petrographic labels to the final list of 54 labels.
3) The presence of zero values may indicate either the complete absence of certain components in a sample or that components present in trace amounts were not encountered during point counting (Weltje, 2002). Not only are zero values incompatible with log ratio transformation, but variation in the proportions of the missing component may be a passive response to the variations in one of the components that were present yet unmeasured (Weltje, 2012). Under the assumption that missing components in database samples are present in trace amounts within their original sample, zeros within

recalculated proportions were replaced with values of 0.1%, implying one observation of a component for every 1,000 counts made. When the number of zeros is small, this practice is justified (Weltje, 2002; Garzanti, 2016) and will have a minimal influence on the overall data covariance structure (Vermeesch, 2018). However, if the number of zeros is significant, the use of advanced imputation techniques is advised (Martín-Fernández et al., 2003; Hron et al., 2010, Vermeesch, 2018). Some recalculated Q-F-L proportions have up to 9% of their values as zeros, but for some octonary proportions this value was much higher (up to 68.5%, Supplemental Table E). The choice of imputation method may yield a greater influence on recalculated proportions with a high number of zero values.

4) The database contains 317 Pleistocene samples from marine depositional environments, amounting to approximately 62.9% of all marine samples and 9.9% of all database samples. The inclusion of Pleistocene samples allows improved representation of marine environments. However, these samples are older than modern samples reported from continental depositional settings. Because climate has varied widely during Pleistocene and Holocene time, we use a simplifying assumption that modern values of global temperature and precipitation are representative of average climate conditions operating on sediment production, transport, and deposition.
5) There is a degree of uncertainty added from the sample georeferencing workflow. For fluvial samples, georeferencing was informed by each sample's source publication. However, the locations of smaller, locally named drainages were difficult to identify in some instances (e.g., Garzanti et al., 2005). The best candidate drainage was usually deduced from internet searches and annotated online maps (e.g., Google Maps). More subjective was the placement of manually drawn pour point polygons to calculate littoral

and marine sample catchment boundaries. Littoral cell boundaries were inferred from the source publication, prevailing wind direction in the locality of the sample, and the samples position with respect to discrete physical boundaries that inhibit the longshore transport of sediment including promontories, submarine canyons, and inlets. Marine sample catchment boundaries were inferred predominantly from littoral cell boundaries, especially for nearshore shelf samples. Boundaries for slope and deepwater sample catchments were further informed by bathymetric contours as well as proximity to local submarine channels.

6) Due to the availability of published datasets, the GloPrSM model was largely trained on a fluvial dataset (71% of samples) supplemented by samples from offshore (15.7%) and littoral (13.3%) environments. A prediction was then applied to HydroSHEDS BasinATLAS level 8 terrestrial catchments, including aeolian and glaciated environments where fluvial processes are absent or where their extent is limited. The GloPrSM model thus is designed to represent the composition of fluvially transported sediments and may not be representative of other sediment transport types (e.g., aeolian sediment transport in dune fields).
7) The independent variables used in the GloPrSM model to predict sand mineralogy, including source lithology, climate, transport history, and tectonic setting, are not the only factors that are known to influence sand composition (Johnsson, 1993). Other potentially important factors that are unaccounted for in the GloPrSM model include vegetation (Jackson and Keller, 1970; Knoll and James, 1987), grain size (Ingersoll et al., 1984; Tolosana-Delgado and von Eynatten, 2009), recycling (Garzanti et al., 2013a; Garzanti et al., 2016), sorting and mixing effects within the depositional environment

(Weltje et al., 1996; Garzanti et al., 2013b), residence time (Dosseto et al., 2006; Maher, 2010), and anthropogenic factors. The addition of these variables to the model may result in an improved prediction and reduce spatial uncertainty.

8) The use of first level GLiM lithologies presents a major limiting factor when establishing a relationship between source rock composition and sand mineralogy. For example, the undivided metamorphic class (MT) includes foliated and nonfoliated rocks across a wide variety of metamorphic grades (e.g., slate to gneiss) and mineral assemblages (e.g., felsic to mafic; Hartmann and Moosdorf, 2012). Moreover, many assumptions were made in the compilation of GLiM itself; for instance, if metamorphism was weak then the protolith was used when defining lithological class (Hartmann and Moosdorf, 2012). Incorporation of second and third level attributes of the GLiM dataset with higher levels of detail such as grain size information for sedimentary lithologies may improve future model predictions.
9) The GloPrSM model prediction was generated using a random forest algorithm. While RFs are generally efficient and accurate, their performance is contingent upon the datasets they are trained from, thus, predictive power can change considerably with small changes in the training dataset. Other machine learning approaches, such as boosting algorithms or neural networks, may better constrain the nonlinear relationship between predictor variables and sand mineralogy.
10) Although sample coverage is global, certain regions on Earth have sparse coverage (Fig. 3.1). For example, the continent of Australia is characterized by only 30 samples. Other poorly characterized regions include southeast Asia, Canada, western Africa, and western Russia (Fig. 3.1). Predictions of sand mineralogy are inferred to be less reliable in

settings that are unconstrained by the dataset, particularly if those regions represent a unique combination of independent variables that are not represented in the training dataset (see section 4.2).

11) The GloPrSM model was developed to predict the Q-F-L proportions of siliciclastic sand following the aforementioned classification schemes (Supplemental Table F). Thus, the model is not applicable for sands composed of non-Q-F-L components (e.g., bioclastic or intrabasinal grains).

4.2. Spatial Uncertainty

In general, the GloPrSM model predicts low Q-F-L variance along the course of many large rivers, including the Mississippi, Missouri, Mackenzie, Yenisei, Lena, Ob, Indus, Niger and Congo rivers (Fig. 3.8). However, some large, well-sampled rivers show moderate to high variance along their reaches, particularly where their upstream catchments display high variance in predicted Q-F-L. These rivers include the Lena at the confluence of the Aldan, the White Nile at the confluence of the Blue Nile, and the entirety of the main Amazon River (Fig. 3.8). Better characterization of high-variance regions that these rivers drain may improve model predictions. There are three types of settings where prediction uncertainty (inner 95th percentile) is particularly high:

1. Large Igneous Provinces (LIPs)

Mafic igneous lithologies are the least well represented lithologic classes in database samples (Supplemental Fig. A) and are some of the least abundant lithologic classes in the GLiM dataset (Fig. 2.1, Supplemental Fig. A; Hartmann and Moosdorf, 2012). In catchments draining mafic volcanic lithologies, the GloPrSM model displays high variance in predicted Q-F-L abundances, depending on latitude (Fig. 3.8). In low

latitudes, high uncertainty for Q, F and L can be seen in the Paraná-Etendeka flood basalts, the East Africa Rift, and the Deccan Traps (Fig. 3.8). In these settings, the model struggles to consistently determine the relative proportions of Q and L that should be derived from mafic volcanic sources. These high variances also highlight the complexities that arise when generating a prediction in LIPs where chemical weathering is moderate to intense and the rate of L depletion and corresponding Q enrichment is poorly constrained by the database samples used to train the model.

In settings where chemical weathering is limited, predicted L-variance in LIPs is generally low (Fig. 3.8c). In arid LIPs, such as the western coastal escarpment of the Arabian Peninsula (Hejaz region), uncertainty is more pronounced in Q and F relative to L, suggesting the GloPrSM model may be more confident at predicting L abundances in settings where chemical weathering is less intense. This is further supported by low L variance within middle and high latitude LIPs, such as in the Columbia River Basalt Group, Iceland, and the Siberian traps (Fig. 3.8b). Instead, high uncertainty in these settings is primarily related to Q and F abundance (Fig. 3.8a, b). While dataset coverage may be higher for some tropical LIPs (e.g., Paraná-Etendeka, East Africa; Fig. 3.2), very few analogs exist in the database for high-latitude and arid LIPs. Additional data from these settings will improve the model's predictive power and may help reduce uncertainty.

2. Sedimentary Terranes

Sedimentary lithologies (SM+SS+SU and EV+SC) are the most abundant GLiM lithologic group at Earth's surface (Fig. 2.1; Hartmann and Moosdorf, 2012). Despite sedimentary lithologies being the best represented lithologic class in database samples

(Supplemental Fig. A), several regions with sedimentary bedrock display very high uncertainty from the GloPrSM model (Fig. 3.8). The highest variance in Q and L is observed in small to mid-sized catchments draining the sedimentary terranes of the Appalachian foreland of the eastern United States (Figs. 3.8a, c). While the GloPrSM model was trained on data from catchments draining the Appalachian foreland, samples are located along the course of the Mississippi River and its major tributaries, and thus only the largest catchments in this region are represented (Fig. 3.2). Additional settings where Q and L variances are high include the sedimentary terranes of western Europe, the mixed sedimentary rocks of the Southern Arctic Islands, the Great Dividing Range of Australia, and the vast expanse of siliciclastic sedimentary rocks and unconsolidated sediments in northern Siberia (Figure 3.8a, c). Meanwhile, Q, F, and L variances are high in catchments draining the siliciclastic sedimentary rocks of the Karoo Supergroup in southern Africa, the unconsolidated sediments of the Buenos Aires province of Argentina, and the unconsolidated sediments of the North China Plain near the Yellow Sea.

High variance in sedimentary terranes can most likely be attributed to the nature of the GLiM lithologic database. At the first level of GLiM, siliciclastic sedimentary lithologies (SS) and unconsolidated sediments (SU) are not subdivided by grain size (Hartmann and Moosdorf, 2012). This lack of specificity may affect the predicted abundance of (sub)compositions produced by these lithologies, such as Q (Qm) derived from sandstone and L (Ls) derived from shale and siltstone, as well as the rate at which these (sub)compositions are enriched or depleted during transport and alluvial storage. Furthermore, the Q-F-L composition of SS and SU is not known despite older sediments

and sedimentary rocks likely being just as compositionally diverse as modern sands. Lastly, the mixed sedimentary class (SM) represents a mixture of siliciclastic and carbonate sedimentary rocks which can produce varying amounts of Q or L.

3. Orogens

Orogenic settings (modern and ancient) that display locally high Q-F-L variance include the Sunda Arc of Indonesia, the Verkhoyansk range of northeastern Siberia, the Great Dividing Range of Australia, the Japan Arc, the Sierra Madre Occidental range of Mexico, and the Andes Mountains (Fig. 3.8). A vast range of lithologies are exposed at the surface in many of the world's orogens (Fig. 2.1a), complicating the GloPrSM model's ability constrain the influence of source lithology on sand composition in these settings. For example, Lm is predicted at relatively high abundance in catchments where MT is absent (Fig. 3.10h, Supplemental Fig. F), suggesting significant quantities of Lm may be produced from rocks that are weakly metamorphosed.

To improve upon the predictive power of the GloPrSM model and reduce uncertainty, more data should be collected from LIPs, sedimentary terranes, and orogens. In addition, a higher resolution lithologic dataset may improve upon modeled relationships between source lithology and sand composition. This issue can, in part, be resolved by employing the second- and third-level information contained by the full GLiM database (Hartmann and Moosdorf, 2012).

4.3. Factors Controlling Sand Mineralogy

4.3.1. Elevation and Slope

According to partial dependence (PD) analysis, increasing elevation tends to produce a modest degree of progressive quartz enrichment and lithic depletion, with this effect most

pronounced between 0 – 1 km and above 5 km (Fig. 4.1a). This prediction is in contrast with some previous observations that silicate weathering rates are elevation-independent (Drever and Zobrist, 1992). One explanation for the modeled result relates to sedimentary (SM+SS+SU and EV+SC), metamorphic, and felsic to intermediate plutonic (PA+PI) lithologies, which tend to be Q prone (Supplemental Fig. F) and comprise some of the highest elevation catchments in database samples (Supplemental Figs. B, E) and in the world (Supplemental Figs. C, F).

It has been suggested that slope angle exerts a significant control on the composition of sediments (Basu, 1985; Johnsson, 1993). Slope is ranked as the single most important predictor feature for both log-ratio models (Fig. 4.2). Slope largely determines whether a chemical weathering regime is transport- or kinetic-limited, with low slopes corresponding to long fluid and sediment residence times that allow for the near-complete alteration or removal of framework proportions stored in alluvial environments (Johnsson, 1993; Maher, 2010). As slopes increase, so too does the predicted proportion of L until slopes reach 25º, above which there is a modest shift towards Q enrichment (Fig. 4.1b). The abundance of sedimentary lithologies in many of the Earth's highest mountains may account for the increase in predicted quartz abundance in slopes greater than 25º. Approximately 7% of database samples contain catchment averaged slopes greater than 25º (Supplemental Fig. A), most of which are derived from the sedimentary and metasedimentary terranes of the Himalaya and Alps (Garzanti et al., 2005; Garzanti et al., 2007; Garzanti et al., 2010; Vezzoli et al., 2017). Future sampling efforts could target steeply sloping regions underlain by a more diverse array of lithologies.

4.3.2. Precipitation and Temperature

Precipitation is the fourth most important variable for the ln(L/Q) model and ranks even lower for the ln(F/Q) model (Fig. 4.2). Modest quartz enrichment and feldspar depletion are

predicted as precipitation increases from approximately 200 – 1,000 mm/yr, and between 1,000 – 2,000 mm/yr lithic enrichment occurs (Fig. 4.1d). This pattern suggests that, although water is a critical component of chemical weathering (Maher, 2010), high precipitation alone is not enough to enrich quartz in sand. Quartz enrichment in moderate rates of precipitation (200 –1,000 mm/yr) might reflect a lack of chemical weathering in arid regions and high rates of denudation associated with a kinetically-limited regime in areas where precipitation is very high (Potter, 1978; Johnsson, 1993). Only 11% of dataset samples have precipitation above 2,000 mm/yr or below 200 mm/yr, suggesting that the model is less well constrained for the driest and wettest regions at Earth's surface (Supplemental Fig. A).

Temperature is predicted to have a strong influence on Q-F-L proportions with substantial quartz enrichment above 15ºC (Fig. 4.1e). Temperature ranks as the 2nd highest contributor to model accuracy for the best ln(L/Q) model but has low influence on the best ln(F/Q) model (Fig. 4.2). Because samples with catchment averaged temperatures between -2.5ºC and 27.5ºC are well represented in the sample dataset (Fig. 4.1e), the relationship between temperature and sand composition at 15ºC is well constrained. Furthermore, many studies show a positive, approximately exponential increase in weathering rates with increasing temperature (e.g., White et al., 1999; Oliva et al., 2003; West et al., 2005). The effects of increasing temperature on lithic reduction may be explained by a higher proportion of unstable minerals in these grains and a higher proportion of reactive surfaces with respect to grain volume due to their aphanitic nature (White, 2002; White and Buss, 2003; Israeli and Emmanuel, 2018). The lack of an appreciable influence of temperature on the F:L ratio may reflect both of these constituents being chemically unstable, and thus experiencing concomitant reductions in abundance within hot, wet climates.

4.3.3. Basin Area

Increasing basin area is predicted to yield relative enrichment of Q at the expense of L and with little change in F abundance (Fig. 4.1c). As basin area is a proxy for sediment transport distance (Somme et al., 2009), it is thus inferred that increasing transport distance appears to favor quartz enrichment and lithic dilution. This finding is agreeable with studies that show a reduction in the proportion of unstable lithic fragments during transport (Cameron and Blatt, 1971; Shukis and Ethridge, 1975). Further, the rate at which the Q:F ratio increases with basin area (Supplemental Fig. D) is consistent with other studies that quantify feldspar depletion during sediment transport (Mackie, 1896; Martens, 1931; Plumley, 1948; Hayes, 1962; McBride et al., 1996). As basin area increases to $2x10^6$ km^2, quartz proportions increase gradually. At and above $2x10^6$ km^2 there is a sharp increase in Q and F with a corresponding decrease in L (Fig. 4.1c). These largest catchments account for approximately 8% of database samples and are predominantly from low latitudes (Supplemental Fig. B) where chemical weathering rates are likely higher but the effects mechanical and tropical weathering are difficult to separate from one another (Johnsson, 1993). Moreover, mechanical comminution of softer grains during transport may lead to their redistribution in different size fractions (Cather and Folk, 1991) and accelerated chemical weathering following an inverse-proportional relation between grain size (i.e., mineral surface area) and weathering rates (White, 2002; White and Buss, 2003).

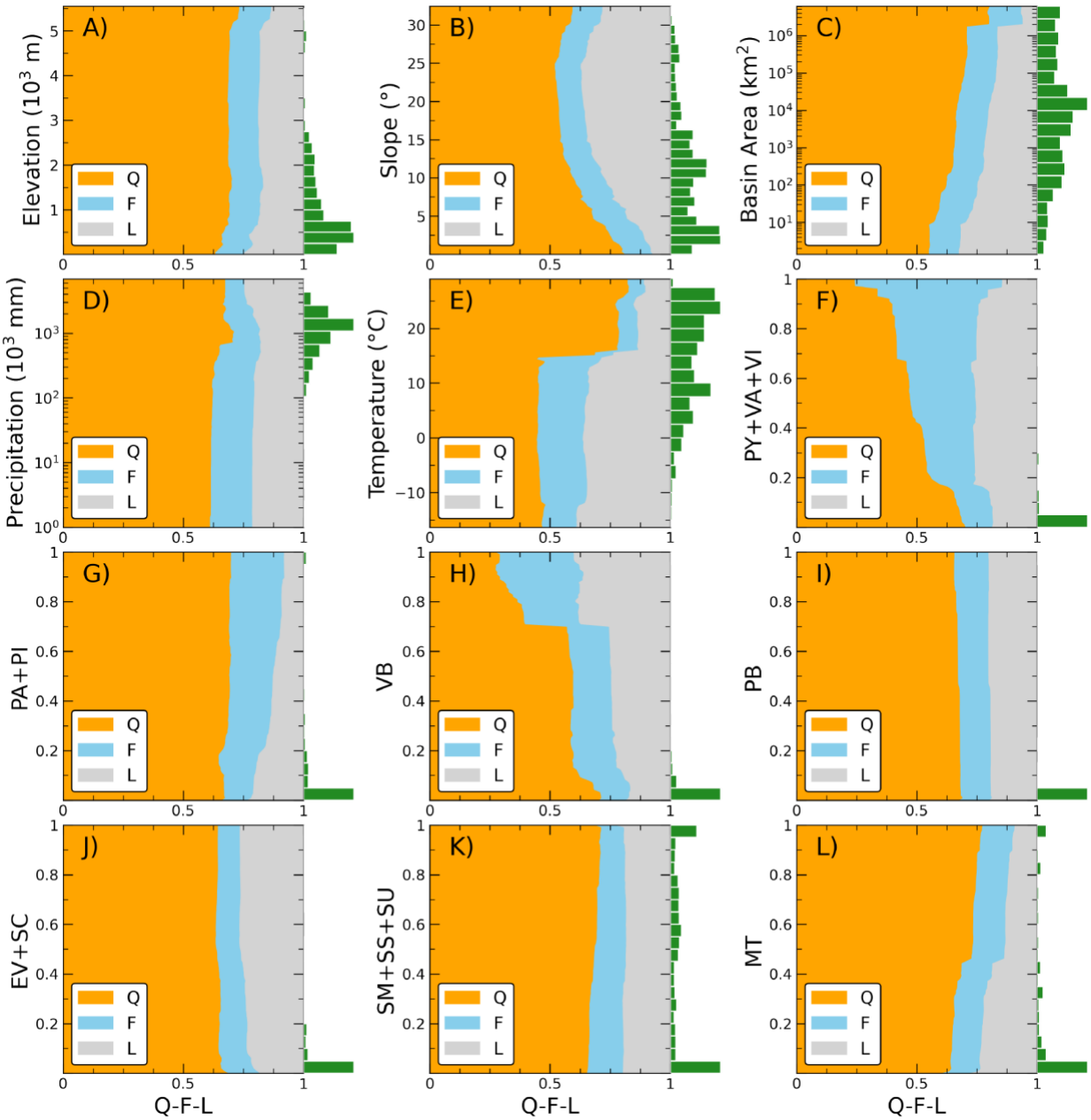

Figure 4.1. Stack plots of independent variable partial dependence for the highest-scoring ln(F/Q) and ln(L/Q) models. Histograms to the right of each variable's stack plot show data density of the training dataset.

4.3.4. Lithology

Of all lithologic variables, the felsic to intermediate volcanic and pyroclastic lithologic group (PY+VA+VI) is predicted by PD analysis to exert the greatest control on Q-F-L proportions, with increasing PY+VA+VI predicted to lead to the rapid enrichment of feldspar

and depletion of quartz (Fig. 4.1f). It should be noted that PY+VA+VI are among the least abundant lithologic classes globally and only compose 3.3% of the Earth's unglaciated surface area (Hartmann and Moosdorf, 2012). Thus, the spatially limited extent may explain the low permutation importance of PY+VA+VI, ranking 8th and 11th for ln(L/Q) and ln(F/Q) models, respectively (Fig. 4.2). Meanwhile, the MT lithologic class is among the most important variables for both log-ratio models, ranking 2nd most important for ln(FQ) and 3rd for ln(L/Q) (Fig. 4.2). Increasing MT abundance corresponds to a decrease in L proportions (Fig. 4.2l) while the F:Q ratio is predicted to hold constant (Supplemental Fig. D), however, high values of MT are similarly underrepresented in database samples (Supplemental Fig. A).

Other lithologies that exert significant control on Q-F-L proportions include the mafic volcanic (VB) class whose increasing abundance is predicted to increase F and L proportions at the expense of Q with rapid lithic enrichment above 70% VB (Fig. 4.1h). Data are scarce for this lithologic class as well as for the PA+PI group whose increasing abundance is predicted to increase F proportions at the expense of L while holding Q abundance constant (Fig. 4.1g). PD analysis suggests that the remaining lithologic variables have little influence on sand mineralogy (e.g., mafic plutonic lithologies [PB] and the evaporites + carbonate sedimentary group [EV+SC], Fig. 4.1i, j). Although SM+SS+SU is very abundant globally, the lack of its predicted influence on Q-F-L compositions may reflect the undifferentiated nature of this lithologic class. In other words, siliciclastic sediments and sedimentary rocks are likely diverse in terms of their Q-F-L composition and thus can produce a wide variety of Q-F-L sand compositions that are ultimately averaged in the GloPrSM model prediction (Fig. 4.1k). A more detailed lithologic category that differentiates quartzose, feldspathic, and lithic siliciclastic sediments and sedimentary rock may thus improve the GloPrSM prediction.

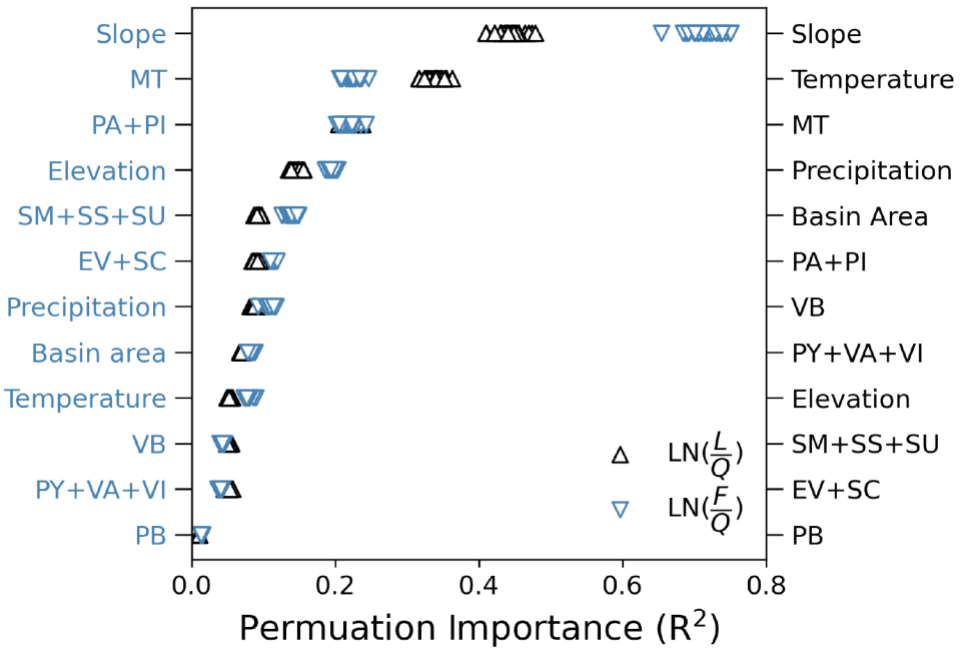

Figure 4.2. Results of permutation importance analysis for the highest-scoring ln(F/Q) (y-axis left) and ln(L/Q) (y-axis right) GloPrSM models.

4.4. Synthesis

The influence of source lithology on sand mineralogy is more prominent in middle to high latitudes where the effects of tropical weathering are reduced (Fig. 4.3). For example, the proportion of predicted F may correspond to an increase in PA+PI abundance between 35˚N to 70˚N and 35˚S to 55˚S, and an increase in predicted Ls may also correspond to an increase in SM+SS+SU north of 70ºN (Fig. 4.3a–b). Increased Lv abundance between 35ºS and 50ºS may correspond to the increase in volcanic lithologies (VB and PY+VA+VI) seen at these latitudes. Lastly, increased Lm at around 40ºS may result from increased MT at this latitude (Fig. 4.3a–b).

In low latitudes (35ºS to 35ºN), however, Q-F-L proportions cannot be easily explained by lithological variability (Fig. 4.3a–b). Transitioning from high to low latitudes, average Q abundance is predicted to increase from 30-40% to 80% of Q-F-L proportions (Fig. 4.3a). This exceptional enrichment cannot be explained by any single variable's influence (Fig. 4.1) but by the combined effects of several factors. Large drainage areas and high temperatures observed in

low latitudes (Fig. 4.3c) are predicted to increase Q proportions (Fig. 4.1a, c, e). Globally, low average slopes are also found in these latitudes (Fig. 4.3c), yielding transport-limited weathering regimes and amplifying the rate of Q generation and L depletion (Fig. 4.1b). Despite the seemly unimportant influence of precipitation on Q-F-L proportions (Fig. 4.1d), high average precipitation in low latitudes not only provides a medium for chemical work to be enacted on sediments, but also transports sediments from high elevation, high slope catchments in a kinetically-limited regime to large, low elevation, gently sloping basins where the residence times are long and transport-limited weathering dominates.

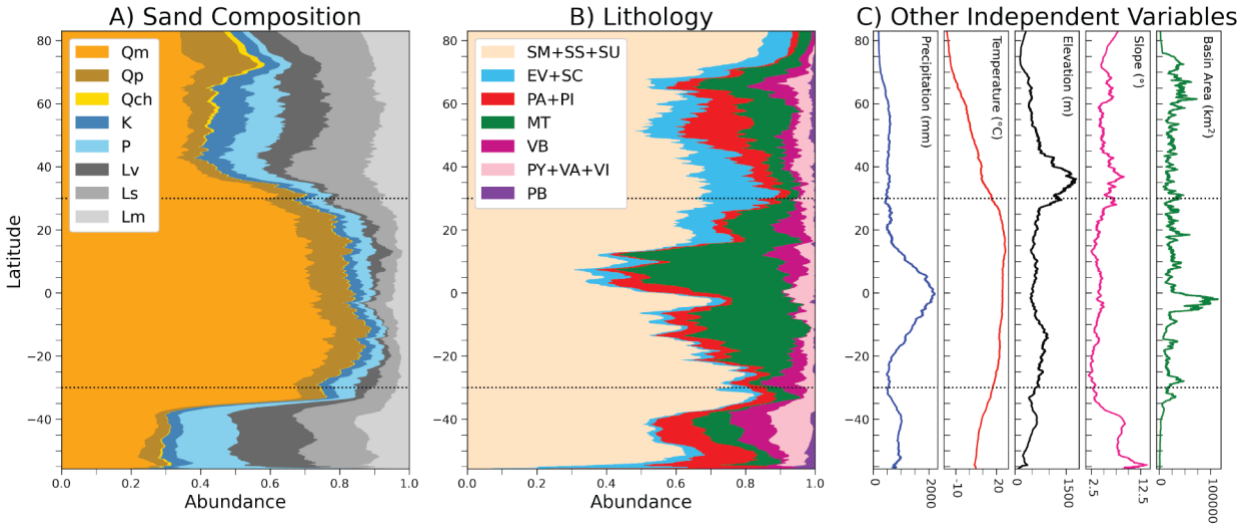

Figure 4.3. Comparison of the GloPrSM Q-F-L prediction moving average to the moving average of independent variables as a function of latitude with the positions of 30ºN and 30ºS shown as black dotted lines. a) Stack plot of GloPrSM Q-F-L abundance displayed as octonary proportions. b) GLiM proportions of recalculated lithologic groupings. c) Climatic and topographic independent variables.

Conclusions

This study introduces a framework for predicting sand mineralogy at a global scale based on fundamental controls. The GloPrSM model is a data driven, forward sediment generation model trained to predict Q-F-L proportions from source lithology, climate (precipitation and temperature), and topographic factors including slope, elevation, and basin area. The model was extended to HydroSHEDS BasinATLAS level 8 fluvial catchments. The median of 100 global predictions shows quartz enrichment occurs in low latitudes (35ºS to 35ºN) and in large, gently sloping basins characterized by high precipitation and temperature. Feldspar enrichment occurs in mid to high latitude catchments draining coarse-crystalline bedrock where chemical weathering processes are suppressed. Lithic enrichment occurs near the world's active and ancient orogens and within intraplate volcanic provinces. Variance within the inner 95% (2.5^{th} to 97.5^{th} percentile) of predictions reveals spatial uncertainty is greatest in large igneous provinces, sedimentary terranes, and orogens where unique combinations of some fundamental controls are underrepresented in dataset samples.

The influence of fundamental controls of mineralogic proportions in sand can be quantified using model interrogation techniques such as permutation importance and partial dependence analysis. Slope is the most important feature contributing to model accuracy and largely determines if weathering occurs in a transport-limited or kinetically-limited regime; steep slopes favor lithic enrichment and low slopes favor quartz enrichment. Temperature is the most influential climate factor on Q-F-L proportions with rapid quartz enrichment occurring in catchments with average temperatures greater than 15ºC. Some lithologic variables are predicted to exert a significant control on Q-F-L proportions, including PY+VA+VI, VB, and MT.

The expression of the Q-F-L composition of sand at Earth's surface cannot be attributed to a solitary factor, but rather the interplay and feedbacks between multiple factors. Where these factors are well understood, it is possible to predict how their unique combinations will manifest in mineralogy of sand, and ultimately within the global sedimentary archive. Although the GloPrSM model represents a first step towards predicting the Q-F-L composition of modern sand globally, directed future sampling, improved characterization of bedrock lithology, and the incorporation of additional factors that control sand mineralogy will improve the reliability and spatial resolution of the global prediction.

References

Aagaard, P., Egeberg, P.K., Saigal, G.C., Morad, S., and Bjørlykke, K., 1990, Diagenetic albitization of detrital K-feldspars in Jurassic, Lower Cretaceous, and Tertiary Clastic reservoir rocks from offshore Norway, II. Formation water chemistry and kinetic considerations: Journal of Sedimentary Petrology, v. 60, p. 575-581.

Aitchison, J., 1982, The Statistical Analysis of Compositional Data: Journal of the Royal Statistical Society. Series B (Methodological), v. 44, p. 139-177.

Amante, C., and Eakins, B.W., 2009, ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis: NOAA Technical Memorandum NESDIS NGDC-24. National Geophysical Data Center, NOAA. Doi:10.7289/V5C8276M [September 2021].

Basu, A., 1985, Influence of climate and relief on compositions of sands released at source areas, in Zuffa, G. G., ed., Provenance of arenites: Dordrecht, D. Reidel, p. 1-18.

Behrens, R., Bouchez, J., Schuessler, J.A., Dultz, S., Hewawasam, T., and von Blanckenburg, F., 2015, Mienralogical transformations set slow weathering rates in low-porosity metamorphic bedrock on mountain slopes in a tropical climate: Chemical Geology, v. 411, p. 283-298.

Berner, R. A., and Berner, E. K., 1997, Silicate weathering and climate, *in* Ruddiman, W. F., ed, Uplift and Climate Change: Springer, Boston, MA, p. 353-364.

Berner, R.A., Lasaga, A.C., Garrels, R.M., 1983, The carbonate-silicate geochemical cycle and its effect on atmospheric carbon dioxide over the past 100 million years: American Journal of Science, v. 283, p. 641-683.

Bhatia, M.R., 1983, Plate tectonics and geochemical composition of sandstones: Journal of Geology, v. 91, p. 611-627.

Bhatia, M.R., and Crook, K.A.W., 1986, Trace element characteristics of graywackes and tectonic setting discrimination of sedimentary basins: Contributions to Mineralogy and Petrology, v. 92, p. 181-193.

Bjølykke, K., 1984, Formation of secondary porosity: How important is it? *in* McDonald, D. A., and Surdam, R. C., eds., Clastic diagenesis: American Association of Petroleum Geologists Memoir 37, p. 277-286.

Blatt, H., 1982, Sedimentary Petrology. W.H. Freeman, New York, 514 pp.

Bloch, S., 1994, Effect of detrital mineral composition on reservoir quality, in M. D. Wilson Ed., Reservoir Quality Assessment and Prediction in Clastic Rocks. Society of Economic Paleontologists and Mineralogists Short Course, 30, p. 161-182.

Boles, J.R., 1982, Activate albitization of plagioclase, Gulf Coast Tertiary: American Journal of Science, v. 282, p. 165-180.

Boles, J.R., 1984, Secondary porosity reactions in the Stevens Sandstone, San Joaquin Valley, California, *in* McDonal, D.A., and Surdam, R.C., eds., Clastic diagenesis: American Association of Petroleum Geologists Memoir 37, p. 217-224.

Brady, P.V., and Carroll, S.A., 1994, Direct effects of CO_2 and temperature on silicate weathering: Possible implications for climate control: Geochimica et Cosmochimica Acta, v. 58, p. 1853-1856.

Brantley, S.L., and Lebedeva, M., 2011, Learning to read the chemistry of regolith to understand the critical zone: Annual Review of Earth and Planetary Sciences, v. 39, p. 387-416.

Brantley, S.L., Goldhaber, M.B., and Ragnarskottir, K.V., 2007, Crossing disciplines and scales to understasnd the Critical Zone: Elements, v. 3, p. 307-314.

Breiman, L., 2001, Random Forests: Machine Learning, v. 45, p. 5-32.

Cameron, K. L., and Blatt, H., 1971, Durabilities of sand-size schist and "volcanic" rock fragments during fluvial transport, Elk Creek, Black Hills, South Dakota: Journal of Sedimentary Petrology, v. 41, p. 565-576.

Cather, S. M., and Folk, R. L., 1991, Pre-diagenetic sedimentary fractionation of andesitic detritus in a semi-arid climate: An example from the Eocene Datil Group, New Mexico, in Fisher, R. V., and Smith, G. A., eds., Sedimentation in volcanic settings: Tulsa, Oklahoma, SEPM Special Publication 45, p. 211-226.

Chamov, N.P., and Murdmaa, I.O., 1995, Coarse fraction minerals of sands in Cascadia margin sediments: Proceedings of the Ocean Drilling Program, Scientific Results, Vol. 146, p. 33-43.

Chayes, F., 1956. Petrographic Modal Analysis. Wiley, New York, 113 pp.

Chayes, F., 1960, On correlation between variables of constant sum: Journal of Geophysical Research, v. 65, p. 4185-4193.

Chilingar, G. V., 1964, Relationship between porosity, permeability and grain size distribution of sands and sandstones, in Van Straaten, L. M. J. U, eds, Deltaic and Shallow Marine Deposits: Elsevier, New York, p. 71-75.

Colbourn, G., Ridgwell, A., and Lenton, T.M., 2015, The time scale of the silicate weathering negative feedback on atmospheric CO_2: Global Biogeochemical Cycles, v. 29, p. 583-596.

Costa, J.E., 1975, Effects of agriculture on erosion and sedimentation in the Piedmont Province, Maryland: Geological Society of America Bulletin, v. 86, p. 1281-1286.

Crook, K.A.W., 1960, Classification of arenites: American Journal of Science, v. 258, p. 419-428.

Dalal, R.C., and Probert, M.E., 1997, Soil nutrient depletion, *in* Clarke, A.L., and Wylie, P.B., eds., Sustainable crop production in the subtropics: an Australian perspective, p. 42–63.

Dickinson, W.R., 1970, Interpreting detrital modes of graywacke and arkose: Journal of Sedimentary Petrology, v. 40, p. 695-707.

Dickinson, W.R., 1985, Interpreting provenance relations from detrital modes of sandstones, *in* Zuffa, G.G., ed., Provenance of Arenites: Dordrecht, D.R., p. 333-362.

Dickinson, W.R., and Suczek, C.A., 1979, Plate tectonics and sandstone compositions: American Association of Petroleum Geologists Bulletin, v. 63, p. 2164-2182.

Dickinson, W.R., Beard, S.L., Brakenridge, G.R., Erjavec, J.L., Ferguson, R.C., Inman, K.F., Knepp, R.A., Linkberg, F.A., and Ryberg, P.T., 1983, Provenance of North American Phanerozoic sandstones in relation to tectonic setting: Geological Society of America Bulletin, v. 94, p. 222-235.

Döll, P., Kaspar, F., and Lehner, B., 2003, A global hydrological model for deriving water availability indicators: model tuning and validation: Journal of Hydrology, v. 270, p. 105-134.

Doran, L.M., Marsaglia, K.M., and Browne, G.H., 2020, Reconnaissane composition of river sand from northern South Island, New Zealand: a modern analogue for southern Taranaki Basin: New Zealand Journal of Geology and Geophysics, v. 63, p. 35-57.

Dosseto, A., Bourdon, B., Gaillardet, J., Allègre, C.J., and Filizola, N., 2006, Time scale and conditions of weathering under tropical climate: Study of the Amazon basin with U-series: Geochimica et Cosmochimica Acta, v. 70, p. 71-89.

Drever, J.I., and Zobrist, J., 1992, Chemical weathering of silicate rocks as a function of elevation in the southern Swiss Alps

Dryden, A.L., 1931. Accuracy in percentage representation of heavy mineral frequencies. Proc. Natl. Acad. Sci. U. S. A. 17, 233– 238.

Ferrier, K.L., Kirchner, J.W., Riebe, C.S., and Finkel, R.C., 2010, Mineral-specific chemical weathering rates over millennial timescales: Measurements at Rio Icacos, Puerto Rico: Chemical Geology, v. 277, p. 101-114.

Folk, R.,L., 1966, A review of grain-size parameters: Sedimentology, v. 6, p. 73-93.

Franks, S.G., and Forester, R.W., 1984, Relationships among secondary porosity, pre-fluid chemistry and carbon dioxide, Texas Gulf Coast, *in* McDonald, D.A., and Surdam, R.C., eds., Clastic Diagenesis: American Association of Petroleum Geologists Memoir 37, p. 63-79.

Franzinelli, E., and Potter, P.E., 1983, Petrology, chemistry and texture of modern river sands, Amazon River system: Journal of Geology, v. 91, p. 23-39.

Gabet, E.J., 2007, A theoretical model coupling chemical weathering and physical erosion in landslide-dominated landscapes: Earth and Planetary Science Letters, v. 264, p. 259-265.

Gabet, E. J., and Mudd, S. M., 2009, A theoretical model coupling chemical weathering rates with denudation rates: Geology, v. 37, p. 151-154.

Gaillardet, J., Dupré, B., Louvat, P., and Allègre, C.J., 1999, Global silicate weathering and CO_2 consumption rates deduced from the chemistry of large rivers: Chemical Geology, v. 159, p. 3-30.

Galloway, W.E., 1974, Deposition and diagenetic alteration of sandstone in northeast Pacific arc-related basins: Implications for graywacke genesis: Geological Society of America Bulletin, v. 85, p. 379-390.

Garzanti E., Andò, S., and Vezzoli, Giovanni, 2009, Grain-size dependence of sediment composition and environmental bias in provenance studies: Earth and Planetary Science Letters, v. 277, p. 422-432.

Garzanti, E., 2016, From static to dynamic provenance analysis – Sedimentary Petrology upgraded: Sedimentary Geology, v. 336, p. 3-13.

Garzanti, E., 2019, Petrographic classification of sand and sandstone: Earth-Science Reviews, v. 192, p. 545-563.

Garzanti, E., and Resentini, A., 2016, Provenance control on chemical indices of weathering (Taiwan river sands): Sedimentary Geology, v. 336, p. 81-95.

Garzanti, E., Vezzoli, G., Andò, S., Paparella, P., and Clift, P., 2005, Petrology of Indus River sands: a key to interpret erosion history of the Western Himalayan Syntaxis: Earth and Planetary Science Letters, v. 229, p. 287-302.

Garzanti, E., Doglioni, C., Vezzoli, G., and Andò, S., 2007, Orogenic belts and orogenic sediment provenance: Journal of Geology, v. 115, p. 315-334.

Garzanti, E., Resentini, A., Vezzoli, G., Andò, S., Malusà, M.G., Padoan, M., and Paparella, P., 2010, Detrital fingerprints of fossil continental-subduction zones (axial belt provenance, European Alps): Journal of Geology, v. 118, p. 341-362.

Garzanti, E., Limonta, M., Resentini, A., Bandopadhyay, P.C., Nahman, Y., Andò, S., and Vezzoili, G., 2013a, Sediment recycling at convergent plate margins (Indo-Burman Ranges and Andaman-Nicobar Ridge): Earth-Science Reviews, v. 123, p. 113-132.

Garzanti, E., Vermeesch, P., Andò, S., Vezzoli, G., Valagussa, M., Allen, K., Kadi, K.A., and Al-Juboury, A.I.A, 2013b, Provenance and recycling of Arabian desert sand: Earth-Science Reviews, v. 120, p. 1-19.

Garzanti, E., Resentini, A., Ando, S., Vezzoli, G., Pereira, A., and Vermeesch, P., 2015a, Physical controls on sand composition and relative durability of detrital minerals during ultra-long distance littoral and aeolian transport (Namibia and southern Angola): Sedimentology, v. 62, p. 971-996.

Garzanti, E., Andò, S., Padoan, M., Vezzoli, G., El Kammar, A., 2015b, The modern Nile sediment system: Processes and products: Quaternary Science Reviews, v. 130, p. 9-56.

Garzanti, E., Al-Jubouri, A.I., Zoleikhaei, Y., Vermeesch, P., Jotheri, J., Akkoca, D.B., Obaid, A.K., Allen, M.B.,, Andó, S., Limonta, M., Padoan, M., Resentini, A., Rittner, M., and Vezzoli, G., 2016, The Euphrates-Tigris-Karun river system: Provenance, recycling and dispersal of quartz-poor foreland-basin sediments in arid climate: Earth-Science Reviews, v. 162, p. 107-128.

Gazzi, P., 1966, Le arenaria del flysch sopracretaceo dell' Appennino Modenese: Correlazioni con il flysch di Monghidoro: Mineralogica et Petrographica Acta, v. 12, p. 69-97.

Gillies, S., and others, 2019, Rasterio: Geospatial raster I/O for Python Programmers, https://github.com/mapbox/rasterio.

Girty, G.H., 1987, Sandstone provenance, Point Loma Formation, San Diego, California: Evidence for uplift of the Peninsular Ranges during the Laramide Orogeny: Journal of Sedimentary Petrology, v. 57, p. 839-844.

Graham, S.A., Ingersoll, R.V., and Dickinson, W.R., 1976, Common provenance for lithic grains in Carboniferous sandstones from Ouachita Mountains and Black Warrior Basin: Journal of Sedimentary Petrology, v. 46, p. 620-632.

Graham, R.C., Schoeneberger, P.J., Anderson, M.A., Sternberg, P.D., and Tice, K.R., 1997, Morphology, porosity, and hydraulic conductivity of weathered granitic bedrock and overlying soils: Soil Science Society of America Journal, v. 61, p. 516-622.

Griffiths, J.C., 1958, Petrography and porosity of the Cow Run Sand, St. Marys, West Virginia: Journal of Sedimentary Petrology, v. 28, p. 15-30.

Gue S.-S., and Tan, Y.-C., 2006, Landslides: Abuses of the prescriptive method: International Conference on Slope 2006, 13 p.

Hartmann, J., and Moosdorf, N., 2012, The new global lithological map database GLiM: A representation of rock properties at the Earth surface: Geochemistry, Geophysics, Geosystems, v. 13, Q12004.

Hayes, J.R., 1962, Quartz and feldspar content in South Platte, Platte, and Missouri River sands: Journal of Sedimentary Petrology, v. 32, p. 793-800.

Heins, W.A., 1992, The effect of climate and topography on the composition of modern, plutoniclastic sand [Ph.D. thesis]: Los Angeles, University of California, 495 p.

Heins, W.A., and S. Kairo, 2007, Predicting sand character with integrated genetic analysis: Geological Society of America Special Paper, v. 420, p. 345–379.

Helmold, K.P., and van de Kamp, P.C., 1984, Diagenetic Mineralogy and Controls on Albitization and Laumontite Formation in Paleogene Arkoses, Santa Ynez Mountains, California, *in* McDonald, D.A., and Surdam, R.C., eds., Clastic diagenesis: American Association of Petroleum Geologists Memoir 37, p. 239-276.

Hessler, A.M., Zhang, J., Covault, J., and Ambrose, W., 2017, Continental weathering coupled to Paleogene climate changes in North America: Geology, v. 45, p. 911-914.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., and Jarvis, A., 2005, Very high resolution interpolated climate surfaces for global land areas: International Journal of Climatology, v. 25, p. 1965-1978.

Hilley G.E., and Porder, S., 2010, A framework for predicting global silicate weathering and CO_2 drawdown rates over geologic time-scales: Proc. Natl. Acad. Sci. U.S.A. v. 105, p. 16855–16859.

Hilley, G.E., Chamberlain, C.P., Moon, S., Porder, S., and Willett, S.D., 2010, Competition between erosion and reaction kinetics in controlling silicate-weathering rates: Earth and Planetary Science Letters, v. 293, p. 191-199.

Hron, K., Templ, M., and Filzmoser, P., 2010, Imputation of missing values for compositional data using classical and robust methods: Computational Statistics and Data Analysis, v. 54, p. 3095-3107.

Hunter, R.E., 1967, The petrography of some Illinois Plesitocene and Recent sands: Sedimentary Geology, v. 1, p. 57-75.

Ingersoll, R.V., 1988, Tectonics of sedimentary basins: Geological Society of America Bulletin, v. 100, p. 1704-1719.

Ingersoll, R.V., 1990, Actualistic sandstone petrofacies: Discriminating modern and ancient source rocks: Geology, v. 18, p. 733-736.

Ingersoll, R.V., Bullard, T.F., Ford, R.L., Grimm, J.P., Pickle, J.D., and Sares, S.W., 1984, The effect and grain size on detrital modes: A test of the Gazzi-Dickinson point-counting method: Journal of Sedimentary Petrology, v. 54, p. 103-116.

Ingersoll, R.V., Kretchmer, A.G., and Valles, P.K., 1993, The effect of sampling scale on actualistic sandstone petrofacies: Sedimentology, v. 40, p. 937-953.

Israeli, Y., and Emmanuel, S., 2018, Impact of grain size and rock composition on simulated rock weathering: Earth Surface Dynamics, v. 6, p. 319-327.

Jackson, T.A., and Keller, W.D., 1970, A comparative study of the role of lichens and "inorganic" processes in the chemical weathering of recent Hawaiian lava flows: American Journal of Science, v. 269, p. 446-466.

Jett, G.A., and Heller, P.L., 1988, Tectonic significance of polymodal compositions in mélange sandstones, Western Mélange Belt, North Cascade Range, Washington: Journal of Sedimetnary Petrology, v. 58, p. 52-61.

Johnsson, M.J., 1993, The system controlling the composition of clastic sediments, *in* Johnsson, M. J., and Basu, A., eds., Processes Controlling the Composition of Clastic Sediments: Geological Society America Special Paper 284, p. 1-19.

Johnsson, M.J., Stallard, R.F., and Lundberg, N., 1991, Controls on the composition of fluvial sands from a tropical weathering environment: Sand of the Orinoco drainage basin, Venezuela and Colombia: Geological Society of America Bulletin, v. 103, p. 1622-1647.

Johnsson, M.J., Stallard, R.F., and Meade, R.H., 1988, First-cycle quartz arenites in the Orinoco River basin, Venezuela and Colombia: Journal of Geology, v. 96, p. 263-277.

Jordahl, K., and others, 2020, GeoPandas v0.8.1, https://doi.org/10.5281/zenodo.3946761.

Knoll, M.A., and James, W.C., 1987, Effect of the advent and diversification of vascular land plants on mineral weathering through geologic time: Geology, v. 15, p. 1099-1102.

Krevor, S.C., and Lackner, K.S., 2009, Enhancing process kinetics for mineral carbon sequestration: Energy Procedia, v. 1, p. 4867-4871.

Krynine, P. D., 1942, Differential sedimentation and its products during one complete geosynclinal cycle, in Primer Congresso Panamericano de Ingenieria de Minas y Geologia: Santiago, Chile, p. 537-561.

Krynine, P.D., 1950, Petrology, stratigraphy and origin of the Triassic rocks of Connecticut: Connecticut Geological Survey Bulletin 73, 239 p.

Lackner, K.S., Wendt, C.H., Butt, D.P., Joyce, E.L., Jr., and Sharp, D.H., 1995, Carbon dioxide disposal in carbonate minerals: Energy, v. 20, p. 1153-1170.

Land, L.S., and Milliken, K.L., 1981, Feldspar diagenesis in the Frio Formation, Brazoria County, Texas Gulf Coast: Geology, v. 9, p. 314-318.

Lander, R.H., and Walderhaug, O., 1999, Predicting porosity through simulating sandstone compaction and quartz cementation: American Association of Petroleum Geologists, v. 83, p. 433-449.

Lasaga, A.C., Soler, J.M., Ganor, J., Burch, T.E., and Nagy, K.L, 1994, Chemical weathering rate laws and global geochemical cycles: Geochimica et Cosmochimica Acta, v. 58, p. 2361-2386.

Lehner, B., Verdin, K., and Jarvis, A., 2008, New global hydrography derived from spaceborne elevation data: Eos, v. 89, p. 93-94.

Linke, S., Lehner, B., Dallaire, C.O., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F., and Thieme, M., 2019, Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution: Scientific Data, v. 6, 283, https://doi.org/10.1038/s41597-019-0300-6.

Longstaffe, F.J., 1984, The role of meteoric water in diagenesis of shallow sandstones: Stable isotope studies of the Milk river aquifer and gas pool, southeastern Alberta, *in* McDonald, D.A., and Surdam, R.C., eds., Clastic diagenesis: American Association of Petroleum Geologists Memoir 37, p. 91-98.

Mackie, W., 1896, Sand and sandstones of the eastern Moray: Edinburgh Geological Society Transactions, v. 7, p. 148-172.

Maher, K., 2010, The dependence of chemical weathering rates on fluid residence time: Earth and Planetary Science Letters, v. 294, p. 101-110.

Markert, J.C., and Al-Shaieb, Z., 1984, Diagenesis and evolution of secondary porosity in Upper Minnelusa Sandstones, Powder River Basin, Wyoming, *in* McDonald, D.A., and Surdam, R.C., eds., Clastic diagenesis: American Association of Petroleum Geologists Memoir 37, p. 367-389.

Marsaglia, K.M., 1992, Petrography and provenance of volcaniclastic sands recovered from the Izu-Bonin Arc, Leg 126: Proceedings of the Ocean Drilling Program, Scientific Results, v. 126, p. 139–154.

Marsaglia, K.M., and Ingersoll, R.V., 1992, Compositional trend in arc-related, deep-marine sand and sandstone: A reassessment of magmatic-arc provenance: Geological Society of America Bulletin, v. 104, p. 1637-1649.

Marsaglia, K.M., Torrez, X.V., Padilla, I., and Rimkus, K.C., 1995a, Provenance of Pleistocene and Pliocene sand and sandstone, ODP Leg 141, Chile margin: Proceedings of the Ocean Drilling Program, Scientific Results, v. 141, p. 133-151.

Marsaglia, K.M., Rimkus, K.C., and Behl, R.J., 1995b, Provenance of sand deposited in the Santa Barbara Basin at Site 893 during the last 155,000 years: Proceedings of the Ocean Drilling Program, Scientific Results, v. 146, p. 61-75.

Martens, J.H.C., 1931, Persistence of feldspars in beach sand: American Mineralogist, v. 16, p. 526-531.

Martin-Fernández, J.A., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 2003, Dealing with zeros and missing values in compositional data sets using nonparametric imputation: Mathematical Geology, v. 35, p. 253-278.

Masch, F.D., and Denny, K.J., 1966, Grain size distribution and its effect on the permeability of unconsolidated sands: Water Resources Research, v. 2, p. 665-677.

Maynard, J.B., Valloni, R., and Yu, H.S., 1982, Composition of modern deep-sea sands from arc-related basins, in Leggett, J.K., ed., Trench-forearc geology: Geological Society of London Special Publication 10, p. 551-560.

McBride, E.F., Abel-Wahab, A., and McGilvery, T.A., 1996, Loss of sand-size feldspar and rock fragments along the South Texas Barrier Island, USA: Sedimentary Geology, v. 107, p. 37-44.

Merten, G.H., and Minella, J.P.G., 2013, The expansion of Brazilian agriculture: Soil erosion scenarios: International Soil and Water Conservation Research, v. 1, p. 37-48.

Mestanza-Ramón, C.M., Pranzini, E., Anfuso, G., Botero, C.M., Chica-Ruiz, J.A., and Mooser, A., 2020, An attempt to characterize the "3S" (Sea, Sun, and Sand) parameters: Application to the Galapagos Islands and Continental Ecuadorian beaches: Sustainability, v. 12, p. 1-19, doi:10.3390/su12083468.

Molinaroli, E., Blom, M., and Basu, A., 1991, Methods of provenance determination tested with discriminant function analysis: Journal of Sedimentary Petrology, v. 61, p. 900-908.

Molnar, C., 2020, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Accessed: October 2020. [Online]. Available: https://christophm.github.io/interpretable-ml-book/.

Moncure, G.K., Lahann, R.W., and Siebert, R.M., 1984, Origin of secondary porosity and cement distribution in a sandstone/shale sequence from the Frio formation (Oligocene), *in* McDonald, D.A., and Surdam, R.C., eds., Clastic diagenesis: American Association of Petroleum Geologists Memoir 37, p. 151-161.

Montgomery, D.R., 2007, Soil erosion and agricultural sustainability: Proc. Natl. Acad. Sci. U.S.A. v. 104, p. 13268-13272.

Morad S., Bergan, M., Knarud, R., and Nystuen, P., 1990, Albitization of detrital plagioclase in Triassic reservoir sandstones from the Snorre field, Norwegian North Sea: Journal of Sedimentary Petrology, v. 60, p. 411-425.

Morton, A.C., 1986, Dissolution of apatite in North Sea Jurassic sandstones: Implication for the generation of secondary porosity: Clay Minerals, v. 21, p. 711-733.

Nedkvitne, T., and Bjørlykke, K., 1992, Secondary porosity in the Brent Group (Middle Jurassic), Huldra field, North Sea: Implication for predicting lateral continuity of sandstones: Journal of Sedimentary Petrology, v. 62, p. 23-24.

Nesbitt, H.W., Young, G.M., McLennan, S.M., and Keays, R.R., 1996, Effects of chemical weathering and sorting on the petrogenesis of siliciclastic sediments, with implications for provenance studies: Journal of Geology, v. 104, p. 525-542.

Nickel, E., 1973, Experimental dissolution of light and heavy minerals in comparison with weathering and intrastratal solution: Contributions to Sedimentology, v. 1, p. 1-68.

Nilsen, T.H., Taylor, F.A., and Brabb, E.E., 1976, Recent landslides in Alameda Country, California (1940-71): An estimate of economic losses and correlations with slope, rainfall, and ancient landslide deposits: US Geological Survey Bulletin 1398.

Oliva, P., Viers, J., and Dupré, B., 2003, Chemical weathering in granitic environments: Chemical Geology, v. 202, p. 225-256.

Osborne, R.H., Bomer, E.J., III, Wang, Y.-C., and Lu, Y., 1993, Application of a tumbler experiment using granodioritic grus to examine the character of quartz-grain fracture in high-gradient streams, *in* Johnsson, M. J., and Basu, A., eds., Processes Controlling the Composition of Clastic Sediments: Geological Society America Special Paper 284, p. 211-234.

Pedregosa, F., and others, 2011, Scikit-learn: Machine Learning in Python: Journal of Machine Learning Research, v. 12, p. 2825-2830.

Pettijohn, F.J., Potter, P.E., Siever, R., 1972, Sand and Sandstone: Berlin, Heidelberg, New York, Springer-Verlag, 618 p.

Pettijohn, F.J., 1954, Classification of sandstones: Journal of Geology, v. 62, p. 360-365.

Pimentel, D., Allen, J., Beers, A., Guinand, L., Linder, R., McLaughlin, P., Meer, B., Musonda, D., Perdue, D., Poisson, S., Siebert, S., Stoner, K., Salazar, R., and Hawkins, A., 1987, World agriculture and soil erosion: BioScience, v. 37, p. 277-283.

Pittman, E.D., 1969, Destruction of plagioclase twins by stream transport: Journal of Sedimentary Petrology, v. 39, p. 1432-1437.

Potter, P. E., 1978, Petrology and chemistry of modern big river sands: Journal of Geology, v. 86, p. 423-449.

Potter, P.E., 1986, South America and a few grains of sand: Part 1: Beach sands: Journal of Geology, v. 94, p. 301-319.

Potter, P. E., 1994, Modern sands of South America: composition, provenance and global significance: Geol. Rundschau, v. 83, p. 212-232.

Potter, P.E., Huh, Y., and Edmond, J.M., 2001, Deep-freeze petrology of Lena River sand, Siberia: Geology, v. 29, p. 999-1002.

Plumley, W.J., 1948, Black Hills terrace gravels: A study in sediment transport: Journal of Geology, v. 56, p. 526-577.

Qasim, Q., Harahap, I.S.H., and Osman, S.B.S., 2013, Causal factors of Malaysian landslides: A narrative study, Research Journal of Applied Sciences, Engineering and Technology, v. 5, p. 2303-2308.

Riebe, C.S., Kirchner, J.W., and Finkel, R.C., 2004, Erosional and climatic effects on long-term chemical weathering rates in granitic landscapes spanning diverse climate regimes: Earth and Planetary Science Letters, v. 224, p. 547-562.

Rittner, M., Vermeesch, P., Carter, A., Bird, A., Stevens, T., Garzanti, E., Andò, S., Vezzoli, G., Dutt, R., Xu, Z., and Lu, H., 2016, The provenance of Taklamakan desert sand: Earth and Planetary Science Letters, v. 437, p. 127–137.

Robinson, N., Regetz, J., and Guralnick, R.P., 2014, EarthEnv-DEM90: A nearly-global, void-free, multi-scale smoothed, 90m digital elevation model from fused ASTER and SRTM data: ISPRS Journal of Photogrammetry and Remote Sensing, v. 87, p. 57-67, doi: 10.1016/j.isprsjprs.2013.11.002.

Roser, B.P., and Korsch, R.J., 1986, Determination of tectonic setting of sandstone-mudstone suites using SiO_2 content and K_2O/Na_2O ratio: Journal of Geology, v. 94, p. 635-650.

Russell, R.G., 1989, Correlation of permeability and grain size: Ground Water, v. 27, p. 633-638.

Sclater, J.G., and Christie, P.A.F., 1980, Continental stretching: An explanation of the post-mid-Cretaceous subsidence of the Central North Sea basin: Journal of Geophysical Research, v. 85, p. 3711-3739.

Sharman, G.R., and Malkowski, M.A., 2020, Needles in a haystack: Detrital zircon U–Pb ages and the maximum depositional age of modern global sediment: Earth-Science Reviews, v. 203, 103109, doi:10.1016/j.earscirev.2020.103109.

Shukis, P. S., and Ethridge, F. G., 1975, A petrographic reconnaissance of sand size sediment upper St. Francis River, southeastern Missouri: Journal of Sedimentary Petrology, v. 45, p. 115-127.

Siebert, R.M., Moncure, G.K., and Lahann, R.W., 1984, A theory of framework grain dissolution in sandstones, *in* McDonald, D.A., and Surdam, R.C., eds., Clastic diagenesis: American Association of Petroleum Geologists Memoir 37, p. 163-175.

Sklar, L.S., Riebe, C.S., Marshall, J.A., Genetti, J., Leclere, S., Lukens, C.L., and Merces, V., 2017, The problem of predicting the size distribution of sediment supplied by hillslopes to rivers: Geomorphology, v. 277, p. 31-49.

Sømme, T.O., Helland-Hansen, W., Martinsen, O.J., and Thurmond, J.B., 2009, Relationships between morphological and sedimentological parameters in source-to-sink systems: a basis for predicting semi-quantitative characteristics in subsurface systems: Basin Research, v. 21, p. 361–387.

Stallard, R.F., 1988, Weathering and erosion in the humid tropics, *in* Lerman, A., and Meybeck, M., eds., Physical and chemical weathering in geochemical cycles: Dordrecht, Kluwer, p. 225-246.

Stallard, R.F., and Edmond, J.M., 1983, Geochemistry of the Amazon: 2. The influence of the geology and weathering environment on the dissolved load: Journal of Geophysical Research, v. 88, p. 9671-9688.

Straume, E. O., Gaina, C., Medvedev, S., Hochmuth, K., Gohl, K., Whittaker, J. M., Abdul Fattah, R., Doornenbal, J.C., and Hopper, J.R.,2019, GlobSed: Updated total sediment thickness in the world's oceans: Geochemistry, Geophysics, Geosystems, v. 20, p. 1756–1772. https://doi.org/10.1029/2018GC008115.

Surdam, R.C., Boese, S.W., and Crossey, L.J., 1984, The chemistry of secondary porosity, *in* McDonald, D.A., and Surdam, R.C., eds., Clastic diagenesis: American Association of Petroleum Geologists Memoir 37, p. 127-161.

Suttner, L. J., Basu, A., and Mack, G., 1981, Climate and the origin of quartz arenites: Journal of Sedimentary Petrology, v. 51, p. 1235-1246.

Tan, Z.X., Lal, R., and Wiebe, K.D., 2005, Global soil nutrient depletion and yield reduction: Journal of Sustainable Agriculture, v. 26, p. 123-146.

Tarolli, P., and Sofia, G., 2016, Human topographic signatures and derived geomorphic processes across landscapes: Geomorphology, v. 255, p. 140-161.

Tolosana-Delgado, R., and von Eynatten, H., 2009, Grain-size control on petrographic composition of sediments: Compositional Regression and Rounded Zeros: Mathematical Geosciences, v. 41, p. 869-886.

Valloni, R., and Maynard, J.B., 1981, Detrital modes of recent deep-sea sands and their relation to tectonic setting: A first approximation: Sedimentology, v. 28, p. 75-83.

Van Andel, T.H., 1958, Origin and Classification of Cretaceous, Paleocene and Eocene Sandstones of Western Venezuela: American Association of Petroleum Geologists Bulletin, v. 42, p. 734-763.

Van der Plas, L., and Tobi, A.C., 1965, A chart for judging the reliability of point counting results: American Journal of Science, v. 263, p. 87-90.

Van Grinsven, M., and Marsaglia, K.M., 2019, Sand provenance form source to sink in the Santa Monica Basin, California Borderlands: Significance of Calleguas Creek input: SEPM Special Publication, v. 110, p. 102-116.

Verma, S.P., and Armstrong-Altrin, J.S., 2013, New multi-dimensional diagrams for tectonic discrimination of siliciclastic sediments and their application to Precamprian basins: Chemical Geology, v. 355, p. 117-133.

Vermeesch, P., 2018, Statistical models for point-counting data: Earth and Planetary Science Letters, v. 501, p. 1-7.

Vermeesch, P., 2019, Exploratory analysis of provenance data using R and the Provenance Package: Minerals, v. 9, 193, doi:10.3390/min9030193.

Vezzoli, G., Lombardo, B., and Rolfo, F., 2017, Petrology of the Tista and Rangit river sands (Sikkim, India): Italian Journal of Geosciences, v. 136, p. 103-109.

von Eynatten, H., Tolosana-Delgado, R., and Karius, V., 2012, Sediment generation in modern glacial settings: source-rock and grain-size control on sediment composition: Sedimentary Geology, v. 280, p. 80–92.

von Eynatten, H., Tolosana-Delgado, R., Karius, V., Bachmann, K., and Caracciolo, L., 2016, Sediment generation in humid Mediterranean setting: Grain-size and source-rock control on sediment geochemistry and mineralogy (Sila Massif, Calabria): Sedimentary Geology, v. 336, p. 68-80.

Walderhaug, O., 1994, Precipitation rates for quartz cement in sandstones determined by fluid-inclusion microthermometry and temperature-history modeling: Journal of Sedimentary Research, v. 64, p. 324-333.

Walker, T.R., 1984, Diagenetic albitization of potassium feldspar in arkosic sandstones: Journal of Sedimentary Petrology, v. 54, p. 3-16.

Weltje, G.J., 2002, Quantitative analysis of detrital modes: statistically rigorous confidence regions in ternary diagrams and their use in sedimentary petrology: Earth-Science Reviews, v. 57, p. 211-253.

Weltje, G.J., 2006, Ternary sandstone composition and provenance: an evaluation of the 'Dickinson model', *in* Buccianti, A., Mateu-Figueras, G., Pawlowsky-Gahn, V. eds. , Compositional Data Analysis in the Geosciences: From Theory to Practice: Geological Society of London, Special Publications, v. 264, p. 79–99.

Weltje, G.J., 2012, Quantitative models of sediment generation and provenance: State of the art and future developments: Sedimentary Geology, v. 280, p. 4-20.

Weltje, G.J., and Prins, M.A., 2003, Muddled or mixed? Inferring palaeoclimate from size distribution of deep-sea clastics: Sedimentary Geology, v. 162, p. 39-62.

Weltje, G.J., van Ansenwoude, S.O.K.J., and de Boer, P.L., 1996, High-frequency detrital signals in Eocene fan-delta sandstones of mixed parentage (south-central Pyrenees,

Spain): A reconstruction of chemical weathering in transit: Journal of Sedimentary Research, v. 66, p. 119-131.

Weltje, G.J., Meijer, X.D., and de Boer, P.L., 1998, Stratigraphic inversion of siliciclastic basin fills: a note on the distinction between supply signals resulting from tectonic and climatic forcing: Basin Research, v. 10, p. 129-153.

Wentworth, C.K., 1922, A scale of grade and class terms for clastic sediments: Journal of Geology, v. 30, p. 377-392.

West, A.J., Galy, A., Bickle, M., 2005, Tectonic and climatic controls on silicate weathering: Earth and Planetary Science Letters, v. 235, p. 211-228.

White, A.F., 2002, Determining mineral weathering rates based on solid and solute weathering gradients and velocities: application to biotite weathering in saprolites: Chemical Geology, v. 190, p. 69-89.

White, A.F., and Blum, A.E., 1995, Effects of climate on chemical weathering in watersheds: Geochimica et Cosmochimica Acta, v. 59, p. 1729-1747.

White, A.F., and Brantley, S.L., 2003, The effect of time on the weathering of silicate minerals: why do weathering rates differ in the laboratory and field?: Chemical Geology, v. 202, p. 479-506.

White, A.F., and Buss, H.L., 2003, Natural Weathering Rates of Silicate Minerals, *in* Holland, H. D., eds, Treatise on Geochemistry: Elsevier Ltd, v. 7, p. 115-155.

White, A.F., Blum, A.E., Bullen, T.D., Vivit, D.V., Schulz, M., and Fitzpatrick, J., 1999, The effect of temperature on experimental and natural chemical weathering rates of granitoid rocks: Geochimica et Cosmochimica Acta, v. 63, p. 3277-3291.

Yang, L., Xu, T., Keyu, L., Peng, B., Yu, Z., and Xu, X., 2017, Fluid-rock interactions during continuous diagenesis of sandstone reservoirs and their effects on reservoir porosity: Sedimentology, v. 64, p. 1303-1321.

Yerino, L.N., and Maynard, J.B., 1984, Petrography of modern marine sands from the Peru-Chile Trench and adjacent areas: Sedimentology, v. 31, p. 83-89.

Yoo, K., and Mudd, S.M., 2008, Discrepancy between mineral residence time and soil age: Implications for the interpretation of chemical weathering rates: Geology, v. 36, p. 35-38.

Yoo, K., Mudd, S.M., Sanderman, J., Amundson, R., and Blum, A., 2009, Spatial patterns and controls of soil chemical weathering rates along a transient hillslope: Earth and Planetary Science Letters, v. 288, p. 184-193.

Zeiglar, D.L., and Spotts, J.H., 1978, Reservoir and source-bed history of Great Valley, California: American Association of Petroleum Geologists Bulletin, v. 62, p. 813-826.

Zuffa, G. G., 1985, Optical analyses of arenites: Influence of methodology on compositional results, *in* Zuffa, G. G., ed., Provenance of arenites: Dordrecht, D. Reidel, p. 165-190.

Zuffa, G.G., 1980, Hybrid arenites: their composition and classification: Journal of Sedimentary Petrology, v. 50, p. 21-29.