How to evaluate sediment fingerprinting source apportionments

P. V. G. Batista, J. P. Laceby, O. Evrard

Journal of Soils and Sediments
https://doi.org/10.1007/s11368-022-03157-4

SEDIMENTS, SEC 3 • HILLSLOPE AND RIVER BASIN SEDIMENT DYNAMICS • RESEARCH ARTICLE

How to evaluate sediment fingerprinting source apportionments

P. V. G. Batista[1] · **J. P. Laceby**[2] · **O. Evrard**[3]

Received: 27 September 2021 / Accepted: 31 January 2022

Abstract

Purpose Evaluating sediment fingerprinting source apportionments with artificial mixtures is crucial for supporting decision-making and advancing modeling approaches. However, artificial mixtures are rarely incorporated into fingerprinting research and guidelines for model testing are currently lacking. Here, we demonstrate how to test source apportionments using laboratory and virtual mixtures by comparing the results from Bayesian and bootstrapped modeling approaches.

Materials and methods Laboratory and virtual mixtures ($n = 79$) with known source proportions were created with soil samples from two catchments in Fukushima Prefecture, Japan. Soil samples were sieved at 63 µm and analyzed for colorimetric and geochemical parameters. The MixSIAR Bayesian framework and a bootstrapped mixing model (BMM) were used to estimate source contributions to the artificial mixtures. In addition, we proposed and demonstrated the use of multiple evaluation metrics to report on model uncertainty, residual errors, performance, and contingency criteria.

Results and discussion Overall, there were negligible differences between source apportionments for the laboratory and virtual mixtures, for both models. The comparison between MixSIAR and BMM illustrated a trade-off between accuracy and precision in the model results. The more certain MixSIAR solutions encompassed a lesser proportion of known source values, whereas the BMM apportionments were markedly less precise. Although model performance declined for mixtures with a single source contributing greater than 0.75 of the material, both models represented the general trends in the mixtures and identified their major sources.

Conclusions Virtual mixtures are as robust as laboratory mixtures for assessing fingerprinting mixing models if analytical errors are negligible. We therefore recommend to always include virtual mixtures as part of the model testing process. Additionally, we highlight the value of using evaluation metrics that consider the accuracy and precision of model results, and the importance of reporting uncertainty when modeling source apportionments.

Keywords MixSIAR · Sediment fingerprinting · Sediment tracing · Model testing · Artificial mixtures

1 Introduction

Sediment source fingerprinting is increasingly used to estimate sources of particulate material in riverine, lacustrine, and coastal systems (Jalowska et al. 2017; Lavrieux et al. 2019; Gibbs et al. 2020). This approach capitalizes on differences in physical and biogeochemical parameters, or fingerprints, to model source contributions to target material. Fallout radionuclides, mineral magnetic properties, color parameters, major and trace element geochemistry, and other fingerprints have all been used to apportion sources in end-member mixing models. Several reviews have been published that provide a thorough discussion of sediment source fingerprinting fundamentals (Walling 2005; Koiter et al. 2013; Owens et al. 2016) and its importance in integrated water resource management (Collins et al. 2017; Owens 2020).

Central to the confident apportionment of target material to their sources is the modeling process. In the 1980s, end-member mixing models were first introduced to solve simultaneous equations with mean values of selected source and

Responsible editor: Hugh Smith

✉ P. V. G. Batista
pedro.batista@unibas.ch

[1] Department of Environmental Sciences, University of Basel, Bernoullistrasse 30, 4056 Basel, Switzerland

[2] Alberta Environment and Parks, 3535 Research Rd NW, Calgary, AB T2L 2K8, Canada

[3] Laboratoire des Sciences du Climat et de l'Environnement (LSCE/IPSL), CEA-CNRS-UVSQ, Université Paris-Saclay, Gif-sur-Yvette, France

Published online: 09 February 2022

target fingerprints (Peart and Walling 1986; Yu and Oldfield 1989). Approaches to modeling have significantly advanced over the last 20–30 years, shifting from deterministic optimization procedures (Walden et al. 1997) to stochastic frameworks that rely on Bayesian and/or Monte Carlo methods (Cooper et al. 2014; Nosrati et al. 2014; Laceby and Olley 2015). Importantly, stochastic modeling approaches produce distributions of unmixed source apportionments, to which prediction or credible intervals can be assigned. The analysis of prediction or credible intervals is crucial for understanding and reporting the uncertainty in the model estimates and, therefore, providing robust source information for sediment management programs.

Decision-makers and stakeholders often require accurate (low error) and precise (low uncertainty) sediment source information. While the precision of fingerprinting-estimated source contributions can be assessed with the previously mentioned stochastic approaches, evaluating their accuracy is challenging. Ultimately, this evaluation of model accuracy would require in situ measurements of sediment fluxes from each potential source unit, which is pragmatically unfeasible. Accordingly, artificial mixtures have been used to, at the very least, test the accuracy with which the models can estimate known source proportions. These artificial mixtures can be created in the lab (Martínez-Carreras et al. 2010; Haddadchi et al. 2014), by physically combining known masses of the source material, or virtually, by mathematically generating target tracer values (Laceby et al. 2015; Palazón et al. 2015; Sherriff et al. 2015).

Testing sediment fingerprinting source apportionments against artificial mixtures provides information on the quality of the source discrimination afforded by the tracer suite, the impact of within-source tracer variability, and the structure of the modeling approach (Pulley et al. 2017, 2020; Shi et al. 2021). Moreover, artificial mixtures have been used to assess the influence of corrupt and/or non-conservative tracers on model outputs, as well as the importance of different tracer selection procedures (Sherriff et al. 2015; Cooper and Krueger 2017; Latorre et al. 2021). Although artificial mixtures can provide a powerful tool for evaluating fingerprinting models, their use in sediment tracing papers is relatively rare (Fig. 1). For example, a query on the Web of Science with the term "sediment fingerprinting" AND ("artificial mixtures" OR "virtual mixtures") returned 24 articles for the period of 1987 to 2020. This represents 1.2% of the number of papers returned from the query "sediment fingerprinting" for the same time period (2061 hits).

In addition, models are predominantly evaluated with the mean absolute error (MAE) scoring metric, which is calculated from a single mean modeled relative source contribution and the known source proportion of a given artificial mixture. This approach to model testing counters the fundamental purpose of stochastic modeling, which is to provide a distribution of potential model solutions that reflects uncertainties in the data and model structure. Arguably, it is more important for a fingerprinting model to identify the main sources in a mixture, and for source apportionments to encompass known proportions with the most realistic representation of data and model uncertainty, than for

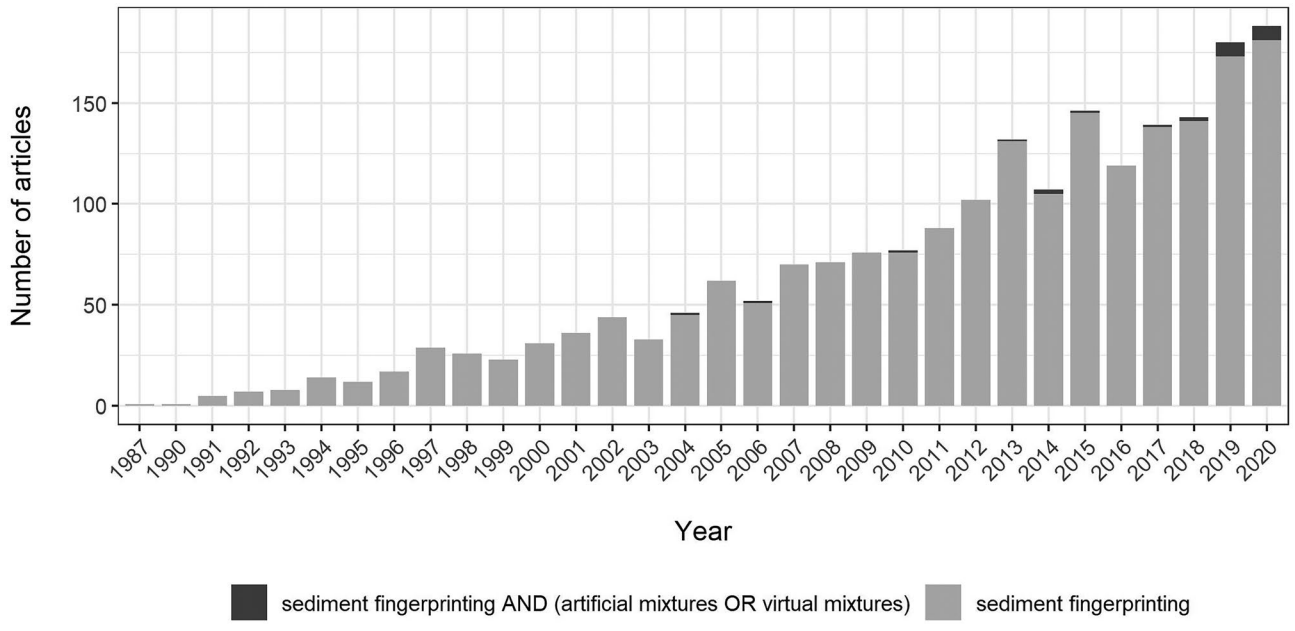

Fig. 1 Number of articles returned by the queries "sediment fingerprinting" and "sediment fingerprinting AND artificial mixtures OR virtual mixtures" in the WoS (1987–2020)

mean model solutions to be highly accurate. To this end, we believe that guidelines and appropriate metrics for evaluating current sediment fingerprinting models against artificial mixtures are somewhat lacking in the literature.

Another fundamental question regarding the evaluation of sediment fingerprinting mixing models relates to the use of laboratory or virtual mixtures for model testing. While the first approach might be more robust in terms of representing analytical error, biases might be introduced due to particle size effects and sample mixing (Collins et al. 2020). Moreover, a limitless number of virtual mixtures can be created without cost, which is particularly advantageous when the measurement of fingerprint properties is expensive and/or time consuming. As laboratory and virtual mixtures have not yet been thoroughly compared, how should researchers decide on which approach is more appropriate for testing their data and models?

Here, we present the results and code for an analysis of 79 artificial laboratory and virtual mixtures to facilitate researchers' comparison of various models and tracer selection procedures. In particular, we compare a bootstrapped mixing model (BMM) (Batista et al. 2019) with the Bayesian MixSIAR framework (Stock et al. 2018) to demonstrate what we can learn from testing fingerprinting source apportionments against artificial mixtures. Furthermore, we propose a set of evaluation metrics and testing guidelines to help researchers and managers to assess model outputs, and to help specify which conclusions can be drawn from their data in support of sediment management programs.

2 Materials and methods

2.1 Study site

The artificial mixtures were created from samples obtained from the Mano (175 km^2) and Niida (275 km^2) catchments (Fig. 2), north of the Fukushima Dai-ichi Nuclear Power Plant (FDNPP) in Fukushima Prefecture, Japan. Rivers in these catchments flow from mountainous plateaus (700–900 m above sea level) downstream across a coastal plain before discharging into the Pacific Ocean. The mountainous plateaus are predominantly forested, with paddy fields and other croplands situated along the main river channel. The coastal plains consist mainly of cropland and developed areas.

Fallout from the FDNPP nuclear accident mainly contaminated the upper regions in these two catchments. For example, ^{137}Cs activities in soils ranged from 20 to 75 kBq kg^{-1} in the upper plateaus resulting in the evacuation of the local population in 2011. To facilitate the return of the local population, the Japanese authorities conducted extensive decontamination of agricultural and residential areas (Evrard et al. 2019). In particular, the top 5-cm layer of the soil profile was removed as it was reported to contain 95–99% of the ^{137}Cs concentrations (Lepage et al. 2015). Light-colored crushed granite extracted from local quarries replaced the ^{137}Cs-contaminated topsoil layer with a new substrate as part of the remediation program (Evrard et al. 2020). The crushed granite was then mixed into the topsoil layer prior to recultivation. The impetus for this research project was to investigate the utility of apportioning sediment sources from decontaminated and non-decontaminated cropland sources in the Fukushima region.

2.2 Sample collection and preparation

Soil samples ($n = 70$) were collected on two different occasions (July 2015 and March 2019). Sources included cropland ($n = 18$), forest ($n = 30$), subsurface ($n = 9$) (i.e., channel banks), and decontaminated soils ($n = 13$). Decontaminated soils were collected after the crushed granite was plowed into the soil. Most samples were obtained from the upper catchment plateau, in close proximity to the river network in areas susceptible to erosion processes. Each soil sample consisted of ~ 10 subsamples taken with a plastic spatula from the top 2 cm of the soil profile.

Samples were dried at 40 °C for 48 h, sieved to 63 µm, and placed into 15-ml plastic boxes prior to analysis. For colorimetric parameters, samples were analyzed with a portable diffuse reflectance spectrophotometer (Konica Minolta CM-700d), calibrated prior to each session with samples analyzed in triplicates. *XYZ* tristimulus values were determined based on the color-matching functions defined by the International Commission on Illumination (CIE 1931). Standardized tristimulus values were transformed into the CIE $L^*a^*b^*$ and CIE $L^*u^*v^*$ Cartesian coordinate systems with 15 colorimetric parameters determined for each sample (L^*, a^*, b^*, c^*, h, x, y, z, L, a, b, u^*, v^*, u', v').

An energy-dispersive X-ray fluorescence spectrometer (ED-XRF; Epsilon 4) was used to determine the chemical composition of the soil samples. Measurements were conducted on a minimum of 0.1 g of material in containers covered with Mylar films showing a 10-mm exposure surface, with all samples analyzed in triplicates. In total, 16 geochemical elements were determined, including Mg, Al, Si, P, K, Ca, Ti, Cr, Mn, Fe, Co, Ni, Cu, Zn, Sr, Ba, and Pb. Four elements (P, Cr, Co, and Ba) had values below the detection limit for multiple samples, and as such, they were excluded from subsequent analyses.

2.3 Artificial mixtures

To generate the laboratory mixtures, we prepared composite samples for each source. An equal amount of the individual

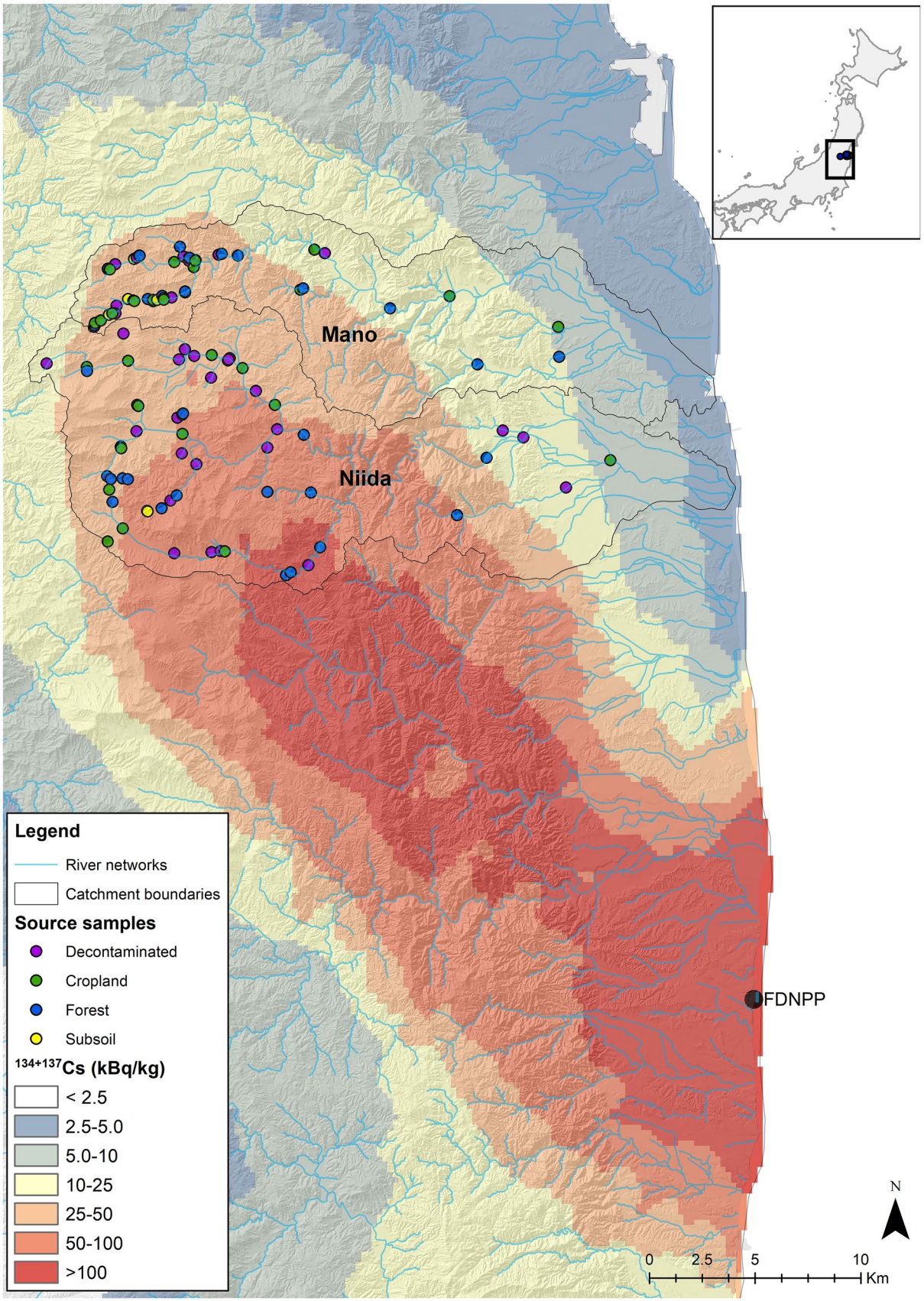

Fig. 2 Location of the Mano and Niida catchments, the Fukushima Dai-ichi Nuclear Power Plant (FDNPP), and soil sampling locations

source samples collected in the field was used to prepare the composite. In total, 79 laboratory mixtures containing variable proportions of the four sources (ranging between 0 and 100% and summing to 100%, Supplementary Material Fig. 1) were generated and prepared in the same containers as the soil samples.

The virtual mixtures for the BMM were generated by calculating multivariate normal distributions for the color and geochemical parameters in each source, in order to represent the uncertainty in the target mixtures and to preserve the correlations between tracers. The values in the multivariate normal distribution were then multiplied by the source contribution used to generate the laboratory mixtures. This was necessary because the model draws target tracer values from multivariate distributions for each iteration of the Monte Carlo simulation. Code to generate the artificial mixtures is appended in the supplemental information. For MixSIAR, the virtual mixtures were produced by simply multiplying mean source tracer values by their proportion in each mixture. This simpler procedure was chosen due to the ability of MixSIAR to handle single mixture data by use of a process error structure, in which mixture tracer values are drawn from a normal distribution, with a same mean, and a variance derived from weighted source variances (Stock et al. 2018).

2.4 Tracer selection and modeling

Fingerprint parameters were selected following the widely employed three-step procedure: (i) a range test for conservative behavior, (ii) a Kruskal–Wallis H-test for group differences, and (iii) a stepwise discriminant function analysis (DFA) (niveau = 0.1) for defining a minimum set of tracers that maximize source discrimination. For the range test, we assumed that the median and the interquartile range (IQR) of a tracer value in the mixture should be bracketed by the medians and the IQRs of the tracer value in the source groups. Of note, we did not test different fingerprint selection approaches, as this goes beyond the scope of our research. However, we encourage others to test alternative tracer screening methods with our dataset, code, and evaluation guidelines.

For the bootstrapping approach, we used the BMM outlined in Batista et al. (2019). The BMM is the open-source evolution of the distribution mixing model (Laceby and Olley 2015) that minimizes the sum of square residuals of the mixing equation for each iteration of a Monte Carlo simulation ($n = 2500$). Tracer values are sampled from multivariate normal distributions, which were constructed based on the log-transformed source and mixture data. For the laboratory mixtures, tracer distributions were created by taking the measured values in the mixture as a mean and assuming 5% standard deviation for each tracer.

The MixSIAR model trials for the laboratory mixtures were carried out with a process error formulation, which is used when single mixture values are available (Stock and Semmens 2016). The same error structure was used for the virtual mixtures. In both cases, we used an uninformative Dirichlet prior, and a very long Markov Chain Monte Carlo chain length with default burn-in and thinning values. Model convergence was assessed by the Gelman-Rubin diagnostic, in which none of the variables had a value greater than 1.01. Prior to modeling, all tracers were log-transformed to enforce a higher degree of normality to the fingerprint distributions. This was performed because MixSIAR assumes that source tracer values are normally distributed, and the removal of non-normally distributed tracers may reduce model accuracy (Smith et al. 2018). According to Laceby et al. (2021), log transformations were able to provide a higher degree of normality for geochemical data used as input for MixSIAR, compared to other transformations.

2.5 Model assessment

Appropriate evaluation metrics should be consistent with the purpose of a model. We believe that stochastic approaches to fingerprinting should provide distributions of source apportionments that encompass the actual source proportion in a mixture, while providing a realistic uncertainty estimation. That is, the distributions of source apportionments should reflect the errors in the data, in the mixing model, and the inherent uncertainty stemming from tracer variability, source sampling, and source grouping. From a management perspective, a mixing model should at least be able to consistently identify the main source in a mixture. That is, if we are unable to pinpoint the dominant sediment source in a catchment, we should not make management decisions based on modeled source apportionments.

Accordingly, we present four types of model assessment criteria, which focus on uncertainty, residuals, performance, and contingency errors. These criteria are outlined below, and a summary of the metrics and equations are provided in Table 1. Of note, all evaluation metrics originally developed for testing point-based estimates (e.g., MAE) were here calculated considering two prediction/credible intervals, namely the 50% (0.25 quantile–0.75 quantile) and 95% (0.025 quantile–0.975 quantile).

For the uncertainty parameters, we calculated the width of the 50 and 95% prediction/credible intervals of the source apportionments (W50 and W95). Wider distributions translated to higher uncertainty in the model solutions. In addition, we determined the proportion of known values in the artificial mixtures encompassed by the 50 and 95% (P50 and P95) prediction/credible intervals of the model outputs. The proportion should be equal to or greater than the interval (i.e., ≥ 0.5 for the 50% interval, ≥ 0.95 for the 95% interval).

Table 1 Model evaluation metrics and criteria

Criteria	Parameter	Equation	Reference
Uncertainty	Interval accuracy (P)	$\frac{encompassed}{total}$	–
	Interval width (W)	$upperquantile - lowerquantile$	–
Residual methods	Mean absolute error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	Bennett et al. (2013)
	Mean error (ME)	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$	Bennett et al. (2013)
Performance	Continuous ranked probability score (CRPS)	$(F_i, y_i) = \int_{-\infty}^{\infty}(F_i(y_i) - H\{y_i \geq \hat{y}_i\})^2 dx$	Matheson and Winkler (1976)
	Nash–Sutcliffe efficiency index	$1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y}_i)^2}$	Nash and Sutcliffe (1970)
Contingency	Critical success index (CSI)	$\frac{hits}{hits+misses+falsealarms}$	Bennett et al. (2013)
	Hit rate (HR)	$\frac{hits}{hits+misses}$	Bennett et al. (2013)

Regarding the residual methods, we calculated the mean absolute error of the 50 and 95% intervals (MAE50 and MAE95). In this case, if the known proportion was within the mentioned intervals, the error was zero. Otherwise, the error was calculated based on the difference between the upper or lower limits of the intervals and the known source values. MAE values range from zero to infinity, with zero being the perfect value. The MAE takes the same unit as the model, and it is not affected by cancelation (Bennett et al. 2013). We also estimated the mean error of the 50 and 95% intervals (ME50 and ME95). An ideal value of zero represents an unbiased model, while negative and positive values indicate under- and overprediction, respectively.

To quantify model performance, we calculated the continuous ranked probability score (CRPS) (Matheson and Winkler 1976), which is used to compare probabilistic model forecasts of continuous variables. The CRPS evaluates both the accuracy and the sharpness (i.e., precision) of a distribution of forecasted values: if the observed value corresponds to a high probability value in the distribution of model outputs, while the probability of other values is minimal, then the CRPS is minimized (Boucher et al. 2009). CRPS values range from zero (perfect deterministic forecast) to infinity, taking the same unit as the continuous variable. Formulae and detailed descriptions are provided in Jordan et al. (2019) and Laio and Tamea (2007). In addition, we calculated the Nash–Sutcliffe efficiency index (NSE) (Nash and Sutcliffe 1970) for the 50 and 95% intervals (NSE50 and NSE95). The NSE compares the performance of the model against the mean of the observed data, which we assumed to be an equal contribution from every source (0.25). Values range from minus infinity to one, with negative values indicating a worse performance than the mean, and one being a perfect model. As with the residual methods, the difference between known and estimated values was nullified if the actual source proportions were within the considered intervals.

For the contingency error metrics, we utilized the critical success index (CSI) for the identification of dominant sources. The CSI measured the fraction of cases in which dominant sources (> 0.5 proportion) were correctly identified within the 50 and 95% intervals (CSI50 and CSI95). The score penalizes both misses and false alarms. We also estimated the hit rate (HR) of the 50 and 95% intervals (HR50 and HR95). The hit rate represented the probability of detection of dominant sources (> 0.5 proportion). Unlike the CSI, the hit rate does not penalize misses or false alarms.

All scores and metrics described above were used both cumulatively to calculate global model results and, at individual source levels, to analyze source discrimination and specific patterns. Model implementation and statistical analyses were performed with the R open-source software and programming language (R Core Team 2021). Raw source and mixture data, as well as model code, are available as supplementary material.

3 Results

3.1 Tracer selection

All analyzed tracers were within the range of the laboratory and virtual mixtures, as expected. Fe and Ni did not display a significant difference according to the Kruskal–Wallis test and were therefore excluded from further analysis. The final tracer selection based on the stepwise DFA included color and geochemical parameters (L, h, y, c^*, K, Mn, Si, Ti, Zn), and yielded a reclassification accuracy of 90% (Supplementary Material Fig. 2).

3.2 Laboratory versus virtual mixtures

Source apportionments from the MixSIAR and BMM approaches for laboratory and virtual mixtures were highly similar. The results from the laboratory mixtures are plotted against the virtual mixtures in Fig. 3. MixSIAR apportionments for the different mixture types had a 0.95 correlation coefficient for the 0.05, 0.25, 0.5, 0.75, and 0.95 quantiles. The BMM results were slightly less correlated for the lower quantiles, showing a steady increase from the 0.05 quantile ($r = 0.91$) to the 0.95 quantile ($r = 0.95$). In addition, the accuracy and precision of estimated source proportions for both MixSIAR and BMM were parallel for the laboratory and virtual mixtures. This is illustrated by the negligible differences between CRPS values in Table 2. The contingency metrics demonstrated the largest difference between mixture types among the evaluation criteria, in particular for MixSIAR, which had superior results for the virtual mixtures.

3.3 Model assessment

There was a greater dissimilarity between MixSIAR and BMM outputs than between the results from the different mixture types. In particular, the uncertainty bands for the BMM apportionments were approximately twice as wide as those generated by MixSIAR (Table 2 and Fig. 4). Accordingly, a larger percentage of known source values was encompassed by the 50 and 95% intervals for the BMM apportionments. Although the 50% interval widths for the BMM calculations were almost identical to the 95% for MixSIAR, the BMM model's 50% interval widths were more accurate, encompassing ~ 70% of the known data, compared to ~ 50% for MixSIAR (Table 2).

Regarding the residual criteria, both models yielded slightly negative ME values (μ MixSIAR $= -0.01$, μ BMM $= -0.03$) for the 50% interval, which demonstrates a small bias toward underpredictions. MAE values per

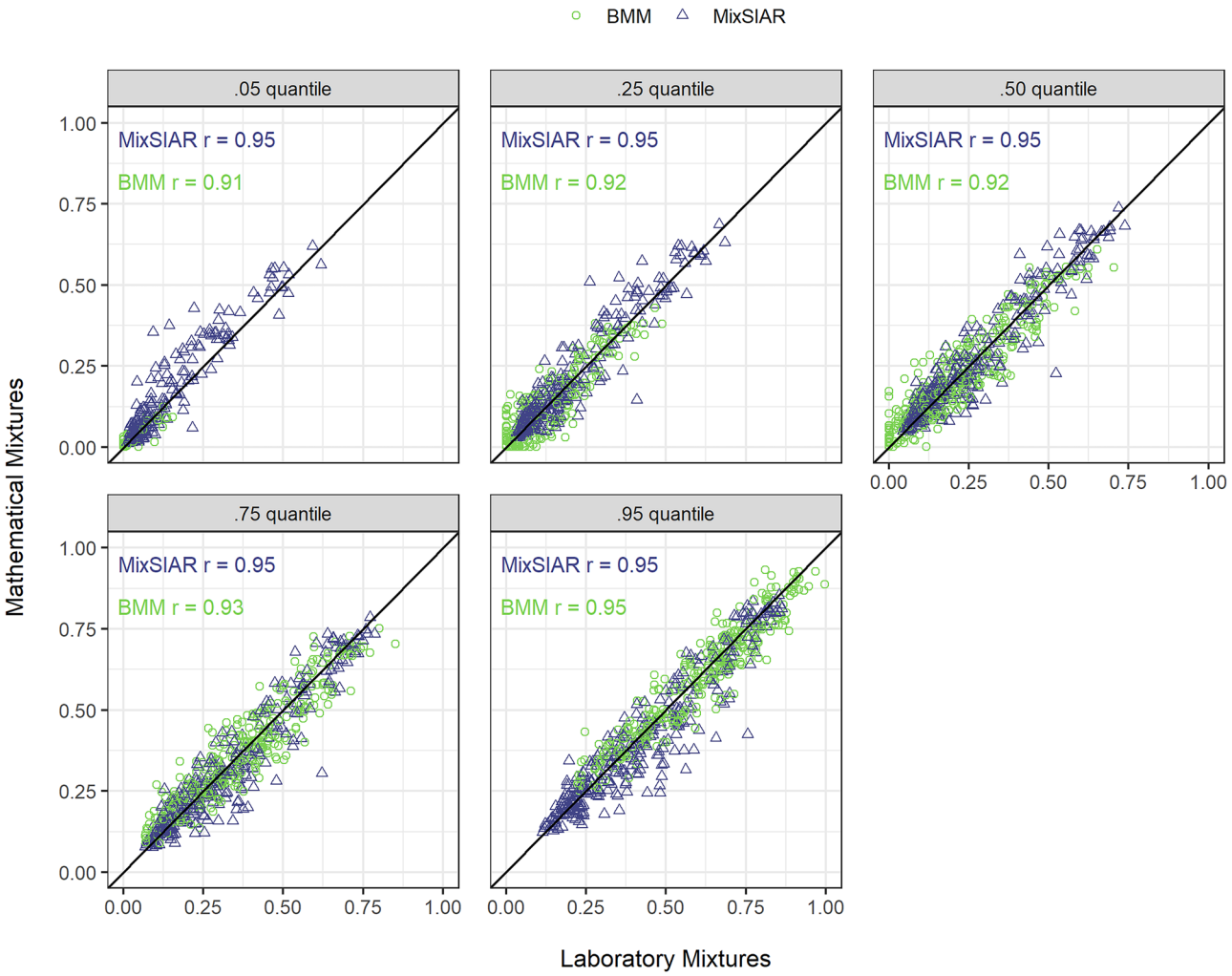

Fig. 3 Comparison between model solutions for the laboratory and mathematical mixtures. The 1:1 line represents a perfect fit. Legend: r = Pearson's correlation coefficient

Table 2 Model evaluation metrics grouped by mixture type and model

Evaluation criteria	Parameter	Mixture type			
		Laboratory		Mathematical	
		Model			
		MixSIAR	BMM	MixSIAR	BMM
Uncertainty	P50*	0.26	0.67	0.30	0.68
	P95*	0.47	0.94	0.49	0.93
	W50*	0.14	0.29	0.12	0.29
	W95*	0.32	0.60	0.29	0.60
Residuals	MAE50	0.07	0.04	0.06	0.04
	MAE95	0.03	0.00	0.02	0.00
	ME50	−0.01	−0.03	−0.01	−0.03
	ME95	0.00	0.00	0.00	0.00
Performance	CRPS*	0.10	0.10	0.09	0.10
	NSE50	0.86	0.92	0.89	0.91
	NSE95	0.96	1.00	0.97	1.00
Contingency	CSI50	0.72	0.73	0.88	0.78
	CSI95	0.58	0.29	0.72	0.30
	HR50	0.81	0.91	0.91	0.91
	HR95	0.98	1.00	0.98	1.00

Mean values per model and mixture type

distribution interval were slightly lower for the BMM model (~0.02), again due to the wider uncertainty bands.

With respect to performance criteria, the BMM model had marginally higher NSE values (between 0.03 and 0.06) for the analyzed intervals (Table 2). The CRPS score, which considers both the precision and accuracy of the model simulations, indicated a similar performance for BMM and MixSIAR. The accuracy and precision of model estimates decreased with increasing source proportions in the mixtures, for both MixSIAR and BMM, as demonstrated by the CRPS curves in Fig. 5. Although residuals were also large when actual source proportions were low (< 0.05), errors and uncertainty increased sharply when known contributions were above 0.75 in the mixtures. In particular, MixSIAR source apportionments were less accurate/precise for lower source proportions, whereas the BMM model displayed higher errors and uncertainty with increasing source contributions.

Regarding the contingency criteria, the CSI metric showed a stronger performance for MixSIAR, specifically for the 95% interval (~100% increase). The BMM displayed a higher hit rate for the 50% interval (~10% increase), and both models encompassed the main sources when the 95% interval was considered (0.98 and 1.0 hit rates for MixSIAR and the BMM, respectively).

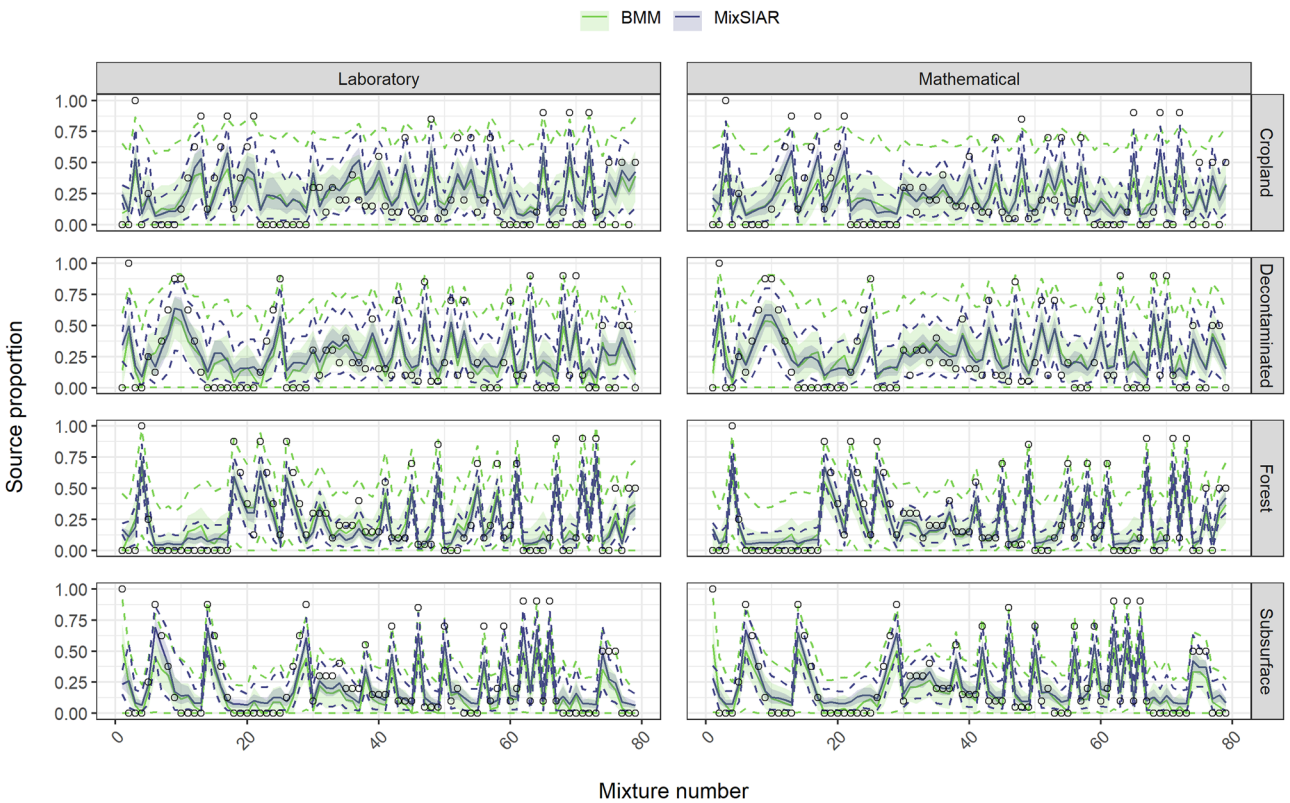

Fig. 4 Summary of the mixing model results by source, mixture type, and mixture number. Solid lines are group medians, shaded areas (ribbons) are the IQR, and dashed lines are the 0.025 and 0.975 quantiles. Circles represent actual mixture proportions

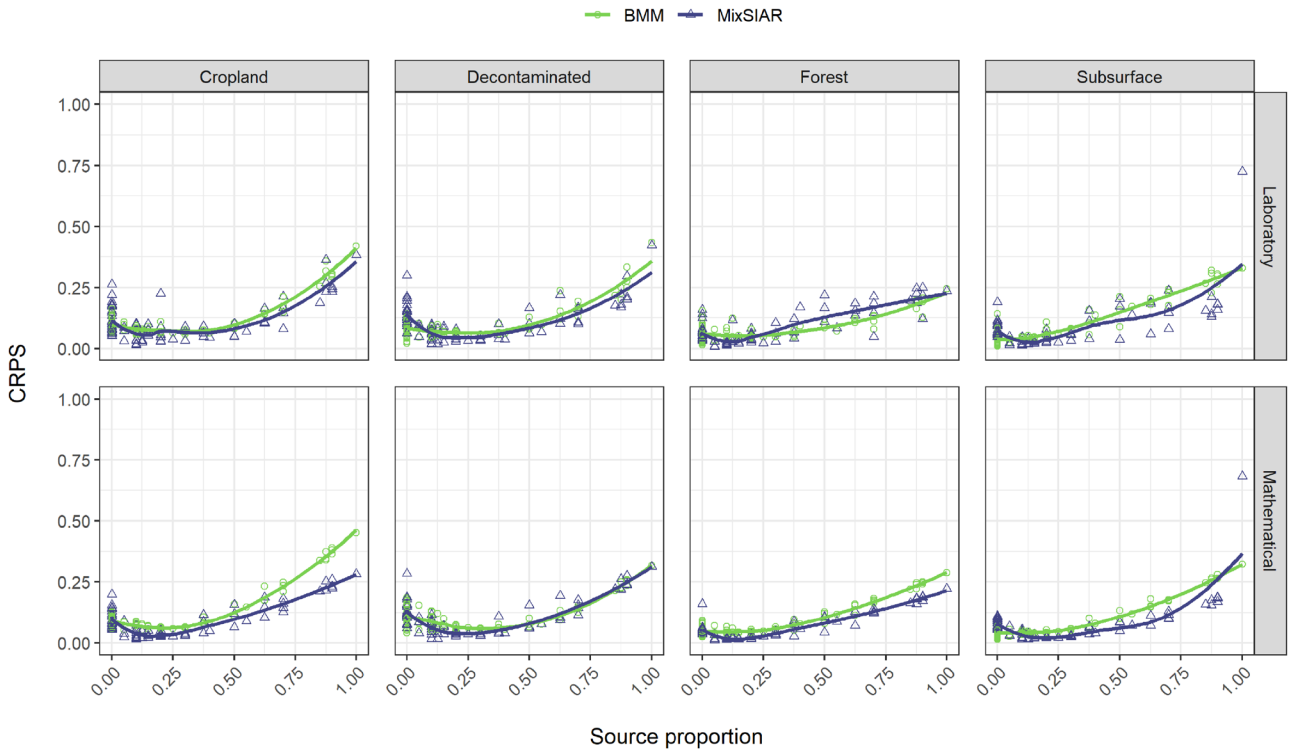

Fig. 5 Relation between artificial mixture source proportions and the CRPS for the different sources, models, and mixture types

Considering each source separately, the model evaluation revealed a higher capability from both MixSIAR and the BMM for estimating the contributions from forests and subsurface material than decontaminated soil and cropland (Fig. 4). This is highlighted by the lower MAE and CRPS values for the forest and subsurface sources (Supplementary Material Table 1). While most evaluation metrics displayed small differences between mixture types when grouped by model, some contrast could be observed when sources were analyzed separely. This is particularly evident when comparing the hit rates of the laboratory and virtual mixtures, for both BMM and MixSIAR (Supplementary Material Table 1). For instance, the HR50 for the subsurface source increased from 0.69 in the laboratory mixtures to 0.88 in the virtual mixtures, for both modeling approaches. Overall, models, mixture types, and evaluation metrics illustrated a lower discrimination power between the decontaminated soils and cropland as sediment sources.

4 Discussion

4.1 Laboratory or virtual mixtures?

The comparison of 79 laboratory and virtual sediment mixtures found limited differences between them, with respect to the ability of the two modeling frameworks to unmix their known source proportions. This indicates that the analytical methods did not introduce significant uncertainty into the model solutions. Moreover, it demonstrates how virtual mixtures are potentially as useful for model testing as laboratory mixtures. Nonetheless, the comparison of virtual to laboratory mixtures does hold value in highlighting situations where analytical methods may introduce errors into the modeling process. Laboratory mixtures might also be required in proof-of-concept experiments which propose novel tracing approaches (see Lake et al. 2022).

Multiple studies have demonstrated how tracer selection procedures can affect fingerprinting source apportionments (Laceby and Olley 2015; Palazón et al. 2015; Smith et al. 2018; Gaspar et al. 2019), and how typical approaches to quantify the quality of source discrimination (e.g., DFA reclassification accuracy) are not always reliable proxies for model accuracy (Smith et al. 2018; Batista et al. 2019; Latorre et al. 2021). Hence, to avoid the use of non-optimal tracer suites and to increase the reliability of sediment fingerprinting approaches, we recommend that researchers and analysts always use multiple virtual mixtures for model testing, and potentially for tracer selection (see Latorre et al. 2021). Although this will not guarantee precise and accurate model results when real-world sediment data is unmixed, it will at least constrain the situations in which data and/or models are non-behavioral, and therefore should not be used for decision-making. As there is no cost involved in the

generation of virtual mixtures, there is no reason not to use them to assess the performance of sediment fingerprinting approaches.

4.2 Model comparison

Much research has been focused on the frequentist (a term which is used loosely to designate anything from deterministic model optimizations to different Monte Carlo and bootstrapping approaches) versus Bayesian debate when it comes to fingerprinting mixing models (see Davies et al. 2018 for a review). Since both frameworks essentially solve similar linear equations, the differences between them are associated to how error structures, covariates, and prior information are incorporated (Stock and Semmens 2016; Cooper and Krueger 2017), and ultimately how uncertainty is represented. Moreover, while the bootstrapping approach optimizes the mixing equation for each model draw, Bayesian models explore all distribution parameters simultaneously (Cooper and Krueger 2017).

For our dataset, the main differences between BMM and MixSIAR outputs were the width and the shape of the distributions of model solutions. While MixSIAR outputs were usually sharp and bell-shaped, BMM model solutions were wide, skewed, or even bimodal (Fig. 6). A similar pattern was reported by Cooper et al. (2014) when comparing Bayesian and bootstrapping fingerprinting approaches. This difference in the distributions of model outputs was reflected in their ability to encompass the known mixture proportions within a given quantile interval, and consequently in some of our evaluation metrics. However, the median values and the general pattern of source apportionments was similar between the tested approaches (Figs. 4 and 6).

Overall, the comparison between the BMM and MixSIAR illustrates a trade-off between accuracy and precision, which was interestingly canceled out in the CRPS values. The score was identical between models for the laboratory mixtures with only a minimal difference for virtual mixtures (i.e., 0.01). Importantly, the ability of the BMM to encompass a larger number of known source contributions seemed to be related to the skewed distributions of the model solutions (Fig. 6). That is, the lower quantiles of the outputs were often zero, which was not the case for MixSIAR. As many mixtures contained zero contributions from a given source, this favored some of the evaluation metrics for the BMM (perhaps spuriously). In any case, the low interval accuracy

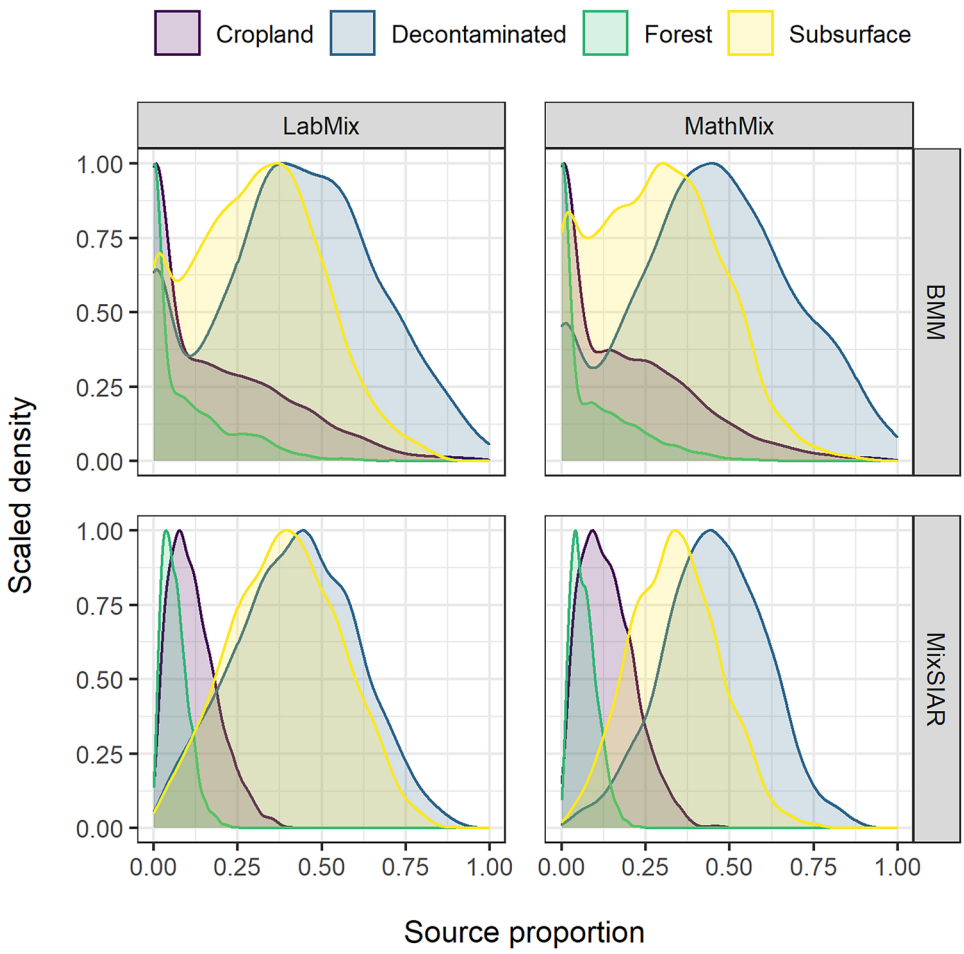

Fig. 6 Probability density functions of model solutions for mixture ten (decontaminated: 0.625, forest: 0.0, cropland: 0.0, subsurface: 0.375)

(P50, P95) observed for MixSIAR indicated a certain degree of overconfidence in the source apportionments, and potentially overly optimistic uncertainty bands. However, it should be highlighted that MixSIAR outputs are sensitive to choices in the error structure and the number of mixture samples (Smith et al. 2018). Hence, we recommend for different structures to be tested in the future.

Ultimately, both mixing model frameworks provided converging solutions, which mutually corroborates their ability to provide reliable source apportionments for this dataset. We speculate there will not be a "best" model framework for all sediment fingerprinting purposes. Instead, researchers and analysts should decide on which approach to take on a case-by-case basis and substantiate their choices with data.

4.3 Beyond mean errors

As we have demonstrated, many evaluation metrics can be used to assess the uncertainty and performance of sediment fingerprinting source apportionments. All scores operate under the explicit knowledge that mixing model outputs consists of distributions of possible source contributions and should be tested as such.

For sediment management, we found the contingency metrics particularly useful as they quantify the ability of the models to correctly identify major sediment sources in a mixture. The region affected by fallout from the FDNPP accident is an example where identifying main sediment sources is of particular interest to managers. Importantly, the CSI demonstrated how the lower uncertainty in MixSIAR apportionments minimized the confusion in the identification of main sources. On the other hand, the wide uncertainty bands from the BMM apportionments led to a higher hit rate, which might be desirable for identifying rare events (e.g., when a single source has a disproportional dominance).

For more general model testing purposes, and specifically for scientists, the CRPS provided a useful metric for assessing both the accuracy and precision of the mixing model source apportionment. The CRPS should be particularly useful for comparing models or tracer selection approaches. In addition, the NSE helped to identify whether the sediment fingerprinting approach was at all useful: if the value had been negative, one would be better off guessing that all sources contribute the same to the mixture. Such situations may arise when a model severely misclassifies the contributions from a given source.

Researchers should choose the best-suited evaluation metrics for their sediment fingerprinting models based on the purpose of their application. We find it unlikely that a single score will be sufficient for characterizing model performance. However, we recommend that researchers should always provide figures and evaluation metrics that characterize not only the model accuracy, but also their uncertainty. The current generation of mixing models provides distributions of source apportionments, and this invaluable information needs to be reported more regularly.

4.4 How good is good enough?

A common difficulty in evaluating environmental models is defining what is good enough, or what are acceptable limits of model error (Beven 2009). A potential solution involves quantifying the error in the testing data, and then assuming that models cannot be expected to be more accurate than our ability to measure a system response. For sediment fingerprinting models being tested against artificial mixtures, it would be possible to define such limits based on the quantification of analytical errors. The NSE can also provide a clear threshold for model failure, as mentioned above.

Another option is to define clear purposes for the sediment fingerprinting application, and then scrutinizing the model solutions for flaws, which would impede it from fulfilling the stated objectives. This would not require fixed limits of acceptability, but rather an exploration of the potential errors and uncertainties in the approach. If the model results will be used for sediment management programs, the definition of what is tolerable error can be discussed with end users. This might include evaluating the capacity of the sediment fingerprinting approach to identify large contributions from a single source, or just highlighting the major sources.

4.5 Model evaluation, limitations, and future perspectives

Ultimately, models are rarely tested to their core: what we evaluate is a combination of data, models, auxiliary hypotheses, and potentially subjective choices made by the modeler. Within our dataset and assumptions, we identified that both the BMM and MixSIAR approaches could not provide accurate and precise apportionments in mixtures for which relative contributions from a single source were higher than 0.75. In addition, BMM apportionments were in some cases too uncertain to provide meaningful insight, while MixSIAR outputs more often failed to encompass known source values due to what seems to be unwarranted certainty. Both frameworks were, however, able to represent the overall trends in the mixtures and to identify the major sources in them—although to a lesser extent for croplands and decontaminated soils. For such purposes, we would not reject using either of the approaches with real sediment data from the Mano and Niida catchments.

Importantly, acceptable results from sediment fingerprinting mixing models tested against laboratory or virtual mixtures will not always translate into accurate catchment

source apportionments. The mixtures represent a best-case scenario, in which all sources are known, samples are representative of the source material, particle size selectivity is negligible, and all tracers are within source range. Open systems are much more complicated, and testing fingerprinting apportionments requires an investigative approach, in which multiple sources of data are used to evaluate the consistency, coherence, and consilience of models as hypotheses of the system (see Baker 2017). This type of "soft" evaluation might entail incorporating processual understanding of source signal development into tracer selection approaches (Koiter et al. 2013), analyzing fingerprinting apportionments against any kind of field observations, and comparing unmixed source contributions to the results from erosion and sediment transport models (Laceby 2012; Wilkinson et al. 2013). The latter option seems attractive; however, as it essentially relies on a model comparison, it can at best provide mutually corroborative evidence. When results are contrasting, it might be difficult to identify which model is more wrong (Batista et al. 2021). When the results support one another, it provides multiple lines of evidence that models are able to simulate the behavior of the system.

5 Conclusions

Here, we presented a comparison between laboratory and virtual mixtures used for evaluating sediment fingerprinting source apportionments. In addition, we compared the bootstrapping BMM model with the MixSIAR Bayesian framework, while providing guidelines, scores, and metrics to improve model evaluation and the communication of the uncertainty in model outputs. Our results demonstrate that virtual mixtures can be as useful as laboratory mixtures for model testing, at least when analytical errors are negligible. As a large number of virtual mixtures can be easily created at no cost, we recommend that sediment fingerprinting studies strive to provide a model testing section in their results. We hope our data and code can incentivize others to include virtual mixtures in their work, and that model evaluation will become standard practice in sediment fingerprinting research.

As most models currently employed use stochastic frameworks, their modeled source apportionments should be tested as distributions, instead of point-based estimates. Among the different scores and metrics employed for model evaluation in the current research, we found the CRPS particularly useful for scientists who wish to compare modeling approaches or tracer selection methods, as this score quantifies both the accuracy and precision of the source apportionments. The modified NSE should also be useful for defining limits of acceptability of model error, and the contingency metrics might be interesting for managers who want to know if the fingerprinting approach can, at the very least, identify the major source in a catchment.

Finally, our model comparison illustrated a trade-off between accuracy and precision, with the BMM outputs being often too uncertain to provide robust source estimates, and MixSIAR frequently failing to encompass the known source values. Moreover, model performance decreased with increasing source proportions. Both models were, however, able to identify major sources and to represent the general trends in the data. We therefore understand that a bootstrapped model like the BMM might be useful for scanning the uncertainties in the data. On the other hand, MixSIAR presents several advantages, such as the ability to incorporate covariates, prior information, concentration dependency, and different residual structures. Modelers can thus, on a case-by-case basis, decide which framework is more appropriate for their purposes and corroborate their model selection with data and/ or research context.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11368-022-03157-4.

Acknowledgements The authors are grateful to Roxanne Durand, who prepared the experimental mixtures during her MSc traineeship.

Author contribution All of the authors contributed to the conceptualization of the study. PVGB and JPL both drafted the manuscript with editorial contributions from OE. JPL and PVGB wrote the model code and PVGB performed the modeling. OE and JPL sampled source soils, and OE supervised the laboratory analyses.

Funding Open access funding provided by University of Basel. This research was funded by the AMORAD project supported by the French National Research Agency (Agence Nationale de la Recherche [ANR], Programme des Investissements d'Avenir; grant no. ANR-11-RSNR-0002). The support of the Centre National de la Recherche Scientifique (CNRS, France) is also gratefully acknowledged (CNRS International Research Project – IRP –MITATE Lab).

Data and code availability All raw data and code for running models, and summarizing and plotting results, are presented in the supplementary material.

Declarations

Competing interests JPL and OE are subject editors for JSSS. PVGB has no competing interests to declare.

References

Baker VR (2017) Debates - hypothesis testing in hydrology: pursuing certainty versus pursuing uberty. Water Resour Res 53:1770–1778. https://doi.org/10.1002/2016WR020078.Received

Batista PV, Laceby JP, Silva MLN et al (2019) Using pedological knowledge to improve sediment source apportionment in tropical environments. J Soils Sediments 19:3274–3289. https://doi.org/10.1007/s11368-018-2199-5

Batista PVG, Laceby JP, Davies J et al (2021) A framework for testing large-scale distributed soil erosion and sediment delivery models : dealing with uncertainty in models and the observational data. Environ Model Softw 137. https://doi.org/10.1016/j.envsoft.2021.104961

Bennett ND, Croke BFW, Guariso G et al (2013) Characterising performance of environmental models. Environ Model Softw 40:1–20. https://doi.org/10.1016/j.envsoft.2012.09.011

Beven KJ (2009) Environmental modelling: an uncertain future. Routledge, Oxon

Boucher MA, Perreault L, Anctil F (2009) Tools for the assessment of hydrological ensemble forecasts obtained by neural networks. J Hydroinformatics 11:297–307. https://doi.org/10.2166/hydro.2009.037

Collins AL, Blackwell M, Boeckx P et al (2020) Sediment source fingerprinting: benchmarking recent outputs, remaining challenges and emerging themes. J Soils Sediments 20:4160–4193. https://doi.org/10.1007/s11368-020-02755-4

Collins AL, Pulley S, Foster IDL et al (2017) Sediment source fingerprinting as an aid to catchment management: a review of the current state of knowledge and a methodological decision-tree for end-users. J Environ Manage 194:86–108. https://doi.org/10.1016/j.jenvman.2016.09.075

Commission Internationale de l'Eclairage (CIE) (1931) CIE Proceedings. Cambridge University Press, Cambridge

Cooper RJ, Krueger T (2017) An extended Bayesian sediment fingerprinting mixing model for the full Bayes treatment of geochemical uncertainties. Hydrol Process 31:1900–1912. https://doi.org/10.1002/hyp.11154

Cooper RJ, Krueger T, Hiscock KM, Rawlins BG (2014) Sensitivity of fluvial sediment source apportionment to mixing model assumptions: a Bayesian model comparison. Water Resour Res 50:9031–9047. https://doi.org/10.1002/2014WR016194

Davies J, Olley J, Hawker D, Mcbroom J (2018) Application of the Bayesian approach to sediment fingerprinting and source attribution. Hydrol Process 3978–3995. https://doi.org/10.1002/hyp.13306

Evrard O, Durand R, Nakao A et al (2020) Comptes Rendus Géoscience. Comptes Rendus Géoscience—Sciences la Planète 352:199–211

Evrard O, Laceby JP, Nakao A (2019) Effectiveness of landscape decontamination following the fukushima nuclear accident: a review. Soil 5:333–350. https://doi.org/10.5194/soil-5-333-2019

Gaspar L, Blake WH, Smith HG et al (2019) Testing the sensitivity of a multivariate mixing model using geochemical fingerprints with artificial mixtures. Geoderma 337:498–510. https://doi.org/10.1016/j.geoderma.2018.10.005

Gibbs M, Leduc D, Nodder SD et al (2020) Novel application of a compound-specific stable isotope (CSSI) tracking technique demonstrates connectivity between rerrestrial and deep-sea ecosystems via submarine canyons. Front Mar Sci 7:608. https://doi.org/10.3389/fmars.2020.00608

Haddadchi A, Olley J, Laceby JP (2014) Accuracy of mixing models in predicting sediment source contributions. Sci Total Environ 497–498:139–152. https://doi.org/10.1016/j.scitotenv.2014.07.105

Jalowska AM, Laceby JP, Rodriguez AB (2017) Tracing the sources, fate, and recycling of fine sediments across a river-delta interface. CATENA 154:95–106. https://doi.org/10.1016/j.catena.2017.02.016

Jordan A, Krüger F, Lerch S (2019) Evaluating probabilistic forecasts with scoring rules. J Stat Softw 90:1–37. https://doi.org/10.18637/jss.v090.i12

Koiter AJ, Owens PN, Petticrew EL, Lobb DA (2013) The behavioural characteristics of sediment properties and their implications for sediment fingerprinting as an approach for identifying sediment sources in river basins. Earth-Science Rev 125:24–42. https://doi.org/10.1016/j.earscirev.2013.05.009

Laceby JP (2012) The provenance of sediment in three rural catchments in South East Queensland. Griffith University, Australia

Laceby JP, Batista PVG, Taube N et al (2021) Tracing total and dissolved material in a western Canadian basin using quality control samples to guide the selection of fingerprinting parameters for modelling. CATENA 200:105095. https://doi.org/10.1016/j.catena.2020.105095

Laceby JP, McMahon J, Evrard O, Olley J (2015) A comparison of geological and statistical approaches to element selection for sediment fingerprinting. J Soils Sediments 15:2117–2131. https://doi.org/10.1007/s11368-015-1111-9

Laceby JP, Olley J (2015) An examination of geochemical modelling approaches to tracing sediment sources incorporating distribution mixing and elemental correlations. Hydrol Process 29:1669–1685. https://doi.org/10.1002/hyp.10287

Laio F, Tamea S (2007) Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrol Earth Syst Sci 11:1267–1277. https://doi.org/10.5194/hess-11-1267-2007

Lake NF, Martínez-Carreras N, Shaw PJ, Collins AL (2022) High frequency un-mixing of soil samples using a submerged spectrophotometer in a laboratory setting—implications for sediment fingerprinting. J Soils Sediments 22:348–364. https://doi.org/10.1007/s11368-021-03107-6

Latorre B, Lizaga I, Gaspar L, Navas A (2021) A novel method for analysing consistency and unravelling multiple solutions in sediment fingerprinting. Sci Total Environ 789:147804. https://doi.org/10.1016/j.scitotenv.2021.147804

Lavrieux M, Birkholz A, Meusburger K et al (2019) Plants or bacteria? 130 years of mixed imprints in Lake Baldegg sediments (Switzerland), as revealed by compound-specific isotope analysis (CSIA) and biomarker analysis. Biogeosciences 16:2131–2146. https://doi.org/10.5194/bg-16-2131-2019

Lepage H, Evrard O, Onda Y et al (2015) Depth distribution of cesium-137 in paddy fields across the Fukushima pollution plume in 2013. J Environ Radioact 147:157–164. https://doi.org/10.1016/j.jenvrad.2015.05.003

Martínez-Carreras N, Udelhoven T, Krein A et al (2010) The use of sediment colour measured by diffuse reflectance spectrometry to determine sediment sources: application to the Attert River catchment (Luxembourg). J Hydrol 382:49–63. https://doi.org/10.1016/j.jhydrol.2009.12.017

Matheson JE, Winlker RL (1976) Scoring rules for continuous probability distributions. Manag Sci 22:1087–1096. https://doi.org/10.1287/mnsc.22.10.1087

Nash E, Sutcliffe V (1970) River flow forecasting through conceptual models Part I - a discussion of principles. J Hydrol 10:282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nosrati K, Govers G, Semmens BX, Ward EJ (2014) A mixing model to incorporate uncertainty in sediment fingerprinting. Geoderma 217–218:173–180. https://doi.org/10.1016/j.geoderma.2013.12.002

Owens PN (2020) Soil erosion and sediment dynamics in the Anthropocene: a review of human impacts during a period of rapid global environmental change. J Soils Sediments 20:4115–4143. https://doi.org/10.1007/s11368-020-02815-9

Owens PN, Blake WH, Gaspar L et al (2016) Fingerprinting and tracing the sources of soils and sediments: Earth and ocean science, geoarchaeological, forensic, and human health applications. Earth-Science Rev 162:1–23. https://doi.org/10.1016/j.earscirev.2016.08.012

Palazón L, Latorre B, Gaspar L et al (2015) Comparing catchment sediment fingerprinting procedures using an auto-evaluation approach with virtual sample mixtures. Sci Total Environ 532:456–466. https://doi.org/10.1016/j.scitotenv.2015.05.003

Peart MR, Walling DE (1986) Fingerprinting sediment source: the example of a drainage basin in Devon, UK. Drainage basin sediment delivery: proceedings of a symposium held in Albuquerque, NM., 4–8 August 1986

Pulley S, Collins AL, Laceby JP (2020) The representation of sediment source group tracer distributions in Monte Carlo uncertainty routines for fingerprinting: an analysis of accuracy and precision using data for four contrasting catchments. Hydrol Process 34:2381–2400. https://doi.org/10.1002/hyp.13736

Pulley S, Foster I, Collins AL (2017) The impact of catchment source group classification on the accuracy of sediment fingerprinting outputs. J Environ Manage 194:16–26. https://doi.org/10.1016/j.jenvman.2016.04.048

R Core Team (2021) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sherriff SC, Franks SW, Rowan JS et al (2015) Uncertainty-based assessment of tracer selection, tracer non-conservativeness and multiple solutions in sediment fingerprinting using synthetic and field data. J Soils Sediments 15:2101–2116. https://doi.org/10.1007/s11368-015-1123-5

Shi Z, Blake WH, Wen A et al (2021) Channel erosion dominates sediment sources in an agricultural catchment in the Upper Yangtze basin of China: evidence from geochemical fingerprints. CATENA 199:105111. https://doi.org/10.1016/j.catena.2020.105111

Smith HG, Karam DS, Lennard AT (2018) Evaluating tracer selection for catchment sediment fingerprinting. J Soils Sediments 18:3005–3019. https://doi.org/10.1007/s11368-018-1990-7

Stock BC, Jackson AL, Ward EJ et al (2018) Analyzing mixing systems using a new generation of Bayesian tracer mixing models. PeerJ e5096. https://doi.org/10.7717/peerj.5096

Stock BC, Semmens BX (2016) Unifying error structures in commonly used biotracer mixing models. Ecology 97:2562–2569. https://doi.org/10.1002/ecy.1517

Walden J, Slattery MC, Burt TP (1997) Use of mineral magnetic measurements to fingerprint suspended sediment sources: approaches and techniques for data analysis. J Hydrol 202:353–372. https://doi.org/10.1016/S0022-1694(97)00078-4

Walling DE (2005) Tracing suspended sediment sources in catchments and river systems. Sci Total Environ 344:159–184. https://doi.org/10.1016/j.scitotenv.2005.02.011

Wilkinson SN, Hancock GJ, Bartley R et al (2013) Using sediment tracing to assess processes and spatial patterns of erosion in grazed rangelands, Burdekin River basin, Australia. Agric Ecosyst Environ 180:90–102. https://doi.org/10.1016/j.agee.2012.02.002

Yu L, Oldfield F (1989) A multivariate mixing model for identifying sediment source from magnetic measurements. Quat Res 32:168–181