Response to reviewer 1

We would like to thank this anonymous reviewer for his/her detailed reading of our manuscript and the very useful comments. One main concern raised by the reviewer is about the statistical validity of this approach, specifically the usage of importance sampling instead of an Markov chain Monte Carlo (MCMC) approach to sample the posterior distribution of the mixing ratio. As we discuss hereafter, we will implement an MCMC sampler to have a complete comparison with the method proposed. Responses to the specific comments are below.

P2L30: Are there really "n" linear Equations? Since k is the number of tracers, I would assume that we have a system of k linear equations. Then, if k=n we have "n" equations and "n" unknown variables. Why "n-1"?

Assuming we have "*k*" tracers, there will be "*k*+1" linear equations. The first "k" equations are:

$$\boldsymbol{\rho_1 S_1^k + \rho_2 S_2^k + \cdots + \rho_n S_n^k = Y^k}. \quad 1$$

The last equation corresponds to the aggregation of different sources in the mixture, i.e. $\sum_{i=1}^{n} \rho_i = 1$, which we mention in P1L30. In order to solve this system of linear equations, "n-1" different tracers are required. As also proposed by reviewer 2, we will state this clearly in the revised manuscript and move the equation part to the methods section.

P3L14: The authors state that it is a major shortcoming of traditional mixing models that the source concentrations are assumed to come from standard statistical distributions, which are described by some parameters. The authors should acknowledge that this can be a useful approach to account for the fact that the measurements of tracer concentrations in the same source are related to each other in some way, which is a very reasonable assumption. It is a priori not entirely clear that omitting such an assumption is beneficial to solving the mixing problem, since we might neglect reasonable prior knowledge in that case.

We thank the reviewer for this very important comment. The reviewer correctly points out that the traditional approach of fitting a statistical distribution to the source concentrations reflects a priori knowledge, which may be useful in a lot of cases. We will mention this in the 'Limitations and Opportunities' Section, where we think this discussion is best located. The main contribution of this paper is in instances where a priori knowledge about the source concentration is limited. These are mostly cases where very few measurements of source contributions are available, with limited a priori knowledge where the usage of HydroMix is more appropriate.

Furthermore, the authors do not directly compare the performance of their approach to the approaches they criticize (e.g. in a synthetic case study). The added value remains therefore rather vague.

We first would like to point out that we did not mean to criticize existing approaches in our manuscript. We will carefully revise the text to downgrade such critical statements since our approach is supposed to address mixing problems that are not straightforward to address with existing methods. We indeed do not compare our approach to existing approaches in detail here, since our objective is to present a new method for data sparse situations in which existing methods are not appropriate. Instead, we simulate synthetic case studies to evaluate the results of HydroMix. As these are statistical tests, the correct answer is known a priori, and we evaluate if HydroMix converges to the correct results. Given that the paper is already rather long, as also pointed out by reviewer 2, we decided not to increase the complexity and length of the paper.

P3L20-25: Arguments 1) and 2) seem to be very similar, if not the same. The authors should either provide two more distinct wordings, or combine the two arguments into one. Argument 3) is not entirely clear. Of course, the true mean and variance of the entire population can only be estimated with high uncertainty from a small sample. But this can be formally considered and should not pose a fundamental problem.

The reviewer refers here to the following statements:

1) The mean and variance may not accurately reflect the statistical properties of the source composition. 2) If there is a large amount of information on the source composition, the mean and variance may be an unnecessary simplification of its variability. 3) If the source compositions have a low number of samples, then the mean and variance estimates may be poorly constrained.

We agree that the stated arguments look very similar. This is because we argue mostly based on the very commonly assumed normality condition for the pdf of the different source concentrations, obtained empirically with small sample sizes. We will combine the three points into one sentence in the revised manuscript.

P3L37: Please specify what is exactly meant by the "above limitations". The list 1)-3)? Or additional things in the text above?

We were referring to the previous two paragraphs, we will clarify this in the revised manuscript.

P3L37-46: The authors claim that the most important advantage of HydroMix is that there is no assumption about the distribution of the source tracer concentrations, if I understand correctly. While technically true, I think that this argument is misleading. Also HydroMix makes an assumption about the probabilistic nature of the model, namely that the residuals are normally distributed with mean zero and constant variance. It is not clear to me why this should be a milder / better assumption than the one about the observations of the source concentrations being realizations of e.g. a normal distribution. All the uncertainty is just treated in a lumped way, by epsilon, which implicitly contains the deviations of the observed and the true source concentrations. Therefore, the assumptions on the source concentration distribution are not really avoided, they are just all "hidden away" in epsilon.

This is an important comment, which we will carefully address in the revised version to avoid "hiding away" the statistical assumptions. Thanks for pointing this out. As well summarized by the reviewer, the main difference of HydroMix is that it parameterizes the error function instead of the statistical distribution of the source concentrations. Parameterizing source concentrations requires large sample sizes, which is seldom the case in tracer hydrology. Error parameterization offers a useful alternative and is interesting since our methodological set-up allows increasing the error sample size compared to the input observation sample size. Furthermore, since the model error results from the aggregation of a large number of simplifications and observational errors, i.e. it aggregates a large number of random variables, the error can reasonably be assumed (central limit theorem) to follow a normal distribution. Despite this point, and as in any comparable inference approach (based on model residuals), assessing a posteriori the validity of the underlying distributional assumptions is a key step (we will include such an assessment in the revised version).

In hydrological modeling, this model residual assessment step might typically reveal that the residuals do not have zero mean (i.e. there is a bias). It is noteworthy that the normality assumption (and namely the zero mean assumption) for the model error is not required and could be modified. And, as a side note, Bayesian parameter inference assuming a wrong error distribution has been shown to be surprisingly robust for hydrological modelling (Schaefli & Kavetski, 2017).

In addition, as the reviewer correctly states above, the proposed approach lumps all error sources into a single error term (i.e. input observational uncertainties, model uncertainties, output observational uncertainties, parameter uncertainties). This is standard in hydrological model inference (e.g. Schaefli et al., (2007)) but there is indeed a large body of literature that discusses how this could be relaxed (e.g. via explicit accounting for observational uncertainties (Kavetski et al., 2006)), with the inherent limit that we do not know how to account for model structural uncertainties (e.g. a missing source in our case). We will mention this limitation in the revised version.

P3L44-46: unclear what the authors mean here

The reviewer refers to the following sentence:
"*An advantage of this approach is that additional model parameters can be incorporated in the framework to describe how the source tracer concentrations might be modified according to specific hydrologic processes that can be decided and explored by the user*."

In HydroMix, the model parameter (mixing ratio) is inferred in an inverse modeling setup. Theoretically, multiple parameters can be inferred in this framework. We demonstrate this in case study 3.5, where isotopic lapse rate is inferred along with the mixing ratio. We will make this clearer in the revised manuscript.

P5Eq4: I don't understand why Eq. 4 is a solution to limitation ii)

This was a typo, correct would have been "to limitation iv)".

Besides, we also remove the statement "*Section 3.5 introduces such an example and proposes a solution in these cases*." Since this cuts the flow of text.

P5L11: Timestep of what? Explain what you mean by "assuming a timestep". Isn't the timestep given by the times at which the samples were taken (observed)? I don't see why "tau" can be neglected for short and long timesteps.

The formulation was not clear in the paper. *t* is indeed the sampling time. We assume however, that all observed samples are samples from the long term (>> 1 year) source and target compositions, i.e. the modelling time step is much larger than the lag, which can thus be removed from the equation. The difference between sampling time step and modelling time step will be made clear in the revised version. We will also remove the sentence that *tau* could also be short. This was misleading.

P5Eq5: What do you mean exactly by time-integrated processes? S is a state, not a process, I believe. Please clarify. You might need to provide an equation to clarify which quantity is integrated and what the lower and upper limits of integration are.

We use here the term process in a generic sense to designate a stochastic process. The explanation in the text was not clear, we simply use different symbols for stochastic processes with different temporal support. This will be clarified.

P5Eq6: how do "i" and "j" relate to "t"? "t" seems to disappear in the following equations.

We did indeed not explicitly mention that we omit the time stamp of each sample and simply number them (thus the shift from *t* to *i*). This happens because we do not consider our samples as being a time series but simply as being a set of samples from a random variable, the source composition. This shift from a temporal process to a time-averaged random variable was indeed not clear in the manuscript. Given the questions of this and the 2nd reviewer, we consider to re-formulate this entire section. It is probably not necessary to present the temporal process viewpoint before switching to the source composition viewpoint.

Is it reasonable to compare all the samples of the sources to all the samples of the target mixture? The authors should expand on this. When does it make sense and when not?

HydroMix does not take into account the time at which a source was sampled exactly because we assume that each sample provides a sample of a random variable, which is the source composition. To build the model residuals, we need to compare the model output (modeled target compositions) with observed samples. Assuming that all available source samples are independent, building all possible model simulations from the available source samples is the most natural solution. It is noteworthy that even if the source and target samples are taken at the same time, this does not mean a priori that they are dependent since the target samples result from water that has travelled a certain time in the catchment, i.e. target sample from a time step *t* do not result from the source composition at the same time *t*. If in a system we have instantaneous mixing, then the source and target samples taken at the same moment in time will necessarily be strongly correlated. In this case the assumption of independent samples and thus the combination of all samples would not make sense. We will specify all this in the revised version and try to come up with an example where the assumption of independent samples would not make sense.

P6L1: I believe that the citation provided here is not justified. The cited paper has nothing to do with mixing problems, it is about spectral domain likelihoods for modeling streamflow. However, it would be important to have a reference here that justifies the assumption of the Gaussian distribution for the errors in the specific case of how it is applied in this study (comparing all the measurements of source to all measurements of the mixture concentrations with normal errors). An alternative would be to check the statistical characteristics of the resulting "epsilon" and see if the normal assumption was justified.

The cited paper is indeed about spectral calibration, which is insensitive to the distribution of errors in the time-domain. Therefore it compares spectral-domain to time-domain calibration and an important result is that time-domain calibration is very robust under wrong distributional assumptions, and in particular much more robust than commonly assumed in hydrological modelling. We will discuss why Gaussianity is justified and we will assess how well the normality assumption holds in our examples.

P6L29-30: Please provide the original reference for importance sampling and Metropolis algorithm.

We will include them in the revised manuscript.

P6L31-33: This sentence is not entirely correct and can be omitted

We will remove this in the revised manuscript.

My most important comment refers to P6L42 – P7L4. The approach that the authors chose to sample from the posterior distribution is not a valid approach. Random sampling of the parameter space with retaining the best X % of the likelihood function does not yield the correct posterior distribution. Instead, the authors obtain some arbitrary measure of spread of the parameters and neither the parameter range nor the predictions done with them have any probabilistic interpretation. This is also one of the potential reasons why they do not manage to reproduce the known "rho" in Figure 2. The true value of "rho" should be inside the confidence limits also for the overlapping (high-variance) case. Also, the higher uncertainty of the mixing ratio "rho" should be visible in the high variance case, but the estimated "rhos" seem to have the same or even less uncertainty associated to them in the high variance case than in the low variance case. This seems odd to me. I would recommend that the authors implement a proper MCMC sampler (e.g. Metropolis) to obtain an actual sample from the posterior distribution. Also, the convergence of the chains needs to be checked, either by visual assessment or by convergence tests. Only converged results should be reported, otherwise meaningful conclusions are not possible.

We agree with the comments of the reviewer and will implement a MCMC sampler to sample the posterior function in the revised manuscript.

P8Eq15: Does this consider rain on snow events? This might not be so important but could be mentioned

In the hydrologic model used, heat exchange within the snowpack from rain-on-snow events are not considered as that lies beyond the scope of this paper. We will mention this in the revised manuscript.

P9L26: When are samples taken in the model? How many of them are taken?

In this case study, isotopic ratios are simulated and all the rain and snowmelt isotopic ratios are used with HydroMix. The number of rainfall, snowfall and snowmelt samples are 39, 24 and 107, as mentioned in the labels of Figures 6 and 7. We will also mention this in this section in the revised manuscript.

Section 4.1: From the design of the case study, it is not clear if HydroMix is a statistically coherent framework. The authors should provide a proof of concept instead of or in addition to Section 4.1. The basic design of the experiment could be similar to 4.1, but it should be done with a large number of samples, to demonstrate that HydroMix converges to the correct solution in that case. This should also be possible for the high variance case, I think. As I mentioned before, this should be done via proper MCMC sampling, and convergence needs to be checked. For a large number of samples, the uncertainty intervals should contain the true value of "rho", also in the high variance case.

Thank you for this very valuable comment. The current section 4.1 will be modified into a proof of concept test. In the current version of the manuscript, we have shown that the error in estimation of the mixing ratio reduces with increasing sample size (Figure 3a). In addition, we will add a test that evaluates the convergence of HydroMix for the high variance case with a large number of samples. This we believe will test the statistical validity of the approach.

As it stands, conclusion 1 is favorable and it would indicate that HydroMix is a statistically valid method, but it is not supported by the results. If one replaces the word "uncertainty" by "bias", then the conclusion is supported by the results, but it is not a favorable conclusion anymore and indicates major deficiencies of the used approach. The authors should aim to obtain results that support conclusion 1.

We do not fully understand the point that the reviewer is trying to make here. Conclusion 1 reads as:
"*HydroMix gives reliable results for mixing applications with small sample sizes. As expected, the variance in source tracer composition and the ensuing composition overlap determines the uncertainty in the mixing ratio estimates. The uncertainty in mixing ratio estimates increases with increasing variance in source tracer compositions. Mixing ratio estimates improve (in terms of lower error) with increasing number of source samples.*"

We believe that the proof of concept test will prove the reliability and validity of HydroMix. We will revise the text according to the new results with the MCMC sampler.