In the following text please find the corrections and comments to the referee's response (for better understanding, comments from the referees were copied are black and our comments in blue).

Replies to Referee #1 (Markus Hrachowitz)

General comment: The manuscript "Understanding mean transit times in Andean tropical montane cloud forest catchments: combining tracer data, lumped parameter models and uncertainty analysis" by Timbe et al. explores differences in transport processes in different parts of the hydrological system. Let me upfront say that in spite of the comparatively simplistic (i.e. time-invariant) modelling approach I do really like the approach taken by the authors in the manuscript under review as it provides and analyses an interesting tracer data set from different hydrological components such as soil, springs, etc. and provides some level of experimental insight on the differences in transport processes between these components. I do have, however, some comments that I would encourage the authors to address in detail. The major points being that firstly I think that the manuscript is too much focused on the choice of models themselves rather than on what these models can tell us about the underlying processes and the way water is routed through the system. Secondly, I found that the methods section is kept very superficial and requires some more attention and detail.

General reply: We are grateful for the valuable comments provided and therefore thank the reviewer. As recommended, we have condensed the analyses of results and discussion on the models, and included an analysis of processes that each retained model represents. Besides, we expanded the method section.

Specific comments:

1) **p.15872, l.26ff:** this is quite an unusual definition of what a TTD is. And I am not entirely sure it is correct.

Definition has been changed to:

"Whereas the MTT describes the average time it takes for any given water parcel to leave the catchment, the transit time distribution function (TTD) describes the retention behavior of all those water parcels as a frequency function over time (McGuire and McDonnell, 2006)."

2) **p.15873, l.15ff:** where do the 12 years come from? This seems a bit too specific. In addition, there is a wide range of other methods than carbon dating for older waters. Thus maybe rather say "[…], while, for example, carbon isotopes are employed […]

Sentence has been change to:

"For longer MTTs of up to 200 years (Stewart et al., 2010), tritium radioisotopes are used to analyze the storage and flow behavior in surface water and shallow groundwater systems (Kendall and McDonnell, 1998), while, for example, carbon isotopes are employed for analyzing the dynamics of deep groundwater with ages of hundreds to thousands of years (Leibundgut et al., 2009)."

The following reference has been added:

“Stewart, M. K., Morgenstern, U. and McDonnell, J. J.: Truncation of stream residence time: how the use of stable isotopes has skewed our concept of streamwater age and origin, Hydrol. Process., 24, 1646–1659, doi:10.1002/hyp.7576, 2010.”

3) p.15873, l.20: I do not think that these methods can be called “traditional” when it comes to tracer routing. These are quite recent developments, really, compared to the use of the convolution integral technique used here. The first and for a long time only ones who did it were to my knowledge Barnes and Bonell (1996)**.**

Sentence has been changed to:

“Since Barnes and Bonell (1996), researchers in tracer hydrology use quasi distributed and…”

Besides, the following reference has been added:
“Barnes, C. J. and Bonell, M.: Application of unit hydrograph techniques to solute transport in catchments, Hydrol. Process., 10, 793–802, doi:10.1002/(SICI)1099-1085(199606)10:6<793::AID-HYP372>3.3.CO;2-B, 1996.”

4) p.15873, l.22-28: Good point! But not the MTT itself is of primary importance here (and elsewhere in the manuscript) as it is just a very reductive metric. It is rather the shape of the TTD that is of interest as it gives information about the underlying mixing processes and the way water is routed through the system.

We acknowledge this fact and revised the relevant sections in the manuscript.

5) p.15874, l.3: the terms “more recently” and “new” seem a bit out of place for a paper that has been published almost one and a half decades ago.

We have changed the sentence to:

“Since almost one and a half decades ago, other lumped models are …”

6) p.15874, l.3-8: it would be good if this could be put more into context of actual hydrological function. Why are TPLR and GM more flexible? What can they do better? For example: they allow the representation of different mixing processes in different system components, such as soil and groundwater. In contrast, EM-based models assume instantaneous and complete mixing over the entire model domain, which is only likely in few, if any, surface water systems (see e.g. Hrachowitz et al., 2013).

The following sentence was added to the section:

“The advantage of the latter functions relies on that they allow the representation of different mixing processes in different system components, such as soil and groundwater. In contrast, simpler models assume instantaneous and complete mixing over the entire model domain (Hrachowitz et al., 2013).”

And the following reference has been added:

“Hrachowitz, M., Savenije, H., Bogaard, T. A., Tetzlaff, D. and Soulsby, C.: What can flux tracking teach us about water age distribution patterns and their temporal dynamics?, Hydrol. Earth Syst. Sci., 17, 533–564, doi:10.5194/hess-17-533-2013, 2013.”

7) p.15874, l.17-21: seems to better fit into the methods section

We agree with the reviewer and moved the respective sentence to the method section (Section 2.5).

8) p.15875, l.7: In science, except mathematics, it is close to impossible to verify or confirm hypotheses (Popper, 1959). In addition, as hydrology is an inherently inexact science it may frequently also prove difficult to reject hypotheses simply due to inadequate, i.e. scarce or erroneous data (e.g. Beven et al., 2012). I would thus suggest to replace "confirm or reject" by the more neutral "test"

Sentence was modified as follows:

"Translated into hypotheses the study reported in this paper aimed to test if"

9) p.15875, l.9-20: not sure this is correct. How did you test if tracers are conservative?? How did you test that there are no stagnant waters? How did you test that stationary conditions are dominant? The use of lumped equations does tell you very little about that. They can also be fit to a non-stationary system, trying to get the best average fit. It seems to me as if in this paragraph the authors mixed assumptions with hypotheses they wanted to test.

To clarify, we now differentiate between hypothesis we want to test and the assumptions we have made in order to conduct this study, additionally we include references for each assumption.

Hypotheses:

1. "the diversity of the sampling sites allows evaluating the spatial variability in catchment hydrology, identifying the dominant processes, and screening the performance of the TTD models;"

2. "the multi-model approach and the identifiability of their parameters enable identification of the respective TTDs and MTTs."

Assumptions:

1. "the used tracers are conservative, there are no stagnant flows in the system, and the tracer mean transit time τ represents the MTT of water (e.g. McGuire and McDonnell 2006);"

2. "stationary conditions are dominant in the basin and lumped equations based on linear or quasi-linear behaviors are applicable (Heidbüchel et al., 2012);"

3. "from insights derived of related studies (Soulsby et al., 2010; McGuire and McDonnell 2006, Rodgers et al., 2005), considering the drainage areas, the steepness of the topography and the shallow depth of the soil layers, the transit times of the sampling sites are less than 5 yr, making it possible to use δ^2H and $\delta 1^8O$ as tracers."

10) p.15877, l.5: should read as "Major"

Change performed

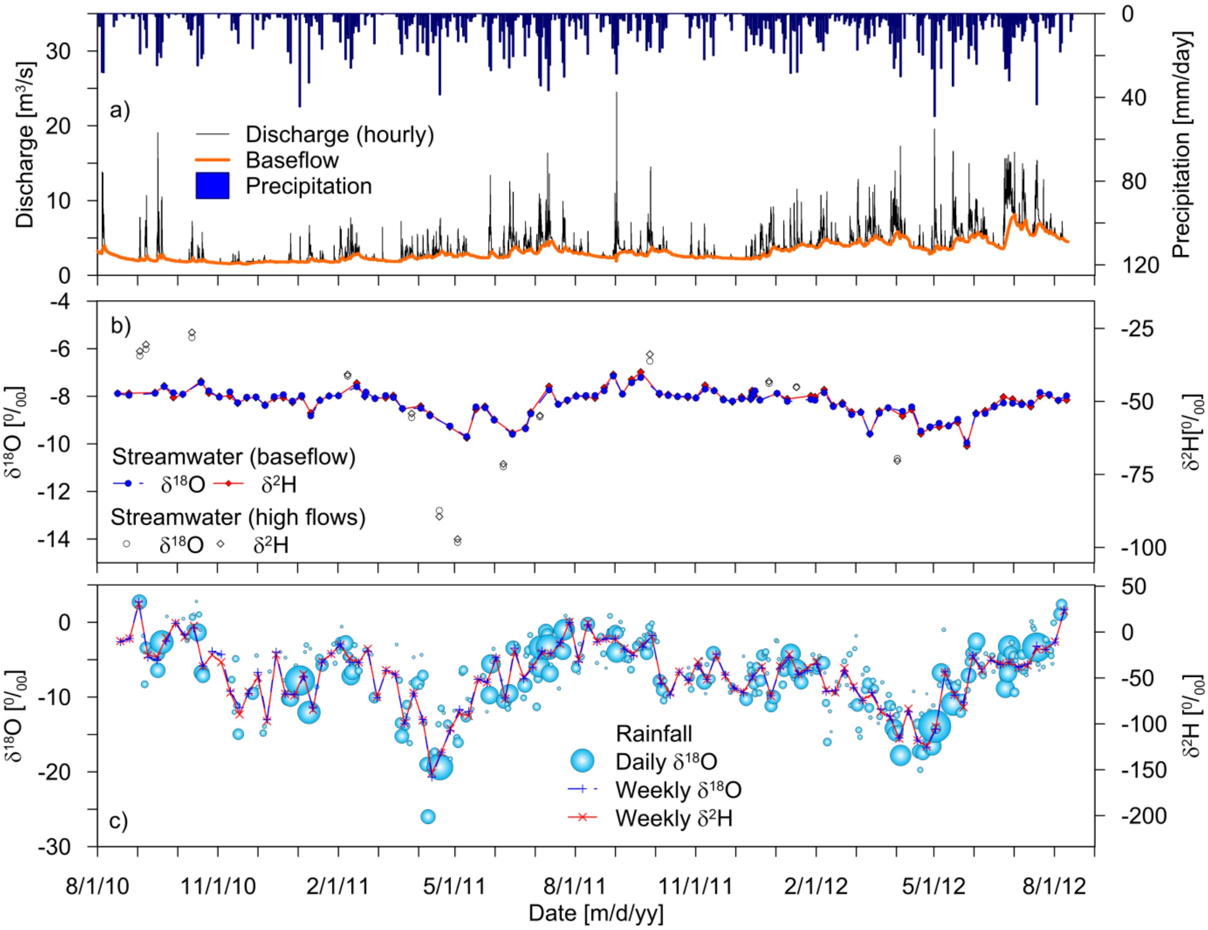
11) p.15877, l.16ff: there is an entire paragraph about streamflow observations. It is however not clear what it is needed for in this study except to define base flow conditions. Can be largely condensed.

The referred paragraph has been condensed accordingly to what is needed to define base flow conditions. Now the paragraph reads as follows:

"The San Francisco catchment was subdivided into seven sub-catchments with areas ranging between 0.7 and 34.9 km^2, characterized by different land uses varying from pristine forest and sub-páramo to pasture areas (Fig. 1 and Table 1). In order to define baseflow conditions, each sub-catchment was equipped with a water level sensor (mini-diver, Schlumberger Water Services, Delft, NL). Reference discharge measurement, using the salt dilution method, where made frequently during the time of sampling. However, due to the high variability of the river bed for the sites QP, QZ and QR, only continuous records for sub-catchments FH, QN, QM, QC, and for the main outlet PL were considered as reliable to calculate stage-discharge curves and the hydrographs, as shown in Fig. 2a for PL. For the remaining sites, discharge measured at the moment of sampling was used."

Please notice that Figures and Tables have been checked for their correct sequential order (as requested by Referee#2). In this sense, figures 2a and 2c are now 2c and 2a respectively, while Table 1 is now 2 and former Table 2 is now 1.

New version of figure 2:



*"**Fig. 2. (a)** Time series of rainfall for ECSF meteorological station, hourly discharge and baseflows at the catchment outlet (PL); **(b)** weekly $\delta^{18}O$ and $\delta^{2}H$ of streamwater at PL for baseflow and high flow conditions; and **(c)** weekly $\delta^{18}O$ and $\delta^{2}H$ at the ECSF rainfall sampling collector; light blue bubbles indicate daily $\delta^{18}O$ and relative volume of daily rainfall."*

12) p.15878, l.21-23: how were the rain event samples obtained? Automatic sampler? An eager student who changed the sampled bottle after each event? Does this also mean that there were samples that spanned for example 2 hour periods, 13 hr periods or, whatever, 4.32 day periods?? What did you do if there were 2 or more events during one day? Conversely, what did you do when there was a 2 day rain event? Please provide details.

We revised section 2.3, not only to explain the present comment but also the comment #13. Now Section reads as follows (find in the last paragraph the information for this comment):

"Weekly water samples for isotope analysis were collected manually in the main river (Fig. 2b), its tributaries, creeks and springs in the period August 2010 to mid-August 2012 and later for soil water starting in September/November 2010 (Table 2), using 2 mL amber glass bottles. Soil water sampling was performed along two altitudinal transects covered by forest and pasture (Table 2), in 6 sites (Fig. 1) and 3 depths (0.10, 0.25 and 0.40 m) using wick-samplers. Wick-samplers were designed and installed as described by Mertens et al. (2007). Woven and braided 3/8 fiberglass wicks (Amatex Co. Norristown, PA, US) were unraveled over a length of 0.75 m and spread over a 0.30 m × 0.30 m × 0.01 m square plastic plate. The plate enveloped with fiberglass was covered with fine soil particles of the parent material and then set in contact with the undisturbed soil, respectively at the bottom of the organic horizon (0.10 m below surface), a transition horizon (0.25 m below surface) and a lower mineral horizon (0.40 m below surface). The low constant tension in the wick-samplers guarantees sampling of the mobile phase of soil water, avoiding isotope fractionation (Landon et al., 1999)."

"Along with the weekly sampling, event based rainfall samples for isotope analyses were collected manually in 1 L bottles using a Ø 25 cm funnel at 1900 m a.s.l. (Fig. 1). After every event, the sample bottles were covered with a lid and stored for analysis within a week in amber glass bottles 2 mL. Only sample volumes > 2 mL were suitable for permanent storage and measurements. Events with a sample volume below 2 mL were discarded. The end of a single rainfall event was marked by a time span of 30 min without rainfall, whereby a total of 946 samples were collected with an average duration of 3.2 h (varying from 0.25 to 19 h with up to 11 events per day). Since the solving of the convolution equation needs a continuous time step of input data (Maloszewski and Zuber, 1982), the time resolution of the input series was set to 7 days (Fig. 2c). In this sense, weekly mean isotopic signatures for smaller rainfall events during longer dry periods (only 5 among 104 weeks had no rainfall event > 2 mL sampling volume) were interpolated using antecedent and precedent measurements."

" The final isotope signature used for the models represents:

- *for rainfall water, the weighted mean of all events during each week (Sundays to Saturdays) using the rainfall data recorded at the nearby meter logical station (400 m to ECSF),*
- *for soil water samples, the weekly average isotope signal for each soil depth, and*
- *for stream, creek and spring water samples, an instantaneous isotopic concentration in time. These samples were not flux-weighted. For stream waters, only isotope samples from designated baseflow conditions were considered (see Section 2.5)."*

13) p.15879, l.1-12: as tracer input "concentrations" always need to be flux weighted, I was wondering which precipitation amount was used for each elevation zone. Did you use the catchment average rain for each elevation zone? Or did you rather use some kind of elevation corrected precipitation for each zone? That could make quite a difference in the catchment averaged input signal!

We used the altitude gradient calculated by Windhorst et al. (2013) to extrapolate the isotope signals from samples collected at ECSF. The samples collected at ECSF were weighted taking the weekly rainfall amounts collected at ECSF meteorological station (See section 2.3). In other words, no correction according to variation of rainfall amounts along the catchments was performed. This approach was adopted based in the following:

- Although rainfall volumes vary between meteorological stations, there is a high correlation (at a daily or weekly time basis) among volumes registered between stations. Meaning that even if we had used the rainfall amounts from another station, the final weighted isotopic values would have remained similar to the ones that were weighted with ECSF station data.

- The gradient calculated by Windhorst et al. (2013) states that only the altitude effect is significant and that in this factor there is no influence of temperature, relative humidity and precipitation amount or intensity.

In order to clarify, we modified the first paragraph of Section 2.4. Now it reads as follows:

"Throughout the catchment, the recorded rainfall time series from meteorological stations are correlated (based on weekly precipitation data r^2 was at least 0.6). As the models in question are only driven by the isotope signal and not the actual amount of incoming precipitation on site, a flux weighting based on a single station within the catchment (ECSF) was sufficient. Given the large altitudinal gradient in the San Francisco basin, it is to be expected that the input isotopic signal of rainfall for every sub-catchment varies according to its elevation (Dansgaard, 1964). In this regard, Windhorst et al. (2013) estimated this variation for the main transect of the catchment: -0.22‰ $\delta^{18}O$, -1.12‰ $\delta^{2}H$ and 0.6‰ deuterium excess per 100 m elevation gain. Applying this altitude gradient to the flux weighted isotope signal under the assumption that the incoming rainfall signal is the sole source of water, thereby excluding any unlikely source of water from outside the topographic catchment boundaries with a different isotope signal, it was possible to derive the recharge elevation and localized input signal in each sub-catchment. The derived recharge elevations were used to crosscheck that they are inside the topographic boundaries of every sub-catchment and comparable to their mean elevations."

We also added the following paragraph:

"The justification to adopt only the mentioned gradient to extrapolate the isotope signals, was based in previous studies on spatial and temporal variation of stable isotopes of rainfall in the same catchment, which revealed that, only the altitude effect is significant and that in this factor there is no influence of temperature, relative humidity and precipitation amount or intensity (Windhorst et al, 2013)."

14) p.15880, l.1-2: please justify in a bit more detail.

We changed the referred sentence by the following in which explanation now is more detailed, it reads as follows:

"In contrast, all spring and creek water samples were included in the analysis since their isotopic signatures were less influenced by particular rain events (as inferred from the smooth shape of the observed isotope signal) in the San Francisco catchment. In regard to soil water, we considered all samples, since each sample represents a volume weighted weekly average signature (isotopic signatures of particular high rainfall events are smoothed at a weekly time span)."

15) p.15881, l.18ff: this is the greatest mystery for me in this manuscript. Why would the authors choose to dismiss an interesting high-resolution data set to aggregate the available observed daily input data to weekly values??? That is quite an amazing waste of valuable information. Even if rainfall events spread over two or more days, a uniform input distribution over this period can be assumed. I am pretty sure that the uncertainty introduced by that assumption is easily more than compensated for by the gain of additional information.

The time scale adopted for this study is now justified in the Section 2.3, please see reply to comment#12 (last part of the second paragraph).

16) p.15881, l.22-23: it does not matter which signatures are used in precipitation free periods as the input tracer signal needs to be weighed by the respective precipitation volume which per definition of precipitation free periods equals 0.

See comment 12 (explanation now included in Section 2.3).

17a) p.15882, l.1: I would be glad if you could add the reference Hrachowitz et al. (2011) as an overview of different methods was provided therein.

Suggested reference (see below) has been placed in the referred sentence.

"Hrachowitz, M., Soulsby, C., Tetzlaff, D. and Malcolm, I. A.: Sensitivity of mean transit time estimates to model conditioning and data availabilityle, Hydrol. Process., 25, 980–990, doi:10.1002/hyp.7922, 2011."

18a) p.15881, l.16ff and table 3: The methods are not described concretely enough. The equations are fine, but how exactly was the stream concentration computed? Concentrations are measured weekly (and instantaneously in the stream but as a volume weighted average in precipitation and soil(?)) but water fluxes are measured more frequently, so what was done to distribute concentrations over time?

Explanations about computation of stream concentrations are now given in Section 2.3. Besides we have added the following paragraph at the end of Section 2.6:

"Modelled output results are available for the weekly time span chosen for the input function (an average signal of rainfall was distributed for every week at Wednesdays 12:00). These results were interpolated in order to perform statistical comparisons with instantaneous observed data. For soil waters, direct comparisons were performed between predictions and observed data."

… In addition, how was the gamma function integrated (since it goes to infinity at t=0 when alpha is less than one)?

Regarding to the prior parameter distributions, uniformly distributed pseudo random numbers were generated using the RAND function in Matlab, as this function describes numbers in the open interval between 0 and 1, they were scaled to a predefined range according to the approach mentioned below (e.g., to avoid convergence problems of the gamma function close to 0 the parameter range was limited from 0.0001 to 10, see Table 3, that now includes the wide ranges we took for a first simulation stage). The calculation of weighted quantiles were performed using R, the script used for this step is available and fully described at www.paramo.be (under: Software/Hydrological data analysis and modelling/Uncertainty analysis/GLUE analysis). Using these limits, a final simulation was performed in Matlab (at this stage the 10,000 simulations were allowed to vary only for these final solution ranges). For the mentioned approach we did not use any specific NSE limit. The limit of NSE > 0.45 was only used later in the analysis of results (for comparison between sites and models).

We added the following paragraph to the Section 2.7 in order to clarify the used approach:

"When looking for the optimum parameter range, we first set a wide range (maybe even unrealistic) to be sure to cover all possible solutions (Table 3). By checking the plots of these preliminary results we were able to identify the convergence of model solutions (we used NSE as the objective function for all model parameters), thereby making it possible, for a second simulation, to narrow down the parameter range for each variable. Once the variation ranges were identified and bounded, according to the largest solution peak for every site and for every variable, all the solutions 5% below the top NSE efficiency were selected. For these behavioral efficiencies, weighted quantiles between 0.05 and 0.95 (90% prediction limits) were calculated in order to refine limits of behavioral solutions for every variable. Using these limits, a final simulation for each site and model was performed (at this stage the 10,000 simulations were allowed to vary only for the corresponding final solution ranges). Results are shown in Tables 4 and 5, as well as in Annexes 1 and 2.

We also modified the second paragraph in Section 2.7:

"The before mentioned approach is based on the Generalized Likelihood Uncertainty Estimation (GLUE, Beven and Freer, 2001). The GLUE approach considers that several likely solutions are valid as long as efficiency of a particular simulation is above a pre-set, but subjective threshold. In this sense, considering the large number of sites and models used, no specific lower limit was set to discriminate predictions, but (as explained earlier) a range that depended on the top efficiency for each case. Only for the analysis of results and for intercomparison between predictions, we considered that a prediction was poor for NSE < 0.45."

As mentioned, table 3 has been modified to show the initial range of variability for the model parameters:

*“**Table 3.** Lumped parameter models used for the calculation of the transit time distribution.”*

Model	Transit time distribution g(τ)	Parameter(s) range
Exponential Model (EM)	$\frac{1}{\tau}\exp\left(\frac{-t}{\tau}\right)$	τ [1-400]
Linear Model (LM)	$\frac{1}{2\tau}$ for $t \leq 2\tau$ 0 for $t > 2\tau$	τ [1-400]
Exponential Piston flow Model (EPM)	$\frac{\eta}{\tau}\exp\left(-\frac{\eta}{\tau}+\eta-1\right)$ for $t \geq \tau(1-\eta^{-1})$ 0 for $t < \tau(1-\eta^{-1})$	τ [1-400] η [0.5-4]
Linear Piston flow Model (LPM)	$\frac{\eta}{2\tau}$ for $\tau-\frac{\tau}{\eta} \leq t \leq \tau+\frac{\tau}{\eta}$ 0 for other t	τ [1-400] η [0.5-4]
Dispersion Model (DM)	$\left(\frac{4\pi D_p t}{\tau}\right)^{-1/2} t^{-1}\exp\left[-\left(1-\frac{t}{\tau}\right)^2\left(\frac{\tau}{4D_p t}\right)\right]$	τ [1-400] D_p [0.5-4]
Gamma Model (GM)	$\frac{\tau^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)}\exp^{-\tau/\beta}$	α [0.0001-10] τ [1-400] $\beta = \alpha/\tau$
Two Parallel Linear Reservoirs (TPLR)	$\frac{\phi}{\tau_f}\exp\left(-\frac{t}{\tau_f}\right)+\frac{1-\phi}{\tau_s}\exp\left(\frac{-t}{\tau_s}\right)$	τ_s [1-400] τ_f [1-40] ϕ [0-1]

τ = tracer’s mean transit time; η = parameter that indicates the percentage contribution of each flow type distribution; D_p = fitting parameter; α and β = shape parameters; τ_s, τ_f = transit time of fast and slow flows, φ= flow partition parameter between fast and slow flow reservoirs. Units for parameters and their respective ranges are a-dimensional except for τ, which has units of time (for our case it is given in weeks).

18b) p.15882, l.9ff: Methods again. It was stated that the model performance was evaluated on basis of 10000 MC realizations within the GLUE framework. That is fine. Quite some important information is missing however. What were the prior parameter distributions for the models under consideration (ranges, uniform or informed,…)? GLUE requires the definition of some sort of likelihood measure to weight the solutions and to construct the uncertainty bounds (see e.g. Freer et al., 1996). Yet, no mention of that is made. What likelihood measure was used? Nash-Sutcliffe efficiency? Were the solutions not likelihood weighted at all (implying that ALL solutions retained as feasible, i.e. NSE > 0.45 (?), were assumed as equally likely)? Please specify, justify and reference this part in more detail.

Comments referred to the prior parameter distribution for the models under consideration and about the GLUE approach adopted are now fully explained in the Section 2.7 (please see reply to comment 18a).

19) sections 3 and 4: As mentioned above, I really like the general set-up of the study. The result that MTT in soils is by 1 order of magnitude below that of streams and springs is extremely interesting and in addition it lends considerable experimental support to the hypothesis put up by Hrachowitz et al. (2013), that different system components can exhibit substantially similar transport patterns, i.e. TTDs. However, apart from that the results section (but also the discussion) is too much centered on the models themselves. Bear in mind that models are to be seen as mere tools. Thus the tools themselves are discussed. However, it would be much more instructive if more emphasis was given to "what" the use of these different tools can actually tell us about how the different catchments and (more importantly) the different compartments of the system (soils, springs, streams) function. As a very first step I would thus recommend the authors to largely condense the results and discussions of which model performs best, in favour of showing what is the difference between them. In other words, it would be highly interesting to see the actual TTDs of say the best 2 or 3 models for each compartment. In how far are the shapes of the TTDs similar or dissimilar when comparing one compartment to the other. Do for example the TTDs in the soil show different general shapes than in the stream (e.g. delayed peaks as in EPM, DM or GM models with alpha >1)? No matter if the answer is yes or no, it would tell us something quite fundamental about the characteristics of transport and water routing processes in the different components. I would therefore be very glad if the authors would consider adding such an aspect in the results and discussion by showing the shapes of different TTDs in streams, soils and springs and carefully interpret the shapes and discuss in respect with amongst others the results of Botter et al. (2011), Hrachowitz et al.(2013) and Stewart et al. (2010). Please also note that the results concerning soils are fundamentally different from springs and streams, as the soil data characterize the age of resident water stored in the catchment (often termed "residence time distribution") and spring and stream data give the age of water in fluxes (often termed: "transit time distribution). See Hrachowitz et al. (2013) and Botter et al. (2011) for detailed characterizations.

Sections 3 and 4 have been changed accordingly to suggestions. Analysis of TTD have been integrated in 'Results' and 'Discussion' sections and corresponding figures for the retained models are now shown in Figs. 8 and 9 for soil waters and Figs. 14 and 15 for stream, creek, and spring waters. Besides, 'Results' and 'Discussion' sections have been condensed (in regard to which model perform best). As the changes of these sections are extensive, please check the complete modified sections below.

Please notice that the inclusion of new figures in the Results and Discussion sections (Figures 8, 9, 14 and 15) changes the numbering of figures of the previous version of the manuscript (Fig. 7 is now 10; Fig. 8 is now 7; Figs. 9, 10 and 11 are now 11, 12 and 13 respectively). Additionally figures and tables have been checked for their correct sequential order (as requested by reviewer#2), in this sense, figures 2a and 2c are now 2c and 2a respectively, while Table 1 is now 2 and former Table 2 is now 1.

In the new version of the Section 4, the following references have been added:

"Botter, G., Bertuzzo, E. and Rinaldo, A.: Catchment residence and travel time distributions: The master equation, Geophys. Res. Lett., 38, L11403, doi:10.1029/2011GL047666, 2011."

"Hrachowitz, M., Savenije, H., Bogaard, T. A., Tetzlaff, D. and Soulsby, C.: What can flux tracking teach us about water age distribution patterns and their temporal dynamics?, Hydrol. Earth Syst. Sci., 17, 533–564, doi:10.5194/hess-17-533-2013, 2013."

“Roa-Garcia, M. C. and Weiler, M.: Integrated response and transit time distributions of watersheds by combining hydrograph separation and long-term transit time modeling, Hydrol. Earth Syst. Sci., 14, 1537–1549, doi:10.5194/hess-14-1537-2010, 2010.”

“Stewart, M. K., Morgenstern, U. and McDonnell, J. J.: Truncation of stream residence time: how the use of stable isotopes has skewed our concept of streamwater age and origin, Hydrol. Process., 24, 1646–1659, doi:10.1002/hyp.7576, 2010.”

Modified sections:

3 Results

3.1 Soil water

“Of all predictions the best matches of the models, with respect to the NSE objective function, ranged between 0.64 and 0.91 (Fig. 5a). When only the best goodness of fit was considered, the GM and the EPM models performed best in most of the sampled sites (13 from 18), followed by the DM, LM and LPM models (Fig. 5b). Only these models were considered for further mutual comparison. Even when the derived MTT values were similar among the models that best fitted the objective function (Fig. 6a, Table 4 and Annex 1), the LPM model performed best taking into consideration additional selection criteria, as shown in Figs. 6b and 6c. Fig. 7 depicts, for the LPM model applied to site C2, the uncertainty and the range of behavioral solutions for the two model parameters.”

“Considering results from the LPM model (Table 4), differences between observed and predicted values described by the RMSE are up to 1.72‰ and the larger absolute bias accounts for 0.181‰ (Table 4). Bearing in mind the ranges of behavioral solution, MTT results were between 2.3 to 6.3 weeks for pastures soils and between 3.7 to 9.2 weeks for forested soils, while parameterizations for η (ratio of the total volume to the volume in which linear flow applies) ranged from 0.84 to 2.23 and from 0.76 to 1.61 respectively.”

“Regarding to the shapes of the distribution functions, Fig. 8 shows the best matching results for two representative and comparable sampling sites (C2 for pastures and E2 for forest) for each lumped model (results for LM model are not included since best matching results for LPM were achieved with $\eta \approx 1$, see Table 4). These probability (PDF) and cumulative density functions (CDF) depict how water is routed through the system. In this sense, pasture sites generally show a faster and higher response of the tracer peak when compared to forest sites. The CDF (Figs. 8b and d) of all models are quite similar for the major part of the flows, even including the linear function LPM that averages the shape of the peaks described by the other models. Models based on exponential functions (EPM, DM, or GM in Figs. 8b and d) predict a small portion of the flow with an exponentially delayed tail, which is larger for forested sites than for pastures. Best distribution function results (based on highest NSEs) for all sampled sites, according to the type of land cover, are shown in Figs. 9a and b for the LPM and GM models applied to pasture sites, and in Figs. 9c and d for forest sites. Considering the range of possible or behavioral solutions (e.g., shaded area represents range of solutions for C2 site in Figs. 9a and b, and for E2 in Figs. 9c and d), distributions functions for each type of model and land cover are very similar between each sampled site.”

3.2 River and tributaries

"Considering all sites and models the criteria NSE > 0.45 was exceeded in 41 of the 63 predictions (9 sites per 7 models, Fig. 5a). Among the analyzed sites the TPLR model yielded the best matches for PL, SF, FH, QZ, QN, QM and QC, while the EPM model for the QR and QP sites (Fig. 5b). The GM model reached closest efficiencies when compared to the best match for every site. Consequently only the TPLR, EPM and GM models were further considered. Differences between MTT predictions for all sites are depicted in Fig. 10a and results from retained models in Table 5 and Annex 2. Although MTT results according to the best NSEs were reached using the TPLR model, compared to the GM or the EPM, these predictions also showed the largest uncertainties (Fig. 10b) and at the same time depicted the lowest number of observations inside the predicted range of behavioral solutions (Fig. 10c). Considering these additional selection criteria, EPM performed better. For stream water at the main outlet, Figs. 11-13 show the parameter uncertainties and behavioral solutions for the TPLR, GM and EPM models, respectively."

"Considering results from the EPM model (Table 5, Fig. 10a), the fitting efficiencies reached a maximum NSE of 0.56 for the main stream, and NSEs between 0.48 and 0.58 for the main tributaries (Fig. 5a). The predicted MTT at catchment outlet was 2.0 yr with a η parameter of 1.84 (a similar value was estimated for the main river at the SF sampling site, MTT = 2.0 yr and η = 1.85) and varied from 2.0 (QM, η = 1.85) to 3.9 yr (QC, η = 1.97) for the main tributaries. Uncertainties of MTT predictions between sites were similar with a maximum range between 14.1% and 20.4% of the predicted MTT, as derived for the FH and QM sites (Table 5). Similarly, η ranged from 1.61 (QZ) to 2.21 (QP), the average value of η = 1.85 implies a 54% of volume portion of exponential flow and a 46% volume of piston flow; the uncertainty for the η parameter was 25% on average."

"Figures 14a and 14b show the shape of the TTD for the main river outlet (PL), corresponding to the highest NSEs for EPM, GM and TPLR models. The curve for EPM shows a delayed peak that is not accounted in the GM or TPLR models (Fig. 14a), which in turn are very similar between them (at least after a short initial time since GM tends to infinity for times closes to cero). Besides, the latter models show a more delayed flow tail when compared to EPM, which show in general a faster transit time (Fig. 14b). Differences between stream water TTDs from the main sub-catchments considering EPM and GM models are shown in Figs. 15a and b. For comparison of the degree of similarities between sites, these plots include the range of behavioral solutions for the main outlet (PL), thereby being clear that apart from QC or QP, the remaining sites have similar (EPM or GM) transit time distribution functions."

3.3 Springs and creeks

"Of 35 predictions (7 models for 5 sites) the criterion NSE > 0.45 was fulfilled in 20 cases. Sites with reduced isotope signal (small σ) yielded lower efficiencies (Fig. 5a, Table 5 and Annex 2). Apart from TP and QRS, in the remaining sites the criterion NSE > 0.45 was reached at least by 5 models. TP, PLS and SFS sites were best described by using a TPLR model (Fig. 5b). In this regard, GM and EPM were the second and third best models. Figure 10a shows the MTT results predicted by the three models, while detailed information is given in Table 5 and Annex 2. As for stream waters, the EPM model performed best when looking at the uncertainties and the number of observed data inside the range of behavioral solutions (Figs. 10b and c)."

“Considering EPM, MTTs of 4.5 yr (NSE = 0.49, η = 1.74) for TP and 2.1 yr (NSE = 0.65, η = 1.84) for Q3 were estimated; while for springs, 2.0 yr (NSE = 0.69, η = 1.85) for PLS and 3.3 yr (NSE = 0.47, η = 1.42) for SFS. Results for the QRS site showed poor reliability due to the reduced amplitude of $\delta^{18}O$ in the observed data (Table 5), the lowest among the observed sites (σ = 0.17). Estimations of MTTs for this site was larger than 5 yr, and therefore beyond the level of applicability of the method for natural isotopic tracers.”

“Figures 14c and d show the TTD results of EPM, GM and TPLR models, for a representative site with long MTT (creek TP). This site show a distinctive more delayed time to the peak (for EPM model) and longer duration of flow tails compared to stream water (Figs. 14a and b). In Figs. 15c and d, the TTDs for all spring and creek sampled sites are shown for the EPM and GM models. In these figures, it is noticeable that the sites Q3 and PLS show the same patterns described previously for most of the stream waters (Figs. 14a and b), while some differences related to more delayed flow responses can be accounted for SFS, TP or QRS sites (Figs. 15c and d), which are more similar to QP and QC stream waters.”

4 Discussion

“For each soil water site, similar MTT results of a few weeks to months were obtained regardless of the lumped parameter model used (Fig. 6a, Table 4 and Annex 1). Although the LPM model did not yield predictions with the highest efficiencies (Fig. 5a), provided smaller ranges of uncertainty (Fig. 6b) and a larger number of observations inside them (Fig. 6c), advantages that could not be inferred by using only the best matches to NSE, for which GM and EPM models performed better than others (Fig. 5b). Using a LPM model, suitable to describe a partially confined aquifer with increasing thickness (Maloszewski and Zuber, 1982), we found MTTs varying from 2.3 to 6.3 weeks for pastures sites and from 3.7 to 9.2 weeks for forested soils. If we consider that only the top soil horizon was sampled (maximum sampled depth was 0.4 meters), these results are comparable to values between 7.5 and 31 weeks found in 2.0 meter soil columns of typical Bavarian soil using the DM model (Maloszewski et al., 2006). When analyzing the distribution function for soil waters, similarities between model results are evident (Figs. 8 and 9). Considering the range of possible solutions of each site (shaded areas in Figs. 9a-d), it is noticeable that the major part of the flow’s transit can be described similarly by all models, even using the simpler function (LPM). For these sites, when considering exponential models (EPM, GM or DP), a small portion of the flow is depicted as having a delayed tail; however, compared to the magnitude of the total volume, an LPM distribution could still be considered as a reliable method to estimate MTTs.”

“Considering the LPM results for MTTs of soil water from pastures (4.3 weeks on average) and forest sites (5.9 weeks on average) as independent data sets, a two tailed p-value of 0.0075 for a Student’s t-test was calculated, meaning that the difference between the two groups was statistically significant, although physical characteristics, like length, slope and altitude and meteorological conditions of the respective hill slopes were more or less similar. Land use effects, affecting soil hydraulic properties controlling the infiltration and flow of water, were detected in previous studies within the research area (Huwe et al., 2008). Confirming findings in other tropical catchments were published by Zimmermann et al. (2006) and by Roa-Garcia and Weiler (2010), who stated that under grazing the hydraulic conductivity decreased, overland and near surface flows increased, the storage capacity of the soil matrix declined, with feedbacks on the MTT of soil water. Similar insights were found

by Tetzlaff et al. (2007) comparing two small catchments in Central Scotland Highlands of different land use."

"For larger MTTs (> 1 yr), as derived for sampled surface waters and shallow springs, there were differences when predicted results among models were compared (Fig. 10a, Table 5 and Annex 2), especially for sites with strong damped signals of measured $\delta^{18}O$ (e.g. QRS and TP sites). When considering uncertainties, the EPM model performed significantly better when compared to the TPLR or GM models (Figs. 10b and c), although the latter two performed best for most of the sampled surface waters according to the NSE objective function (Figs. 5a and b)."

"When analyzing results from different models, dotty plots of model parameter uncertainty are very useful to display not only the magnitude of uncertainty but also its tendency. Similarly, the uncertainty bands of behavioral solutions can help to account for the sensitivity of the parameter uncertainty on $\delta^{18}O$ modeled results. For example, when predicted results for the PL site are compared, larger parameter uncertainty and skewness are notorious for TPLR than for EPM or GM models (Figs. 11a-c for TPLR; 12a-c for GM; 13a and b for EPM). At the same time EPM shows the highest sensitivity in modeled results (Figs. 11d, 12d, 13c). In order to contrast the signature of the effluent with younger waters such as rainfall, Figs. 11e, 12e, or 13d show the damped observed (and predicted) $\delta^{18}O$ signatures at the main outlet; a characteristic present in all analyzed surface waters. Considering the efficiencies reached by the predictions, we should keep in mind that ranges of behavioral solutions derived from a fixed 5% of the top NSE are generally smaller than a predefined lower limit for all waters, e.g., a predefined lower efficiency limit of 0.30 and 0.45 were used by Speed et al. (2010) and Capell et al., (2012), respectively."

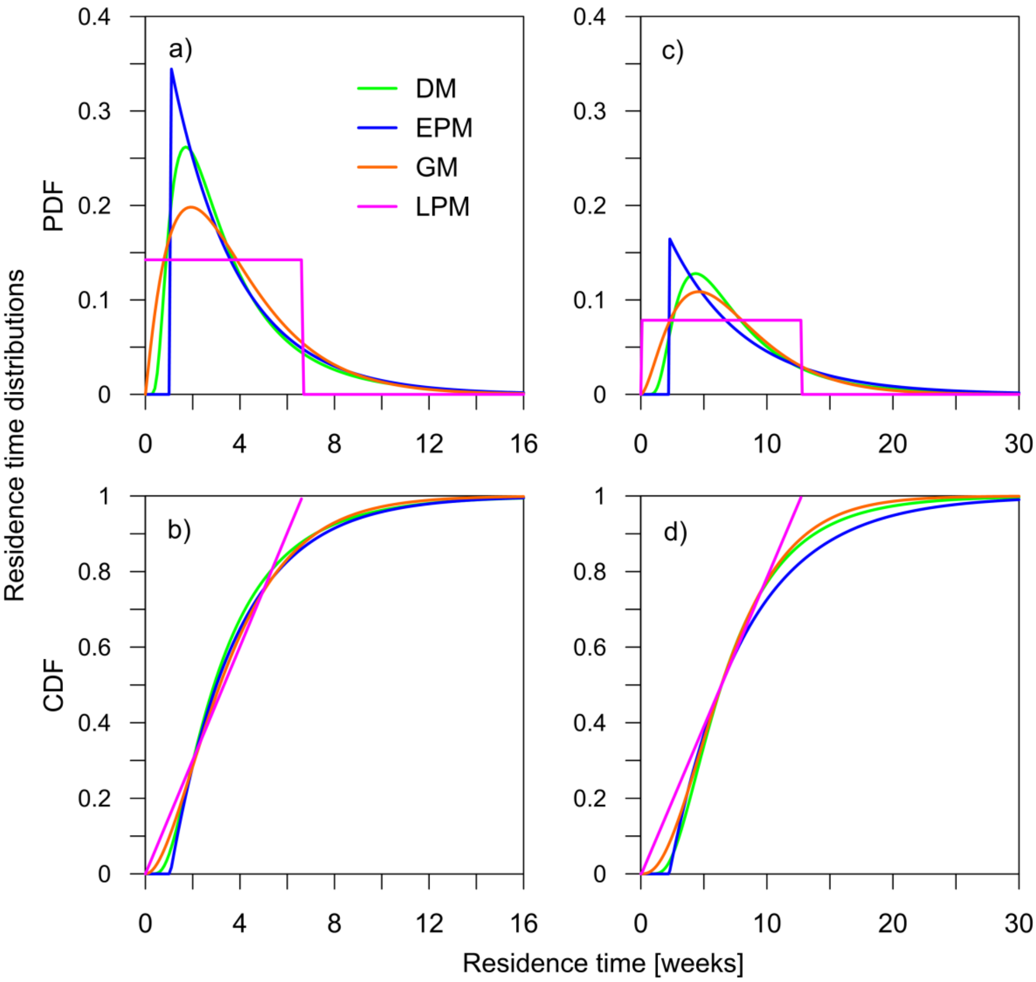
"For stream waters, as for springs and creeks, the main differences between EPM and GM (or TPLR) results consisted first in a delayed response of the tracer signal in the outlet, modeled by a parameter $\eta > 1$ (Table 5), while for GM or TPLR the response of the flow occurred instantaneously after the spread of the tracer along the catchment (Figs. 14 and 15, Annex 2); and secondly by a comparatively smaller exponential flow tails, which also means that in general the flow transport is faster considering EPM than GM or TPLR models. For these cases, regardless of the degree of efficiencies or uncertainties, the decision on which TTD is more reliable would depend on the conceptual knowledge of the functioning of the catchment. For the San Francisco catchment this can be gained through additional field experiments in selected sites or sub-catchments using either higher resolution samples from the effluents in order to analyze non steady conditions (Botter et al., 2011) or considering different mixing assumptions (Hrachowitz et al., 2013). Another approach could be to analyze longer time series of stable isotopes, or even to include radioactive isotopes as tritium, which would help to crosscheck results, as it has been claimed that, in some cases, the inferences of the processes using solely stables isotopes, underestimate the delayed part of the flow (Stewart et al., 2010)."

"Regardless of the used model, efficiencies of MTT for stream waters were lower than for soil waters. This was somehow expected, since the dampening effect on a catchment to sub-catchment scale generates a smoother signal filtering/averaging the heterogeneity observed at a single point along a precise transect. Since for most of the cases MTTs for soil waters showed an increasing trend according to increasing soil depth, longer MTTs corresponding to deeper soil layers are to be expected. Soil water below 0.4 m was not monitored within this study, given the shallow soil depth and the increasing fraction of rock material with depth, preventing the use of wick samplers."

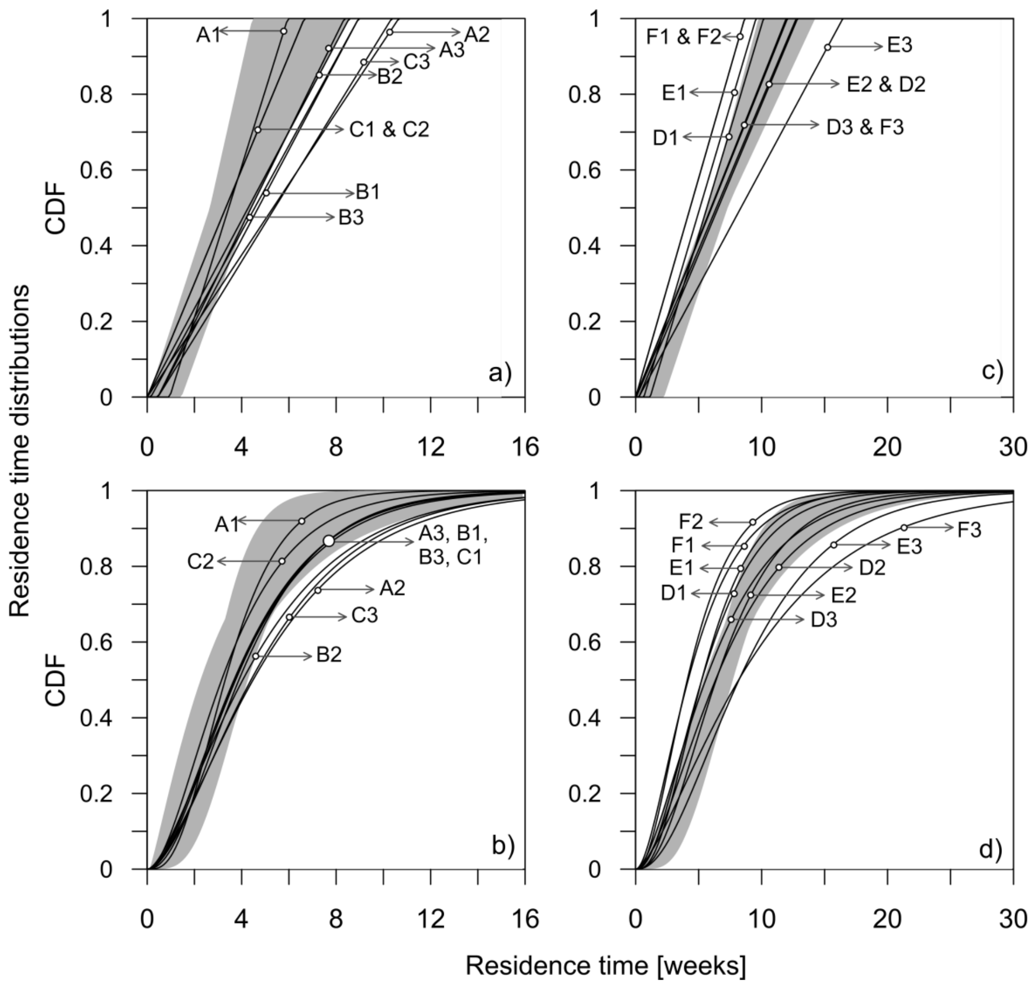
“The similarities and differences between models for sites with MTTs > 1 yr, as for stream and spring waters, gave insights about the importance to account for a proper TTD, defined according to the conceptual knowledge of the catchment’s functioning, before calculating MTT. In this regard, the use of a multi-model approach and uncertainty analysis is believed essential as to be able of defining which functions describes in a better way the parameter identifiability and bounds of behavioral solutions. By considering best matches to NSE for stream waters, best predictions were obtained with the TPLR, EPM and GM models; being more flexible versions of a pure exponential distribution function (i.e. EM model), which help to account for non-linearities of the system. The same distribution functions were identified as good predictors of observed data in a related study by Weiler et al. (2003). When comparing the TPLR to EPM or GM models, the latter two take the non-linearity of the flow without splitting it in two reservoirs with different exponential behaviors into account, therefore yielding more identifiable results. However, findings by Weiler et al. (2003) suggest that the TPLR distribution function could achieve better predictions for runoff events generated by mixed fast and slow flows. In related studies using multiple models, the EPM model yielded the best predictions for surface and spring waters (Viville et al., 2006). Considering this model, in the San Francisco catchment, the average η = 1.85 value for surface waters (similar values were found for creeks: η = 1.79 and springs: η = 1.64) implies that a significant portion of old water (46%) is released previous to the new one (54%). The η value in this study is larger than the η value found in studies for stream water in temperate small headwaters catchments (η = 1.09, Kabeya et al., 2006; η = 1.28, McGuire et al., 2002; η = 1.37, Asano et al., 2002), and close to results published by Katsuyama et al. (2009) for two riparian groundwater systems (η = 1.6 and 1.7).”

“Regarding to the Gamma model, it was also identified as an applicable distribution function in headwater montane catchments with dominant baseflow in temperate climate (Hrachowitz et al., 2009a, 2010; Dunn et al., 2010). For our study area, a characteristic shape parameter α < 1 (e.g. Fig. 12b and Annex 2) was found in all stream and spring sites meaning that an initial peak or a significant part of the flow was quickly transported to the river. Similar results were found recently for mountain catchments of comparable size in Scotland by Kirchner et al. (2010), who also stated the importance for accounting the best distribution shape, which is usually assumed as purely exponential (α = 1). MTTs derived without the use of observed data, using a purely exponential model, frequently led to an overestimation of α and consequently an underestimation of MTTs. The higher flexibility of the GM model permits to account for the non-linearity in the behavior of a catchment system (Hrachowitz et al., 2010).”

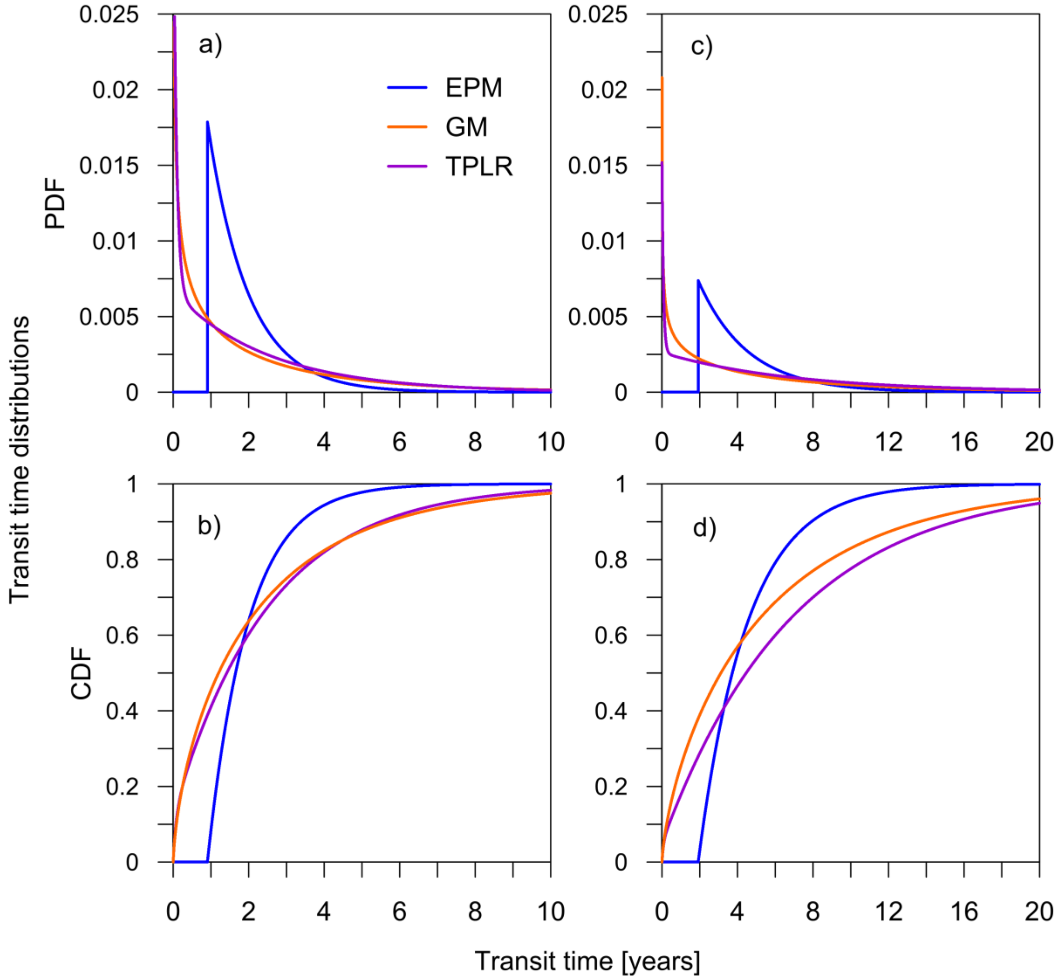
New figures:



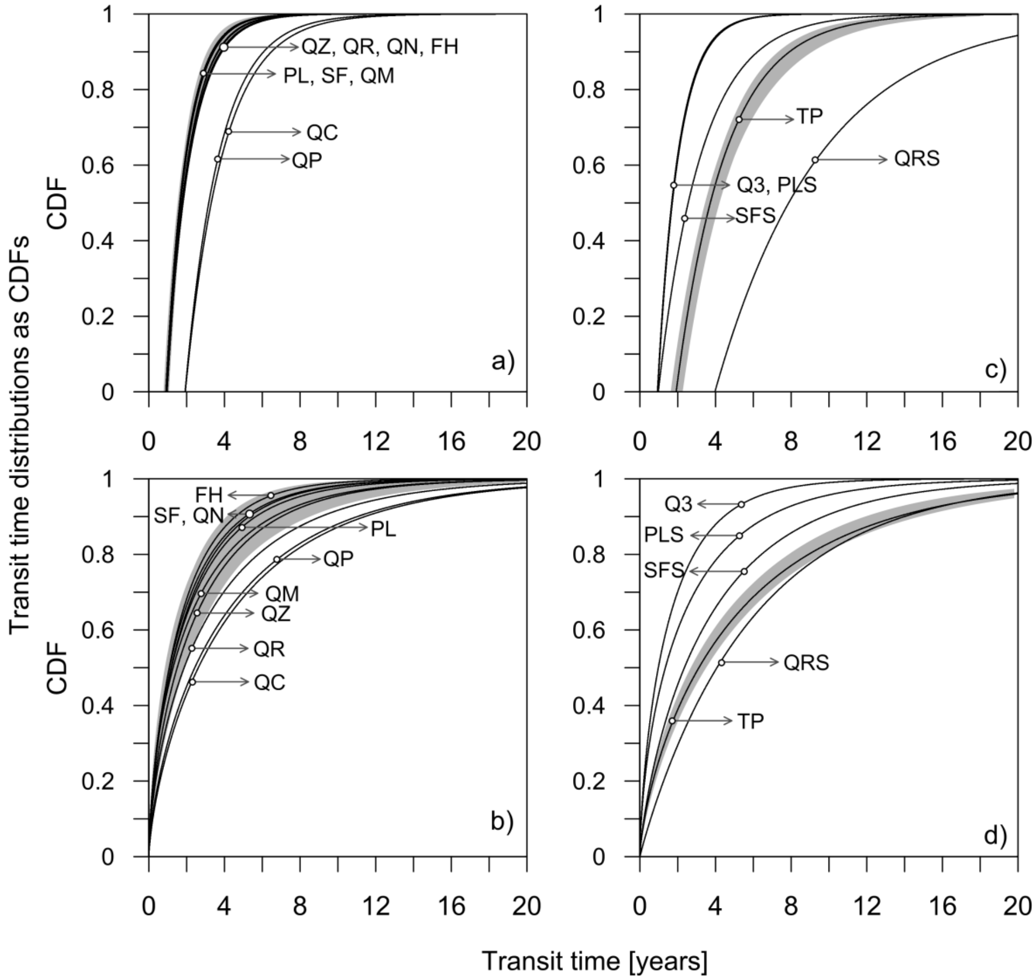
*“**Fig. 8.** Comparative characteristic shapes of residence time distribution functions corresponding to the best NSE using four lumped parameter models (DM, EPM, GM and LPM): **(a)** and **(b)** for the soil site C2 located in a pastures land cover; **(c)** and **(d)** for the soil site E2 located in a forest land cover.”*



*"**Fig. 9.** Comparative results between LPM and GM models of soil water residence time distributions functions corresponding to the best NSE for every sampling site: **(a)** pastures sites using LPM; **(b)** pastures sites using GM; **(c)** forest sites using LPM; **(d)** forest sites using GM. Gray shaded area in each plot corresponds to the range of possible shapes of the distribution function for one of the sampling sites: C2 in sub-plots **(a)** and **(b)**, and E2 in sub-plots **(c)** and **(d)**."*



"Fig. 14. Comparative characteristic shapes of the transit time distribution functions corresponding to the best NSE using three lumped parameter models (EPM, GM and TPLR): (a) and (b) for the stream water sampled at the main outlet PL; (c) and (d) for the small creek TP."



*“**Fig. 15.** Comparative results between EPM and GM models of soil water transit time distributions functions corresponding to the best NSE for every sampling site: **(a)** stream water of main outlet and sub-catchments using EPM, and **(b)** using GM; **(c)** spring waters and creeks using LPM, and **(d)** using GM. Gray shaded area in each plot corresponds to the range of possible shapes of the distribution function for one of the sampling sites: the main outlet (PL) in sub-plots **(a)** and **(b)** and TP creek in sub-plots **(c)** and **(d)**.”*

20) section 3: please provide complete a table with results including information such as the optimum model performances and the 5/95% of model performances of the retained models and the same for all parameter(s) for all sites and components.

Tables of results for retained models (showed below) will be included as Annexes 1 and 2.

"Annex 1. Predicted results of soil waters for the Gamma, Exponential-Piston Flow, Dispersion and Linear lumped models."

***"Table 1.** Best predicted results for the Gamma model parameters (τ, α) and corresponding uncertainty ranges."*

Site	Mean ‰	σ ‰	NSE -	RMSE ‰	Bias ‰	τ weeks	α -
Pastures transect							
A1	-6.74	3.06	0.87	1.33	-0.04	3.6(2.9-4.4)	3.6(2.0-13.4)
A2	-6.72	2.46	0.73	1.72	0.07	5.5(4.5-6.7)	1.8(1.2-3.4)
A3	-7.17	3.18	0.85	1.54	-0.04	4.4(3.5-5.5)	2.0(1.4-8.3)
B1	-6.58	3.01	0.83	1.53	0.27	4.4(3.6-5.3)	2.0(1.4-7.0)
B2	-6.88	2.71	0.80	1.53	0.15	5.0(4.1-6.1)	1.7(1.2-3.5)
B3	-6.72	2.97	0.80	1.51	0.04	4.4(3.6-5.4)	2.1(1.3-5.1)
C1	-6.68	3.15	0.86	1.36	-0.04	3.5(2.7-4.2)	2.4(1.6-9.0)
C2	-7.19	3.11	0.88	1.19	-0.14	3.7(2.9-4.6)	2.1(1.2-5.4)
C3	-6.53	2.56	0.80	1.35	-0.01	5.3(4.5-6.4)	1.9(1.3-3.8)
Forest transect							
D1	-7.26	2.79	0.81	1.35	0.12	6.1(5.1-7.5)	2.4(1.5-4.9)
D2	-7.03	2.35	0.82	1.08	0.03	7.6(6.6-9.2)	1.9(1.3-3.2)
D3	-6.82	2.40	0.82	1.16	-0.02	6.7(5.8-7.9)	1.8(1.2-3.6)
E1	-6.54	2.79	0.82	1.34	0.10	5.9(5.1-6.8)	2.9(1.8-7.1)
E2	-6.52	2.44	0.78	1.37	0.11	7.3(6.4-8.2)	2.7(1.8-5.6)
E3	-6.43	1.97	0.79	1.16	0.02	9.4(8.2-10.7)	2.5(1.8-4.0)
F1	-6.81	2.72	0.90	0.99	-0.06	5.0(4.2-6.1)	1.9(1.3-4.7)
F2	-6.74	2.79	0.90	0.97	-0.29	4.7(3.8-5.7)	2.4(1.4-6.4)
F3	-8.50	1.87	0.69	1.41	-0.41	10.2(8.7-12.5)	1.6(1.2-2.2)

σ= standard deviation, NSE = Nash-Sutcliffe Efficiency, RMSE = Root Mean Square Error

"**Table 2.** *Best predicted results for the Exponential Piston flow model parameters (τ, η) and corresponding uncertainty ranges.*"

Site	Mean	σ	NSE	RMSE	Bias	τ	η
	‰	‰	-	‰	‰	weeks	-
Pastures transect							
A1	-6.88	3.00	0.86	1.38	-0.18	3.7(2.9-4.8)	1.40(1.28-1.59)
A2	-6.91	2.53	0.73	1.71	-0.12	5.7(4.7-7.2)	1.26(1.18-1.34)
A3	-7.31	3.23	0.84	1.57	-0.18	4.5(3.5-5.7)	1.33(1.21-1.48)
B1	-6.99	3.11	0.84	1.49	-0.14	5.1(4.0-6.3)	1.33(1.24-1.43)
B2	-7.14	2.82	0.80	1.51	-0.10	5.5(4.4-6.9)	1.28(1.20-1.36)
B3	-6.82	2.94	0.80	1.52	-0.05	4.7(3.8-6.0)	1.30(1.21-1.40)
C1	-6.75	3.15	0.86	1.38	-0.10	3.7(2.9-4.7)	1.40(1.29-1.57)
C2	-7.15	3.09	0.88	1.18	-0.09	3.8(3.0-5.0)	1.36(1.25-1.51)
C3	-6.59	2.54	0.80	1.36	-0.08	5.5(4.5-6.9)	1.25(1.17-1.33)
Forest transect							
D1	-7.40	2.79	0.83	1.28	-0.02	7.0(5.6-8.6)	1.44(1.33-1.56)
D2	-7.06	2.36	0.82	1.11	0.00	8.5(7.2-10.2)	1.32(1.26-1.39)
D3	-6.84	2.36	0.81	1.19	-0.05	7.2(6.0-8.9)	1.18(1.12-1.23)
E1	-6.67	2.75	0.82	1.34	-0.03	6.6(5.5-8.1)	1.47(1.37-1.63)
E2	-6.69	2.38	0.77	1.40	-0.07	8.2(6.9-9.8)	1.37(1.29-1.46)
E3	-6.54	1.99	0.78	1.21	-0.09	10.3(8.9-12.1)	1.45(1.32-1.58)
F1	-6.88	2.73	0.90	0.97	-0.13	5.2(4.2-6.6)	1.27(1.19-1.36)
F2	-6.61	2.65	0.91	0.95	-0.16	4.8(3.8-6.1)	1.25(1.16-1.37)
F3	-8.14	2.02	0.74	1.30	-0.05	9.6(8.4-11.7)	1.37(1.22-1.47)

σ= standard deviation, NSE = Nash-Sutcliffe Efficiency, RMSE = Root Mean Square Error

*“**Table 3.** Best predicted results for the Dispersion model parameters (τ, D_p) and corresponding uncertainty ranges.”*

Site	Mean $^0/_{00}$	σ $^0/_{00}$	NSE -	RMSE $^0/_{00}$	Bias $^0/_{00}$	τ weeks	D_p -
Pastures transect							
A1	-6.77	3.10	0.86	1.33	-0.07	3.6(3.1-4.7)	0.13(0.07-0.53)
A2	-6.63	2.45	0.72	1.76	0.16	5.7(4.7-7.8)	0.33(0.22-0.99)
A3	-7.15	3.23	0.84	1.59	-0.02	4.5(3.6-6.1)	0.22(0.11-0.97)
B1	-6.56	3.08	0.82	1.55	0.29	4.6(3.8-5.8)	0.21(0.11-0.78)
B2	-6.79	2.77	0.79	1.57	0.24	5.1(4.3-7.2)	0.31(0.21-1.06)
B3	-6.64	2.94	0.80	1.52	0.12	4.5(3.8-6.5)	0.28(0.17-0.87)
C1	-6.61	3.17	0.86	1.37	0.03	3.5(2.9-5.1)	0.19(0.10-0.85)
C2	-7.00	3.06	0.88	1.19	0.06	3.7(3.2-5.8)	0.29(0.17-0.97)
C3	-6.46	2.53	0.79	1.38	0.06	5.5(4.7-7.5)	0.32(0.20-0.84)
Forest transect							
D1	-7.24	2.68	0.83	1.28	0.14	6.7(5.7-9.2)	0.31(0.18-0.64)
D2	-6.99	2.33	0.82	1.08	0.07	8.4(7.2-11.7)	0.34(0.23-0.76)
D3	-6.77	2.40	0.81	1.19	0.03	7.2(6.2-10.0)	0.32(0.19-0.82)
E1	-6.55	2.75	0.82	1.33	0.10	6.3(5.4-7.8)	0.21(0.12-0.46)
E2	-6.51	2.45	0.77	1.39	0.11	7.6(6.8-9.4)	0.20(0.13-0.43)
E3	-6.41	2.00	0.78	1.19	0.03	9.8(8.7-11.8)	0.22(0.15-0.39)
F1	-6.72	2.72	0.90	1.00	0.04	5.2(4.3-7.1)	0.29(0.18-0.83)
F2	-6.66	2.73	0.91	0.95	-0.21	4.8(3.9-6.7)	0.29(0.15-0.73)
F3	-8.49	1.90	0.70	1.39	-0.40	11.6(9.8-14.6)	0.41(0.29-0.75)

σ= standard deviation, NSE = Nash-Sutcliffe Efficiency, RMSE = Root Mean Square Error

*"**Table 4.** Best predicted results for the Linear Model parameter (τ) and corresponding uncertainty ranges."*

Site	Mean $^0/_{00}$	σ $^0/_{00}$	NSE -	RMSE $^0/_{00}$	Bias $^0/_{00}$	τ weeks
Pastures transect						
A1	-6.85	3.06	0.86	1.37	-0.15	3.5(2.8-4.5)
A2	-6.87	2.63	0.73	1.72	-0.08	5.4(4.5-6.2)
A3	-7.32	3.30	0.86	1.46	-0.19	4.4(3.5-5.2)
B1	-6.89	3.19	0.83	1.52	-0.04	4.3(3.3-4.9)
B2	-7.03	3.02	0.78	1.57	0.00	4.4(3.8-5.2)
B3	-6.77	3.03	0.79	1.54	0.00	4.4(3.4-4.9)
C1	-6.72	3.17	0.84	1.44	-0.07	3.5(2.5-4.1)
C2	-7.10	3.16	0.87	1.27	-0.04	3.5(2.9-4.5)
C3	-6.54	2.71	0.80	1.36	-0.02	4.9(4.4-5.9)
Forest transect						
D1	-7.31	2.91	0.76	1.50	0.07	5.4(4.8-6.2)
D2	-6.97	2.56	0.78	1.19	0.09	6.6(5.9-7.1)
D3	-6.74	2.61	0.80	1.22	0.05	6.0(4.9-6.6)
E1	-6.65	2.84	0.80	1.41	0.00	5.4(4.8-6.1)
E2	-6.64	2.55	0.78	1.37	-0.01	6.4(5.8-7.1)
E3	-6.48	2.14	0.76	1.24	-0.04	8.1(7.3-9.2)
F1	-6.79	2.90	0.89	1.05	-0.03	4.5(4.0-5.5)
F2	-6.52	2.79	0.89	1.03	-0.08	4.6(3.9-5.6)
F3	-8.42	2.37	0.64	1.51	-0.33	7.2(7.1-8.2)

σ= standard deviation, NSE = Nash-Sutcliffe Efficiency, RMSE = Root Mean Square Error

"Annex 2. Predicted results of stream, creek and spring waters for the lumped models Gamma and Two Parallel Linear Reservoirs."

"Table 1. *Best predicted results for the Gamma model parameters (τ, α) and corresponding uncertainty ranges."*

Site	*Mean* $^0/_{00}$	*σ* $^0/_{00}$	*NSE* -	*RMSE* $^0/_{00}$	*Bias* $^0/_{00}$	*τ* *yr*	*α* -
Stream							
PL	*-8.16*	*0.42*	*0.61*	*0.34*	*0.0909*	*2.2(1.6-3.2)*	*0.62(0.55-0.71)*
SF	*-8.03*	*0.43*	*0.62*	*0.34*	*0.0836*	*2.0(1.5-3.1)*	*0.63(0.56-0.72)*
Streamwater tributaries							
FH	*-8.21*	*0.42*	*0.58*	*0.36*	*0.0765*	*1.8(1.5-2.9)*	*0.71(0.60-0.78)*
QZ	*-8.35*	*0.36*	*0.58*	*0.31*	*0.0596*	*2.7(2.0-3.9)*	*0.63(0.57-0.72)*
QN	*-8.21*	*0.40*	*0.64*	*0.30*	*0.0681*	*2.1(1.6-3.2)*	*0.66(0.58-0.75)*
QR	*-7.86*	*0.16*	*0.45*	*0.35*	*0.0915*	*3.5(2.6-4.4)*	*0.60(0.56-0.67)*
QP	*-8.04*	*0.26*	*0.54*	*0.23*	*0.0240*	*4.3(3.3-5.4)*	*0.65(0.62-0.73)*
QM	*-7.74*	*0.44*	*0.60*	*0.37*	*0.0706*	*2.5(1.8-3.7)*	*0.57(0.51-0.64)*
QC	*-7.57*	*0.24*	*0.53*	*0.21*	*0.0508*	*4.5(3.7-5.4)*	*0.68(0.64-0.74)*
Creeks							
TP	*-7.63*	*0.20*	*0.45*	*0.18*	*0.0249*	*5.5(4.8-5.9)*	*0.68(0.64-0.73)*
Q3	*-7.66*	*0.45*	*0.68*	*0.30*	*0.0126*	*1.7(1.3-2.8)*	*0.65(0.55-0.74)*
Springs							
PLS	*-7.94*	*0.43*	*0.69*	*0.28*	*0.0945*	*2.6(1.9-3.7)*	*0.58(0.53-0.66)*
SFS	*-7.57*	*0.23*	*0.56*	*0.19*	*0.0432*	*3.9(3.0-4.9)*	*0.74(0.68-0.81)*
QRS	*-7.78*	*0.09*	*0.25*	*0.14*	*0.0146*	*6.0(5.3-6.5)*	*0.94(0.91-1.00)*

σ= standard deviation, NSE = Nash-Sutcliffe Efficiency, RMSE = Root Mean Square Error

*“**Table 2.** Best predicted results for the Two Parallel Reservoir model parameters (τ_s, φ) and corresponding uncertainty ranges. A fixed range from 4 to 4.5 weeks was maintained for τ_f in all cases.”*

Site	Mean $^0/_{00}$	σ $^0/_{00}$	NSE -	RMSE $^0/_{00}$	Bias $^0/_{00}$	τ_s yr	φ -
Stream							
PL	-8.24	0.44	0.66	0.32	0.0176	2.5(1.9-5.6)	0.622(0.554-0.706)
SF	-8.10	0.44	0.64	0.33	0.0117	2.1(1.6-4.3)	0.631(0.555-0.721)
Streamwater tributaries							
FH	-8.24	0.43	0.60	0.34	0.0383	2.0(1.5-3.1)	0.708(0.605-0.782)
QZ	-8.41	0.37	0.60	0.30	0.0000	2.5(1.9-4.7)	0.632(0.570-0.717)
QN	-8.27	0.41	0.67	0.29	0.0141	2.2(1.6-3.6)	0.660(0.582-0.749)
QR	-7.93	0.23	0.52	0.33	0.0280	4.6(3.1-7.0)	0.603(0.562-0.672)
QP	-8.09	0.24	0.54	0.23	-0.0207	3.6(2.8-6.5)	0.653(0.620-0.728)
QM	-7.84	0.48	0.63	0.36	-0.0307	2.7(2.1-8.3)	0.565(0.506-0.636)
QC	-7.60	0.23	0.59	0.19	0.0183	5.2(3.8-6.8)	0.685(0.642-0.741)
Creeks							
TP	-7.65	0.17	0.51	0.17	0.0054	7.0(5.7-7.8)	0.680(0.642-0.726)
Q3	-7.71	0.43	0.67	0.31	-0.0428	1.7(1.3-2.7)	0.648(0.554-0.742)
Springs							
PLS	-8.04	0.44	0.78	0.24	-0.0045	4.0(2.6-8.0)	0.581(0.526-0.659)
SFS	-7.58	0.23	0.59	0.19	0.0255	3.6(2.8-5.2)	0.735(0.684-0.813)
QRS	-7.79	0.09	0.25	0.14	0.0119	6.1(5.3-6.6)	0.945(0.911-0.997)

σ= standard deviation, NSE = Nash-Sutcliffe Efficiency, RMSE = Root Mean Square Error

21) p.15888, l.1-4: maybe include Roa-Garcia and Weiler (2010) as reference here

Suggested reference has been included in the referred text, now it reads (please see reply to comment 19 for the new version of Section 4):

“Confirming findings in other tropical catchments were published by Zimmermann et al. (2006) and by Roa-Garcia and Weiler (2010), who stated…”

22) p.15889, l.5 and elsewhere in the manuscript: this should not come as a surprise. Rule of thumb: more parameters = more uncertainty, simply by the additional degrees of freedom in a model, allowing for different parameter combinations giving the same results (equifinality)

We agree and deleted the sentence.

23) Table 1: not sure that the SI units for “site code” is [m a.s.l.] and for “altitude” [weeks]. Just saying… ;-)

Corrected.

24) Table 3: symbols need to be defined somewhere in the manuscript

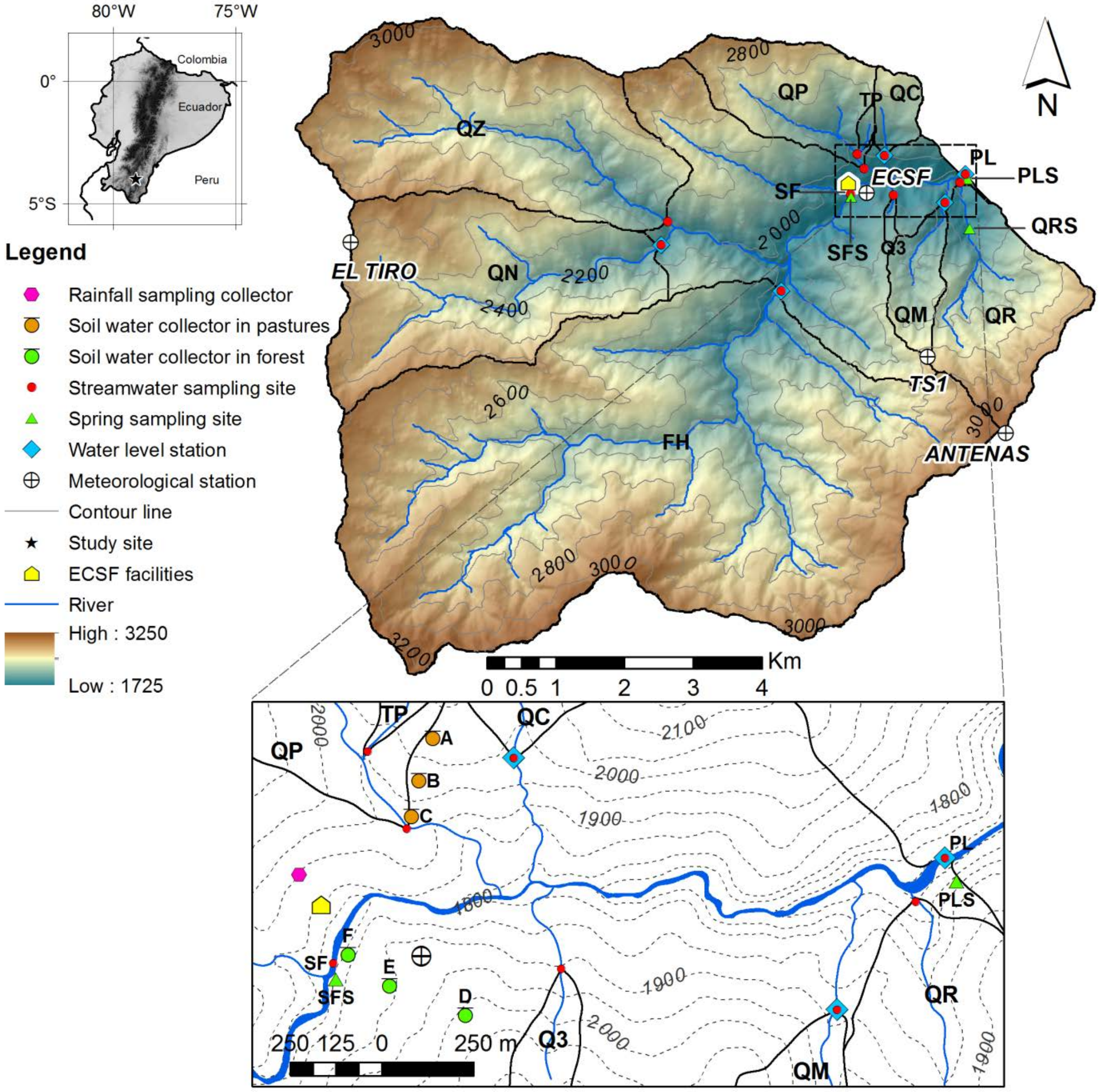
Description of symbols (model parameters) are now included in Table 3. Please see new version of Table 3 in the reply to comment 18a.

25) Table 4: it would be nice to also provide and discuss a figure with the transects and the respective MTT and/or TTD depth profiles therein.

Similarities between soil sites according to their respective distribution functions (for two representative models) are now analyzed in the 'Discussion' section (See reply to comment 19). A more in depth analysis, than the one shown in the mentioned section, has been avoided accordingly to the suggestions of referee #2.

26) Figure 1: please add a zoom-in to better show the transects

As showed below, figure 1 has now a zoomed area for the lower part of the catchment.



*"**Fig. 1.** San Francisco catchment with sampling locations and delineation of drainage area. Acronyms in bold are defined in Table 1. Framed image shows the zoomed area of the lower part of the catchment."*