This manuscript investigates the temporal variability of the young water fractions (Fyw) based on a 4.5-year time series of δ18O in precipitation and streamwater in the German Weiherbach catchment. For this, the authors fit sine curves to the entire 4.5-year data set to estimate the long-term average Fyw. Then, they cut out 189 individual 1-year 18O time series from the 4.5-year data set (i.e., shifting each 1-year period by 1 week) and fit individual sine curves to these 1-year periods to estimate 189 Fyw-values. The goodness-of-fit of the sine curves to the δ18O data was quantified through adjusted R2 values. Three hypotheses were tested: "(1) Fyw estimates do not change over time (time-invariance) (2) Short-term changes in the start of a tracer sampling campaign do not influence the Fyw estimate (sampling-invariance) (3) Fyw estimates are similar for a given calendar month of different years (seasonal-invariance)" (P3L13-15). By applying a Fyw-threshold value of 2%, the authors reject all hypotheses and conclude that Fyw-values based on 1-year isotope data sets can be highly variable over time. This time-variably of Fyw hampers catchment-comparison studies that utilize tracer data of different time series lengths or time periods. I find the critical evaluation of new metrics (young water fraction Fyw) interesting and useful as this allows to better plan sampling campaigns or to use existing data sets more efficiently for a robust estimation of Fyw. In that regard, I consider the testing of hypothesis 2 the most useful scientific contribution of this study (Sect. 3.4) because it quantifies how much a 4-weeks delay can change the Fyw-values that are estimated from 1-year data sets. Unfortunately, these changes in Fyw do not correlate with any of the tested hydro-meteorological variables (Sect. 4.4) so that the authors cannot provide any suggestions about the optimal sampling strategy.

We thank reviewer #2 for evaluating our study and the helpful comments.

We now conducted an extended analysis of hydro-meteorological data and it points to possible unique meteorological conditions during summer 2015 that might have influenced Fyw. Please also refer to our answer given to the next reviewer comment:

Besides the testing of hypothesis 2, I find it difficult to identify a clear motivation and the novel scientific contribution of this study. The fact that Fyw responds to changes in precipitation, discharge and/or catchment wetness has already been established (e.g., Kirchner, 2016b; Lutz et al., 2018; von Freyberg et al., 2018; Wilusz et al., 2017), and thus it can be expected that Fyw changes over time in a catchment with a variable hydro-climatic regime such as the Weiherbach. Thus, the temporal changes in Fyw that have been identified in the present study are likely related to the hydro-climatic conditions at the site, however, the scientific analysis of these relationships often remains too superficial (which is surprising, given that the Weiherbach catchment is an intensively studied research site). As such, this study does not teach us something new about the catchment but rather shows that different tracer time series provide different young water fractions. Although it is interesting to quantify these temporal differences in Fyw, it remains to be tested how these findings for the 0.385km2 Weiherbach catchment are transferable to other landscapes and climates.

First, to avoid any misunderstandings we would like to make clear that our study area is the Wüstebach catchment and not the Weiherbach catchment as indicated by the reviewer. The primary focus of this study was not to identify catchment influences on Fyw but to investigate the extent and significance of temporal Fyw changes using variable 1-year time series of isotope tracer data. We primarily aim to improve the robustness of the Fyw method and not to analyze specific Wüstebach influences on Fyw. This is essential information for planning future sampling strategies in other catchments and for the layout of catchment comparison studies.

We extended the analysis of hydro-meteorological data and found evidence of a possible unique meteorological situation in summer 2015. Without going into too much detail, presently several data (hydrologic, meteorological and isotopic) indicate this special situation. The analysis is not yet finalized. The finished version will be incorporated into the manuscript and if possible, we will give guidelines for a more robust Fyw estimation.

Even if in the end no clear recommendation for a sampling time can be given, a recommendation was given: estimating Fyw with data from a single year is not enough (page 12, 25f); the time-variant Fyw for a catchment should be calculated to understand the behavior and uncertainty for a given location (page 13, lines 2ff).

The reviewer already mentioned that previous studies found Fyw reacts to changes in precipitation, discharge, and other factors. Thus it is safe to assume that other catchments also have a time-varying Fyw and applying our method would yield more information about the Fyw behavior and uncertainty in each catchment. We highly suggest conducting the same study for other catchments in other landscape or climatic units to be able to generalize the findings.

We will add these discussion points to a revised version of the manuscript.

Major comments:
We first answer major comment #6 as it is of great importance:

6. At the very end of the Discussion section the authors state that a previous analysis has been carried out that used a 3-year isotope times series from the Weiherbach catchment. This previous study already showed that the Fyw values differed substantially between three 1-year periods (Stockinger et al., 2017, data in the supplement). In the present study, the authors simply repeat this analysis with a 4.5-year isotope data set from the same site knowing that their hypothesis "(1) Fyw does not deviate more than _2% from the mean of all Fyw results indicating long-term invariance [: : :]" will likely be rejected. I was surprised to read about a very similar previous study at the very end of the current manuscript and wondered why is it necessary to repeat the analysis when the result (rejection of hypothesis 1) is already known?

The focus of Stockinger et al., 2017 was to correct canopy-induced isotope changes in throughfall for the complete time series. 3 individual years were cut out as a test without going into further detail. We extended this test into its own study to focus on investigating the extent and dynamics of the time-variance of Fyw, which was not the focus of Stockinger et al., 2017 and would have been out of scope for the previous study.

We had no possibility to know the results beforehand since for hypothesis 1 more than 90% of Fyw results must be within ±2% (page 6, line 10). It is not possible to estimate from 3 out of 189 results if approximately 19 results (10%) would be outside the ±2%.

1. One of my largest concern is that the presented analysis did not provide any information on the uncertainties of the individual Fyw estimates. Only the adjusted R2 values of the sine fits are presented. Previous studies that calculated young water fractions for several other catchment reported uncertainties in Fyw between 1% and 41% (e.g., Jasechko et al., 2016; Stockinger et al., 2016; von Freyberg et al., 2018). Thus, it should be tested whether the individual young water fractions that were calculated from the 1-year time series are indeed statistically significantly different from each other when their uncertainties are considered. Looking at the low adjusted R2 values for the July 2014-October 2015 period (e.g. Figure 4), I would expect the uncertainties of the 1-year Fyw values to be rather large. However, instead of analyzing the uncertainties in Fyw, the authors mainly focus on the time-variability of the individual 1-year Fywvalues and conclude rather boldly (P12L22) "The obtained Fyw could be a potential outlier, a larger value or part of the Fyw baseline". I would argue that the uncertainty in Fyw (e.g., expressed as standard error) would allow us to objectively judge whether we can believe our Fyw estimates or not. Such an analysis is, however, missing here. In fact, knowing the uncertainties of the individual 1-year Fyw values would allow a more informative analysis of how the Fyw-uncertainty (not Fyw itself) is controlled by hydroclimatic conditions. Such an analysis might provide concrete guidelines for planning targeted sampling campaigns to robustly estimate Fyw.

We thank the reviewer for pointing out this very important issue. It is the aim of this study to present a generic method to analyze Fyw for time-variance and thus improve the robustness of the Fyw method. We added uncertainty estimates of Fyw using Gauß error propagation (preliminary Figure R1):

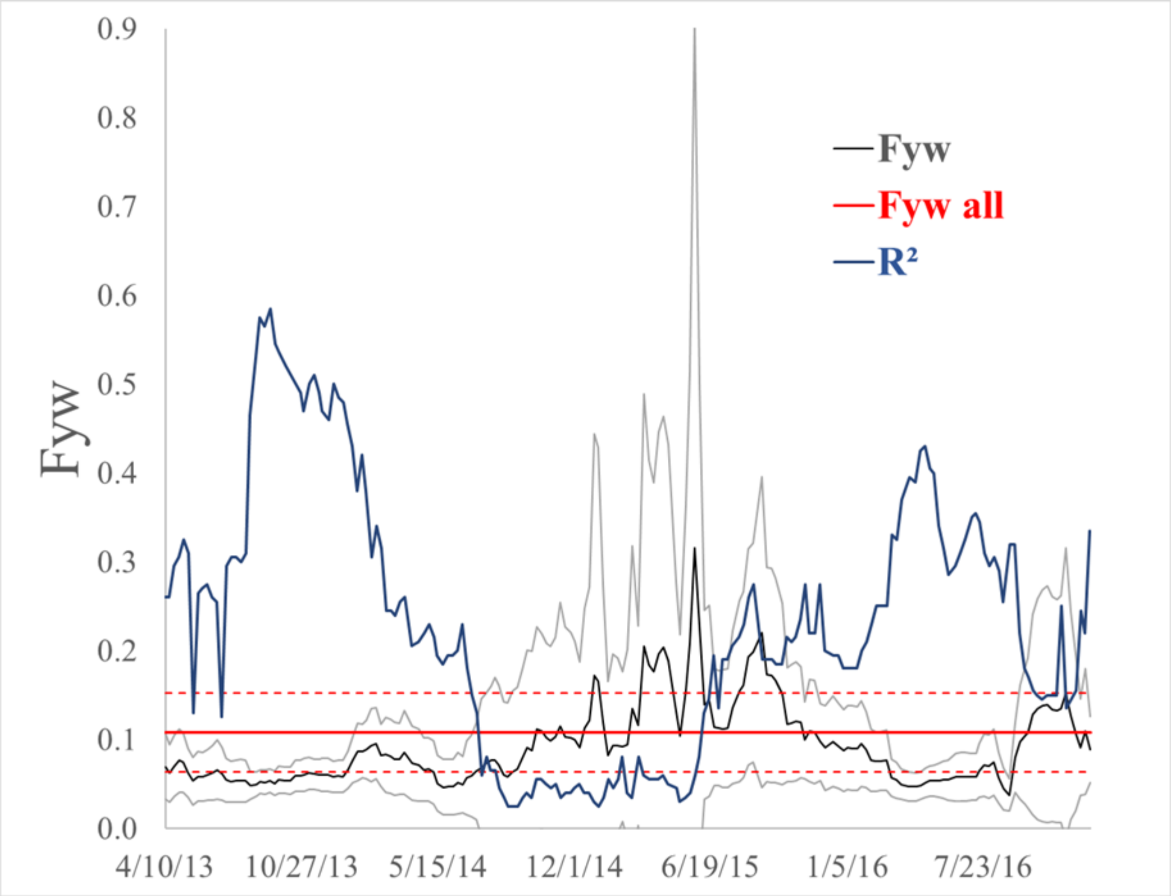

Figure R1. 189 Fyw results (black) and uncertainty (grey) compared to Fyw for all data (red, solid line) and respective uncertainty (red, dashed line). Additionally plotted is the adjusted R^2 (blue).

The following can be said from this result:

a) with a drop in R^2 below approx. 0.2 the uncertainty increases drastically. This, together with the strongly fluctuating Fyw results (page 11, lines 6-8), indicates that in the Wüstebach an R^2 of at least 0.2 should be reached. We highly recommend conducting similar studies in different catchments to test whether different R^2 threshold values exist in other catchments.
b) Fyw of all data (Fyw all in Figure R1) had an uncertainty of appr. ±4%. We will use this new data-driven value instead of the ±2% for re-evaluating our hypotheses.
c) the Fyw results become highly uncertain during 2014/2015. From our preliminary analysis mentioned above it seems that special meteorological conditions during summer 2015 are responsible.

2. Furthermore, given that the uncertainties in Fyw values can potentially be much larger than 2% (as it was shown in the previous studies cited above), to me the 2% threshold seems too low and the authors' justification for that 2% threshold is not convincing.

We will use 4% based on the uncertainty estimation of using all data, see comment above.

3. The hydro-meteorological conditions in the Weiherbach catchment were highly variable during the 4.5-year study period. For instance, only the winter 2013/2014 was snow-free in contrast to the other

winters when a snowpack built up (Sect. 3.1). In addition, 21% "…of the forest were clear-cut in August/September 2013: : :" (P3L27), which significantly altered the streamflow regime of the Weiherbach creek (Wiekenkamp et al., 2016). In Fig. 8d (P24) we find that the runoff coefficients for the Weiherbach catchment ranged between roughly 0.8 and 1.25, suggesting that hydro-climatic conditions at the site varied considerably over time. The authors do not, however, provide any data or figures that present the hydro-climatic conditions during the study period except for the scatter plots in Figure 8, which contrasts 1-year averages of four hydroclimatic metrics with the respective Fyw-values. Despite the highly variable streamflow regime of the catchment and the authors citing another study where flow weighting of the streamwater isotope values resulted in "…significant changes in Fyw…" (P2L4), the authors should more thoroughly investigate how catchment wetness might control Fyw. Why was streamflow-weighting not done here? Why was there no further analysis of potential factors that may control the large variability in 1-year Fyw values, particularly in the period July 2014-October 2015? It seems likely, that individual storm events may have had strong effects on the discharge of young water, so it may be useful to investigate extreme events rather than average behavior.

In a revised version we will present the hydroclimatic data, including catchment wetness conditions expressed as a mean soil water content, either as a figure or as a supplementary. As mentioned above, we already started a more extensive analysis of hydrometeorological conditions and their possible impact on Fyw and its uncertainty. It is likely that special conditions during summer 2015 influenced Fyw. The investigation is still ongoing.

Precipitation and streamflow weighing were both done for Fyw calculation, but unfortunately not clearly mentioned in the submission.

4. Sect. 3.5 and Figure 9: It is not clear to me how the Fyw values for testing hypothesis 3 (seasonal invariance) were determined. As far as I understood, Fyw-values were calculated for 189 1-year periods (Sect. 2.3). How were month-specific Fyw-values extracted from these annual Fyw-values? Wouldn't each 1-year Fyw-value be affected by the isotope values of all 12 months that comprise this 1-year period? If so, I doubt that the analysis presented in Sect. 3.5 and Figure 9 provides useful information.

No monthly Fyw were extracted from the data. Each of the 189 Fyw results was assigned to the date that lies in the middle of the calculation period. For example, Fyw was calculated from 1 January to 31 December and the corresponding result was assigned to 1 July.

We then grouped all Fyw results according to the month they were assigned to. All results in Fig 9 are still 1-year calculation results. Should a seasonal trend be observable, one could argue that e.g., a 1-year sampling campaign centered around July would lead to higher/lower Fyw estimates compared to when it is centered on March.

5. Part of the analysis presented in "4.2 Fraction of young water" is not valid. First, the authors calculated Fyw from the entire 4.5-year data set (Fyw,4.5=10.8%) and compared this to the average of the 186 1-year Fyw values (9.3%), concluding that both values are similar with regard to their 2% threshold. A second comparison was carried out with Fyw,4.5 and the average of a much smaller number of 1-year Fyw values that neglects the Fyw values from the period July 2014-October 2015 (7.5%). This second comparison should, however, use another Fyw value as a reference based on the same isotope data set (i.e., 4.5 years minus the period July 2014-October 2015)- otherwise the authors compare apples with oranges.

We thank the reviewer for pointing this out. This is absolutely correct and we will change it accordingly.

Minor comments:

P3L8-9: “However, it remains to be tested how sensitive the Fyw method is towards the timing and the length of the available data.” Why does this need to be tested? Can you provide an example of where the length and the timing of the isotope data resulted in different Fyw values? Otherwise, a clear motivation for your analysis is missing.

We will cite Stockinger et al., 2017 here (see also answer to major comment #6) where a multi-year Fyw as well as three individual years Fyw were calculated and differences found.

P4L25: “Because of this on average 43 isotope values were available for precipitation compared to 53 values for streamflow.” Does this average refer to a 1-year period? Please clarify. It would also be nice to provide the total number of streamwater and precipitation samples of the entire 4.5-year period.

Yes, this refers to the 1-year calculation periods. The total number of P and Q samples (156 and 195, respectively) is mentioned on page 7, line 21 but we will move it here.

P5L22-23: I would suggest to move these two sentences to the beginning of the chapter to make clear where the number “189” comes from.

We agree.

P5L13: 24*365.25 is 8766 not 1/8766

We will rephrase this “i.e., if CP(t) and CS(t) are calculated in hourly time steps then the frequency f is 1/8766; once per 24 x 365.25 hours).”

P5L31-32: What do you mean with “the timing of peaks and the individual amplitudes”? Do you refer to the isotope time series or to the fitted sine functions?

We referred to the fitted sine functions and suggest the following change to the sentence

“We also calculated Fyw for the whole time series with one sine wave and compared its peak timing and amplitude to the timing of peaks and amplitudes of the 189 sine waves.”

P6L3: Here you switch units of Fyw (0.02 and 2%). Also, in the text you express Fyw in percent, whereas in the figures you use the scale from zero to one. Please be consistent throughout the manuscript.

We will consistently use e.g., 0.02 instead of 2%.

P7L13-14: Please be more specific about what water isotopes you are talking about, e.g. add δ18O.

Page 7 line 13 already featured “δ18O” at the end of the first sentence:

“Precipitation isotope ratios ranged from -3.04 to -17.80‰, spanning a range of 14.76‰ in δ18O values.”

P8L28-30: Please provide some metrics for the strength of these correlations (e.g., Pearson correlation coefficients).

We will add statistical information to the text, e.g.:

The equation for the whole data set (including the low R^2) is
Runoff Coefficient = -1.24 * Fyw + 1.13, with an adjusted R^2 of 0.30 and p-value of 3E-16

The equation for the limited data set (low R^2 excluded) is
Runoff Coefficient = -2.11 * Fyw + 1.19, with an adjusted R^2 of 0.23 and p-value of 2E-7

P8L29: Was the runoff coefficient calculated with catchment-average precipitation or throughfall? I would suggest to add the runoff coefficients to Fig. 3 since the relationship between Q/P and the sine wave fits to the isotope data are discussed in Sect. 4.1.

The runoff coefficient was calculated with throughfall but we plan to change it to open precipitation to enable comparability to other studies. Since we started an extensive analysis of hydrometeorological data that features the runoff coefficient but is not yet finished (see also mentions above), we will present the runoff coefficient more prominently in the revised manuscript, but not necessarily as part of Figure 3.

P9L16-20: You suddenly present groundwater isotope data without providing information about the source (location, sampling procedure, number of samples etc.) of these data. Please include this information into Sect. 2.2.

This was on oversight on our part and we will add it in a revised manuscript version.

P10L13: "The double-peak in precipitation of autumn 2015 was not found in streamflow (Figure 3)." Do you refer to the δ18O in precipitation and streamflow or to the sine fits to the isotope data?

To the sine fits, we adapted the sentence:

"The sine wave double-peak in precipitation of autumn 2015 was not found in streamflow (Figure 3)."

P11L33: "Thus, during the 4.5-years Fyw never fell below the baseline of 5% […]" This statement is incorrect. Figures 6 and 7 clearly show that Fyw fell below 5% on several occasions, such as around June 2014 and September 2016.

We will adapt the sentence "Thus, during the 4.5-years Fyw seldom fell below the baseline of 5% […]"

P12L5: "The variability in Fyw of this study could not be explained by most meteorological or hydrometric variables". Could a lack of correlation be explained by the large distance (3km) of the meteorological station to the study site? What about median values of the hydro-climatic variables or metrics that describe extreme events?

Correlations of precipitation amounts (R^2 = 0.95), temperature (R^2 = 0.99) and relative humidity (R^2 = 0.94) of the 3 km distant climate station with the respective climate data from the clear-cut area of the Wüstebach catchment showed good R^2. We did not use the on-site climate station for our study as its data does not cover the full study period.

We are currently in the process of re-evaluating and extending our analysis of hydrometeorological data (as mentioned above). Several newly analyzed data point to influencing Fyw during summer 2015. We will incorporate and test median values and extreme event metrics.

P12L9: "…the different sampling periods of all mentioned studies…". This contradicts a previous statement: "…Lutz et al. [2018] used the same sampling period for precipitation and streamflow for all 24 investigated catchments." (P11L25).

We will correct this to "most of the mentioned studies".

P12L23: “As the violation of hypothesis 2 did not correlate with any meteorological or hydrometric data : : :”. How can a violation correlate with anything? Please clarify.

We referred here to the timing of the violation of hypothesis 2: if the timing could be connected to hydrometric or meteorological data.

We adapted:

“As the timing of the violation of hypothesis 2 did not[…]”

Figures: The date formats in all figures are confusing. Does 4/10/13 mean 4th October 2013 or 10 April 2013? Also, I would suggest to have each tick mark at the first of the month and to have consistent date axes in all figures.
4/10/13 refers to April 10th, 2013. We will adapt the figures to uniformly start on 4/1/13 with the exception of Figure 2 (just a theoretical example) and Figure 3 (showing the input data and starting on a different date than the Fyw result figures; see also explanation below).

Figure 4: This figure misses a proper legend (e.g., What does “Mean” stand for?). The unit and numbers of Fyw on the right vertical axes don’t match. Do panels a and b share the same legend? Why are the shown time series much shorter than 4.5 years?

We will change the legend entries to “Mean R^2”, “TF R^2” and “Q R^2”.
We are not sure about the reviewer comment regarding the right axis having mismatching units and numbers for Fyw. The only right-hand axis is in Figure 4b and shows Amplitudes (in permille) and not Fyw (in percent). In a revised version we would remove [%] from the Fyw-axis to avoid the misleading conclusion of “0.3% Fyw”.

Panels a and b share the same legend.
The time series are shorter than 4.5 years since each Fyw result was placed in the middle of the year it was calculated for. The time series starts on 10/10/12, thus the first Fyw result is placed on 4/10/13. Doing this cuts off the first half year and the last half year of the complete time series, explaining the shortening.