Long-term monitoring reveals convergent patterns of recovery from mining contamination across 4 western US watersheds

William H. Clements[1,5], David B. Herbst[2,6], Michelle I. Hornberger[3,7], Christopher A. Mebane[4,8], and Terry M. Short[3,9]

[1]Department of Fish, Wildlife and Conservation Biology, Colorado State University, Fort Collins, Colorado 80523 USA
[2]Sierra Nevada Aquatic Research Laboratory, 1016 Mount Morrison Road, Mammoth Lakes, California 93546 USA; and Institute of Marine Sciences, University of California, Santa Cruz, California 95064 USA
[3]United States Geological Survey, Water Mission Area, Earth Systems Processes Division, 345 Middlefield Road, Menlo Park, California 94025 USA
[4]United States Geological Survey, Idaho Water Science Center, 230 Collins Road, Boise, Idaho 83702 USA

Abstract: Long-term studies of stream ecosystems are essential for assessing restoration success because they allow researchers to quantify recovery trajectories, gauge the relative influence of episodic events, and determine the time required to achieve clean-up objectives. To quantify responses of benthic macroinvertebrate assemblages to stream remediation, we integrated results of 4 long-term (20–29 y) assessments of mining-impacted watersheds that were broadly distributed across the western US (California, Colorado, Idaho, Montana). Using a before–after control–impact (BACI) study design, we observed substantial reductions in metal concentrations and corresponding improvements of benthic assemblages following remediation. Recovery rates were relatively consistent, and streams typically recovered within 10 to 15 y after remediation was initiated (mean = 10.25 y), although episodic events changed trajectories at some sites. Differences in recovery among watersheds were likely determined by a number of factors, including the severity of contamination, effectiveness of remediation, proximity to upstream sources of colonization, and hydrologic variation. We also observed considerable variation in the rate and extent of recovery among assemblage metrics. For example, total abundance and richness recovered rapidly at most sites, but the composition of benthic macroinvertebrate assemblages remained substantially altered compared with reference sites. Using piecewise linear regression, we estimated a threshold response of Ephemeroptera, Plecoptera, and Trichoptera (EPT) species richness at ~1 cumulative criteria unit (CCU), which is the sum of the fractions of chronic water-quality criteria for metals measured, suggesting this value was protective of benthic assemblages. However, EPT richness was reduced by ~20% at 2× this CCU value, indicating that moderate exceedances of water-quality criteria could substantially affect stream biodiversity. Non-metric multidimensional scaling analyses identified common sets of species trait states across the 4 watersheds that were associated with either metal contamination or with recovering and intact reference stream assemblages. Our study illustrates the importance of long-term studies for quantifying responses to stream restoration and the usefulness of BACI designs for demonstrating cause-and-effect relationships between restoration treatments and community recovery. Because these 4 watersheds were among the most severely polluted sites in the western US, our study demonstrates the value of these investments in watershed restoration and the potential for success under the most extreme conditions.

Key words: benthic macroinvertebrates, biomonitoring, cumulative criterion units, long-term, recovery, mining contamination, species traits

Predicting the time required for damaged ecosystems to recover following remediation and restoration activities is a central challenge in applied ecology. Depending on the type and magnitude of the disturbance and the specific ecosystem affected, recovery times can range from decades (Jones and Schmitz 2009) to centuries or even millennia (Dobson et al. 1997, Foley et al. 2005). Release of metals from abandoned mines is a disturbance that can persist for many decades

E-mail addresses: [5]william.clements@colostate.edu; [6]herbst@lifesci.ucsb.edu; [7]mhornber@usgs.gov; [8]cmebane@usgs.gov; [9]tmshort@usgs.gov

DOI: 10.1086/714575. Received 22 October 2020; Accepted 9 March 2021; Published online 4 May 2021; Associate Editor, Christoph Matthaei.
Freshwater Science. 2021. 40(2):407–426.

after mining has ceased, the most extreme example being historic Roman mine workings that have continued discharging metals for up to 2000 y (INAP 2009, Lefcort et al. 2010). Although metal release from abandoned mines is generally considered an issue of local significance, mining pollution can have regional consequences (Hudson-Edwards 2016). For example, ~6% of English and Welsh streams are contaminated by discharge from abandoned mines (Jones et al. 2013), and 23% of streams in Colorado's central Rocky Mountains in the US are degraded by metals from historical mining disturbance (Clements et al. 2000).

Our ability to quantify the effectiveness of restoration programs in aquatic ecosystems is often limited because of poor study designs and the failure to account for ecological theory (Bernhardt et al. 2005, Palmer et al. 2014). This issue is especially problematic given the high cost of restoration, which in the continental US exceeds $1 billion annually (Bernhardt et al. 2005). Beyond these initial costs, abandoned mines often require perpetual on-site treatment, which is estimated to cost over $60 million/y in the US (Gestring and Sumi 2013). Continued public support for these restoration projects likely depends on demonstrating their success. A review of sediment remediation projects at the United States Environmental Protection Agency (USEPA) Superfund megasites (defined as locations where expenditures >$50 million) reported that it was not possible to evaluate their success, primarily because of inadequate post-restoration monitoring (NRC 2007). Finally, a lack of broadly accepted criteria that define restoration success has hindered progress (Palmer at al. 2014). Some studies have quantified ecological responses to restoration treatments; however, even the removal of a stressor or simply demonstrating habitat improvement is often considered evidence of success. Furthermore, apparent recovery of certain variables, such as species abundance or species richness, may occur despite persistent alterations in community composition or loss of functional redundancy (Berumen and Pratchett 2006). These results demonstrate the importance of investigating multiple indicators and expanding the definition of ecosystem recovery beyond traditional community metrics.

The limited number of long-term (e.g., >10 y) studies conducted to quantify responses to remediation or restoration treatments has also impeded our ability to assess recovery, despite broad agreement for the importance of long-term research. Because recovery trajectories following the removal of a stressor are not necessarily linear (Scheffer and Carpenter 2003, Folke et al. 2004, Clements et al. 2010, Khan et al. 2012), a long-term perspective is critical for quantifying restoration effectiveness. In a comprehensive synthesis of factors that determine ecosystem recovery, Jones and Schmitz (2009) concluded that studies failing to show recovery were often limited by insufficient study duration. This issue is especially problematic for benthic macroinvertebrate studies, where the median duration of long-term research is ~9 y (Jackson and Füreder 2006). Many of the factors that are likely to influence responses of aquatic ecosystems to restoration (e.g., land-use changes, regional climate, hydrologic alterations) operate at much longer time scales. Long-term assessments of aquatic ecosystems are especially important for gauging responses to restoration and the concomitant effects of climate change. Interactions between climate change and water quality are well documented in the literature (Clements et al. 2008, Noyes et al. 2009, Moe et al. 2013) and may involve complex feedback mechanisms. An understanding of these potential interactions is of critical importance because climate change may offset the benefits of improved water quality or habitat after remediation (Barbour et al. 2010, Floury et al. 2013, Van Looy et al. 2016).

Viewed as either a natural or management experiment (Diamond 1983, Clements et al. 2010), assessing long-term responses to restoration treatments provides a unique opportunity to test ecological theory. The failure of some systems to recover following elimination of a stressor has been attributed to the loss of ecological resilience, entrenchment of disturbance-tolerant taxa, and shifts to alternative stable states (Scheffer and Carpenter 2003, Folke et al. 2004, Wolff et al. 2019). These long-term changes in community composition, which are triggered by both natural and anthropogenic disturbances, may be very difficult to reverse, even after the initial stressor has been removed. An emerging paradigm in restoration ecology predicts that some regime shifts are permanent and that the resulting novel communities may never return to pre-disturbance conditions (Hobbs et al. 2006). The mechanisms responsible for regime shifts in aquatic ecosystems are diverse, but the concept of novel communities has important implications for how we define and quantify restoration success (Hobbs et al. 2009).

The longitudinal connectivity of stream ecosystems and the proximity of tributaries for recolonization are important factors that influence the rate of recovery. The historical depiction of streams as isolated linear reaches has been replaced by a more modern representation of watersheds as complex, dendritic networks structured by local and regional processes (Benda et al. 2004). Because tributaries are critical sources of recolonization that contribute sensitive taxa to downstream reaches (Pond et al. 2014, Mebane et al. 2015), the branching pattern and architecture of a stream will likely influence the rate of recovery following the removal of a stressor. For example, recovery of isolated headwater streams is predicted to be slower because of their limited potential for recolonization (Smith et al. 2011).

Recovery of macroinvertebrate assemblages will also be influenced by the specific traits (e.g., relative sensitivity to metals, life-history characteristics, dispersal ability) of organisms that survive impaired conditions. For example, organisms that are especially sensitive to contaminants (e.g., some Ephemeroptera), or are long-lived and relatively poor dispersers (e.g., many Plecoptera), will likely require more

time to recover (Clements et al. 2010, Smith et al. 2011, Mebane et al. 2015, Herbst et al. 2018). Quantifying responses of these broad taxonomic groups across geographic regions is possible, but assessing effects at lower levels of taxonomic resolution is challenging because of biogeographic variation in community composition. The use of species traits that are directly linked to life history and other characteristics known to influence recovery is an effective alternative to traditional community metrics for assessing responses to restoration across geographic regions (Verberk et al. 2013).

The primary objective of this study was to integrate results of 4 long-term assessments of mining-impacted streams to quantify responses to remediation treatments. Because most sites received both remediation (the process of reducing pollution) and restoration (the process of improving of habitat), for simplicity, we will use these terms interchangeably. Different treatments were used in each watershed, but all were designed to control and reduce metal loadings. The watersheds were broadly distributed across the western US (California, Colorado, Idaho, Montana) and included 4 USEPA Superfund sites considered to be among the most contaminated streams in the region. Study designs in the 4 watersheds were similar and consisted of a treatment period when restoration was being implemented followed by a post-restoration assessment of recovery, allowing us to use a before–after control–impact (BACI) approach. Because all streams were affected by a similar set of stressors (primarily metals), we were especially interested in assessing biotic and abiotic factors that determined responses to restoration and the timing of recovery across these geographic regions. Because of large differences in species composition of macroinvertebrate assemblages among regions, we focused primarily on relatively coarse assemblage metrics (e.g., total number of taxa, number of Ephemeroptera) and species traits to quantify post-restoration responses. Finally, we calculated threshold responses of Ephemeroptera, Plecoptera, and Trichoptera (EPT) taxa richness to determine if aquatic life criteria for metals were protective of the ecological integrity of these watersheds.

METHODS

Detailed descriptions of the 4 watersheds, remediation/restoration treatments, and sampling methods have been published previously (Hornberger et al. 2009, Clements et al. 2010, Mebane et al. 2015, Herbst et al. 2018), so we have limited the following description to a brief overview. We sampled wadeable, cobble-bottom, 2nd- to 4th-order streams in Colorado, Idaho, Montana, and California that ranged in elevation from 1448 to 2898 m a.s.l. (Table 1, Fig. S1). Remediation included various combinations of active water treatment, construction of containment ponds, removal of metal-contaminated soils, revegetation of riparian areas, and other habitat improvements. Routine water-quality characteristics (pH, temperature, specific conductance, water hardness), metal concentrations, and macroinvertebrate assemblage structure were measured annually before and after completion of remediation at metal-impacted, downstream, and reference sites for 20 to 29 y (for physicochemical raw

Table 1. Watershed characteristics, study duration, metals of concern, and primary restoration treatments used in 4 western watersheds where long-term responses to mining restoration were measured. NA = not available, CO = Colorado, MT = Montana, CA = California, ID = Idaho. See Table 2 for metals abbreviations.

Watershed	Elevation (m)	Watershed area (km^2)	Stream order	Distance to upstream colonization source (km)	Study duration (y)	Pre-treatment sampling (y)	Metals of concern	Primary restoration and remediation treatments
Arkansas River (CO)	2898	256	4th	0.2	29	12	Cd, Cu, Zn	Active water treatment, removal of tailings, and revegetation of riparian areas
Clark Fork (MT)	1448	1699	3rd	NA	24	10	Cu, Cd, Pb, Zn	Containment ponds, removal of bank and floodplain soils, bank stabilization and revegetation
Leviathan Creek(CA)	1912	55	2nd	NA	20	9	Al, As, Co, Cu, Fe, Mn, Ni, Se, Zn	Containment ponds, liming, microbial bioreactor (sulfate-reducing bacteria)
Big Deer Creek (ID)	1560	119	3rd	0.3	24	12	Cu, Co	Water diversions, active treatment

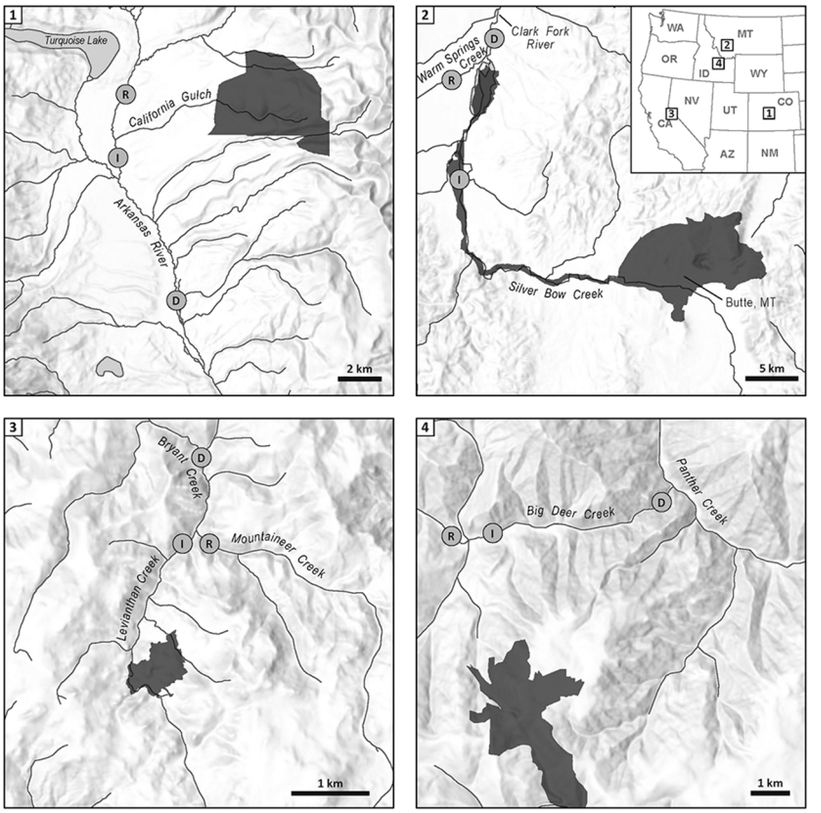

Figure 1. Map of sampling locations showing reference (R), impacted (I), and downstream (D) sites in each watershed. Note the different scales in each panel. Shaded areas indicate locations of the major mining activities. Inset shows the location of each watershed in the western United States. CO = Colorado, MT = Montana, CA = California, ID = Idaho.

data, see figshare data repository: https://doi.org/10.6084/m9.figshare.13889885). In each watershed, a single reference site was located either immediately upstream of the impacted site or in an adjacent drainage (Fig. 1). A single impacted site was located near the mine source (0–4 km below) and a single downstream site, also with elevated metal concentrations, was located 2 to 19 km further downstream in each watershed. Stream order varied among the 4 watersheds, but stream size, depth, and substrate composition of reference sites were similar to those at impacted sites within each watershed, regardless of whether they were located immediately upstream or on a nearby tributary. We used quantitative (Hess samples) or semi-quantitative (D-frame nets) sampling techniques to collect macroinvertebrates and generally identified them to genus or species (except for Arkansas River, Colorado, where we identified chironomids to tribe). Water samples were collected on each sampling occasion and we used either atomic adsorption spectrophotometry (flame or furnace), inductively coupled plasma mass spectrometry, or inductively coupled argon plasma emission spectroscopy to analyze water samples for dissolved metals. Metal concentrations were elevated at the Arkansas River reference station for a 3-y period early in the study and prior to remediation of a separate source of contamination (Clements et al. 2010). However, we retained these data in our analysis because 1) these 3 y constitute a relatively small portion of the pre-remediation record for this stream, and 2) remediation had already begun at the impacted site during this period. Excluding these data would eliminate a critical part of the recovery trajectory at the impacted site when differences with the reference site were greatest.

Long-term changes in stream hydrology and metal exposure

We calculated several variables to characterize differences in the hydrologic regime among the 4 watersheds. We limited analyses of hydrologic characteristics to impacted sites or to similar nearby streams owing to the lack of discharge records for all sites. We determined flow extremes as described in Clausen and Biggs (2000), where low-flow (Q_{90}) and high-flow (Q_{10}) metrics represent flows that were exceeded 90 and 10% of the time, respectively, relative to median discharge (Q_{50}). We calculated flow variability (F_v) as the difference between the 90th and 10th percentile flows divided by Q_{50}, with higher F_v values indicating higher flow variability (Sheldon and Thoms 2006). Because hydrologic conditions can influence both the composition of benthic

communities and discharge of metals, we compared streamflow variables before and after completion of restoration and examined the relationship between streamflow and metal concentrations in each region.

Because each watershed was contaminated by a different set of metal mixtures, we estimated exposure of benthic macroinvertebrate assemblages to metals on each sampling occasion by using cumulative criterion units (CCUs) (Clements et al. 2000). Water samples used to estimate CCUs were limited to those collected in late summer or early autumn to match the period of biological sampling in each watershed. We calculated CCUs as $CCU = \sum M_i/C_i$, where M_i is the measured concentration of each metal of interest, and C_i is a chronic aquatic life criterion value that is intended to protect freshwater communities from adverse effects. Because water quality standards differed among study areas and because regulatory criteria have not been established for all metals, we selected criteria values used in CCU calculations to provide a common measure of exposure across the 4 watersheds (Table 2).

Factors that modify the bioavailability and toxicity of metals, such as pH, dissolved organic carbon (DOC), and major ions (e.g., calcium and magnesium), were not routinely measured in all watersheds. We estimated missing values by either using data from nearby sites or, assuming that water chemistry was relatively stable year to year, using data collected during other years. Other than DOC, estimated values constituted about 11% of the data (206/1910 values) and were used most commonly for minor constituents. Data for DOC were sparse and were estimated as described at https://doi.org/10.6084/m9.figshare.13889885.

Long-term changes in macroinvertebrate assemblages

We evaluated long-term responses of benthic macroinvertebrate assemblages to restoration treatments by using assemblage metrics (e.g., abundance, richness), and we used their similarity at impacted and downstream sites to those at reference sites to quantify recovery. Regional differences in taxonomic composition among watersheds made comparisons at lower levels of taxonomic resolution challenging. To analyze differences among sites and over time, we selected species traits derived from published literature (Thorp and Covich 2001, Poff et al. 2006, Vieira et al. 2006, Merritt et al. 2008, Andersen et al. 2013) that were most likely associated with metal exposure, toxicological effects, and recolonization rates. We examined differences in 27 trait states related to feeding habits (collector–gatherers, collector–filterers, grazers, shredders, predators), respiration mode (cutaneous; spiracles; thoracic, abdominal, or anal gills), drift propensity (rare, common, abundant), behavioral habits (burrowers, climbers, sprawlers, clingers, swimmers), life cycle (multivoltine, univoltine, semivoltine), developmental rate (fast, slow, nonseasonal), and body size (small <9 mm, medium 9–16 mm, large >16 mm).

Data analyses

We expressed long-term changes in basic assemblage metrics (number of taxa, total abundance, mayfly richness and abundance) at impacted and downstream sites relative to mean values at reference sites in each watershed. We calculated functional dispersion, a functional diversity index based on species traits (Laliberté and Legendre 2010). We used the Bray–Curtis (BC) similarity index to compare similarity of impacted and downstream assemblages to the long-term mean of reference assemblages within each watershed. Using long-term means of reference assemblages was necessary because of large year-to-year variation in assemblage composition at some reference sites and because of an 11-y gap in sampling at the Clark Fork, Montana, reference site. We concluded that metrics had recovered if

Table 2. Summary of aquatic life criteria values used to calculate cumulative criterion units to estimate metals exposure. For criteria that varied as a function of toxicity-modifying factors, values were calculated for pH 7, dissolved organic C (DOC) 1 mg/L, and water hardness 50 mg/L.

Metal	Factors that modify criteria values	Criteria value (μg/L)	Intended form	References
Aluminum (Al)	pH, DOC, hardness	340	Total (unfiltered)	USEPA 2018
Arsenic (As)	None	150	Total (unfiltered)	USEPA 1984b
Cadmium (Cd)	Hardness	0.43	Dissolved (filtered)	USEPA 2016a
Cobalt (Co)	None	7.1	Dissolved (filtered)	Stubblefield et al. 2020
Copper (Cu)	pH, DOC, hardness	1.3	Dissolved (filtered)	Brix et al. 2017
Iron (Fe)	None	251	Total (unfiltered)	Cadmus et al. 2018
Manganese (Mn)	Hardness	1310	Dissolved (filtered)	Stubblefield and Hockett 2000
Nickel (Ni)	Hardness	29	Dissolved (filtered)	USEPA 1986, 1996
Selenium (Se)	None	3.1	Dissolved (filtered)	USEPA 2016b
Zinc (Zn)	Hardness	66	Dissolved (filtered)	USEPA 1980, 1996

they were within or greater than the 95% confidence intervals of mean reference-site values. For metrics that did not reach this threshold, we estimated the proportion of reference condition achieved after 2 to 3 consecutive y at maximum values.

We used 2-way general linear models (GLM) (SAS version 9.4; SAS Institute, Cary, North Carolina) to test for effects of site (reference, impacted, downstream), restoration treatment (before vs after restoration), and the site × treatment interaction on all macroinvertebrate metrics in each watershed. In this analysis, a strong site × treatment interaction indicated that differences among sites varied with treatment, a critical expectation of a BACI design. Although replicate samples were collected from each site on each sampling occasion, we based all statistical analyses on the means of individual replicates; thus, sample sizes (years sampled × sites) for the 4 watersheds were: Arkansas River, 87; Clark Fork, 59; Big Deer Creek, Idaho, 56; Leviathan Creek, California, 55. Data were log-transformed to satisfy assumptions of normality and homogeneity of variance, and we verified that data met these assumptions by inspection of residual plots. We used least-squares regression (GLM; SYSTAT, version 13.0; Systat Software, San Jose, California) to examine the influence of flow properties on metal concentrations (as CCU).

We used non-metric multidimensional scaling (NMDS) analyses (PC-ORD version 7.0; Wild Blueberry Media, Corvallis, Oregon) to characterize changes in the composition of species traits in response to restoration treatments in each watershed. For these analyses, we included only taxa occurring with >5% frequency in each region. The sum of trait states for each species trait was weighted by the relative abundance of taxa possessing that trait state (these values will sum to 1 for each trait in a sample). Using indicator species analysis and substituting trait states for species, we identified which trait states were associated with the low metals at reference sites vs elevated metals at mining-polluted sites and before vs after completion of restoration. NMDS uses an environmental matrix to identify correlations with the ordination axes, showing the strength and direction of these correlations. We used an environmental matrix consisting of sample year (time), water quality (hardness and specific conductivity), CCU, individual metal concentrations, and flow variables in NMDS to identify temporal, chemical, and hydrological associations with the trait state ordinations.

To identify a potential threshold response to metals, we used piecewise regression to examine the relationship between CCU and species richness of EPT (Toms and Lesperance 2003, Erickson 2015). Piecewise regression can be used to estimate a substantial change in the slope of a concentration–response relationship and can be used to separate a no-effect concentration from concentrations that cause progressively more severe effects (Khan et al. 2012, Mebane 2015). We chose the EPT metric to examine responses across watersheds because the level of taxonomic resolution for these groups was consistent among regions and across years. We did not include the EPT metric in the other long-term analyses of assemblage responses described above because it was highly correlated with other richness metrics (total species richness $r = 0.76$ for the overall dataset and $r = 0.90$–0.96 for the 4 datasets individually) and therefore considered redundant.

RESULTS

Long-term changes in stream hydrology and metal concentrations

Median monthly discharge was generally low (<2.0 m^3/s) but highly variable among watersheds, ranging from 0.06 m^3/s at Leviathan Creek to 1.97 m^3/s at Arkansas River (Fig. S2). Mean discharge during the summer months (June, July, and August) was highest for Arkansas River (4.08 m^3/s), similar for Clark Fork and Big Deer Creek (1.26 and 1.24 m^3/s, respectively), and lowest for Leviathan Creek (0.03 m^3/s). Except for a few high-flow years at Leviathan Creek, differences between the 10^{th} and 90^{th} percentile flows were greatest for Arkansas River and Big Deer Creek. These differences in flow extremes accounted for much greater F_v at Arkansas River (6.7) and Big Deer Creek (6.5) compared with Clark Fork (1.9) and Leviathan Creek (2.0). Overall, hydrologic properties related to mean discharge, high- and low-flow extremes, and F_v showed few differences between pre- and post-remediation periods (Table S1).

We observed dramatic declines in metal concentrations in each watershed during and after completion of remediation (Fig. 2). Metal concentrations at impacted sites in the 2 most contaminated streams (Leviathan Creek and Big Deer Creek) decreased by 82 to 96% within 7 y, whereas 60 to 76% declines occurred in the 2 less contaminated watersheds. Prior to remediation, CCU values at all impacted sites ranged from 10.5 to 197.6, but CCU values were reduced to 2.2 to 4.1 CCU after remediation was completed. Proportional contributions of different metals to CCUs varied over time and among watersheds (Fig. S3). CCU values were not strongly related to any streamflow variables in Arkansas River or Big Deer Creek but were positively correlated with low-flow extremes for Clark Fork (Table S1). Because CCU values at Clark Fork's downstream site were relatively low prior to remediation, we did not observe additional decreases in metal concentrations after remediation.

Long-term changes in macroinvertebrate assemblages

Total number of taxa increased following initiation of restoration activities in each watershed, with the greatest improvements generally occurring early during restoration (Fig. 3A). Rapid and complete recovery of the total number of taxa at impacted sites was observed within 10 to 14 y in Arkansas River, Leviathan Creek, and Big Deer Creek. These changes in species richness reflected long-term

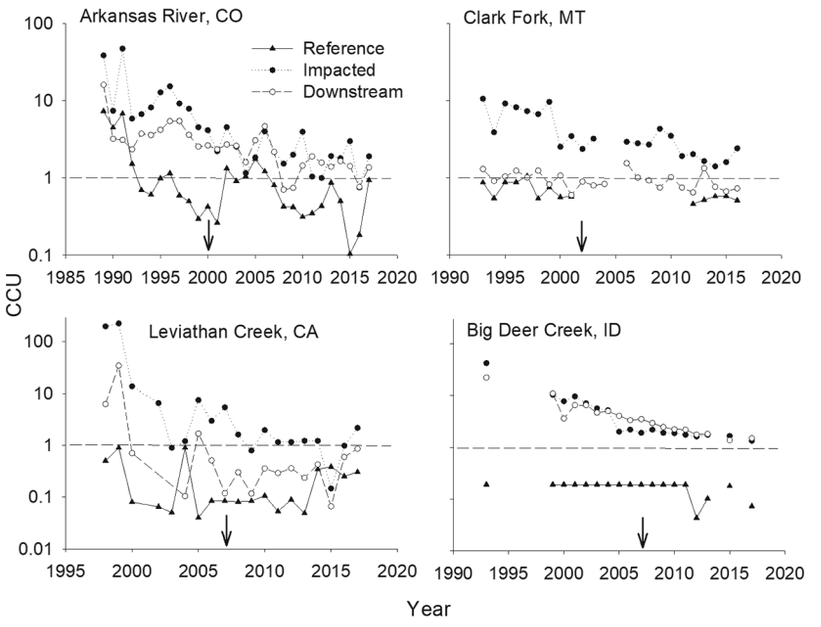

Figure 2. Long-term changes in metal concentrations (as cumulative criterion units [CCU]) at reference sites, mining-impacted sites, and sites located downstream from sources of mining contamination in each watershed. Dashed horizontal lines show where CCU = 1, the theoretical value that should be protective of most species. Arrows indicate the year when major restoration activities were completed. CO = Colorado, MT = Montana, CA = California, ID = Idaho.

decreases in metal concentrations, but much of this recovery occurred before metal concentrations approached 1 CCU. In contrast to these results, we did not observe complete recovery of species richness at Clark Fork's impacted site, despite achieving relatively low metal concentrations. The total number of taxa at Clark Fork gradually increased over the study period but decreased in the last 2 y of the study. Long-term changes in species richness at downstream sites in each watershed were more complex and either followed patterns observed at impacted sites or were similar to those at reference sites. Interestingly, in Big Deer Creek, the number of taxa at the impacted site recovered faster than at the downstream recovery site.

Unlike measures of species richness, which rapidly reached reference conditions in most streams, BC similarity at impacted and downstream sites never fully recovered (Fig. 3B). BC similarity to reference sites in each watershed steadily increased, but the trajectories of these responses varied among streams and between sites. Except for Arkansas River, where similarity of the impacted assemblage approached reference conditions, the other impacted assemblages reached thresholds ~50 to 60% of reference.

In addition to increases in species richness and similarity to reference sites likely resulting from improvements in water quality (see below), we also observed responses to episodic events and other anthropogenic disturbances. For example, the large decreases in richness and BC similarity at impacted sites in Leviathan Creek (2005–2006) and Arkansas River (2013) resulted from pulses of metals associated with containment pond overflows during high runoff years and disturbance resulting from additional (post-restoration) habitat improvements, respectively.

We evaluated long-term changes in number of taxa and BC similarity to reference sites before and after remediation by using a BACI analysis (Table 3, Fig. 4A, B). Results of GLM-analyses testing for effects of site, restoration treatment, and the site × treatment interaction on total number of taxa differed among the 4 watersheds. Site and treatment effects were substantial for all watersheds and showed that the number of taxa was consistently lower at impacted sites and generally increased at both impacted and downstream sites following restoration. There were strong site × treatment interactions for Clark Fork, Leviathan Creek, and Big Deer Creek, indicating that improvements in number of taxa were likely a result of restoration. There was no detectable site × treatment interaction for Arkansas River, suggesting that improvements in number of taxa at impacted and downstream sites may not be directly attributed to restoration. The lack of an interaction effect at Arkansas River resulted from the moderate differences in number of taxa among sites before restoration and the relatively small increases at impacted and downstream sites after restoration.

We observed substantial increases in BC similarity of impacted sites to reference sites after remediation in all watersheds (Table 3, Fig. 4B). Differences between impacted and downstream sites were also substantial, but these

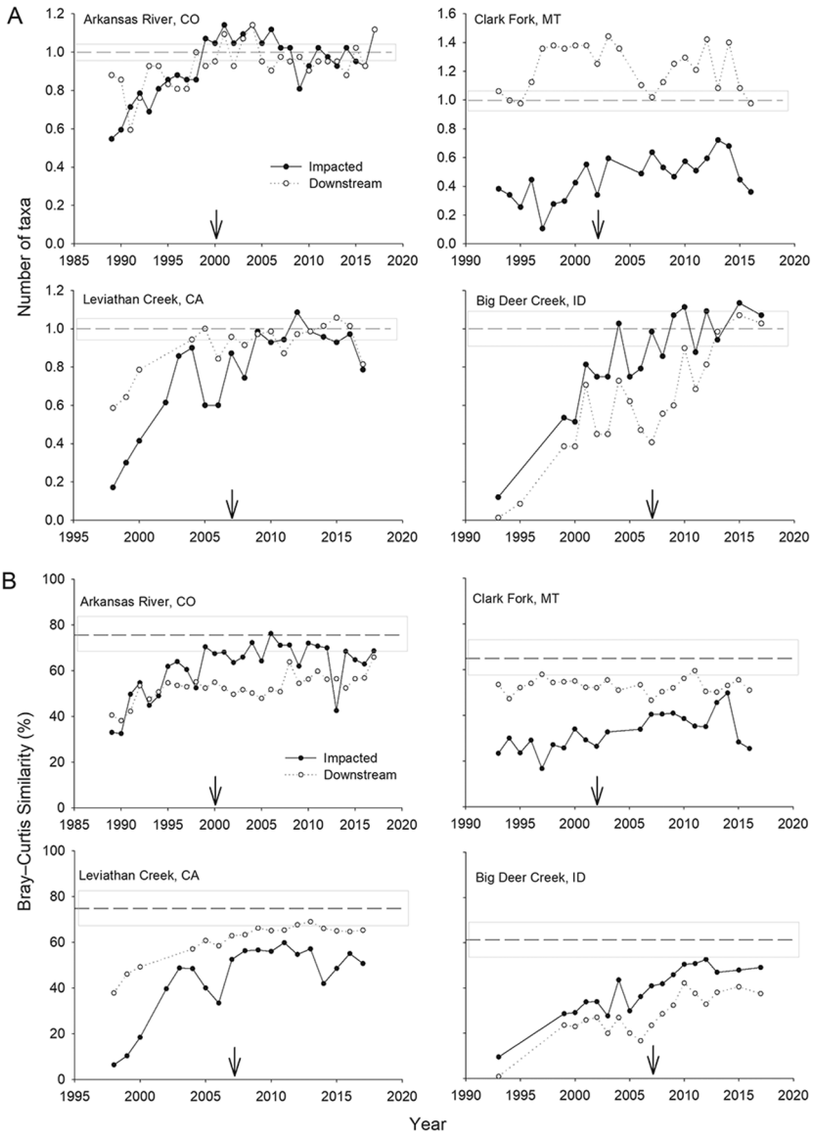

Figure 3. Long-term changes in total number of macroinvertebrate taxa (A) and Bray–Curtis similarity (%) (B) at mining-impacted sites and sites located downstream from the source of mining contamination in each watershed. Data are expressed relative to the reference sites. Dashed horizontal lines and boxes show means ±95% confidence intervals at reference sites based on either long-term averages (total number of taxa) or mean year-to-year variation in assemblage composition (Bray–Curtis similarity). Arrows indicate completion of restoration activities in each watershed. CO = Colorado, MT = Montana, CA = California, ID = Idaho.

differences varied among regions. In the 2 watersheds where sources of colonization were located immediately upstream (Arkansas River and Big Deer Creek), the site effects resulted from lower similarity of downstream sites to reference sites. In contrast, downstream sites were more similar to reference sites in Clark Fork and Leviathan Creek, watersheds where upstream sources of colonization were either not present or limited to very small tributaries. Because BC similarity in this analysis is a measure of similarity to the reference site, interpretation of the site × treatment interaction term is different than for the other metrics. The site × treatment interaction was substantial for Arkansas River and Clark Fork, indicating that increases in BC similarity after restoration were greater at impacted sites than at downstream sites. The lack of an interaction effect at both Leviathan Creek and Big Deer Creek indicated that increases in similarity to the reference site did not differ between impacted and downstream sites.

Other benthic metrics that we examined, including total macroinvertebrate abundance, mayfly richness, and mayfly abundance, increased at most impacted and downstream sites as metal concentrations decreased (Figs S4–S7). Most

Table 3. Results of 2-way general lineal models (GLM) showing the effects of treatment (before, after remediation of mining pollution), site (reference, impacted, downstream), and the site × treatment interaction term for all macroinvertebrate metrics in each watershed. Total (corrected) degrees of freedom for the Arkansas River, Clark Fork, Big Deer Creek, and Leviathan Creek GLMs were 86, 58, 55, and 54, respectively.

		Arkansas River			Clark Fork			Leviathan Creek			Big Deer Creek		
Macroinvertebrate metric	Source	R^2	*F*-value	*p*-value	R^2	*F*-value	*p*-value	R^2	*F*-value	*p*-value	R^2	*F*-value	*p*-value
No. of taxa	Model	0.44	12.8	0.001	0.85	59.1	0.001	0.48	8.9	0.001	0.42	7.3	0.001
	Treatment		50.9	0.001		4.5	0.040		17.5	0.001		12.1	0.001
	Site		5.1	0.008		135.9	0.001		9.4	0.004		7.6	0.001
	Site × treatment		1.9	0.162		11.2	0.001		4.3	0.019		2.9	0.066
Total abundance	Model	0.39	10.2	0.001	0.68	22.6	0.001	0.63	17.0	0.001	0.47	8.9	0.001
	Treatment		39.1	0.001		5.0	0.030		11.4	0.001		15.5	0.001
	Site		2.5	0.087		44.7	0.001		34.7	0.001		11.3	0.001
	Site × treatment		4.1	0.020		10.2	0.002		3.1	0.055		1.7	0.191
Mayfly richness	Model	0.57	21.1	0.001	0.85	60.6	0.001	0.65	18.5	0.001	0.71	24.0	0.001
	Treatment		56.0	0.001		3.7	0.062		15.1	0.003		22.4	0.001
	Site		24.2	0.001		142.0	0.001		30.4	0.001		36.2	0.001
	Site × treatment		1.3	0.268		6.2	0.004		9.3	0.0004		6.7	0.003
Mayfly abundance	Model	0.35	8.9	0.001	0.84	56.1	0.001	0.70	23.1	0.001	0.50	10.1	0.001
	Treatment		24.5	0.001		7.1	0.010		14.3	0.001		7.5	0.009
	Site		4.2	0.019		121.1	0.001		48.6	0.001		12.0	0.001
	Site × treatment		6.4	0.003		17.1	0.001		1.8	0.175		6.3	0.004
Bray–Curtis similarity	Model	0.43	13.6	0.001	0.85	76.0	0.001	0.59	15.4	0.001	0.60	15.7	0.001
	Treatment		19.7	0.001		13.0	0.001		21.5	0.001		31.7	0.001
	Site		14.0	0.001		197.3	0.001		20.4	0.001		15.1	0.001
	Site × treatment		4.0	0.052		16.2	0.001		1.5	0.208		0.0	0.897
Functional diversity	Model	0.23	4.8	0.001	0.45	8.54	0.001	0.35	5.17	0.001	0.24	2.96	0.021
	Treatment		17.4	0.001		0.33	0.566		16.69	0.001		7.66	0.008
	Site		3.1	0.052		20.85	0.001		2.45	0.097		1.48	0.237
	Site × treatment		0.5	0.583		0.19	0.829		1.9	0.161		1.7	0.194

of these metrics recovered and were similar to or greater than reference values by the time restoration was completed. Similar to total taxonomic richness, species richness of mayflies was especially sensitive to remediation, as indicated by the large amount of variation explained (57–85%) by the GLM analyses (Table 3). In contrast to measures of species richness, total macroinvertebrate abundance and abundance of mayflies showed considerable annual variation at all sites. Despite this greater variability, treatment effects, site effects, and the effects of the site × treatment interaction on the abundance metrics were substantial in most watersheds, consistent with the hypothesis that long-term changes were a result of restoration and resulting lower metal concentrations. Functional trait diversity also recovered rapidly and approached reference conditions at most sites, but this metric showed considerable annual variability, especially at impacted sites (Fig. S7). Consequently, the total amount of variation explained by the GLM analyses and the magnitude of site, treatment, and interaction effects was generally lower for functional trait diversity than for other metrics.

Averaged across all metrics we examined, the mean recovery time at impacted sites ranged from ~10 to 14 y, and the mean proportion of reference conditions achieved ranged from 0.67 to 0.95 (Table 4). With the exception of Big Deer Creek, less contaminated downstream sites recovered faster than impacted sites for most metrics. Across the 4 watersheds, the mean proportion of reference conditions achieved at impacted sites was highly variable among metrics. Total number of taxa and functional trait diversity showed the greatest resemblance to reference conditions after restoration (mean = 0.92–1.0), whereas BC similarity showed the least (mean = 0.54).

Annual (year-to-year) variation in the composition of reference assemblages based on BC similarity was relatively high in some watersheds (Fig. S8). Because our measures of recovery were based on similarity to reference conditions, this variation likely increased the time required for impacted and downstream sites to recover and reduced the potential for achieving reference conditions. For example, across all metrics, recovery times were generally longer and the similarity

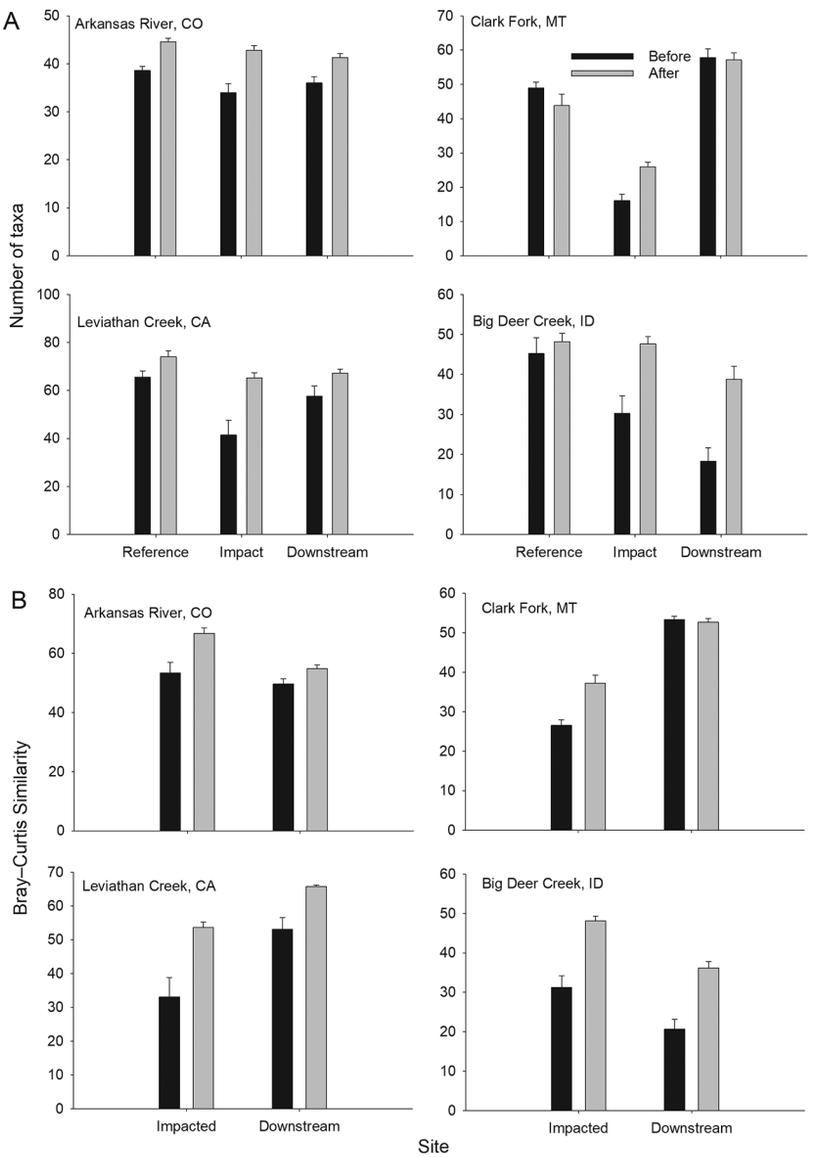

Figure 4. Mean (±SE) number of macroinvertebrate taxa (A) and Bray–Curtis similarity (%) to reference assemblages (B) at mining-impacted and sites located downstream from sources of mining contamination before and after completion of remediation in each watershed. Note that Bray–Curtis similarity values are not shown for reference sites because these values were calculated based on similarity to reference assemblages. Details of statistical analyses showing effects of site, restoration treatment, and the site × treatment interaction are shown in Table 3. CO = Colorado, MT = Montana, CA = California, ID = Idaho.

to reference conditions was lower for Clark Fork and Big Deer Creek, the 2 watersheds that showed the greatest interannual variation in reference assemblages.

Results of piecewise regression analysis showed a distinct threshold relationship between EPT species richness and CCU (Fig. 5; Table S2). The estimated threshold for this relationship occurred at ~1 CCU, and a 20% reduction of EPT richness occurred at 2.1 CCUs (95% confidence interval = 1.4–3.2). Approximately 50% of all EPT taxa were eliminated from these streams at CCU levels >15.5.

Analysis of species traits

Results of NMDS analyses based on species traits showed distinct spatial and temporal separation, with reference sites generally clustering together in each panel (Fig. 6A). Based on NMDS results of indicator traits analysis, we identified 6 trait states (grazers, semivoltine taxa, organisms using cutaneous respiration, organisms common in the drift, clingers, and organisms with nonseasonal development) that were primarily associated with reference assemblages and recovering streams (Fig. 6B). We also found that assemblages

Table 4. Estimates of the proportion of reference conditions achieved (Prop. reference) and the time required to achieve maximum similarity to reference or equilibrium conditions (Years) for each macroinvertebrate metric at impacted and downstream sites in all watersheds after remediation of mining impacts was initiated.

Metric	Recovery indicator	Arkansas River		Clark Fork		Leviathan Creek		Big Deer Creek	
		Impacted	Downstream	Impacted	Downstream	Impacted	Downstream	Impacted	Downstream
Total taxa	Prop. reference	1.00	1.00	0.67	1.00	1.00	1.00	1.00	1.00
	Years	10	12	10	0	11	9	14	20
Total abundance	Prop. reference	1.00	1.00	1.00	1.00	0.59	0.70	0.60	0.60
	Years	10	3	10	0	14	8	16	22
Mayfly richness	Prop. reference	1.00	0.80	0.31	1.00	0.83	0.92	1.00	0.56
	Years	13	11	13	3	9	6	15	17
Mayfly abundance	Prop. reference	1.00	1.00	0.65	1.00	0.20	1.00	0.60	0.77
	Years	10	4	20	0	9	14	14	15
Bray–Curtis similarity	Prop. reference	0.68	0.53	0.41	0.51	0.57	0.65	0.51	0.37
	Years	10	6	14	5	11	9	17	17
Functional diversity	Prop. reference	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
	Years	7	7	21	0	4	2	10	10

altered by metal contamination before restoration was complete were dominated by small collector–gatherers and fast-developing, multivoltine taxa having strong drift propensity, sprawling habits, and anal gills (many Chironomidae). The most important variables separating sites and treatments across all watersheds were metal concentrations (as CCU) and year (Fig. 6C).

The proportional abundance of trait states at reference sites in each watershed was much more consistent than the proportional abundance of the dominant macroinvertebrate groups (Fig. S9). All major groups were represented in each watershed, but there was considerable variation in their relative abundance. For example, mayflies, chironomids, and riffle beetles (Elmidae) dominated reference sites at Arkansas River, Clark Fork, and Leviathan Creek, respectively, whereas non-insect taxa (e.g., oligochaetes) were considerably more common at Big Deer Creek. Despite these large differences in taxonomic composition among reference sites, proportional abundance of the dominant trait states that characterized reference conditions showed little variation among sites.

DISCUSSION

The primary goals of this study were to compare the responses of 4 mining-impacted western US watersheds to long-term improvements in water quality and demonstrate the value BACI designs for quantifying restoration success. Despite variation in watershed characteristics, primary metals of concern, and remediation practices, we observed convergent responses to improvements in water quality among the 4 watersheds. Across all macroinvertebrate assemblage metrics, mean recovery times and the extent of recovery were relatively consistent, although episodic events changed recovery trajectories. Differences in recovery among watersheds were likely determined by a number of factors, including the severity of contamination, effectiveness of remediation, habitat quality, proximity to upstream sources of colonization, and hydrologic variation. We also observed considerable variation in the rate and extent of recovery among metrics, which was best illustrated by differences between total species richness and BC similarity. Consistent with previous studies (De Laender et al. 2012, Dornelas et al. 2014, Mori et al. 2018), species richness recovered rapidly at most sites, but assemblage composition remained very different from reference sites.

Long-term changes in metals and stream hydrology

Metal concentrations at impacted and downstream sites responded to remediation in all watersheds and approached theoretically protective concentrations (CCU = 1) soon after treatments were completed. The rate at which metals decreased was likely determined by the degree of contamination,

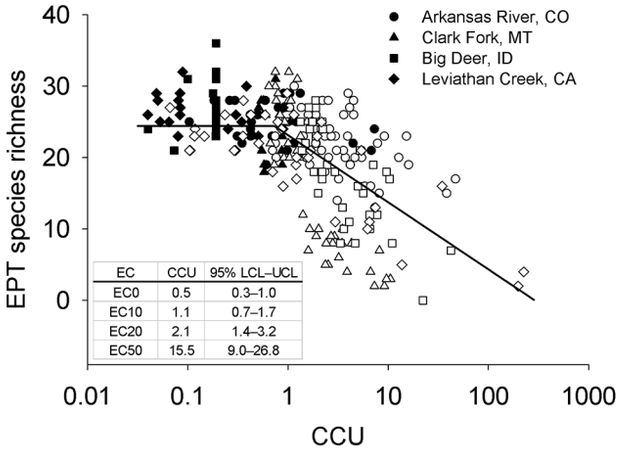

Figure 5. Results of piecewise linear regression analysis showing the relationship between metal concentration (cumulative criterion units [CCU]) and Ephemeroptera, Plecoptera, and Trichoptera (EPT) species richness in the Arkansas River, Colorado (CO), Clark Fork, Montana (MT), Leviathan Creek, California (CA), and Big Deer Creek, Idaho (ID). Closed symbols are reference sites, and open symbols indicate mining-impacted sites and sites located downstream from sources of mining contamination. Inset shows estimated CCU levels with lower 95% confidence intervals (LCL) and upper 95% confidence intervals (UCL) that would result in 0, 10, 20, and 50% reductions in EPT species richness (EC).

specific remediation treatments used, and the intrinsic hydrological and physical properties of each watershed.

The relationship between metal loading and stream discharge can be difficult to quantify because of unmeasured seasonal variation. To be consistent among watersheds, metal concentrations and other physicochemical data reported here were collected in late summer or early autumn. Consequently, hydrologic characteristics measured during these baseflow conditions had relatively little influence on metal concentrations. However, in all 4 watersheds, elevated metal concentrations have been reported during periods of high spring runoff (Clements et al. 2010, Sando et al. 2014, Mebane et al. 2015, Herbst et al. 2018). In addition, brief but intense summer rainstorms in mining-disturbed watersheds can lead to metal pulses that are sufficiently elevated to cause direct toxicity to aquatic organisms (Table 2). Episodic increases in metal concentrations following rainstorms have been observed in Clark Fork (Nagorski et al. 2003, Balistrieri et al. 2012) and Big Deer Creek (Mebane et al. 2015). These results suggest that limiting sampling to late summer low-flow conditions may underrepresent metals exposure and miss potentially important episodic events.

We also observed considerable annual variation in metal concentrations at most of the impacted and downstream sites. This variation resulted from annual changes in hydrologic characteristics (stream discharge, precipitation), residual metals in sediments and riparian areas, and impacts from ongoing remediation activities. Metal concentrations in Clark Fork and Leviathan Creek were greater during low-flow years, likely as a result of reduced dilution (Bird 1987, Runkel et al. 2013) and increased influx of metals from groundwater sources (Gandy et al. 2007, Hudson et al. 2018). Changes in streamflow before and after remediation could also result in corresponding changes in metal concentrations and confound linkages between metals and biological responses. However, with the exception of Big Deer Creek, we observed few differences in streamflow, suggesting that hydrology was probably not a key factor contributing to differences in metal exposure before and after remediation.

Benthic assemblage responses to remediation

Uncertainty regarding the time required for an ecosystem to recover from disturbance and the composition of post-restoration communities are significant issues in restoration ecology (Hobbs et al. 2009). Numerous biotic and abiotic factors determine recovery of disturbed ecosystems, but insufficient duration of post-restoration monitoring is the most common reason that many studies fail to document a return to pre-disturbance conditions (Jones and Schmitz 2009). Our study of these 4 western watersheds continued for 9 to 17 y post-restoration, providing sufficient time to evaluate recovery trajectories. Most sites recovered by the time restoration was completed, although there was variation among individual metrics and watersheds. On several occasions, metric values were actually greater at impacted or downstream sites compared with those at reference sites. This pattern was especially evident for abundance metrics after remediation, which we attribute to their greater annual variability compared to measures of species richness. These findings are consistent with other studies that have demonstrated rapid recovery of benthic communities following the removal of a stressor (Jones and Schmitz 2009, Gergs et al. 2016).

Maintaining and enhancing biodiversity has been a primary focus of global conservation efforts for many years. In addition to measuring changes in species richness, we believe monitoring programs should also assess species replacement or changes in β-diversity at disturbed sites. Biotic homogenization, defined as the replacement of rare species by more cosmopolitan taxa (Dornelas et al. 2014, Magurran 2016), can occur without declines in species richness (Mori et al. 2018) and can mask effects of contaminants and other stressors (De Laender et al. 2012). Biotic homogenization may also result in the loss of ecosystem services and reduce the resilience of communities to novel perturbations (Paine et al. 1998, Dornelas et al. 2014, Magurran 2016). Maintaining species diversity theoretically improves ecological resilience, but only if species vary in their sensitivity to environmental perturbations (Folke et al. 2004). As ecological resilience is reduced, even small perturbations can trigger regime shifts and lead to the persistence of homogenized communities. For example, mesocosm experiments conducted with reference and downstream macroinvertebrate assemblages from

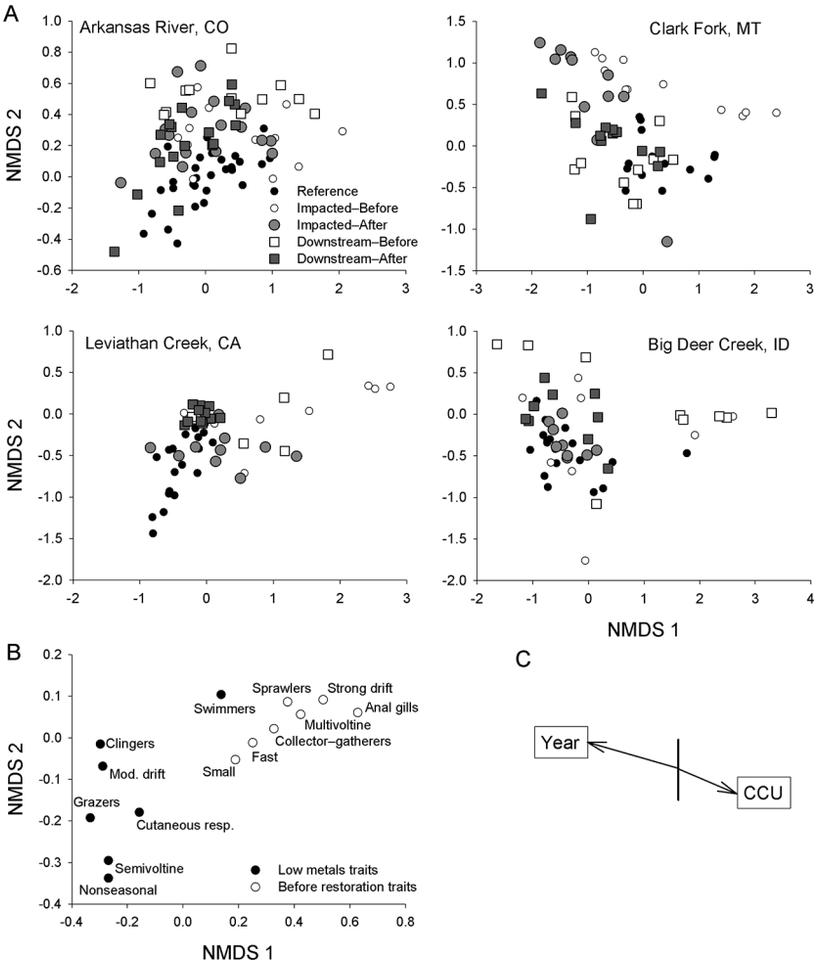

Figure 6. Nonmetric multidimensional scaling (NMDS) ordination of stream sites (reference, mining-impacted, and those located downstream from sources of mining contamination) and restoration treatments (before, after) based on macroinvertebrate species traits in each watershed (A). Insets show the trait states that were primarily responsible for separating sites and treatments based on indicator traits analysis (B) and the major environmental vectors related to the ordinations (early to later years right to left and cumulative criterion units [CCU] increasing left to right (C). Year = year of study. CO = Colorado, MT = Montana, CA = California, ID = Idaho, Mod. = moderate, resp. = respiration.

Arkansas River demonstrated that metal-tolerant communities were substantially more sensitive to other stressors, including acidification, UV-B radiation, and petroleum hydrocarbons (Wolff et al. 2019). These results suggest that, in addition to traditional community metrics, measures of restoration success should also include recovery of community composition and estimates of community resilience (Clements et al. 2010).

Differences in the rate and extent of recovery among assemblage metrics have important implications for how we characterize spatiotemporal patterns of aquatic insect assemblages. A recent meta-analysis of 166 long-term studies (including sites in the present study) concluded that aquatic insect abundance has increased globally by ~11%/decade (van Klink et al. 2020). In our study, abundance and taxa richness recovered rapidly at most sites, but the composition of assemblages remained substantially altered, and their similarity to reference sites was relatively low (mean BC similarity after 10 y = 0.53). Furthermore, including results of restoration studies in assessments of long-term changes in aquatic insect abundance or diversity may provide an overly optimistic perspective because these studies are designed specifically to document post-restoration responses of depleted communities. Although these studies provide encouraging results for localized cases, they indicate only that assemblages are recovering where we invest the effort in restoration and do not demonstrate that broader-scale diversity is actually increasing.

Using species traits to assess responses to metals

In addition to assessing long-term responses of assemblage metrics to restoration, we investigated spatial and temporal changes in species traits. Because these traits are

mechanistically linked to critical ecological processes, preserving their functional diversity may be more important than maintaining the actual number of species, especially in the face of anthropogenic disturbances (Cadotte et al. 2011). We observed clear separation among sites and between restoration treatments based on trait states that were shown to be negatively associated with metals, including abundance of grazers, semivoltine taxa, organisms using cutaneous respiration, organisms common in the drift, clingers, and those with nonseasonal development. The proportional abundance of these trait states at impacted and downstream sites increased after restoration, suggesting they are useful indicators of recovery in metal-polluted streams. Common patterns among the disparate taxa in each region showed that small, fast-developing collector–gatherers with multiple generations each year and high propensity for drift (primarily midges) characterized the most polluted sites.

In addition to providing a more mechanistic understanding of assemblage responses to metals, we analyzed species traits for practical reasons. Because our 4 study sites were distributed across a broad geographic region, each with a very different species pool, we expected that some taxonomic metrics would be ineffective for assessing recovery. Even at relatively coarse levels of taxonomic resolution (e.g., macroinvertebrate orders), the composition of reference assemblages was highly variable among watersheds. In contrast, the proportional abundance of species trait states at reference sites was remarkably consistent, highlighting the usefulness of species traits for assessing effects of contaminants across regions. The sensitivity of mayflies, particularly heptageniids (Clements et al. 2000, Mebane et al. 2015, Herbst et al. 2018), to metals also illustrates the value of using species traits to quantify restoration effectiveness. We believe the response of heptageniids to metals likely has little to do with their taxonomic classification but is primarily the result of a set of unique traits that determines metal exposure (grazers, medium body size, respiration by gills), recolonization ability (weak-to-moderate drift), and behavioral habit (clingers). Experimentally verifying the mechanistic links between species traits and responses to metals will improve our ability to predict effects on aquatic insects and help reconcile the vast differences in sensitivity reported between field studies and laboratory experiments (Brix et al. 2011, Clements et al. 2013, Poteat and Buchwalter 2014).

Factors that influenced recovery

There was variation in recovery trajectories among watersheds and among metrics in our study, although benthic macroinvertebrate assemblages in each of the streams we sampled approached reference conditions. Our study provided a unique opportunity to quantify this variation because all streams were impacted by the same type of stressors and because reference assemblages were generally dominated by organisms with a similar set of species traits.

Landscape features of the 4 watersheds likely influenced the rate of recovery. Because recolonization after disturbance in lotic ecosystems occurs primarily from upstream sources, the network structure of a watershed can influence recovery rates, especially in mining-polluted streams (Kitto et al. 2015). In the 2 watersheds where impacted sites were located immediately downstream from uncontaminated sources of recolonization (Arkansas River and Big Deer Creek), assemblage similarity to reference conditions was actually greater at the impacted sites located closer to the pollution source compared with the more distant downstream sites. Conversely, species richness at the impacted site in Clark Fork, where an upstream recolonization source was absent, did not fully recover, and these assemblages showed low similarity to reference conditions after 14 y (BC similarity = 0.41). These results suggest that proximity to upstream, uncontaminated colonization sources likely influenced the rate and completeness of recovery in these watersheds, although we cannot completely exclude other potential explanations (e.g., differences in habitat, residual metals).

Because we characterized recovery based on comparisons to reference sites within each watershed, these assessments were influenced by natural spatial and temporal variability. Although reference sites in our study were not identical to impacted sites, the major habitat characteristics that structure benthic communities (e.g., stream size, substrate composition, current velocity) were similar. Reference sites in our study also showed considerable year-to-year variation in BC similarity, but this variability was greatest in Big Deer Creek and Clark Fork. This high annual variability creates a moving target for recovery and may explain the overall lower similarity of impacted and downstream assemblages to reference conditions in these watersheds. These results suggest that recovery of mining-contaminated streams depends not only on the severity of disturbance and characteristics of impacted communities but also on the natural variation of reference communities.

Hydrologic characteristics, which differed greatly among watersheds, and episodic events may partially explain differences in recovery rates. For example, the high F_v observed in Arkansas River may have selected for taxa capable of rapid recolonization, thereby increasing recovery rate of this watershed. We also observed substantial episodic changes in benthic assemblages, some of which can be explained, whereas specific explanations for others are more speculative. The dramatic decrease in several assemblage metrics in Leviathan Creek in 2005 to 2006 likely resulted from a spike in metal concentrations associated with high stream discharge and flooded containment ponds. Similarly, the large decreases in total abundance, mayfly abundance, and BC similarity at Arkansas River's impacted site in 2013 resulted from physical disturbances associated with large-scale habitat treatments. Moderately elevated metal concentrations and flood events just prior to sampling Clark Fork in 2015 and 2016 contributed to the decline in number of taxa in this system,

whereas the high variability at Big Deer Creek's reference site early in the study likely resulted from a large wildfire and associated flooding. These results demonstrate that, in addition to the well-documented biotic and abiotic factors that influence recovery, stochastic events and other sources of disturbance should be considered. The relative effects of these episodic events on recovery trajectories could not have been identified in our study without a long-term perspective.

Relationship of our findings to water-quality criteria for metals

An obvious question arising from this research is whether chronic water-quality criteria for metals, either individually or in combination, are protective of aquatic insect assemblages. Piecewise regression analysis of the relationship between EPT richness and CCU suggests that the aquatic criteria values used here were near the concentration that caused substantial declines in taxa richness. Our estimated effect concentrations resulting in 10 and 20% taxa loss were ~1 and 2 CCUs, respectively (Fig. 5). These calculations are consistent with those reported from standard laboratory toxicity tests; however, we believe a field-based effect concentration resulting in a 20% loss of taxa is a more ecologically significant effect than is a 20% reduction in growth or survival of laboratory test species. Furthermore, regulatory criteria for metals and the cleanup objectives for these watersheds are applied individually and do not account for the potential combined effects of different metals. Because each study area had at least 2 metals of concern (Fig. S3), it follows that water-quality objectives would be considered achieved in these systems when concentrations were reduced to ~2 CCUs. Results of our study indicate that compliance allowed at this level could result in a substantial loss of aquatic insect biodiversity.

We used CCUs to characterize metal effects because, unlike laboratory studies where the relative toxicity of individual metals can be quantified, CCUs are a practical necessity for field studies of metal mixtures. However, there are several important assumptions associated with the use of CCUs to estimate toxicological effects. First, we assumed that an individual criterion unit (CU) has equal toxicity to benthic macroinvertebrates, regardless of the specific metal. Establishing criterion values for metals involves compiling aquatic toxicity data for a variety of laboratory test species, developing a species sensitivity distribution that is expected to protect 95% of these tested species, and assuming this value will also protect natural communities (Stephan et al. 1985). Results of mesocosm experiments have generally shown consistent sensitivity rankings of taxa among metals (Mebane et al. 2020), suggesting that the assumption of equitoxicity of individual CUs is reasonable. The use of CCUs also assumes that dissolved metals were primarily responsible for the observed toxicological effects. However, diet may be a dominant route of uptake for some metals (Kim et al. 2012), and its toxicological significance varies among taxa and with concentration (Balistrieri et al. 2020). Regardless of the specific mechanisms of toxicity, we assumed CCUs are appropriate for estimating exposure because concentrations in dietary sources (e.g., periphyton, benthic organic matter) are often correlated with dissolved metals (Kiffney and Clements 1993, Hickey and Clements 1998, Mebane et al. 2020). We also recognize that, in addition to dissolved metals, other stressors, such as physical habitat alterations (e.g., deposition of Fe oxide precipitates; Cadmus et al. 2016) or indirect effects associated with alterations in food chains (Niyogi et al. 2001, Carlisle and Clements 2005), impact benthic communities in mining-contaminated streams. Finally, we assumed that toxicity of metal mixtures is approximately additive and that individual CUs can be combined into a cumulative measure of effects. In other words, a no-effect concentration of a single metal combined with a no-effect concentration of a 2nd metal can produce a toxic mixture (Meyer et al. 2015, Versieren et al. 2016, Mebane et al. 2020). This assumption is consistent with the present results, which showed that 1 CCU was generally protective of species richness, but 2 CCUs corresponded with a substantial loss of taxa richness. If the potential for increased toxicity of 2 or more chemicals, each at or below their individual criteria limits, is ignored, then criteria will provide less protection than intended by USEPA guidelines.

Importance of long-term BACI studies for assessing restoration effectiveness

The necessity of long-term studies for identifying factors that structure the distribution and abundance of freshwater organisms is well established (Jackson and Füreder 2006, Armitage et al. 2007, Hornberger et al. 2009, Clements et al. 2010, Smith et al. 2011, Mebane et al. 2015, Herbst et al. 2018). However, long-term studies are surprisingly uncommon in the literature, and very few have been conducted across broad geographic regions. Furthermore, many long-term studies consist only of a few snapshots in time rather than continuous monitoring, making it difficult to assess responses to other natural or anthropogenic changes. The 4 watersheds in the current study were distributed across a broad geographic region and were sampled continuously over a 20- to 29-y period, providing sufficient opportunity to evaluate long-term changes in benthic assemblages and alterations in hydrologic regimes, metal loading, and responses to restoration.

Although long-term studies can improve our ability to understand natural variation in aquatic communities, separating this variation from effects of restoration (or other) treatments remains a significant challenge. BACI designs, in which control and restoration sites are sampled on multiple occasions before and after restoration, may be the only way to definitively quantify restoration success (e.g., Kotalik et al. 2021). A unique advantage of BACI study designs is

that they provide strong evidence for a direct cause-and-effect relationship between improvements in water quality or habitat and community responses. However, because BACI designs are susceptible to annual variation, they are also influenced by low statistical power (Christie et al. 2019), making it difficult to detect interactions between site and treatment. In our study, high annual variability was responsible for several site × treatment interaction terms having no statistically detectable effect, even though inspection of long-term trajectories showed likely treatment effects. In this situation, we believe that it is more appropriate to rely on graphical representations of BACI results rather than strict adherence to somewhat arbitrary *p*-values (Carpenter et al. 1995, Murtaugh 2002). Furthermore, because a site × treatment interaction can also occur when temporal changes at control sites are greater than those at restoration sites (Chevalier et al. 2019), additional information is required when interpreting BACI interaction terms. In our study, increases in most metrics after restoration were consistently greater at impacted and downstream sites compared with reference sites. These findings provide strong evidence that the improvements in benthic assemblages across these 4 watersheds after treatment were a direct result of restoration.

Defining restoration success

Regulatory guidelines that identify restoration success are typically based on improvements in water quality or habitat and often do not include assessments of ecological condition. Other measures of restoration success may include biological responses, such as the recovery of sensitive or recreationally important species. For example, trout populations have returned to Arkansas River (now classified as a Gold Medal Trout Stream) and Big Deer Creek (Clements et al. 2010, Mebane et al. 2015), which could be considered indicators of success, despite persistent alterations in macroinvertebrate assemblage composition. Given that recovery rates vary among metrics and may depend on a complex set of biotic and abiotic variables, defining restoration success based on a single, pre-defined threshold is often unrealistic.

In the current study, we used comparable reference streams in each watershed as benchmarks for recovery and defined restoration success as achieving reference or near reference conditions for a diverse set of biological metrics that was sustained over time. Based on the dramatic improvements in water quality and benthic assemblages observed across a broad geographic region, we conclude that remediation of these systems has been generally successful. Despite there being variation in defining what constitutes complete restoration because of differences among metrics, using multiple lines of evidence provided quantifiable measures of the timing and completeness of recovery relative to reference conditions. Because these 4 watersheds were among the most severely polluted sites in the western US, our study demonstrates the value of these investments in watershed restoration and the potential for success under the most extreme conditions.

ACKNOWLEDGEMENTS

Author contributions: All authors provided original data, conducted data analyses, and contributed to preparation of the final manuscript.

Comments by Chris Kotalik, Ian Waite, and Dave Mount significantly improved an earlier draft of this manuscript and are greatly appreciated. All data for the Arkansas River (Colorado) project were collected by researchers at Colorado State University with funding provided primarily by the United States Environmental Protection Agency (USEPA), the National Institutes of Health, and Colorado Parks and Wildlife. We are especially grateful to >150 undergraduates (and counting) that assisted in the field and diligently spent thousands of hours sorting benthic samples over the past 3 decades. Primary funding for the Clark Fork River (Montana) project was provided by USEPA Region 8, with additional support provided by the Earth Systems Processing Division of the United States Geological Survey (USGS) Water Mission Area and the USGS Environmental Health Mission Area. Macroinvertebrate assemblage data for the Clark Fork and tributary sites were based on collections by Daniel L. McGuire of McGuire Consulting. All discharge and water chemistry data were collected by the USGS Wyoming–Montana Water Science Center in Helena, Montana. John Lambing, Kent Dodge, Steve Sando, Chris Ellison, and Greg Clark were instrumental in providing the water quality data during the course of this study. The Big Deer Creek (Idaho) biological data were primarily collected by Robert Eakins, Brian Fraser, the late Paul McKee, and their associates at EcoMetrix Incorporated, Mississauga, Ontario, Canada. Water chemistry data were primarily collected by Golder Associates, Redmond, Washington. Data collection was primarily funded by the Blackbird Mine Site Group, which in turn is funded by a group of mining companies and the US government. Further details are given in Mebane et al. (2015). Leviathan Creek (California) sample processing has been ably accomplished by Sierra Nevada Aquatic Research Laboratory staff, including Bruce Medhurst, Scott Roberts, Mike Bogan, Bruce Hammock, Ian Bell, Sandi Roll, Matt Wilson, Jeff Kane, and others. Ned Black of USEPA collected and collated water chemistry data. Funding has come from the USEPA, Atlantic Richfield, Lahontan Water Quality Control Board, and the United States Army Corps of Engineers. Tom Suk, Daniel McMindes, and Julie Sullivan provided project coordination. Thanks to Wood Environment and Infrastructure for on-site assistance.

LITERATURE CITED

Andersen, T., P. S. Cranston, and J. H. Epler (editors). 2013. Chironomidae of the Holarctic region: Keys and diagnoses, Part 1: Larvae. Insect systematics and evolution supplement 66. Lund, Sweden.

Armitage, P. D., M. J. Bowes, and H. M. Vincent. 2007. Long-term changes in macroinvertebrate communities of a heavy metal polluted stream: The river Nent (Cumbria, UK) after 28 years. River Research and Applications 23:997–1015.

Balistrieri, L. S., C. A. Mebane, and T. S. Schmidt. 2020. Time-dependent accumulation of Cd, Co, Cu, Ni, and Zn in natural

communities of mayfly and caddisfly larvae: Metal sensitivity, uptake pathways, and mixture toxicity. Science of the Total Environment 732:139011.

Balistrieri, L. S., D. A. Nimick, and C. A. Mebane. 2012. Assessing time-integrated dissolved concentrations and predicting toxicity of metals during diel cycling in streams. Science of the Total Environment 425:155–168.

Barbour, M. T., B. G. Bierwagen, A. T. Hamilton, and N. G. Aumen. 2010. Climate change and biological indicators: Detection, attribution, and management implications for aquatic ecosystems. Journal of the North American Benthological Society 29:1349–1353.

Benda, L., N. L. Poff, D. Miller, T. Dunne, G. Reeves, G. Pess, and M. Pollock. 2004. The network dynamics hypothesis: How channel networks structure riverine habitats. BioScience. 54:413–427.

Bernhardt, E. S., M. A. Palmer, J. D. Allan, G. Alexander, K. Barnas, S. Brooks, J. Carr, S. Clayton, C. N. Dahm, J. Follstad-Shah, D. L. Galat, S. Gloss, P. Goodwin, D. R. Hart, B. Hassett, R. Jenkinson, S. L. Katz, G. M. Kondolf, P. S. Lake, R. Lave, J. L. Meyer, T. K. O'Donnell, L. Pagano, B. Powell, and E. Sudduth. 2005. Synthesizing U.S. river restoration efforts. Science 308: 636–637.

Berumen, M. L., and M. S. Pratchett. 2006. Recovery without resilience: Persistent disturbance and long-term shifts in the structure of fish and coral communities at Tiahura Reef, Moorea. Coral Reefs 25:647–653.

Bird, S. C. 1987. The effect of hydrological factors on trace metal contamination in the river Tawe, South Wales. Environmental Pollution 45:87–124.

Brix, K. V., D. K. DeForest, and W. J. Adams. 2011. The sensitivity of aquatic insects to divalent metals: A comparative analysis of laboratory and field data. Science of the Total Environment 409:4187–4197.

Brix, K. V., D. K. DeForest, L. M. Tear, M. Grosell, and W. J. Adams. 2017. Use of multiple linear regression models for setting water quality criteria for copper: A complementary approach to the Biotic Ligand Model. Environmental Science & Technology 51:5182–5192.

Cadmus, P., S. F. Brinkman, and M. K. May. 2018. Chronic toxicity of ferric iron for North American aquatic organisms: Derivation of a chronic water quality criterion using single species and mesocosm data. Archives of Environmental Contamination and Toxicology 74:605–615.

Cadmus, P., W. H. Clements., J. L. Williamson, J. F. Ranville, J. S. Meyer, and M. J. G. Gines. 2016. The use of field and mesocosm experiments to quantify effects of physical and chemical stressors in mining-contaminated streams. Environmental Science & Technology 50:7825–7833.

Cadotte, M. W., K. Carscadden, and N. Mirotchnick. 2011. Beyond species: Functional diversity and the maintenance of ecological processes and services. Journal of Applied Ecology 48:1079–1087.

Carlisle, D., and W. H. Clements. 2005. Leaf litter breakdown, microbial respiration and shredder production in metal-polluted streams. Freshwater Biology 50:380–390.

Carpenter, S. R., S. W. Chisholm, C. J. Krebs, D. W. Schindler, and R. F. Wright. 1995. Ecosystem experiments. Science 269:324–327.

Chevalier, M., J. C. Russell, and J. Knape. 2019. New measures for evaluation of environmental perturbations using Before-After-Control-Impact analyses. Ecological Applications 29: e01838.

Christie, A. P., T. Amano, P. A. Martin, G. E. Shackelford, B. I. Simmons, and W. J. Sutherland. 2019. Simple study designs in ecology produce inaccurate estimates of biodiversity responses. Journal of Applied Ecology 56:2742–2754.

Clausen, B., and B. J. F. Biggs. 2000. Flow variables for ecological studies in temperate streams: Groupings based on covariance. Journal of Hydrology 237:184–197.

Clements, W. H., M. L. Brooks, D. R. Kashian, and R. E. Zuellig. 2008. Changes in dissolved organic material determine exposure of stream benthic communities to UV-B radiation and heavy metals: Implications for climate change. Global Change Biology 14:2201–2214.

Clements, W. H., P. Cadmus, and S. F. Brinkman. 2013. Responses of aquatic insects to Cu and Zn in stream microcosms: Understanding differences between single species tests and field responses. Environmental Science & Technology 47:7506–7513.

Clements, W. H., D. M. Carlisle, J. M. Lazorchak, and P. C. Johnson. 2000. Heavy metals structure benthic communities in Colorado mountain streams. Ecological Applications 10:626–638.

Clements, W. H., N. K. M. Vieira, and S. E. Church. 2010. Quantifying restoration success and recovery in a metal-polluted stream: A 17-year assessment of physicochemical and biological responses. Journal of Applied Ecology 47:899–910.

De Laender, F., D. Verschuren, R. Bindler, O. Thas, and C. R. Janssen. 2012. Biodiversity of freshwater diatom communities during 1000 years of metal mining, land use, and climate change in Central Sweden. Environmental Science & Technology 46:9097–9105.

Diamond, J. M. 1983. Ecology: Laboratory, field and natural experiments. Nature 304:586–587.

Dobson, A. P., A. D. Bradshaw, and A. J. M. Baker. 1997. Hopes for the future: Restoration ecology and conservation biology. Science 277:515–522.

Dornelas, M., N. J. Gotelli, B. McGill, H. Shimadzu, F. Moyes, C. Sievers, and A. E. Magurran. 2014. Assemblage time series reveal biodiversity change but not systematic loss. Science 344:296–299.

Erickson, R. J. 2015. Toxicity Relationship Analysis Program, version 1.30a. United States Environmental Protection Agency, National Health and Environmental Research Laboratory, Mid-Continent Ecology Division, Duluth, Minnesota. (Available from: https://archive.epa.gov/med/med_archive_03/web/html/trap.html)

Floury, M., P. Usseglio-Polatera, M. Ferreol, C. Delattre, and Y. Souchon. 2013. Global climate change in large European rivers: Long-term effects on macroinvertebrate communities and potential local confounding factors. Global Change Biology 19:1085–1099.

Foley, J. A., R. DeFries, G. P. Asner, C. Barford, G. Bonan, S. R. Carpenter, F. S. Chapin, M. T. Coe, G. C. Daily, H. K. Gibbs, J. H. Helkowski, T. Holloway, E. A. Howard, C. J. Kucharik, C. Monfreda, J. A. Patz, I. C. Prentice, N. Ramankutty, and P. K. Snyder. 2005. Global consequences of land use. Science 309:570–574.

Folke, C., S. Carpenter, B. Walker, M. Scheffer, T. Elmqvist, L. Gunderson, and C. S. Holling. 2004. Regime shifts, resilience,

and biodiversity in ecosystem management. Annual Review of Ecology, Evolution, and Systematics 35:557–581.

Gandy, C. J., J. W. N. Smith, and A. P. Jarvis. 2007. Attenuation of mining-derived pollutants in the hyporheic zone: A review. Science of the Total Environment 373:435–446.

Gergs, A., S. Classen, T. Strauss, R. Ottermanns, T. C. M. Brock, H. T. Ratte, U. Hommen, and T. G. Preuss. 2016. Ecological recovery potential of freshwater organisms: Consequences for environmental risk assessment of chemicals. Reviews of Environmental Contamination and Toxicology 236:259–294.

Gestring, B., and L. Sumi. 2013. Polluting the future: How mining companies are contaminating our nation's water in perpetuity. Earthworks. (Available from: https://earthworks.org/publications/polluting_the_future/)

Herbst, D. B., R. B. Medhurst, and N. P. R. Black. 2018. Long-term effects and recovery of streams from acid mine drainage and evaluation of toxic metals threshold ranges for community re-assembly. Environmental Toxicology and Chemistry 37: 2575–2592.

Hickey, C. W., and W. H. Clements. 1998. Effects of heavy metals on benthic macroinvertebrate communities in New Zealand streams. Environmental Toxicology and Chemistry 17:2338–2346.

Hobbs, R. J., S. Arico, J. Aronson, J. S. Baron, P. Bridgewater, V. A. Cramer, P. R. Epstein, J. J. Ewel, C. A. Klink, A. E. Lugo, D. Norton, D. Ojima, D. M. Richardson, E. W. Sanderson, F. Valladares, M. Vilà, R. Zamora, and M. Zobel. 2006. Novel ecosystems: Theoretical and management aspects of the new ecological world order. Global Ecology and Biogeography 15: 1–7.

Hobbs, R. J., E. Higgs, and J. A. Harris. 2009. Novel ecosystems: Implications for conservation and restoration. Trends in Ecology & Evolution 24:599–605.

Hornberger, M. I., S. N. Luoma, M. L. Johnson, and M. Holyoak. 2009. Influence of remediation in a mine-impacted river: Metal trends over large spatial and temporal scales. Ecological Applications 19:1522–1535.

Hudson, E., B. Kulessa, P. Edwards, T. Williams, and R. Walsh. 2018. Integrated hydrological and geophysical characterisation of surface and subsurface water contamination at abandoned metal mines. Water, Air, & Soil Pollution 229:256.

Hudson-Edwards, K. 2016. Tackling mine wastes. Science 352: 288–290.

INAP (International Network for Acid Prevention). 2009. The global acid rock drainage guide (GARD Guide). The International Network for Acid Prevention. (Available from: http://www.gardguide.com/)

Jackson, J. K., and L. Füreder. 2006. Long-term studies of freshwater macroinvertebrates: A review of the frequency, duration and ecological significance. Freshwater Biology 51:591–603.

Jones, A., M. Rogerson, G. Greenway, H. A. B. Potter, and W. M. Mayes. 2013. Mine water geochemistry and metal flux in a major historic Pb-Zn-F orefield, the Yorkshire Pennines, UK. Environmental Science and Pollution Research 20:7570–7581.

Jones, H. P., and O. J. Schmitz. 2009. Rapid recovery of damaged ecosystems. PLoS ONE 4:e5653.

Khan, F. R., W. Keller, N. D. Yan, P. G. Welsh, C. M. Wood, and J. C. McGeer. 2012. Application of biotic ligand and toxic unit modelling approaches to predict improvements in zooplankton species richness in smelter-damaged lakes near Sudbury, Ontario. Environmental Science & Technology 46:1641–1649.

Kiffney, P. M., and W. H. Clements. 1993. Bioaccumulation of heavy metals by benthic invertebrates at the Arkansas River, Colorado. Environmental Toxicology and Chemistry. 12:1507–1518.

Kim, K. S., D. H. Funk, and D. B. Buchwalter. 2012. Dietary (periphyton) and aqueous Zn bioaccumulation dynamics in the mayfly *Centroptilum triangulifer*. Ecotoxicology 21:2288–2296.

Kitto, J. A. J., D. P. Gray, H. S. Greig, D. K. Niyogi, and J. S. Harding. 2015. Meta-community theory and stream restoration: Evidence that spatial position constrains stream invertebrate communities in a mine impacted landscape. Restoration Ecology 23:284–291.

Kotalik, C. J., P. Cadmus, and W. H. Clements. 2021. Before-After Control-Impact field surveys and novel experimental approaches provide valuable insights for characterizing stream recovery from acid mine drainage. Science of the Total Environment 771: 145419.

Laliberté, E., and P. Legendre. 2010. A distance-based framework for measuring functional diversity from multiple traits. Ecology 91:299–305.

Lefcort, H., J. Vancura, and E. L. Lider. 2010. 75 years after mining ends stream insect diversity is still affected by heavy metals. Ecotoxicology 19:1416–1425.

Magurran, A. E. 2016. How ecosystems change. Science 351:448–449.

Mebane, C. A. 2015. In response: Biological arguments for selecting effect sizes in ecotoxicological testing. Environmental Toxicology and Chemistry 34:2440–2442.

Mebane, C. A., R. J. Eakins, B. G. Fraser, and W. J. Adams. 2015. Recovery of a mining-damaged stream ecosystem. Elementa: Science of the Anthropocene 3:000042.

Mebane, C. A., T. S. Schmidt, J. L. Miller, and L. S. Balistrieri. 2020. Bioaccumulation and toxicity of cadmium, copper, nickel, and zinc to aquatic insect communities. Environmental Toxicology and Chemistry 39:812–833.

Merritt, R. W., K. W. Cummins, and M. B. Berg. 2008. An introduction to the aquatic insects of North America. Kendall Hunt Publishing, Dubuque, Iowa.

Meyer, J. S., K. J. Farley, and E. R. Garman. 2015. Metal mixtures modeling evaluation: 1. Background. Environmental Toxicology and Chemistry 34:726–740.

Moe, S. J., K. De Schamphelaere, W. H. Clements, M. T. Sorensen, P. J. Van den Brink, and M. Liess. 2013. Combined and interactive effects of global climate change and toxicants on populations and communities. Environmental Toxicology and Chemistry 32:49–61.

Mori, A. S., F. Isbell, and R. Seidl. 2018. β-diversity, community assembly, and ecosystem functioning. Trends in Ecology & Evolution 33:549–564.

Murtaugh, P. A. 2002. On rejection rates of paired intervention analysis. Ecology 83:1752–1761.

Nagorski, S. A., J. N. Moore, T. E. McKinnon, and D. B. Smith. 2003. Scale-dependent temporal variations in stream water geochemistry. Environmental Science and Technology 37:859–864.

Niyogi, D. K., W. M. Lewis, and D. M. McKnight. 2001. Litter breakdown in mountain streams affected by mine drainage: Biotic mediation of abiotic controls. Ecological Applications 11:506–516.

Noyes, P. D., M. K. McElwee, H. D. Miller, B. W. Clark, L. A. Van Tiem, K. C. Walcott, K. N. Erwin, and E. D. Levin. 2009. The

toxicology of climate change: Environmental contaminants in a warming world. Environment International 35:971–986.

NRC (National Research Council). 2007. Sediment dredging at superfund sites: Assessing the effectiveness. The National Academies Press, Washington, DC.

Paine, R. T., M. J. Tegner, and E. A. Johnson. 1998. Compounded perturbations yield ecological surprises. Ecosystems 1:535–545.

Palmer, M. A., K. L. Hondula, and B. J. Koch. 2014. Ecological restoration of streams and rivers: Shifting strategies and shifting goals. Annual Review of Ecology, Evolution, and Systematics 45:247–269.

Poff, N. L., J. D. Olden, N. K. M. Vieira, D. S. Finn, M. P. Simmons, and B. C. Kondratieff. 2006. Functional trait niches of North American lotic insects: Traits-based ecological applications in light of phylogenetic relationships. Journal of the North American Benthological Society 25:730–755.

Pond, G. J., M. E. Passmore, N. D. Pointon, J. K. Felbinger, C. A. Walker, K. J. G. Krock, J. B. Fulton, and W. L. Nash. 2014. Long-term impacts on macroinvertebrates downstream of reclaimed mountaintop mining valley fills in central Appalachia. Environmental Management 54:919–933.

Poteat, M. D., and D. B. Buchwalter. 2014. Four reasons why traditional metal toxicity testing with aquatic insects is irrelevant. Environmental Science & Technology 48:887–888.

Runkel, R. L., K. Walton-Day, B. A. Kimball, P. L. Verplanck, and D. A. Nimick. 2013. Estimating instream constituent loads using replicate synoptic sampling, Peru Creek, Colorado. Journal of Hydrology 489:26–41.

Sando, S. K., A. V. Vecchia, D. L. Lorenz, and E. P. Barnhart. 2014. Water-quality trends for selected sampling sites in the Upper Clark Fork Basin, Montana, water years 1996–2010. Scientific Investigations Report 2013–5217.United States Geological Survey, Wyoming-Montana Water Science Center, Reston, Virginia. (Available from: https://doi.org/10.3133/sir20135217)

Scheffer, M., and S. R. Carpenter. 2003. Catastrophic regime shifts in ecosystems: Linking theory to observation. Trends in Ecology & Evolution 18:648–656.

Sheldon, F., and M. C. Thoms. 2006. Relationships between flow variability and macroinvertebrate assemblage composition: Data from four Australian dryland rivers. River Research and Applications 22:219–238.

Smith, J. G., C. C. Brandt, and S. W. Christensen. 2011. Long-term benthic macroinvertebrate community monitoring to assess pollution abatement effectiveness. Environmental Management 47:1077–1095.

Stephan, C. E., D. I. Mount, D. J. Hansen, J. H. Gentile, G. A. Chapman, and W. A. Brungs. 1985. Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses. EPA 822-R-85-100. United States Environmental Protection Agency, Office of Research and Development, Environmental Research Laboratories, Duluth, Minnesota. (Available from: https://www.epa.gov/sites/production/files/2016-02/documents/guidelines-water-quality-criteria.pdf)

Stubblefield, W. A., and J. R. Hockett. 2000. Derivation of a Colorado state manganese table value standard for the protection of aquatic life. (Available from: ENSR Corporation, 4303 West LaPoint Avenue, Fort Collins, Colorado, 80521)

Stubblefield, W. A., E. Van Genderen, A. S. Cardwell, D. G. Heijerick, C. R. Janssen, and K. De Schamphelaere. 2020. Acute and chronic toxicity of cobalt to freshwater organisms: Using a species sensitivity distribution approach to establish international water quality standards. Environmental Toxicology and Chemistry 39:799–811.

Toms, J. D., and M. L. Lesperance. 2003. Piecewise regression: A tool for identifying ecological thresholds. Ecology 84:2034–2041.

Thorp, J. H., and A. P. Covich (editors). 2001. Ecology and classification of North American freshwater invertebrates. Academic Press, San Diego, California.

USEPA (United States Environmental Protection Agency). 1980. Ambient water quality criteria for zinc. EPA 440/5-87-003. United States Environmental Protection Agency, Office of Water Regulations and Standards Criteria and Standards Division, Washington, DC. (Available from: https://www.epa.gov/sites/production/files/2018-12/documents/ambient-wqc-zinc.pdf)

USEPA (United States Environmental Protection Agency). 1984b. Ambient water quality criteria for arsenic -1984. NTIS PB85-227445. United States Environmental Protection Agency, Office of Water Regulations and Standards Criteria and Standards Division, Washington, DC. (Available from: https://www.epa.gov/sites/production/files/2019-02/documents/ambient-wqc-arsenic-1984.pdf)

USEPA (United States Environmental Protection Agency). 1986. Ambient water quality criteria for nickel-1986. EPA 440/5-86-004. United States Environmental Protection Agency, Office of Water Regulations and Standards Criteria and Standards Division, Washington, DC. (Available from: https://www.epa.gov/sites/production/files/2019-03/documents/ambient-wqc-nickel-1986.pdf)

USEPA (United States Environmental Protection Agency). 1996. 1995 updates: Water quality criteria documents for the protection of aquatic life in ambient water. EPA 820-B-96-001. United States Environmental Protection Agency, Office of Water Regulations and Standards Criteria and Standards Division, Washington, DC. (Available from: https://www.epa.gov/sites/production/files/2019-03/documents/1995-updates-wqc-protection-al.pdf)

USEPA (United States Environmental Protection Agency). 2016a. Aquatic life ambient water quality criterion for cadmium – 2016. EPA-820-R-16-002. United States Environmental Protection Agency, Office of Water Regulations and Standards Criteria and Standards Division, Washington, DC. (Available from: https://www.epa.gov/sites/production/files/2016-03/documents/cadmium-final-report-2016.pdf)

USEPA (United States Environmental Protection Agency). 2016b. Aquatic life ambient water quality criterion for selenium – Freshwater 2016. 822-R-16-006. United States Environmental Protection Agency, Office of Water Regulations and Standards Criteria and Standards Division, Washington, DC. (Available from: https://www.epa.gov/sites/production/files/2016-07/documents/aquatic_life_awqc_for_selenium_-_freshwater_2016.pdf)

USEPA (United States Environmental Protection Agency). 2018. Final aquatic life ambient water quality criteria for aluminum. EPA-822-R-18-001. United States Environmental Protection Agency, Office of Water Regulations and Standards Criteria and Standards Division, Washington, DC. (Available from: https://www.epa.gov/sites/production/files/2018-12/documents/aluminum-final-national-recommended-awqc.pdf)

van Klink, R., D. E. Bowler, K. B. Gongalsky, A. B. Swengel, A. Gentile, and J. M. Chase. 2020. Meta-analysis reveals declines

in terrestrial but increases in freshwater insect abundances. Science 368:417–420.

Van Looy, K., M. Floury, M. Ferréol, M. Prieto-Montes, and Y. Souchon. 2016. Long-term changes in temperate stream invertebrate communities reveal a synchronous trophic amplification at the turn of the millennium. Science of the Total Environment 565:481–488.

Verberk, W., C. G. E. van Noordwijk, and A. G. Hildrew. 2013. Delivering on a promise: Integrating species traits to transform descriptive community ecology into a predictive science. Freshwater Science 32:531–547.

Versieren, L., S. Evers, K. A. C. De Schamphelaere, R. Blust, and E. Smolders. 2016. Mixture toxicity and interactions of copper, nickel, cadmium, and zinc to barley at low effect levels: Something from nothing? Environmental Toxicology and Chemistry 35:2483–2492.

Vieira, N. K. M., N. L. Poff, D. M. Carlisle, S. R. Moulton II, M. L. Koski, and B. C. Kondratieff. 2006. A database of lotic invertebrate traits for North America. Data Series 187. United States Geological Survey, United States Department of the Interior, Reston, Virginia. (Available from: http://pubs.usgs.gov/ds/ds187/)

Wolff, B. A., S. B. Duggan, and W. H. Clements. 2019. Resilience and regime shifts: Do novel communities impede ecological recovery in a historically metal-contaminated stream? Journal of Applied Ecology 56:2698–2709.