Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-250-AC1, 2020

Hydrology and Earth System Sciences
Open Access
Discussions

Interactive comment on "Technical Note: A data-driven method for estimating the composition of end-members from streamwater chemistry observations" *by* Esther Xu Fei and Ciaran Joseph Harman

Esther Xu Fei and Ciaran Joseph Harman

charman1@jhu.edu

Received and published: 11 August 2020

Thank you for your encouraging comment. We acknowledge that dataset may largely influence the result of identified end-members. At the current stage, we used a well-studied set of tracers from Hooper et al. (1990) as an example to demonstrate the capability of CHEMMA.

References

Hooper, R. P., Christophersen, N., & Peters, N. E. (1990). Modelling streamwater

chemistry as a mixture of soilwater end-membersâĂŤAn application to the Panola Mountain catchment, Georgia, USA. Journal of Hydrology, 116(1-4), 321-343.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-250, 2020.

<u>Response to Referee #1</u>

Xu Fei and Harman present a method to extend classical end-member mixing analysis by deriving end-members purely based on stream concentrations. By repeatedly delineating the convex-hull around stream concentrations and then classifying results using a k-means clustering approach the method includes an uncertainty assessment of the resulting end-members. The method is successfully applied to the Panola Mountain Research Watershed data.

I consider the outlined method a valuable addition to the end-member mixing literature and recommend publication in HESS after revisions are made that are outlined below. I must admit I am not a math expert and have not checked the given formulae.

Thank you for your concise summary and encouraging comment. We aim to provide a method that is both well-established in mathematical foundations and easy to understand for the general hydrology community. We have adopted some of your comments to improve the readability, particularly in the method section.

Current application of end-member mixing approaches involves the disentangling of stream water concentrations based on pre-conceived 'end-members', or origins of stream water. These end- members are sampled, and hopefully span the spread of stream water concentrations, thus enabling the calculation of flow route proportions. Xu Fei and Harman propose a method that calculates possible end-member concentrations from stream water alone. While this can indeed be very useful in practice, it also defies the purpose of an end-member mixing analysis to some extent: as a hydrologist, we are interested in where the water comes from that makes up the stream, not in its concentrations per se. I am sure the authors agree here, but I miss more discussion in the manuscript on the practical use of the proposed method in a hydrological analysis. Given the end-goal of finding water sources instead of end-member concentrations, where does their method come into play? As a first step, defining possible concentration profiles of end-members, after which you take to the field to 'find' these end-members? Or as a check of a more classical end-member mixing application: are end-members missed? Or vice-versa, as the proposed approach cannot handle end-members that are located outside the convex-hull of stream water concentrations.

We appreciate your suggestion in discussing practical applications of CHEMMA. In the revision, we also adopted the comment from Referee #2 and added a paragraph after l200:

"*For most hydrologists, end-member analysis is used to identify the water sources, and toward that purpose CHEMMA may be useful in the following ways: 1. CHEMMA may be used to reduce subjectivity when selecting from field-measured end-member candidates by comparing them to CHEMMA end-members; 2. CHEMMA may identify end-members that have not been sampled in the field, which may serve as a check for missing sources; 3. CHEMMA end-member compositions may help hydrologists ask better questions and provide guidance for field sampling by suggesting source characteristics; 4. CHEMMA can be used in conjunction with the Diagnostic Tool of Mixing Models (DTMM, developed by Hooper (2003)). DTMM is used to assess the tracer conservation, and mixture rank. CHEMMA can be enhanced by using DTMM analysis to select conserved tracers for analysis. The robustness of CHEMMA end-members also serve as a check for DTMM-determined rank of mixture.*"

Further, I would encourage the authors to provide more discussion of the uncertainty calculation of their method (how certain is the calculation of the convex hull), and how it relates to uncertainty in end-member mixing applications (how clearly defined is a single end-member, how time-variant is its concentration). On l197 you allude to some work you did on this, but this did not make it to the manuscript?

Overall, the manuscript would benefit from a thorough spelling and grammar check.

Thank you for your suggestion. The uncertainty analysis within this technical note is limited to discuss the general intrinsic uncertainty introduced by the CH-NMF algorithm. However, much more work is required to dissect the total uncertainty arising from factors like the time-variability of the end-members and the algorithm's response to data uncertainty. We believe this is beyond the scope of a Technical Note introducing the approach. Therefore, we have decided to leave detailed analysis regarding uncertainty in practical applications for future work.

Minor comments:

l35: and each of which accounts. What which? You are referring to the columns of Y? Why not just speak of the principal components?

Thank you for pointing out this. "which" refers to Y_{obs}, which is the observation matrix projected onto the principal component space. The principal components are the rows of the projection matrix P.

l45: refer to diagnostic tools of (Hooper, 2003), as they propose a more formal analysis to assess the rank of the data. Hooper also finds evidence of a fourth end-member by the way… see around l173

Thank you for highlighting this. Hooper did not explicitly identify the composition of the fourth end-member (Hooper, 2003). We have edited the manuscript to reflect his speculation on its existence by adding the following sentence at l172:

"Hooper (2003) also suggested the existence of a fourth end-member."

l46: After thus subjectively determining the number of… Not sure what the purpose of this entire sentence is by the way, seems like it could be skipped altogether.

Here we intended to differentiate the rank of mixture idea from EMMA with the mathematically defined rank *d* (refer to comment l91) from CHEMMA.

l50: also: spatial and temporal variability in end-member concentrations

Thank you for the suggestion. We have added a fourth point to address this comment:

"4) uncertainties introduced by spatial and temporal variability in end-member concentrations cause extra difficulties."

l86: Only the top k PCs are retained? So while you criticize this subjectivity in lines 45-47, your method necessitates the same step? Please elaborate on this.

Thank you for identifying this confusion. We have added a sentence in the manuscript to acknowledge that the subjective selection of the number of end-members is not completely resolved by the proposed method:

"*The CHEMMA algorithm does not entirely avoid this subjective choice of the number of end-members retained, and so does not resolve this criticism of EMMA.*"

l87: Is there a reason why the analysis must be performed successively in 2D, and not in ND?

Thank you for noticing this technical detail. The search of finding the convex-hull is done in *all* 2D PC projections, not necessarily successive and limited to k retained PCs. The 2D search is a way to subsampling the vertices set and it gives the most "greedy" search result because one needs at least 2D to draw lines/planes (or other linear structures in higher dimension) and to find the linear intersections (candidate end-members) that bound the observation. In addition, searching in 2D subspaces not only preserves detailed structure of the convex hull for further simplification, but also is efficient in terms of computation cost (Thural et al., 2011). We currently have not added a detailed explanation of this into the manuscript but the issue is addressed in the literature cited (de Berg et al, 2000) .

l91: algorithm: also define k and d. And: how are they different?

Thank you. d is the rank of the observation covariance matrix, as defined in Step 2 of Algorithm 1, whereas k is the number of eigenvectors to retain (which is stated in l86 as last comment shown). We have added "where" before the definition of d to make it stand out more. We have also changed sentence a) at l93 to *"CH-NMF decomposes the correlation matrix of the observations to obtain at most d PCs (d is the maximum number of linearly uncorrelated variables)"*.

l123: This is quite a central drawback of the method. Can some directions/ideas already be given? Maybe a hybrid between chemma and field sampling for em signatures? Using the time-variance might also provide a way forward, searching for periods where certain end-members dominate and are thus better characterized from stream water concentrations.

Thank you for this insightful comment. It is certainly true that any method that relies exclusively on the stream water samples will face this challenge. We have shortly discussed this drawback both in the new added paragraph in l200 (in response to one of the above comments) and in the discussion part of the manuscript (l202 - l205). We have added a couple of sentences to address these suggestions in the discussion section.

After modified l199: "*The temporal end-member dominance may further deepen our understanding of stream water characterization.*"

After l206: "*4) using individual field end-member measurements to inform CHEMMA.*"

l144: Elaborate on 'uncertainty' versus spread of local minima of convex hull. are these really the same thing?

Thank you for your suggestion. Total uncertainty includes uncertainty due to the spread of local minima of convex hull as well as uncertainty due to sampling. Here we only examine the uncertainty arises from the algorithm. The complete examination of uncertainty is left for further work.

l149: What is a 'reasonable' variance? Is there a suitable metric? Provide guidance.

Thank you for this valuable comment. We have not identified a suitable metric at the stage. The current "reasonable" variance remains a subjective choice. We acknowledge the subjective here and we have modified the sentence in l202 from "*optimize the model complexity*" to "*eliminate the subjective choice of k*" for clarity.

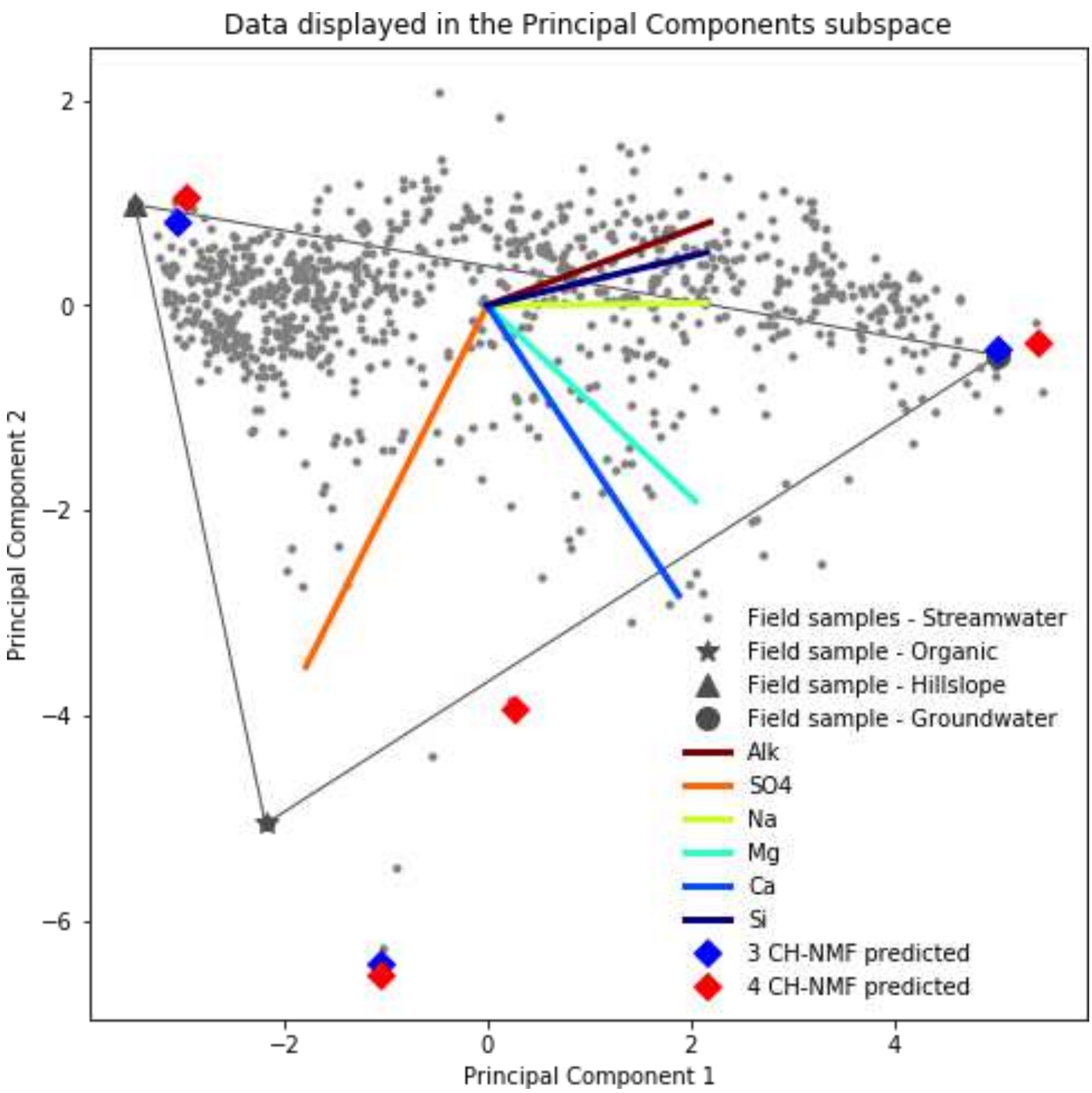
l197: “Fortunately, CHEMMA itself provides a tool for exploring some of these sources of uncertainty. By partitioning the dataset into time periods (or hydrologic state, etc), the temporal variability of end-members could be explored” What is this statement based on?

Thank you. This statement is intended to provide suggestion for ways to explore temporal variability. We have modified the sentence to clarify the intent:

“CHEMMA itself may provide a tool for exploring some of these sources of uncertainty. For example, by partitioning the dataset into time periods (or hydrologic state, etc), the temporal variability of end-members could be explored.”

On the python code:

Consider making your code available as an importable python module through pip and/or conda. Why is Figure 2 different on github code? location 4th endmember? accessed 1/7/2020:



References

Hooper, R. P. (2003). Diagnostic tools for mixing models of stream water chemistry. *Water*

Resources Research, *39*(3), 1–13. https://doi.org/10.1029/2002WR001528

Thank you for your suggestion. We will make a python module in the near future.

In the technical note, Figure 2 is made by using the centroids of the clusters from Figure 3 a and b. The Python code only shows one run of CH-NMF and what you saw is the algorithm find another 4th end-member in the space. By repeatedly run the CH-NMF and using the COP-Kmeans to classify the clusters, this occasional capture of "wrong" convex-hull vertex can be eliminated.

We will add a section in the Jupyter notebook to reproduce Figure 2.

Thank you for your careful reading. We have adopted the following grammatic comments.

l11: streamwater

l30: concentrations: what you mean is the concentration of different solutes are correlated

l33: dividing by the standard deviation. It doesn't require it per se, but yields better results

l34: transforms ~~the~~ from ~~the~~ observation space to ~~the~~ PC space

l55: In spite of EMMA's wide application~~s~~

l81: to ~~the~~ end-member mixing

l89: "convex-hull": why here in quotes? This term has been used throughout the manuscript. Move its definition to its first use.

References

Christophersen, N., & Hooper, R. P. (1992). Multivariate analysis of stream water chemical data: The use of principal components analysis for the end-member mixing problem. Water Resources Research, 28(1), 99-107.

de Berg M, van Kreveld M, Overmars M, Schwarzkopf O (2000) Computational geometry. Springer, Berlin, Heidelberg

Hooper, R. P. (2003). Diagnostic tools for mixing models of stream water chemistry. Water Resources Research, 39(3).

Hooper, R. P., Christophersen, N., & Peters, N. E. (1990). Modelling streamwater chemistry as a mixture of soilwater end-members—An application to the Panola Mountain catchment, Georgia, USA. Journal of Hydrology, 116(1-4), 321-343.

Thurau, C., Kersting, K., Wahabzada, M., Bauckhage, C., 2011. Convex non-negative matrix factorization for massive datasets. Knowl. Inf. Syst. 29, 457–478. https://doi.org/10.1007/s10115-010-0352-6

Referee #2

When I read the manuscript on the first time, I was thrilled, as this is something I have been waiting for many years to come out. If we could determine the chemical composition of end-members using streamflow chemistry alone, end-member mixing analysis would be significantly improved and revived. After I read it for a couple of times, I found the fundamental idea is still intriguing, but the assumptions the main method, namely CHEMMA, is based on may be flawed and cause significant uncertainties on the modeling results. A conceptual set-up of why and how this modeling would work could be strengthened. Readability could be improved as well, particularly in regard to some mathematical details and their connection/implication with/in the hydrologic questions being investigated. Remember that most of readers who are interested in this study are hydrologists not mathematicians.

Thank you for recognizing the value of our work and providing suggestions on improving the quality of this manuscript. We will improve the readability in methodology section to strengthen the conceptual set-up.

Major Comments:

The main approach is to use Convex-Hull Non-negative Matrix Factorization (CH-NMF) to infer possible end-member compositions by searching for a simplex that optimally encloses the stream water observations. The assumption for this is, based on authors, that end-members are located near the most extreme points that bound the observations in "mixing" space. From this assumption, it is clear that a simplex is basically determined by the data structure of observations, in other words, the shape of the sample cloud. What if one or more extreme points are missing in our observations? This could happen if samples are collected sparsely or only on certain hydrologic conditions/seasons that do not contain extreme samples (samples with extreme concentrations for at least one solute). The number of samples could also change how samples are distributed. With the same data set, can similar results (with reasonable uncertainties) be obtained from subsets of samples with varying number of samples that are randomly selected?

Thank you for your insightful comment. This is indeed a drawback, and we have mentioned briefly in the manuscript that CHEMMA can only identify end-members that are well-sampled in the data. We will expand on this point in the revised manuscript by highlighting this issue in both abstract and introduction. We acknowledge that some of the fundamental assumptions could limit the CHEMMA application. Improvements that overcome these limitations are left in future work.

There is a lack of conceptual setup where this study came from and where it goes in relation to existing tools in EMMA, particularly the diagnostic tools of mixing models (DTMM; Hooper, WRR, 2003). In one study, Christopherson and Hooper (WRR, 1992) specifically concluded that “Unambiguous identification of the source solution compositions from the mixture alone is impossible; thus, it is necessary that potential source solutions be derived from independent measurements.” I do not mean this conclusion cannot be challenged, but the rationale must be stated clearly and explicitly, possibly using a conceptual set-up. Also, what is its relation with DTMM? Will the current study be supplemental or a substitute to DTMM in regard to the number of end-members? Can DTMM actually help to enhance CHEMMA and how?

The study used data collected in late 1980s. That is okay but what I am concerned is about the conservativity of all six solutes. How can we be convinced if all six solutes are conservative? If any of those is not conservative, the results of CHEMMA would be different. In my opinion, this is where DTMM may be able to help. Also, isn’t it interesting to compare the number of end-members acquired using CHEMMA to DTMM?

We appreciate your suggestion on improving the understanding of practicing CHEMMA. We agree that the conceptual set-up is not clearly stated in the paper. In the revision, we will modify the last paragraph of the introduction (l55 – l65). We have added a sentence in l65 to clarify the conceptual setup:

"*Christophersen and Hooper (1992) suggested that "[u]nambiguous identification of the source solution compositions from the mixture alone is impossible". In a strict sense this is likely true, since the underlying assumption (streamflow as a conservative mixture of invariant sources) is unlikely to be adhered to in a real watershed. However, we believe there may be utility in developing tools that can seek some insights (perhaps not free of ambiguity) into the potential source solution composition from the observed mixture. We propose CHEMMA as an attempt to push this boundary and to see how far we can get.*"

As we said to Referee #1's comment, we have added a paragraph after l200, and discussed DTMM in point 4:

"For most hydrologists, end-member analysis is used to identify the water sources, and toward that purpose CHEMMA may be useful in the following ways: 1. CHEMMA may be used to reduce subjectivity when selecting from field-measured end-member candidates by comparing them to CHEMMA end-members; 2. CHEMMA may identify end-members that have not been sampled in the field, which may serve as a check for missing sources; 3. CHEMMA end-member compositions may help hydrologists ask better questions and provide guidance for field sampling by suggesting source characteristics; 4. CHEMMA can be used in conjunction with the Diagnostic Tool of Mixing Models (DTMM, developed by Hooper (2003)). DTMM is used to assess the tracer conservation, and mixture rank. CHEMMA can be enhanced by using DTMM analysis to select conserved tracers for analysis. The robustness of CHEMMA end-members also serve as a check for DTMM-determined rank of mixture."

<u>Minor Comments:</u>

L18: Before the first reference, add "e.g.,". Many classical references on EMMA were not actually cited.

Thank you for this suggestion. We have added "e.g.", and also added two new references mentioned in your comment (Liu et al., 2008 and 2017) as applications of EMMA under different climatic settings.

L24: This statement should refer to conservative solutes.

Thank you for bringing up this confusion. We added a word "solute" between "chemical composition" to clarify that the sentence is talking about solute conservation.

L30: "Streamwater concentration are naturally correlated." It is true if you refer to conservative solutes; otherwise it is an ill statement. Use two words "stream water" instead of one word "streamwater". Also, use plural for "concentration".

Thank you for your suggestion. We changed "streamwater" to "stream water" for this manuscript. We also adopted Referee #1's suggestion and changed this sentence to:

"Stream water concentrations of different conservative solutes tend to be correlated."

L28: The second one is no longer a hypothesis or assumption because of the diagnostic tools of mixing models by Hooper (2003); See Liu et al. (WRR, 2008) for a demonstration and how this was addressed.

L45: True traditionally but not after DTMM is developed. See Liu et al. (WRR, 2008, 2017) as examples.

L51-52: Not true with DTMM.

L52-53: True but DTMM can help identify conservative solutes so that users can use only conservative ones. I mention this because I think your study is also based on mixing of conservative solutes. This should be stated/defined earlier in your text.

L186-206: Need to indicate where this modeling will lead to and how it may work together with DTMM.

We would like to response to these five comments collectively. Thank you for your recommendation about DTMM and related application papers. We agree that DTMM workflow is a good complement to both EMMA and CHEMMA. And we added a paragraph to clarify how DTMM and CHEMMA can potentially work together. Please refer to the response to the last major comment above for more details.

L33-35: Multiple issues here. (1) Is Pobs actually eigenvectors? If so, use a parenthesis to annotate so; otherwise explain what it is and how to calculate it. (2) Get rid of the redundant "the". (3) My understanding is that once a standardized data set is used, a correlation matrix is decomposed rather than covariance matrix. Check if this is correct.

Thank you for carefully checking the mathematical details. The rows of P_{obs} are the eigenvectors of the correlation/covariance matrix X_{obs}. We have added a parenthesis segment: *(rows of which are eigenvectors of the correlation matrix)*, and we have deleted the redundant "the" appearing later. Because X_{obs} is standardized observation, the correlation matrix and the covariance matrix are essential the same. Performing eigendecomposition on both matrices yields the same results. We have adopted your comment to change the covariance matrix to correlation matrix to make it clear.

L36: If P are indeed eigenvectors, cite Christopherson and Hooper (1992) for the equation.

Thank you. P are eigenvectors. We have cited Christophersen and Hooper (1992).

Result 2: Eigenvectors and PCs are different. PCs are calculated based on eigenvectors and observed concentrations.

Thank you. We adopted a terminology in this manuscript consistent with usage in applied mathematics literature, such as Jolliffe (2002). In our understanding, eigenvectors are derived from the correlation matrix of the observed concentrations by performing eigendecomposition (as used for this manuscript) or singular value decomposition. Resulting eigenvectors are orthogonal bases as known as Principal Components (PCs) (Jolliffe, 2002). Loadings are the coefficient calculated based on eigenvectors (PCs) and observed concentrations (Jolliffe, 2002), and are referred as contributions (of end-members) in this manuscript.

L93: I still think it is correlation matrix not covariance matrix. Also, what you mean here is eigenvectors not PCs.

Thank you. As we responded before, in the revised manuscript we have changed covariance matrix to correlation matrix for clarity. We used eigenvectors and PCs

interchangeably according to our reference of PCA terminology (Jolliffe, 2002).

Result 3: Is it specified anywhere how to project mathematically?

Thank you for finding this confusing part. Projecting a matrix A to another space through a projection matrix P^T to get projected matrix B is defined as $B = AP^T$, just as Equation 1 and 2 show. We added a parenthesis fragment: *(similar form as Eqn. 1 & 2).*

Result 4: Will the dimension of S differs from one projection plane to another?

Thank you for noticing this technical detail. Yes. S records all boundary points in each projection plane and the number of recorded points at each plane can be different.

Result 5: Is X expression actually [[xem1], [xem2], . . ., [xemk]], as each xemi has a dimension of n by 1?

Thank you for noticing the dimension consistency. Yes, x_{emi} has dimension of n by 1. We have checked the consistency of dimensions in Algorithm 1 a couple of time before submitted the manuscript.

L125: I think “equifinality” is part of your talking here. Why not citing “equifinality” directly? It is a common term that hydrologists are very familiar with.

Thank you for your comment. The concept here is slightly different from equifinality. This paragraph particularly talked about limitation of an optimization problem on minimizing the objective function in Step/Result 5.

Thank you for pointing out style problems. We have adopted the following comments.

L31: Need at least one reference (e.g., Christopherson and Hooper, 1992).

L41-42: Cite Hooper (WRR, 2003).

L60: Need to specify “extreme points”. I think you refer to “extreme points of stream water samples”.

L64: I think you mean “end-members’ composition”.

L94: Spell out PCA as it appears for the first time.

L102: Specify the constraints, each between 0 and 1 with sum of all to be 1.

References

Liu, F., Bales, R. C., Conklin, M. H., & Conrad, M. E. (2008). Streamflow generation from snowmelt in semi-arid, seasonally snow-covered, forested catchments, Valles

Caldera, New Mexico. Water Resources Research, 44(12).

Liu, F., Conklin, M. H., & Shaw, G. D. (2017). Insights into hydrologic and hydrochemical processes based on concentration-discharge and end-member mixing analyses in the mid-Merced River Basin, Sierra Nevada, California. Water Resources Research, 53(1), 832-850.

Jolliffe, I. T. (2002). Mathematical and Statistical Properties of Population Principal Components. In Principal Component Analysis (pp. 8-22). Springer, New York, NY. https://doi.org/10.1007/b98835

Technical Note: A data-driven method for estimating the composition of end-members from ~~streamwater~~ stream water chemistry observations

Esther Xu Fei[1] and Ciaran Joseph Harman[1,2]

[1]Department of Environmental Health and Engineering, Johns Hopkins University

[2]Department of Earth and Planetary Science, Johns Hopkins University

Correspondence: Ciaran Joseph Harman (charman1@jhu.edu)

End-Member Mixing Analysis (EMMA) is a method of interpreting ~~streamwater~~ stream water chemistry variations, and is widely used for chemical hydrograph separation. It is based on the assumption that the ~~streamwater~~ stream water is a mixture of varying contributions from relatively time-invariant source solutions (end-members). These end-members are typically identified by collecting additional measurements of candidate end-members from within the watershed, and comparing these to the observations. This technical note introduces a complementary, data-driven method: Convex-Hull End-Member Mixing Analysis (CHEMMA), to infer the end-member compositions and their associated uncertainties from the ~~streamwater~~ stream water observations alone. The method involves two steps. The first step uses Convex-Hull Non-negative Matrix Factorization (CH-NMF) to infer possible end-member compositions by searching for a simplex that optimally encloses the ~~streamwater~~ stream water observations. The second step uses Constrained K-means Clustering (COP-KMEANS) to classify the results from repeated applications of CH-NMF to analyze the uncertainty associated with the algorithm. In an example application using the 1986 to 1988 Panola Mountain Research Watershed dataset, CHEMMA is able to robustly reproduce the three field-measured end-members found in previous research using only the ~~streawater~~ stream water chemical observations. It also suggests the existence of a fourth end-member. ~~Further~~ In this technical note, we have estimated uncertainties arising from the algorithm itself, but further work is needed to ~~explore the constraints and~~ determine the effect of sampling error and other uncertainties on the capabilities of this approach.

Abstract.

1 Introduction

End-Member Mixing Analysis (EMMA) has been used to interpret observed ~~streamwater~~ stream water chemical concentration profile variability in terms of time-varying contributions from end-member "sources", each supplying water with a constant concentration profile. This method has been applied in many different hydro-climatic and geology settings ~~(Bernal et al., 2006; Hooper et al., 19~~ EMMA has also been used to distinguishing sources of dissolved organic matter in natural streams (Hur et al., 2006; Yang and Hur, 2014), specific conductance (Kronholm and Capel, 2015), and other combinations of ~~streamwater~~ stream water attributes that can be assumed to have conservative mixing (Barthold et al., 2011).

EMMA assumes that the chemical ~~composition of streamwater~~ solute composition of stream water should be explained by the conservative mixing of a finite set of temporally invariant end-members (Hooper et al., 1990). These end-members, therefore, are the most extreme points that define a range within which all ~~streamwater~~ stream water observations are included. End-members are identified by collecting samples of candidate source-water from within the watershed ~~.~~ (i.e. in addition to the 'mixture' samples collected in the stream). Inasmuch as the end-members are identified by candidate sampling, they depend upon hypotheses that 1) identified end-members supply ~~streamwater~~stream water; and 2) identified end-member set ~~may be incomplete.~~ is 'complete' in some sense.

~~Streamwater concentration~~ Christophersen and Hooper (1992) suggested that "[u]nambiguous identification of the source solution compositions from the mixture alone is impossible". In a strict sense this is likely true, since the underlying assumption (streamflow as a conservative mixture of invariant sources) is unlikely to be adhered to in a real watershed. However, recent advances in statistical learning methods suggest there may be some utility in attempting to identify (perhaps not free of ambiguity) potential source solution composition from the observed mixture alone (without additional candidate source-water samples). Here we propose a method (which we will refer to as CHEMMA) as an attempt to push this concept forward.

Stream water concentrations of different conservative solutes are naturally correlated. EMMA utilizes the Principal Component Analysis (PCA) method to convert the naturally correlated ~~streamwater~~ stream water concentrations into a set of linearly uncorrelated variables (Christophersen and Hooper, 1992). Each new variable, which is called Principal Component (PC), is a linear combination of the observed ~~streamwater~~ stream water attributes. For a set of n variables, PCA first requires standardized observations ($\cancel{\mathbf{X}_{obs}}\mathbf{X}_{obs}$) by subtracting the mean and dividing by the standard deviation. Then it calculates a projection matrix $\mathbf{P}_{obs}$ (rows of which are eigenvectors of the correlation matrix), which transforms ~~the from the~~ from observation space to ~~the~~ PC space, by decomposing ~~the covariance matrix of~~ $\cancel{\mathbf{X}_{obs}}$correlation matrix of $\mathbf{X}_{obs}$. The transformed columns of $\mathbf{Y}_{obs}$ (representing the n observations in the PC space) are uncorrelated, and each of which accounts for a portion of total variance (Christophersen and Hooper, 1992):

$$\mathbf{Y}_{obs} = \mathbf{X}_{obs}\,\mathbf{P}_{obs}^{T}. \tag{1}$$

Standardized end-member candidates $\mathbf{X}_{em}$ can be projected into the PC space by the same projection matrix $\mathbf{P}_{obs}$, and be converted in the transformed space as $\mathbf{Y}_{em}$ (Christophersen and Hooper, 1992):

$$\mathbf{Y}_{em} = \mathbf{X}_{em}\,\mathbf{P}_{obs}^{T}. \tag{2}$$

To find the parsimonious subset of appropriate end-members, EMMA then takes the information provided by PCA to determine the approximate dimensionality of the ~~streamwater~~ stream water mixture and to screen end-members (Hooper, 2003). In the PC space, appropriate end-member candidates ($\mathbf{Y}_{em}$) are selected by choosing ones that tightly bound the transformed observations ($\mathbf{Y}_{obs}$) (Christophersen and Hooper, 1992; Hooper et al., 1990; Hooper, 2003). However, the number of retained

PCs is usually determined using a heuristic, such as using the number of PCs that explain at least $\frac{1}{n}$ proportion of the total variance, because of the need to capture the variance ~~. After subjectively determined~~ (Hooper, 2003). After thus subjectively determining the number of PCs, Christophersen and Hooper (1992) mathematically proved that one end-member more than the number of PCs is required to describe the rank of the ~~streamwater~~ stream water observation.

There are limitations to this approach, that can result in spurious or incomplete source identification (Delsman et al., 2013; Hooper, 2003; Valder et al., 2012; Yang and Hur, 2014). Specifically, 1) the composition of a source cannot be determined unless candidate end-member measurements are obtained that are representative of it; 2) determining the number of significant PC is subjective; 3) EMMA is not able to deal with non-conservative mixing if non-linear structure are not provided to replace the current simplex structure (Christophersen and Hooper, 1992); 4) uncertainties introduced by spatial and temporal variability in end-member concentrations cause extra difficulties.

Here we focus on the first of these issues. In spite of EMMA's wide ~~applications~~ application (Ali et al., 2010; Bernal et al., 2006; Burns et al., 2001; Delsman et al., 2013; Hooper and Christophersen, 1992; James and Roulet, 2006; Jung et al., 2009; Li et al., 2019; Lv et al., 2018; Neal et al., 1992; Neill et al., 2011; Valder et al., 2012), there is not a method to characterize missing or unmeasured end-members. In this technical note, we focus on improving the identification of end-member compositions and associated uncertainties. The method depends on the idea inherited from EMMA that the end-members are located near the most extreme points of streamwater samples that bound the observations in "mixing" space. This suggests that we might be able to interrogate the observational data projected in the end-member space to find such extremal points as end-members (even if no individual samples fully represent that end-member). We propose a new EMMA approach, Convex-Hull End-Member Mixing Analysis (CHEMMA), which is a data-driven method that characterizes the ~~end-member chemical compositions~~end-members' chemical composition, as well as an intrinsic uncertainty component associated with the end-member. The ability of the method to infer end-members is demonstrated by an application to the 1986 to 1988 Panola Mountain Research Watershed dataset published in Hooper and Christophersen (1992).

2 Methodology

Convex-Hull End-Member Mixing Analysis (CHEMMA) applies a matrix factorization method, Convex-Hull Non-negative Matrix Factorization (CH-NMF), along with a classification method, Constrained K-means Clustering (COP-KMEANS), to find end-member compositions under EMMA assumptions. The CH-NMF method provides a numerical iterative algorithm to search for end-member compositions that optimally enclose the ~~streamwater~~ stream water observations in the PC space. The CH-NMF algorithm is run many times because each iteration of the search can result in highly non-unique optima. We apply the COP-KMEANS method to classify the CH-NMF numerical outputs into clusters. The centroid of each cluster is assumed to represent our best estimate of an end-member.

2.1 Adaption of CH-NMF to the EMMA problem

The concepts of "convex combination" and "convex hull" connect CH-NMF with the idea of end-member mixing. A convex combination is equivalent to a weighted sum. It is a linear combination of vectors where the weight associated with each vector varies between zero to one, and the weights sum to one. If we construct a simplex, which means a highly dimensional polytope, with some distinct vectors at its vertices, this simplex is a convex hull that encloses points within the hull to be a convex combination of the vertices. Similarly, if we conservatively mixed distinct end-members, the ~~streamwater~~ stream water chemical concentration observations can be a weighted sum of end-members with their contributions. The ideas of "convex combination" and "convex hull" are mathematically identical to ~~the~~ end-member mixing.

The CH-NMF method describes a general methodology of finding the most extreme points (end-members) that form a simplex with ~~k~~ k vertices around the n-dimensional observation data cloud by searching for convex hull that enclose the data when projected into a linear lower dimensional projection subspaces (Thurau et al., 2011). First, the observations are standardized (zero mean and unit variance), the PC vectors are calculated, and the top k PCs are retained as with EMMA. The CHEMMA algorithm does not entirely avoid this subjective choice of the number of end-members retained, and so does not resolve this criticism of EMMA. Next, the standardized data are projected into the 2D subspace spanned by two of the PCs. Qualified points forming a convex hull around the projected data are marked. This is repeated for every pair of PCs. Finally, we interpolate between convex-hull vertices in each subspace to find the vertices of a simplex in a k-dimensional subspace. This simplex forms a ~~"~~convex-hull ~~"~~ such that all the data points can be optimally approximated as convex linear combinations of them. The algorithm is summarized as follows:

Algorithm 1: CH-NMF algorithm (Thurau et al., 2011) adapted to the end-member identification problem given m ~~streamwater~~ stream water observations of n solutes

Result: i^{th} end-member composition $\boldsymbol{x}_{emi}^{n\times 1}$, and its contribution $\boldsymbol{h}_i^{m\times 1}$, $i = 1, 2, ..., k$

1. Subtract the mean ($\mu_{1:n}$) and dividing by standard deviation ($\sigma_{1:n}$) for each solute to obtain standardized observation matrix $\mathbf{X}_{obs}^{m\times n}$
2. Compute d eigenvectors (PCs) $\boldsymbol{e}_1$, ..., $\boldsymbol{e}_d$, where $d = rank(\mathbf{X}_{obs}\mathbf{X}_{obs}^T) \leq n$
3. Project $\mathbf{X}_{obs}$ onto each of the $\binom{d}{2}$ 2D-subspaces spanned by pairs of PCs (similar form as Eqn. 1 & 2)
4. Mark the k convex hull vertices for each projection plane and stored in matrix $\mathbf{S}^{n\times p}$, p is the maximum number of points needed to make a convex hull in one projection plane.
5. Define end-member matrix $\mathbf{X}_{em}^{n\times k} = [\boldsymbol{x}_{em1}, \boldsymbol{x}_{em2}, ..., \boldsymbol{x}_{emk}]$ and let $\mathbf{X}_{em} = \mathbf{SI}$, minimize $\|\mathbf{S} - \mathbf{SI}^{p\times k}\mathbf{J}^{k\times p}\|_F^2$, s.t. $\sum_i \boldsymbol{i}_j = 1, i_{ij} \in [0,1]$, and $\sum_i \boldsymbol{j}_j = 1, j_{ij} \in [0,1]$
6. Minimize $\|\mathbf{X}_{obs} - \mathbf{H}^{m\times k}\mathbf{X}_{em}^T\|_F^2$, s.t. $\sum_j \boldsymbol{h_i} = 1, h_{ij} \in [0,1]$

With given standardized m ~~streamwater~~ stream water samples with n measured attributes $\mathbf{X}_{obs}^{m\times n}$ and desired k end-members (Step 1, Figure 1 a), CH-NMF decomposes the ~~covariance~~ correlation matrix of the observations to obtain at most d PCs (d is the maximum number of linearly uncorrelated variables), which is the same linear orthogonal projection as ~~PCA~~ Principal Component Analysis (PCA) method (Step 2). Instead of immediate dimension reduction as EMMA, CH-NMF examines the distribution of $\mathbf{X}_{obs}$ in all of the subspaces spanned by PC pairs (Step 3, Figure 1 b, light blue points) and marks the most extreme points (Figure 1 b, red crosses) that construct the convex hull (Figure 1 b, red lines) to store in $\mathbf{S}$ (Step 4). Then, a subset of $\mathbf{S}$, $\mathbf{SI} = \mathbf{X}_{em}$, is found as a convex combination of $\mathbf{S}$ (Step 5, Figure 1 c, square vertices of the simplex) that minimizes the Frobenius norm $\|\cdot\|_F^2$ (the entry-wise Euclidean norm of the matrix). Finally, the contribution $\mathbf{H}$ is found by finding the convex combination of end-members that reproduces the data with minimal error (again using the Frobenius norm) (Step 6).

Step 5 is the essential step of the CH-NMF theory, and it is a modification of Convex Nonnegative Matrix Factorization (C-NMF) by adding a convexity constraint on $\mathbf{J}$, which means each component contributes between zero and one with sum of all to be one (Ding et al., 2008; Thurau et al., 2011). In the original setting of C-NMF, the $\mathbf{I}$ and $\mathbf{J}$ are naturally sparse if the vertex search is in PC subspaces (Ding et al., 2008). Adding the convexity constraint on $\mathbf{J}$ makes $\mathbf{J}$ an interpolation between each columns of $\mathbf{SI}$ (i.e. each end-member composition $\boldsymbol{x}_{em}$), however, the sparse nature of $\mathbf{I}$ remains (Thurau et al., 2011).

We could interpret the objective function of Step 5 (minimize $\|\mathbf{S} - \mathbf{SI}^{p\times k}\mathbf{J}^{k\times p}\|_F^2$) in three steps. First , the sparsity of $\mathbf{I}$ results in the end-member composition $\mathbf{X}_{em}$ close to a subset of the extreme observations ($\mathbf{S}$) projected in the PC subspace. Second, $\mathbf{J}$ makes other extreme observations in $\mathbf{S}$ to be expressed as a convex combination (interpolation) of $\mathbf{X}_{em}$. Third, minimizing the Frobenius distance between $\mathbf{S}$ and $\mathbf{X}_{em}\mathbf{J}$ guarantees end-member compositions $\mathbf{X}_{em}$ will be convex hull vertices because all other extreme points can be written as convex combinations of vertices, but not vice versa. As a consequence, a well-supported set of convex hull vertices tightly bound the observations and are as unique as possible, which satisfies the original EMMA assumption of finite set of distinct end-members. The sparse nature of $\mathbf{I}$ helps prevent overfitting because noise will tend to be concentrated on superfluous vertices without degrading identification of the others. The noisy end-members can be identified in the classification step given in the next section.

The constraint requiring that the end-members be a convex combination of the extreme observations implies that CH-NMF may not accurately identify end-members that are not a large fraction of any observation in the dataset. As the synthetic example shown in Figure 1 illustrates, the simplex formed by joining the CH-NMF end-members lies inside the shell formed by connecting the extreme points (red crosses in Figure 1c). If no samples are anywhere close to being 'pure' representatives of an end-member, the apparent end-member identified by CH-NMF may lie closer to the data centroid than the true end-member. Methods to relax the constraint on Step 5 and better identify end-members distant from the data in mixing space will be investigated in future work.

2.2 Quantify the intrinsic uncertainty using COP-KMEANS

Each run of CH-NMF may yield different end-member estimates. This is because the complex structure of the high-dimensional ~~streamwater~~ stream water data result in a rough objective function surface (Step 5). CH-NMF runs with different initial search locations may fall into different local minima.

Depending on the structure of the data cloud, each run's end-members may be nearly identical (if the end-member is well-characterized) or one or more may vary widely. Poor identification of extreme points may result from a lack of sufficient well-defined "vertices" in the data cloud. This may occur if more end-members are sought than the data can support. It may also occur if an end-member is variable in time. Instead of a vertex, the time-varying end-member forms an edge in the data space. Alternatively, the observations may not sample the true mixing space sufficiently to identify an end-member in the space as a convex-hull vertex, perhaps because it never represents more than a small fraction of variance.

Even in the absence of these issues, the variability and uncertainty of the stream concentration observations will contribute to uncertainty in end-member identification. The variation in the CH-NMF-identified end-members can be assessed by running the CH-NMF ~~analysis a~~ a nalysis a large number of times, and then using a clustering algorithm to extract the centroid and spread of areas consistently identified as an end-member. We use the COP-KMEANS variant of the K-means clustering algorithm, which allows us to require that end-members predicted from the same CH-NMF run must not be placed in the same cluster (Wagstaff et al., 2001). This is achieved by assigning a "cannot-link" constraint between every pair of candidate end-members generated by the same CH-NMF run. Apart from the "cannot-link" constraints, COP-KMEANS works identically to normal k-means clustering (Wagstaff et al., 2001). For each cluster identified by COP-KMEANS, we can qualitatively examine the spatial distribution of the associated end-members, and quantitatively calculate the centroid and variance of the cluster.

As the number of end-members increases, the centroid and variance within the cluster may increase or decrease, which provides another way to decide the number of needed end-members for a given observation set. In this technical note, we consider a new cluster to be well-identified as a proper end-member if two conditions are both satisfied: 1) the spread of previously identified clusters remains similar or decreases, and 2) the cluster itself has a reasonable variance.

2.3 Example Python implementation

An example Python implementation of CHEMMA including the application to Panola Mountain data presented in the next section are available in a Jupyter Notebook on GitHub (https://github.com/Estherrrrrxu/CHEMMA). The CH-NMF section uses a Python package, pymf.chnmf, detailed in Thurau et al. (2011). The COP-KMEANS section uses a Python package, COP-Kmeans presented in Babaki (2017).

3 An application of CHEMMA

We applied CHEMMA to a test dataset of 905 samples of six solutes (alkalinity, sulfate, sodium, magnesium, calcium, and dissolved silica) collected from the stream in the Panola Mountain research catchment, Georgia, U.S. and described in Hooper et al. (1990). The six solutes were specifically selected to meet EMMA's assumption that their concentrations vary significantly across the watershed (Hooper et al., 1990). Hooper et al. (1990) found that the stream chemistry could be interpreted as a mixture of hillslope, groundwater, and organic soil end-members, which are identified by sampling within the watershed. Here we ask 1) does CHEMMA recover the same three end-members as Hooper et al. (1990) identified in field-sampling? and 2) does the data support the existence of additional end-members?

We ran CHEMMA for three, four, and five end-member cases ($k = 3, 4, 5$) because two and three PCs account for 94% and 97% of the total variance, respectively . In order to capture the intrinsic uncertainty associated with the identified clusters, we calculated the mean and standard deviation (st.dev) for each case based on 100 CH-NMF runs (Table 1). CHEMMA was able to recover the three field-measured end-members reported by Hooper et al. (1990) (Figure 2, three blue stars). The mean of the three CHEMMA identified clusters (Figure 3 and Table 1) are very similar to the median concentration of the field-measured end-members (Table 2). The median concentration of the hillslope field sample (Table 2) has much lower alkalinity concentration compared with the mean concentration of the CHEMMA identified Green cluster (Figure 3 and Table 1), however, it is still within the cluster spread given in Table 1.

A fourth end-member could be robustly identified (Figure 2, four red stars) that explained more of the data variability. Hooper (2003) also suggested the existence of a fourth end-member. This end-member ~~apperared~~ appeared to be a mixture of hillslope and groundwater in some ways but had relatively high alkalinity and silica concentration compared to those end-members (Figure 2 brown and navy axes). The fourth end-member captures variations along the third PC axis (Figure 3 d), which are not apparent in the 2D view (Figure 3 b).

The spread of all end-member clusters (generated by 100 runs of CH-NMF) was small when four were sought, but a fifth could not be clearly identified. As the number of end-members was increased from three (Figure 3 a) to four (Figure 3 b), the new cluster (cyan Cluster 4) was dense, while the other three clusters (green, blue, and red) remained at similar locations to those clusters identified in the three end-member case. Adding the fourth end-member reduced the spread of the previously identified three clusters in the PC subspace (Figure 3 a and b and Table 1) suggesting they could now be identified with less uncertainty. However, the inclusion of the fifth end-members (Figure 3 c) not only did not further tighten the previously identified clusters, but the fifth cluster was poorly defined (black Cluster 5). Except the cyan cluster has generally decreased within cluster variation, the standard deviations of other clusters increase for both three and four end-member cases (Table 1).

4 Discussion and conclusions

~~As the application results show,~~

4.1 Application results

The results in Figure suggest that identification of end members from the mixture alone may not be as "impossible" as Hooper and Christophersen (1992) assumed. CHEMMA is able to reproduce the three end-members that were ~~verified in the previous study~~ identified in Hooper et al. (1990) as well as a fourth end-member that explains more variation in the data~~(Hooper et al., 1990).The~~ .

This is not to say that the estimates provided by CHEMMA are "unambiguous", or even a complete set of contributing sources. For example, sources that never supply the the plurality of water may not be identified by CHEMMA, since they never produce a 'vertex'-like structure in the data cloud. Further work is needed to determine the limits on end-member identification for a given dataset.

Indeed, the dispersed cluster distributions in Figure 3 c suggests that ~~the~~ a fifth end-member may be ~~finding~~ spurious, and reflects only the noisy edges of the sample space, and so cannot be supported by the data. However, we have not here identified an objective criteria for determining whether an end-member is supported. There are many mathematical methods, such as factor analysis and diffusion map spectral gaps, that could be used in parallel with CHEMMA to estimate data dimensions (Ashley and Lloyd, 1978; Coifman et al., 2008). In addition, we could use k-fold cross validation of CHEMMA itself to try to determine the best number of end-members.

4.2 Uncertainties

Because CHEMMA extracts end-members from the observations, the accuracy of the end-member's composition is influenced by the noise from sample chemical analysis error, how well the collected samples represent the full range of sources in the catchment, and how valid the assumption of conservatively-mixing time-invariant end-members is. The captured variations in PC 3 shown in Figure 3 d may result from temporal variations of the end-member composition. The less concentrated Cluster 3 in Figure 3 b may result from relatively rare contributions from that end-member. Fortunately, CHEMMA itself ~~provides~~ may provide a tool for exploring some of these sources of uncertainty. ~~By~~ For example, by partitioning the dataset into time periods (or hydrologic state, etc), the temporal variability of end-members could be explored. The temporal end-member dominance may further deepen our understanding of stream water characterization.

Besides the uncertainties introduced by temporal variations of the end-member composition, the total uncertainties may also be decomposed into uncertainties associated with stability of the end-members, stream water sampling error, CHEMMA algorithm stability (quantified in the given example application), and structural uncertainties (i.e. choice of the number of end-members k). A full analysis of these errors and their interaction will be addressed in future work.

4.3 Potential applications

For most hydrologists, end-member analysis is used to identify the water sources, and toward that purpose CHEMMA may be useful in the following ways: 1. CHEMMA may be used to reduce subjectivity when selecting from field-measured end-member candidates by comparing them to CHEMMA end-members; 2. CHEMMA may identify end-members that have not been sampled in the field, which may serve as a check for missing sources; 3. CHEMMA end-member compositions may help hydrologists ask better questions and provide guidance for field sampling by suggesting source characteristics; 4. CHEMMA can be used in conjunction with the Diagnostic Tool of Mixing Models (DTMM, developed by Hooper (2003)). DTMM is used to assess the tracer conservation, and mixture rank. CHEMMA can be enhanced by using DTMM analysis to select conserved tracers for analysis. The robustness of CHEMMA end-members also serve as a check for DTMM-determined rank of mixture.

4.4 Future work

Future work refining and applying this method may focus on 1) applying quantitative methods to ~~optimize the model complexity~~eliminate the subjective choice of k, such as the Akaike Information Criterion (AIC), or Bayesian Information Criterion (BIC, or Schwarz criterion); 2) relaxing the constraints on the CH-NMF algorithm (Algorithm 1, Step 5) so that extreme points in $\mathbf{S}$ also lie inside the simplex, allowing the method to better characterize end-members that are never a large fraction of any observations; and 3) exploring the data requirements and uncertainty of the method, including better understanding the relationship between the stability of COP-KMEANS clusters, the temporal variability of end-members, and the number of samples; 4) using individual field end-member measurements to inform CHEMMA.

Code and data availability. Both the example code and data are available in a Jupyter Notebook on GitHub https://github.com/Estherrrrrxu/CHEMMA (Xu Fei, 2020).

Author contributions. Xu Fei and Harman were responsible for conceptualization, methodology, and visualization. Xu Fei was responsible for investigation, formal analysis, and writing (original draft). Harman was responsible for funding acquisition, supervision, and writing (review & editing).

Competing interests. The authors have no competing interests to declare.

Acknowledgements. Panola stream solute chemistry data from Hooper and Christophersen (1992) was collected with the support of the USGS and is available at http://hiscentral.cuahsi.org/pub_network.aspx?n=385. Thanks to Rick Hooper for providing thoughtful feedback on the draft manuscript~~.~~, and to Dr. Joost Delsman and one anonymous reviewer for their careful reading. This work was supported by NSF grant EAR-1654194.

References

Ali, G. A., Roy, A. G., Turmel, M. C., and Courchesne, F.: Source-to-stream connectivity assessment through end-member mixing analysis, Journal of Hydrology, 392, 119–135, https://doi.org/10.1016/j.jhydrol.2010.07.049, 2010.

Ashley, R. and Lloyd, J.: An example of the use of factor analysis and cluster analysis in groundwater chemistry interpretation, Journal of Hydrology, 39, 355–364, 1978.

Babaki, B.: COP-Kmeans version 1.5, https://github.com/Behrouz-Babaki/COP-Kmeans, https://doi.org/10.5281/zenodo.831850, 2017.

Barthold, F. K., Tyralla, C., Schneider, K., Vaché, K. B., Frede, H.-G., and Breuer, L.: How many tracers do we need for end member mixing analysis (EMMA)? A sensitivity analysis, Water Resources Research, 47, 1–14, https://doi.org/10.1029/2011WR010604, 2011.

Bernal, S., Butturini, A., and Sabater, F.: Inferring nitrate sources through end member mixing analysis in an intermittent Mediterranean stream, Biogeochemistry, 81, 269–289, https://doi.org/10.1007/s10533-006-9041-7, 2006.

Burns, D. A., Mcdonnell, J. J., Hooper, R. P., Peters, N. E., Freer, J. E., Kendall, C., and Beven, K.: Quantifying contributions to storm runoff through end-member mixing analysis and hydrologic measurements at the Panola Mountain Research Watershed (Georgia, USA), Hydrological Processes, 15, 1903–1924, https://doi.org/10.1002/hyp.246, 2001.

Christophersen, N. and Hooper, R. P.: Multivariate analysis of stream water chemical data: the use of Principal Components Analysis for the end-member mixing problem, Water Resources Research, 28, 99–107, 1992.

Coifman, R. R., Kevrekidis, I. G., Lafon, S., Maggioni, M., and Nadler, B.: Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems, Multiscale Modeling & Simulation, 7, 842–864, 2008.

Delsman, J. R., Oude Essink, G. H., Beven, K. J., and Stuyfzand, P. J.: Uncertainty estimation of end-member mixing using generalized likelihood uncertainty estimation (GLUE), applied in a lowland catchment, Water Resources Research, 49, 4792–4806, https://doi.org/10.1002/wrcr.20341, 2013.

Ding, C. H., Li, T., and Jordan, M. I.: Convex and semi-nonnegative matrix factorizations, IEEE transactions on pattern analysis and machine intelligence, 32, 45–55, 2008.

Hooper, R. P.: Diagnostic tools for mixing models of stream water chemistry, Water Resources Research, 39, 1055, https://doi.org/10.1029/2002WR001528, 2003.

Hooper, R. P. and Christophersen, N.: Predicting episodic stream acidification in the southeastern United States: combining a long-term acidification model and the end-member mixing concept, Water Resources Research, 28, 1983–1990, https://doi.org/10.1029/92WR00706, 1992.

Hooper, R. P., Christophersen, N., and Peters, N. E.: Modelling streamwater chemistry as a mixture of soilwater end-members - an application to the Panola Mountain Catchment, Georgia, U.S.A., Journal of Hydrology, 116, 321–343, 1990.

Hur, J., Williams, M. A., and Schlautman, M. A.: Evaluating spectroscopic and chromatographic techniques to resolve dissolved organic matter via end member mixing analysis, Chemosphere, 63, 387–402, https://doi.org/10.1016/j.chemosphere.2005.08.069, 2006.

James, A. L. and Roulet, N. T.: Investigating the applicability of end-member mixing analysis (EMMA) across scale: A study of eight small, nested catchments in a temperate forested watershed, Water Resources Research, 42, 1–17, https://doi.org/10.1029/2005WR004419, 2006.

Jung, H. Y., Hogue, T. S., Rademacher, L. K., and Meixner, T.: Impact of wildfire on source water contributions in Devil Creek, CA: evidence from end-member mixing analysis, Hydrological Processes, 23, 183–200, https://doi.org/10.1002/hyp, 2009.

Kronholm, S. C. and Capel, P. D.: A comparison of high-resolution specific conductance-based end-member mixing analysis and a graphical method for baseflow separation of four streams in hydrologically challenging agricultural watersheds, Hydrological Processes, 29, 2521–2533, https://doi.org/10.1002/hyp.10378, 2015.

Li, X., Ding, Y., Han, T., Kang, S., Yu, Z., and Jing, Z.: Seasonal controls of meltwater runoff chemistry and chemical weathering at Urumqi Glacier No.1 in central Asia, Hydrological Processes, 33, 3258–3281, https://doi.org/10.1002/hyp.13555, 2019.

Liu, F., Bales, R. C., Conklin, M. H., and Conrad, M. E.: Streamflow generation from snowmelt in semi-arid, seasonally snow-covered, forested catchments, Valles Caldera, New Mexico, Water Resources Research, 44, 2008.

Liu, F., Conklin, M. H., and Shaw, G. D.: Insights into hydrologic and hydrochemical processes based on concentration-discharge and end-member mixing analyses in the mid-M erced R iver B asin, S ierra N evada, C alifornia, Water Resources Research, 53, 832–850, 2017.

Lv, Y., Gao, L., Geris, J., Verrot, L., and Peng, X.: Assessment of water sources and their contributions to streamflow by end-member mixing analysis in a subtropical mixed agricultural catchment, Agricultural Water Management, 203, 411–422, https://doi.org/10.1016/j.agwat.2018.03.013, 2018.

Neal, C., Robson, A., Reynolds, B., and Jenkins, A.: Prediction of future short-term stream chemistry - a modelling approach, Journal of Hydrology, 130, 87–103, https://doi.org/10.1016/0022-1694(92)90105-5, 1992.

Neill, C., Chaves, J. E., Biggs, T., Deegan, L. A., Elsenbeer, H., Figueiredo, R. O., Germer, S., Johnson, M. S., Lehmann, J., Markewitz, D., and Piccolo, M. C.: Runoff sources and land cover change in the Amazon: An end-member mixing analysis from small watersheds, Biogeochemistry, 105, 7–18, https://doi.org/10.1007/s10533-011-9597-8, 2011.

Thurau, C., Kersting, K., Wahabzada, M., and Bauckhage, C.: Convex non-negative matrix factorization for massive datasets, Knowledge and Information Systems, 29, 457–478, https://doi.org/10.1007/s10115-010-0352-6, 2011.

Valder, J. F., Long, A. J., Davis, A. D., and Kenner, S. J.: Multivariate statistical approach to estimate mixing proportions for unknown end members, Journal of Hydrology, 460-461, 65–76, https://doi.org/10.1016/j.jhydrol.2012.06.037, 2012.

Wagstaff, K., Cardie, C., Rogers, S., and Schroedl, S.: Constrained k-means clustering with background knowledge, Proceedings of the Eighteenth International Conference on Machine Learning, 1, 677–584, 2001.

Xu Fei, E.: Example CHEMMA application code for the technical note, https://github.com/Estherrrrrxu/CHEMMA, https://doi.org/10.5281/zenodo.4116082, 2020.

Yang, L. and Hur, J.: Critical evaluation of spectroscopic indices for organic matter source tracing via end member mixing analysis based on two contrasting sources, Water Research, 59, 80–89, https://doi.org/10.1016/j.watres.2014.04.018, 2014.

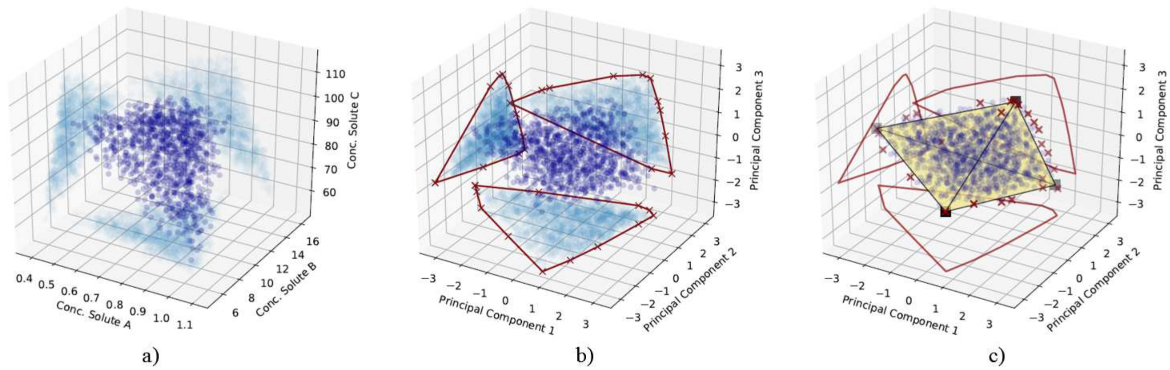


Figure 1. Illustration of the CH-NMF algorithm. a) The standardized observations (dark blue) and its projection (light blue) on the observational space. b) The projected observations (dark blue) and its projection (light blue) on PC subspaces. The red crosses are the marked extreme points ($\mathbf{S}$) that form a convex-hull (the red polygons) in each PC subspaces. c) Find the convex-hull (the black simplex) and its associated vertices (the k vectors $\mathbf{x}_{emi}$) in the PC space, such that the verticies are convex combinations of the extreme points $\mathbf{S}$, and the distance between the simplex and $\mathbf{S}$ is minimized.

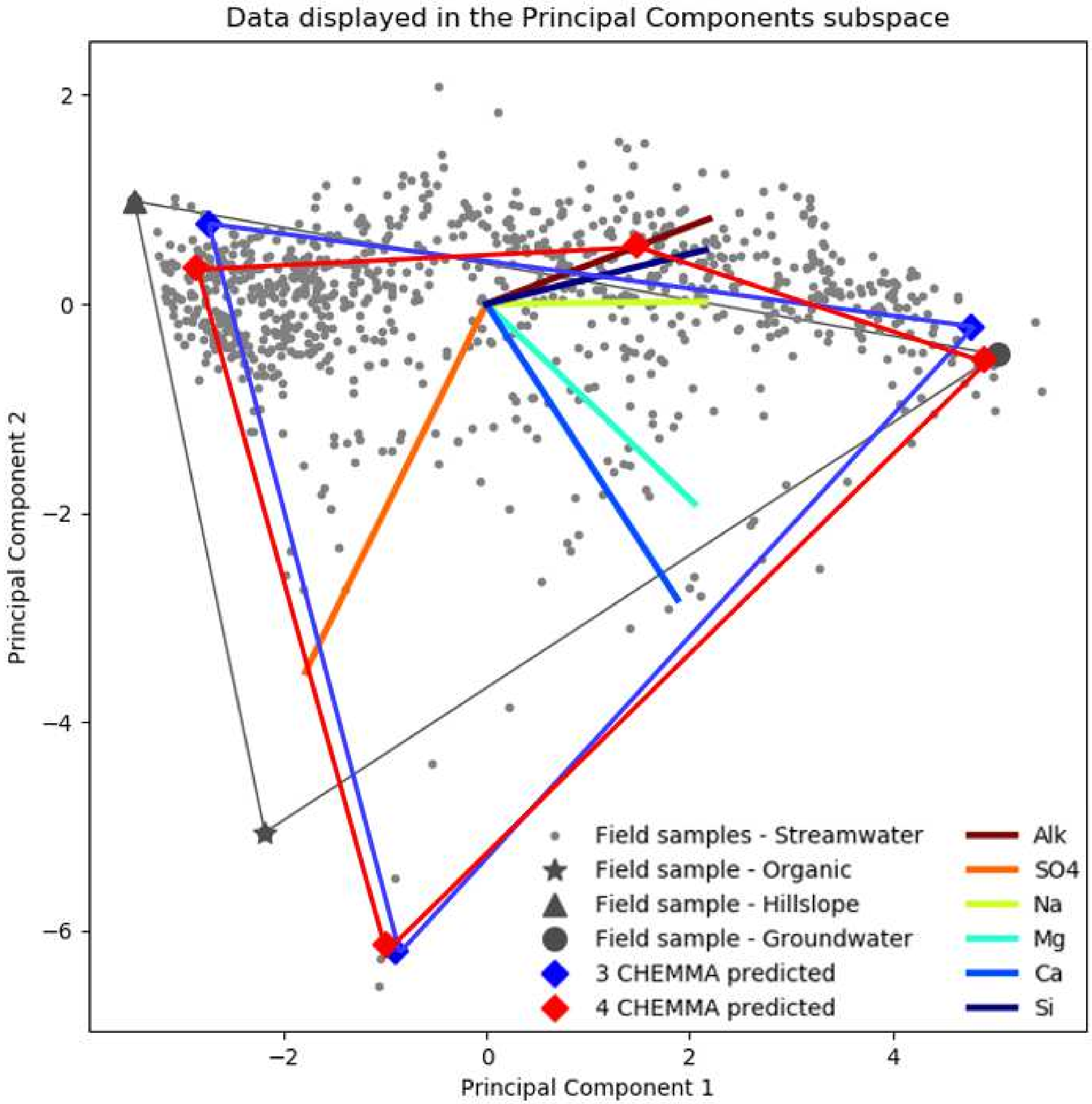


Figure 2. CHEMMA prediction (cluster centroids) for three end-member (blue squares) and four end-member (red squares) cases plotted in the PC2 vs. PC1 subspace. The colored lines that connect those predicted end-members indicate the convex hull formed by those end-members. The observations (grey dots) inside of the convex-hull can be explained as linear combinations of the end-members. The colored lines in the center of the plot are the projected original solute axes in this PC subspace.

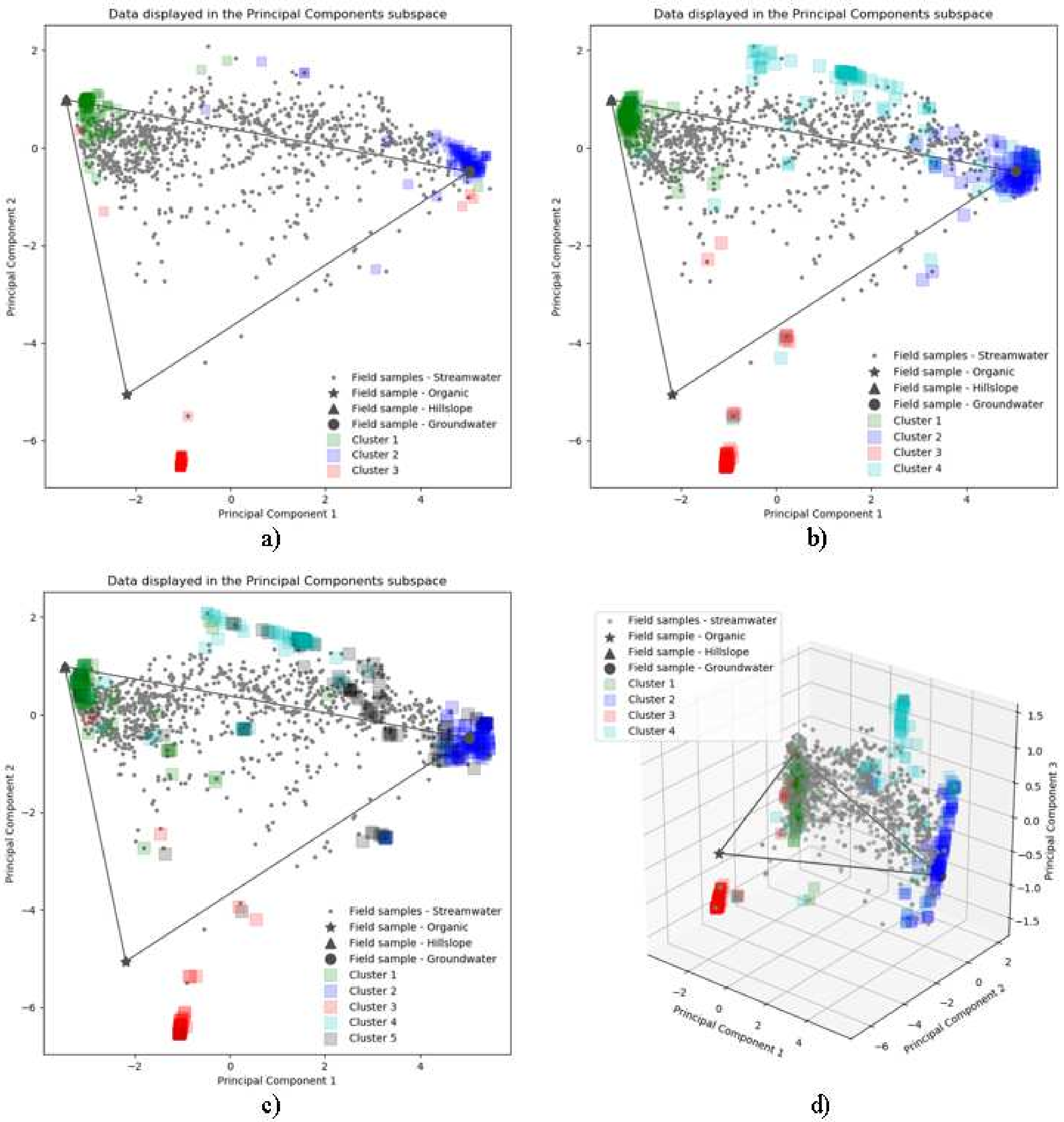


Figure 3. 100 random initialized CH-NMF runs result for three (a), four (b and d), and five (c) end-member cases. a - c are in the 2D PC2 vs. PC1 subspaces. d is in the 3D PC3 vs. PC2 vs. PC1 subspace.

Cluster	Alkalinity		SO_4		Na		Mg		Ca		Si	
	Mean	**St.dev**	**Mean**	**St.dev**	**Mean**	**St.dev**	**Mean**	**St.dev**	**Mean**	**St.dev**	**Mean**	**St.dev**
Red	35.05	27.02	216.75	30.72	48.14	20.28	92.48	7.92	192.37	22.36	90.88	53.51
Blue	348.04	12.16	14.11	2.82	214.87	21.88	90.35	4.64	151.26	9.93	405.86	23.55
Green	33.43	32.27	77.45	12.60	44.70	20.01	32.03	5.84	47.14	10.75	100.34	55.85
Red	32.86	12.33	219.71	17.57	46.66	9.91	93.50	2.44	193.92	15.11	87.25	28.64
Blue	345.01	23.29	15.71	14.91	211.26	26.22	92.02	5.88	157.14	11.86	385.44	50.57
Green	26.80	31.28	85.15	23.04	38.65	13.11	32.83	10.59	54.00	25.65	78.26	28.29
Cyan	207.96	92.01	38.45	40.07	141.51	46.76	61.89	18.02	91.57	42.03	342.13	122.07
Red	38.88	49.76	211.17	41.12	49.60	27.28	91.13	11.34	189.23	29.04	92.71	59.09
Blue	344.76	21.77	15.88	14.39	211.90	30.95	92.44	5.63	158.67	12.07	390.34	40.03
Green	29.62	33.35	85.37	13.38	42.52	17.68	33.40	6.83	52.32	16.99	84.20	29.38
Cyan	171.83	77.99	40.85	33.32	123.60	44.11	54.77	15.08	75.69	29.17	329.06	138.29
Black	253.45	107.65	44.10	47.45	161.55	58.00	75.81	17.47	125.51	38.38	278.05	123.41

Table 1. The mean and standard deviation (st.dev) of each end-member cluster based on 100 random initialized CH-NMF runs. All values are in micromoles per liter. The cluster color indications correspond to Figure 3 a to c.

Field individual samples	Alkalinity	SO_4	Na	Mg	Ca	Si
Organic	37	214	23	78	151	60
Groundwater	370	7	169	97	162	422
Hillslope	9	89	46	22	32	90

Table 2. The median concentration of individual field measured end-members from Hooper and Christophersen (1992). All units are in micromoles per liter.