Using high resolution isotope data and alternative calibration strategies for a tracer-aided runoff model in a nested catchment

C. Tunaley[1], D. Tetzlaff [1], C. Birkel [1,2] and C. Soulsby [1]

[1] Northern Rivers Institute, School of Geosciences, University of Aberdeen, UK

[2] Department of Geography, University of Costa Rica, San Jose, 2060, Costa Rica

Corresponding author: Claire Tunaley (claire.tunaley@abdn.ac.uk)

Abstract

Testing hydrological models over different spatio-temporal scales is important both for evaluating diagnostics and aiding process understanding. High-frequency (6hr) stable isotope sampling of rainfall and runoff was undertaken during 3 week periods in summer and winter within 12 months of daily sampling in a 3.2 km^2 catchment in the Scottish Highlands. This was used to calibrate and test a tracer-aided model to assess the: (1) information content of high resolution data; (2) effect of different calibration strategies on simulations and inferred processes; (3) model transferability to <1 km^2 sub-catchment. The 6-hourly data were successfully incorporated without loss of model performance, improving the temporal resolution of the modelling, and making it more relevant to the time dynamics of the isotope and hydrometric response. However, this added little new information due to old-water dominance and riparian mixing in this peatland catchment.

Time variant results, from differential split sample testing, highlighted the importance of calibrating to a wide range of hydrological conditions. This also provided insights into the non-stationarity of catchment mixing processes, in relation to storage and water ages, which varied markedly depending on the calibration period. Application to the nested sub-catchment produced equivalent parameterisation and performance, highlighting similarity in dominant processes. The study highlighted the utility of high-resolution data in combination with tracer-aided models, applied at multiple spatial scales, as learning tools to enhance process understanding and evaluation of model behaviour across non-stationary conditions. This helps reveal more fully the catchment response in terms of the different mechanistic controls on both wave celerites and particle velocities.

Key words: high resolution isotopes, tracer-aided modelling, parameter transferability, catchment storage, water age, runoff processes

1. Introduction

Incorporation of conservative isotope tracers into conceptual hydrological models has proved insightful in terms of increasing the understanding of integrated catchment functioning, and the partitioning, storage and release of water (Seibert et al., 2003; Weiler, 2003; Dunn et al., 2007; Van Huijgevoort et al., 2016; Alo aho et al., 2017). This understanding is particularly important in headwater catchments, which control downstream water quality and quantity (Freeman et al., 2007; Bishop et al., 2008). A major benefit of using discharge and isotopes in a coupled modelling approach is that it can elucidate storage-flux relationships in a way that reconciles the rapid wave speed

celerity of the hydrological response (e.g., rainfall-runoff transformations) and the slower pore velocities of water particles (i.e., residence times) implied by conservative tracers (Weiler, 2003; McDonnell and Beven, 2014). Therefore, the transit times of water can be estimated along with simulating the hydrograph response (McGuire and McDonnell, 2006; Hrachowitz et al., 2013). Transit times have previously been calculated as a mean or distribution. More recent studies have demonstrated the importance of characterising the time variability of streamwater age (Botter et al., 2010; McMillan et al., 2012; Hrachowitz et al., 2013; Harman, 2015; Rinaldo et al., 2015; Benettin et al., 2017). Tracking water fluxes through the internal stores of semi-distributed models is one approach to improve our understanding of the evolution and non-stationary nature of water ages (Fenicia et al., 2010; Birkel et al., 2015; Soulsby et al., 2015). Through characterising these ages, insights into the non-linearities in catchment storage dynamics and runoff generation processes are revealed. Additionally, incorporating isotope tracers facilitates multi-objective calibration, which provides the opportunity to improve model evaluation and constrain parameter sets potentially reducing equifinality (Beven, 1993; Birkel and Soulsby, 2015; Finger et al., 2015). A potential drawback is increased model complexity introduced through additional mixing parameters for tracers. However, with care it seems the insightful information gained outweighs this negative (Seibert, 2003b).

As conceptual models rely on calibration, a key challenge is obtaining parameter sets that reflect a physically meaningful catchment behaviour (Gharari et al., 2013). A major issue with calibrated parameters is their time dependency, which may limit transferability to conditions different from those during calibration (Gharari et al., 2013; Magand et al., 2015; Thirel et al., 2015a; Yu and Zhu, 2015). Hence, the optimum parameter set for one observation period can be significantly different for another, and models may fail to

provide robust simulations outside of the calibration conditions (Seibert et al., 2003; Beven, 2012). This issue has been increasingly recognised in recent years and studies have used tests, similar to differential split sample tests (DSST), to determine model performance outwith calibration conditions (Seibert, 2003b; Thirel et al., 2015; Yu and Zhu, 2015). Such tests involve calibrating and validating parameter sets over contrasting periods with the aim of obtaining good representation in all conditions (Thirel et al., 2015a). DSSTs help to identify weaknesses in the model structure by investigating parameters that are time variant (Clark et al., 2008; Gharari et al., 2013).

While conditions vary in time, they also vary in space. Hence, an important consideration is whether models developed at one scale can be upscaled or downscaled (Bloschl and Sivapalan, 1995; Didszun and Uhlenbrook, 2008). If the hydrological response units (HRU) remain the same between scales, then models are likely transferable (Flügel, 1995; Didszun and Uhlenbrook, 2008). The proxy-basin test, proposed by Klemeš (1986), is one way to test the spatial transferability of models, whereby, the model is calibrated on one basin and validated on another.

A recent appraisal of tracer-aided modelling by Birkel and Soulsby (2015) identified the need for using high temporal resolution data for model conceptualisation. Previously, isotopes were typically sampled at weekly or daily timescales, where sub-daily variability can be obscured by averaging (McGuire et al., 2005; Rodgers et al., 2005; Tetzlaff et al., 2007). Birkel et al. (2012), found that weekly isotope data did not always capture the daily variability, and daily sampling failed to capture isotope dynamics revealed by 4 hourly sampling. Emergence of improved laser spectroscopy technology now aids high-

frequency capture of data at lower analytical costs, increasing our ability to fully characterise isotope dynamics (Kirchner et al., 2004; Lyon et al., 2009). Previous studies using higher frequency data tended to focus on single event data (Weiler, 2003; Carey and Quinton, 2005; Wissmeier and Uhlenbrook, 2007) which, although produced insightful contributions, did not reveal what happens during low flows or over longer time periods (Birkel and Soulsby, 2015). It is important to assess the insights gained by higher frequency sampling in order to minimise the risk of information loss and also to test the ability of models, developed on coarser resolutions (e.g. daily or weekly), to successfully simulate higher resolution data (e.g. sub-daily) (Kirchner et al., 2004).

In this paper we used a modified version of an existing tracer-aided model within a calibration learning framework to enhance our understanding of hydrological processes and model behaviour across non-stationary conditions. The framework was centred on three specific objectives:

1. To evaluate the additional information content of high temporal resolution (sub-daily) isotope data.

2. To examine the effects of different calibration periods on parameters, model performance, estimated catchment storage and streamwater ages, to aid our process understanding and evaluate model structure.

3. To assess the model transferability to a smaller, nested catchment (< 1 km^2) to determine whether the dominant hydrological processes remain the same.

We incorporated higher resolution 6-hourly isotopes sampled for very wet and dry conditions. The model was calibrated on three periods: a 12 month period using daily isotopes, the wet sub-period and the dry sub-period (6-hour sampling interval). Most previous studies that have assessed the influence of calibration periods have focused on runoff simulations (Coron et al., 2012; Brigode et al., 2015; Kling et al., 2015; Yu and Zhu, 2015). This study goes beyond this to simulate 6-hourly streamflow, deuterium (δ^2H), catchment storage and streamwater age. Finally, for the first time, the model was downscaled from a 3.2 km^2 catchment to a nested < 1 km^2 headwater.

2. Study site

The study focused on the Bruntland Burn (BB), a 3.2 km^2 catchment in the Cairngorms National Park, Scotland. It is a sub-catchment of the Girnock Burn (31 km^2), which drains into the River Dee. Climate is temperate/boreal oceanic with mean annual air temperatures of ~6 °C, ranging between 1 °C in winter to 12 °C in summer. Mean annual precipitation is ~1000 mm and lacks seasonal variability as it is dominated by low intensity events through the year; <10 % of precipitation falls as snow. Annual evapotranspiration is around 400 mm focused on the summer months. Glaciation has formed a valley with steep slopes and a wide bottom (Figure 1a) overlain by glacial till. This till covers 70 % of the catchment and is up to 40 m deep in the valley (Soulsby et al., 2016); resulting in high water storage and a significant contribution of groundwater to flow (Birkel et al., 2011). Soils in the valley bottom are organic-rich peats and peaty gleys (Figure 1b). These remain close to saturation throughout the year and facilitate saturation excess overland flow during rainfall events. Antecedent conditions control the extent of the saturation area, which varies between 2-40 % of the catchment. When the dynamic riparian saturation extent exceeds ~20 %, steeper hillslopes become connected

to the riparian area, facilitating lateral flow of runoff from the hillslopes. These are dominated by more freely draining shallow podzol and rankers which usually facilitate groundwater recharge. Dominant vegetation cover is *Sphagnum* mosses and *Molinia* grass on the peaty soils, and heather (*Calluna*) in the steeper slopes. Forest cover is limited to small areas of Scots Pine (*Pinus sylvestris*) on steeper slopes. More detail is given in earlier work (e.g. Tetzlaff et al., 2014; Geris et al., 2015).

Nested within the BB, is a south-facing 0.65 km^2 sub-catchment (HW1, Figure 1a). This is characterised by an extensive raised (ombrotrophic) riparian peat bog and has a higher percentage of peat soils than BB (15 % compared to 9 %, respectively). Furthermore, it has a higher percentage of peat fringing the stream channel (81 %) compared to BB (53 %). Depressions in the peat allow pools of water to form, which are dynamically connected/disconnected to the stream and peatland drainage network (Lessels et al., 2016). Near-stream peat is constantly connected to the stream, facilitating high baseflow and high dissolved organic carbon concentrations (Tunaley et al., 2017). The surrounding areas of bog receive groundwater from the hillslopes (Lessels et al., 2016).

3. Data and methods

3.1 Hydrological and isotope data

Monitoring occurred between 1 May 2014 and 1 August 2015. Discharge was calculated at 15 minute intervals from stage height measurements at the catchment outlets of BB and HW1 (Figure 1a). Precipitation was measured every 15 minutes from rain gauges within BB and HW1 (Figure 1a). Potential evapotranspiration was estimated using a

modified Penman-Monteith equation (Dunn and Mackay, 1995), based on meteorological data from an nearby automatic weather station (~ 1 km away). Streamwater samples for stable isotope analysis were collected daily at 14:00 from the outlet of both catchments using ISCO 3700 autosamplers. Integrated daily precipitation samples were also taken from the outlet of the BB; given the similar altitude and close proximity they were assumed to be representative of both catchments. In addition, two periods of 6-hourly sampling of both streamwater and precipitation isotopes took place. The first was between 27 October and 20 November 2014 (Nov), representing a wetter, colder period. The second was between 2 and 20 July 2015 (Jul), representing a drier, warmer period. Paraffin was added to bottles to prevent evaporation within the auto-samplers in the field. Samples were analyzed in the laboratory for δ^2H and oxygen-18 ($\delta^{18}O$) using a Los Gatos DLT-100 laser isotope analyzer (precision of +/- 0.4 ‰ for δ^2H; +/-0.1 ‰ for $\delta^{18}O$). Results are expressed in δ notation according to the Vienna Standard Mean Ocean Water standards. Due to the higher precision, we used the δ^2H data in the model. For modelling purposes, the hydrological and isotope data were aggregated into a 6-hourly dataset.

3.2 Modelling approach

3.2.1 Model structure

The coupled flow-tracer model used in this study was developed by Birkel et al. (2010, 2011, 2014, 2015) and Soulsby et al. (2015). A brief overview follows, but readers are referred to these original papers for full details. Figure 2 shows the model structure, the connections between the stores and the basic equations. The model is characterised by three linked reservoirs representing the upper hillslopes, the dynamic riparian saturation area and a groundwater store. These have associated dynamic storage, S_{up}, S_{sat} and $S_{low,}$

respectively. Central to the model is the non-linear streamflow response that conceptualizes the hydrological connectivity of the catchment linking the three conceptual stores. The model uses five calibration parameters to simulate discharge (*a*, *r*, *b*, *k* and α) and a further three additional mixing volumes for each reservoir (upS_p, $satS_p$ and $lowS_p$). The linear rate parameter *a* (6 hr $^{-1}$) controls the hillslope water flux to the saturated area; *r* (6 hr $^{-1}$) controls the groundwater recharge rate; *b* (6 hr $^{-1}$) controls the rate of groundwater discharge to streamflow; *k* (6 hr $^{-1}$) and α conceptualises saturation overland flow and controls the nonlinear runoff from the saturation area to streamflow.

The calibrated mixing volumes (upS_p, $satS_p$ and $lowS_p$, in mm), used to damp out isotope variability, did not affect the dynamic water storage and fluxes, hence allowing for the differences between celerity and velocity to be captured (Birkel et al., 2011). A key feature of the model is the dynamic non-linear variation of the saturation area (dSAT) as a way to generate time-variable mixing volumes (*MV*). A simple antecedent precipitation index-type algorithm was used to derive dSAT (Birkel et al., 2010). dSAT was used both to distribute precipitation inputs between the hillslope and saturation area and also to convert the storage parameters into time-variable mixing volumes (MV). The greater the catchment wetness, the greater the saturation area extent and the potential for mixing (satMV). In the hillslope reservoir, the mixing volume (upMV) decreases as the saturation area expands. Catchment storage was the sum of dynamic storage and the additional storage for isotope mixing. Using daily isotopes to time-stamp and track daily precipitation, as well as input and output fluxes through the reservoirs, the age of the streamwaters could be estimated (see Hrachowitz et al., 2013). Streamwater age was extracted by integrating the time-variant contribution of the differently aged water fluxes from the three stores, giving non-linear mixing at the catchment scale (Birkel et al., 2015).

Previous work in the BB has shown the potential for isotopic fractionation, resulting in surface waters becoming relatively depleted in δ^2H compared to $\delta^{18}O$ and plotting below the local meteoric water line (LMWL) (Lessels et al., 2016; Sprenger et al., 2017). There was evidence of isotopic fractionation at both sites (Figure 3), occurring predominately during the summer and autumn. Therefore, we incorporated evaporative fractionation processes in S_{up} and S_{sat}. The fractionation scheme was based on Gibson (2002) but differs in that it is time variable on a 6-hourly timescale.

$$[1]\ \delta_L(t) = \delta_S - (\delta_S - \delta_0)\exp[-(1+mx)\left(\frac{It}{V}\right)]$$

where δ_L is the change in isotopic composition with time (t), δ_0 is the initial isotopic composition before evaporation, V is the volume of liquid undergoing evaporation, I is the inflow, *x* is the evaporation to inflow ratio, δ_S is the steady-state isotopic composition of the water under constant meteorological conditions (Gonfiantini, 1986) and *m* is the enrichment slope, or rate, of heavy isotope build-up (see Stadnyk et al., 2013 for details).

3.2.2Model calibration and evaluation

The calibration procedure was based on differential split sample tests (DSST). The eight model parameters were calibrated over a one year period (1 August 2014 – 1 August 2015) and two sub-periods: Nov wet period (27 October – 20 November 2014) and Jul dry period (2 – 20 July 2015). These two periods correspond to the high resolution (6 hr) event sampling. The model parameters were optimised using a multi-objective optimisation algorithm (NSGA2 by Deb et al., 2002) that simultaneously optimised the

modified Kling-Gupta efficiency (KGE) for both discharge and isotopes (Kling et al., 2012):

$$[2]\ KGE = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2}$$

where r is the correlation coefficient between simulated and observed values, β is the ratio between the mean simulated and mean observed values and γ is the ratio of the variability between the simulated and observed values. The KGE was used based on a qualitative assessment of the trade-offs between different alternatives and on its use in previous dual calibration studies involving stable isotopes (Birkel and Soulsby, 2016; Soulsby et al., 2015). The widely used Nash-Sutcliffe efficiency (NSE) was not applied as the damped response of isotope data is less well assessed using this metric. In contrast, the KGE uses the Euclidian distance of the three components from an ideal point and is thus well suited to tracer data. The optimisation included 500 parameter sets, which were constrained over 50 iterations. Although no formal uncertainty analysis was undertaken, the simulation ranges of the final 500 best performing parameter sets were used as an indication of parameter variability (Andrews et al., 2011). The optimised parameter sets from the three calibrations were used to simulate discharge and δ^2H over the three different time periods (one year, Nov and Jul) to cross-validate and assess the transferability of different calibrations. This generated nine simulations for which the KGE values were calculated. To assess the model transferability to a smaller scale catchment, the process was repeated for HW1 and a cross basin test was performed, whereby, the parameter set calibrated to HW1 was applied to the BB and vice versa.

4. Results

4.1 Hydrological and isotopic variability

The study began in August 2014, which was wet (Figure 4), with monthly rainfall being 280 % of the 1971−2000 average (see Hannaford et al., 2014). Wetter than average conditions continued during October and November, with monthly rainfall 175 % and 148 % of the 1971−2000 average, respectively (Hannaford et al., 2014, Parry et al., 2014). Thereafter, monthly rainfall was around average at the start of 2015. There was a brief dry period in June 2015, preceding a wet July 2015 when precipitation was 196 % of the long-term average (Parry et al., 2015).

Figure 4 also shows the two high-frequency sampling periods (shaded) in the context of the year-long daily sampling frequency. For the Nov period, total rainfall of the preceding 30 days (P_{30}) were four times higher (117.2 mm) than for the Jul period (25.8 mm). Antecedent precipitation for the one year calibration period had a P_{30} of 73.6 mm (Table 1). These differences were reflected in the maximum and mean discharges during the calibration (Q_{max} and Q_{mean}). During Nov, the Q_{max} was the second highest flow of the year (5.66 mm 6 h^{-1}), whereas, in Jul the Q_{max} was much lower (2.84 mm 6 h^{-1}, Figure 4c). The Nov period also had the highest Q_{mean} (1.47 mm 6 h^{-1}), compared with Jul (0.23 mm 6 h^{-1}) and the year overall (0.46 mm 6 h^{-1}).

Given the close proximity, precipitation inputs for BB and HW1 were very similar (Table 1); though there were subtle differences in flow. Generally BB had larger high flows (Q_5 = 1.37 mm 6 hr^{-1}) than HW1 (Q_5 = 1.31 mm 6 hr^{-1}) and lower low flows (Q_{95} = 0.08 mm

6 hr^{-1}) compared to HW1 (Q_{95} = 0.09 mm 6 hr^{-1}). However, during the one year and November period, HW1 had the highest Q_{peak} (Table 1).

Deuterium in precipitation was highly variable throughout the year (δ^2H range = -145.9 ‰ to -13.9 ‰, CV = 44.7 %) (Figure 4a, Table 1) due to the relatively uniformly distributed wet climate but differing air mass sources. The precipitation most depleted δ^2H occurred in autumn, corresponding to the wettest period. In contrast, streamwater δ^2H was damped (range in the BB = -72.2 ‰ to -51.2 ‰, CV = 5.3 %), reflecting mixing of precipitation with pre-event waters within the catchment (Figure 4b). The most depleted streamwater occurred during autumn and winter events, whilst the most enriched values occurred during summer. Disparities in hydrological conditions between Nov and Jul periods were reflected in the isotopes (Table 1) with Jul having a more enriched range of streamwater δ^2H (-58.5 ‰ to -51.2 ‰) compared to the Nov period (-65.2 ‰ to -57.5 ‰).

Overall, HW1 had a more enriched mean stream δ^2H: -57.7 ‰ compared to -58.7 ‰ for the BB. During Nov (Figure 5), the mean δ^2H of streamwaters were very similar for BB and HW1; they initially depleted during the event and thereafter reflected the precipitation isotope dynamics, albeit very damped (Table 1). Unfortunately the 6-hourly streamwater response to the relatively enriched precipitation at the beginning of the Nov period was missed in HW1 due to technical issues, resulting in a lower sample number (114 in HW1, compared to 131 in BB). In Jul signatures in the HW1 were more enriched than the BB (Figure 6), which reflected the flushing of fractionated peat waters (Sprenger et al., 2017).

Figures 5 and 6 also show the 6-hourly isotope dynamics (open circles) compared to daily sampling (filled circles). For both periods, the general dynamics were well-captured by daily sampling. However, the daily isotope sampling regimes usually missed event peaks which resulted in the range in δ^2H and coefficient of variation (CV) for daily data being lower. For example, for the Nov period the daily range was -64.7 ‰ to -60.2 ‰ and CV 1.3 %, compared to a range for 6-hourly sampling of -65.2 ‰ to -57.5 ‰ and CV 2.4 %. However, for the Jul period this was less evident between 6-hourly and daily sampling with CVs of 2.9 % and 2.7 %, respectively.

4.2 Sensitivity of model parameters to different calibration periods in the BB

Calibration performances of the 500 best parameter sets for each period and the parameter mean and ranges are shown in Table 2. For discharge (Q) the one year calibration performed best (mean KGE of 0.72 compared to 0.64 and 0.60 for Nov and Jul calibration, respectively). For δ^2H, the one year calibration also performed best (KGE 0.75 compared to 0.64 and 0.58 for the Nov and Jul, respectively). The July calibration favoured lower ranges of parameter *b* and a higher non-linear surface water runoff (*k* and α) compared to the other periods, suggesting groundwater fluxes were lower and response from the saturation area was higher. For the mixing volume parameters, higher values of *upSp*, and lower *satSp* and *lowSp*, when calibrated over the Nov period, inferred more relative mixing in the upper hillslopes and limited mixing in both the riparian saturation area and groundwater. The mean *upSp* was lowest for the Jul calibration period, indicating more limited summer contributions from the upper hillslopes.

4.3 Temporal transferability of parameter sets in the BB

We compared the performances (the KGE) of the three parameter sets, when applied to the alternative periods (Figure 7). For discharge, the one year parameters performed reasonably well in the Nov period (KGE = 0.63) but poorly for July (KGE = -0.03). The Nov calibration parameter set performed quite well over the other periods. The Jul parameter set performed badly for both the one year (KGE = 0.25) and the Nov period (KGE = 0.12). This was likely the result of calibrating across a low proportion of the discharge range.

For isotopes, the one year, Nov and Jul parameter sets performed well over the periods they were calibrated, as expected. However, there was substantial deterioration in the KGEs for the different evaluation periods. The Nov parameter set performed particularly badly over one year (KGE = 0.3) and Jul (KGE = -1.5), which likely reflected calibration to a small range of the most depleted streamwater samples. The Jul parameters performed better than the Nov parameter set over the one year period (KGE = 0.51), probably due to better capturing the summer fractionation. However, it performed badly for Nov (KGE = -0.57), again likely the result of being calibrated to a small range of the most enriched isotope samples. Additionally, the one year calibration parameter set performed badly over the Jul (KGE = 0.44) and Nov period (KGE = 0.22). The poor performance over the Nov period was likely due to the modelled isotopes not recovering fully from the large event prior to the Nov sampling period. NSE was also calculated as an additional statistical test and results showed similar relative differences in temporal transferability

for discharge. However for isotope there were no good performances highlighting the inappropriateness of using NSE for isotopes in this study.

Comparison of the median observations and simulations across the three different parameter sets show that the overall dynamics are captured quite well (Figure 8 and 9). However, small peaks and some large events (e.g. autumn) tend to be underestimated for all calibrations which can occur as a result of idiosyncrasies of KGE based calibration. However, this effect is less severe compared to that of NSE (Gupta and Kling, 2009). The one year parameter set did not capture small peaks particularly well, during dry periods (e.g. May and June) and wetting up periods (e.g. July 2015). This set also underestimated baseflows during wet periods, whilst overestimating them slightly during dry periods. However, this model captured the dynamics best out of the three calibrations. The Nov parameter set reproduced baseflow well throughout the whole year, though still underestimated peak flows. Re-wetting in July was reproduced as well as the smaller peaks. The Jul calibration appreciably underestimated winter baseflows, which probably reflects the lower value of the groundwater flux rate (*b*). It also produced a much flashier response, with an over-steep falling limb during wetter periods. During drier periods baseflow and peaks were general well-captured.

The isotope component simulations of each parameter set captured the large damping of δ^2H in streamwater compared to precipitation (Figure 9). The parameter set calibrated to the full year performed well overall and captured the dynamics and range of the observed data until the summer where the simulations under-predicted isotope values and were less responsive to small changes. The Nov parameter set exaggerated the flashiness of the

isotope dynamics during larger events, particularly during the re-wetting period. For the Jul parameter set simulations, depletion during events were over-estimated, predicting more depleted and flashy values than the other parameter sets. The under prediction of δ^2H was evident from all parameter sets during the drier May and June 2015, indicating that the fractionation dynamics were not well-captured.

4.4 Storage and age estimates in the BB

The estimated catchment storage dynamics for the different parameter sets are shown in Figure 10(b). Storage was highest for the Jul parameter set simulation (1492 mm, SD = $\pm$ 111 mm); the one year parameter set yielded a similar value (1400 mm, SD = $\pm$ 41), whilst a substantially lower storage for the Nov parameters (512 mm, SD = $\pm$ 27).

The non-linear interactions between different landscape units with different storage dynamics controlled the non-stationary streamwater ages derived from the three parameter sets. Estimated water ages were linked to the associated storage values, with the parameters sets generating higher storage values resulting in overall older water ages (Table 3). Estimated mean streamwater age derived from the one year calibration was 321 days (± 239 days), similar to the Jul calibration with 353 days (± 263 days), the highest of the three models. The mean streamwater age derived from the Nov calibration was 150 days (± 69 days), the youngest of the three model set ups. Despite considerable uncertainty we focused on the relative differences between the calibration periods and used the mean values, which constrained the likely water ages. The streamwater age was time variant, with younger waters (1-30 days) during events and winter wet periods, whilst older waters (300-800 days) occurred during dry summer periods (Figure 10c).

There was less variability and seasonality in streamwater ages produced from the Nov parameter set.

4.5 Spatial transferability of parameter sets between BB and HW1

To determine whether the dominant hydrological processes remain the same when downscaling, the model was calibrated on HW1 data as well as the BB. The calibration performances (KGEs), for the 500 retained parameter sets for each period, are shown in Table 2. Both catchments had very similar performances across the calibration periods (Figure 7) and the values of the retained parameter sets were very similar. A subtle difference was found between the isotope mixing volume parameters, *upSp* and *satSp*, which were slightly higher for HW1 across all three calibration periods. Conversely, *lowSp* was slightly lower in HW1. Estimated storage and streamwater age for BB and HW1 were also very similar (Table 3). Indeed, the parameter sets for each catchment could provide a reasonable simulation to the other and only small differences in performance occurred (Table 4). When the HW1 parameter set was used to simulate BB discharge and isotopes, the mean KGEs (across all periods) were 0.62 and 0.69, respectively, compared to 0.64 and 0.73 when the BB parameter set was used. When the BB parameter set was used to simulate HW1 discharge and isotopes, the mean KGEs were 0.67 and 0.72, respectively, compared to 0.64 and 0.69 when the HW1 parameter set was used. Hence, using BB slightly improved HW1 simulations, particularly for the Jul calibration period.

5. Discussion

5.1 Importance of high resolution data in tracer-aided modelling

We integrated 6-hourly δ^2H data into a tracer-aided runoff model to evaluate the value of high resolution data on model performance and information gain. The model was previously developed based on weekly and daily data. The need for collecting higher frequency isotope data to improve process representation and modelling capabilities at finer temporal scales was highlighted by McDonnell and Beven (2014), and Birkel and Soulsby (2015). Using sub-daily data in models developed for daily time series one would expect a decrease in model performance due to the increased variability in the input data. However, here, the sub-daily data were successfully incorporated into the model with overall good performance resulting in the simulation of discharge, isotopes, water age and storage on a 6-hourly frequency. Our study, thus, helps bridge the gap of matching the temporal dynamics of the isotopic response to the hydrometric response and, hence, improves our ability to understand the different mechanistic controls on celerities and velocities within a catchment (Kirchner, 2003; McDonnell and Beven, 2014).

In the case of the BB, the high-frequency data provided confirmatory evidence that the 'isostat' behaviour of the riparian peatlands – that is the mixing of different source waters and damping the streamwater isotope signal (Tetzlaff et al., 2014) – is also dominating on sub-daily time scales. Although the higher temporal resolution data provided limited new process insights, it highlighted the dominant role of rapid mixing with older waters within the riparian area in the storm period response (Tetzlaff et al., 2014). Similar dampening and old water dominance has been shown in other environments elsewhere (e.g. Berman et al. 2009). Given the modest increase in the information content of data

gleaned through the 6-hourly sampling in the BB, it is difficult to justify the quadrupling of the resulting logistical and analytical load. However, in other more dynamic environments, such as the wet tropics (Birkel and Soulsby, 2016), catchments affected by snow and glacial melt (Ohlanders et al., 2013; Peralta-Tapia et al., 2016) or urbanised catchments (Jefferson et al., 2015; Soulsby et al., 2015b), high-frequency isotope measurements are likely to yield more significant new insights. Overall, it is important to evaluate the addition of high-frequency data in order to identify the minimal periodicity of isotope sampling required to characterise catchment response (Seibert and Beven, 2009; Seibert and McDonnell, 2015).

5.2 Use of different calibration periods to test process conceptualisation and inform model structure

Differential split sample tests have been widely used in calibration to test a model's skill beyond the calibration conditions (Klemeš, 1986; Seibert, 2003; Chiew et al., 2009; Vaze et al., 2010; Seiller et al., 2012; Brigode et al., 2015; Magand et al., 2015; Thirel et al., 2015b; Zhou et al., 2015). Here we focused on short sub-periods of high-frequency data with different hydrometeorological conditions and used the calibrated parameter values to explore inferences about non-stationary catchment processes (Herbst and Casper, 2008); and assess the implication of the transferability of these parameter sets.

Calibration to a particularly wet period in Nov resulted in the behavioural parameters inferring reduced volumes of storage which dampened the tracer signal within the saturation area, as simulated by a low *satSp* parameter. This parameter behaviour caused

some precipitation to be routed laterally to the stream channel with limited mixing. The majority of mixing occurred in the hillslopes, as inferred by high *upSp* parameter, which subsequently drained water into the riparian saturation area. The transfer of this parameter set to drier periods resulted in an exaggerated isotope response and poor model performances, due to limited mixing and lack of connectivity to the hillslopes. On the other hand, this parameter set simulated discharge quite well over a range of conditions, probably due to the calibration being a wet period that captured a non-linear runoff response over a range of flows. Such differences between discharge and isotope transferability performance highlight the importance of tracer-aided runoff models to reveal more fully the catchment response in terms of the different mechanistic controls on both wave celerites and particle velocities (McDonnell and Beven, 2014). The more rapid water turnover in wetter conditions captured by the Nov parameter set, resulted in the model estimating younger streamwater ages. When catchment wetness increased, the effective storage available for mixing decreased, because more water moves laterally to the stream rather than recharging groundwater, and the age of the water decreased (Birkel et al., 2015; Harman, 2015; Soulsby et al., 2015; Van Huijgevoort et al., 2016). Of course, the total catchment storage is actually higher in these wetter periods, but the process conceptualisation in the model infers a decrease in the storage that is able to mix tracers (Soulsby et al., 2015), a processes termed as the inverse storage effect by Harman (2015). The younger water ages in this wet period are consistent with results derived from a longer term study focusing on the temporal variation in water ages of the main hydrological response units. This study showed that water in the saturated riparian zone was much younger (~1 month), compared to older deeper groundwater (~4 years) (Soulsby et al., 2016).

The parameter set derived from calibration to a wetting up period with dry antecedent conditions (Jul period) resulted in a low groundwater recharge paramater (*b*) for simulating baseflows. Consequently, the transfer of this parameter set to winter resulted in baseflows being significantly underestimated. Furthermore, the small summer discharge peaks encompassed by the calibration resulted in a model failure in simulating higher peak flows (Seibert, 2003; Brigode et al., 2015). The flashy response of both isotopes and discharge simulations during the Jul period was a consequence of the high non-linear surface water sources (higher *k* and α), generating quick simulated runoff from the riparian saturation area combined with limited connectivity to the hillslopes. However, the corollary is that when this parameter set was applied to wet periods, the non-linear surface water runoff generation underestimated mixing, producing an exaggerated isotope response with precipitous discharge recessions. Resulting water age estimations were overall older than those obtained through calibrating on the Nov period due to the higher mixing in the groundwater stores and higher influence of groundwater, consistent with empirical data (Blumstock et al., 2015). The differences in the estimated streamwater age derived from the calibration to 6-hourly data in short wet and dry periods were consistent with those produced by calibration of weekly data in wet (1.1 yr) and dry (1.6 yr) conditions in the larger Girnock catchment (Birkel et al., 2015). Variability in streamwater age was higher for the Jul calibration due to the marked switch from groundwater dominance in dry periods to the younger surface waters produced during small events. This is also consistent with Soulsby et al. (2015) and Tunaley et al. (2016), who showed that youngest waters in the BB were transmitted to the stream during small events with dry antecedent conditions. In contrast, the wet period calibration showed limited age variability due to the lower groundwater influence. The incorporation of water

ages allowed us to test the value of different calibrations on parameter sets and to link this to integrated catchment processes.

Unsurprisingly, calibration over the whole year captured the dynamics of isotopes and flow most effectively, with the best performance statistics as the full range of catchment responses were used. However, focusing on the overall results, and not specific sub-periods, can obscure poor model performance (Guse et al., 2014; Andréassian et al., 2012). For runoff simulations, the one year parameter set performed better during wet periods, than dry periods, as it was less well-able to capture the marked non-linearities (lower k and α) that occurred during the smaller summer events and after dry periods. This weakness was identified in previous versions of the model (Birkel et al., 2014) and likely reflects the spatial heterogeneity of the saturation areas being not fully represented. Field observations have shown that in small events with dry antecedent conditions, connectivity increases to link isolated small pools in the riparian peatland (Lessels et al., 2016). The conceptualisation of this spatial dynamic into a lumped runoff model would require additional parameters and thus, likely increase uncertainty (Birkel et al., 2014).

In addition, the model lacks the skill to capture the summer isotopic enrichment in the stream, due to evaporative fractionation, despite its conceptualisation. This has been identified as a major weakness of the model (Birkel et al., 2011; Soulsby et al., 2015) and the implementation of a new time variable fractionation scheme in the hillslope and saturation area here resulted in only limited improvement. Again, field data imply the fractionation also occurs intensely in small localised areas that connect and disconnect in a non-linear way and more detailed data on the microclimate of these areas may be needed

to improve the modelling (Lessels et al., 2016; Sprenger et al.,2017). Such spatially explicit processes can be incorporated in the recent development of a semi-distributed model structure, which captures smaller scale dynamics in connectivity (Van Huijgevoort et al., 2016). Nevertheless, mean streamwater age estimates from the one year calibration were younger (~ 1 year) compared to estimates by Van Huijgevoort et al. (2016) and Soulsby et al. (2015) who reported ages of ~ 1.6 years and 1.8 years, respectively. These estimates were based on multiyear datasets encompassing some extreme wet and dry periods. The younger ages reported here were likely related to it being a wet year, particularly in the first 6 months.

Recent approaches for identifying how the dominant hydrological processes vary temporally include analysing the temporal dynamics of parameter sensitivity (TEDPAS) (Sieber and Uhlenbrook, 2005; Reusser et al., 2009; Guse et al., 2014). Using the much more highly parameterised SWAT model, Guse et al. (2016) related TEDPAS to specific discharge magnitudes to show, for example, whether high sensitives were related to certain discharge magnitudes and, in turn, demonstrated how the dominant hydrological processes vary depending on discharge. Future work should focus on incorporating these more sophisticated techniques in the search for calibration methods that make better use of the information content of the available data (Wagener et al., 2003). However, here, we have focused on showing how a relatively simple test can provide insights to both model performance of a low parameter model and catchment behaviour.

5.3 Model transferability between spatial scales

Potential difficulties with downscaling models due to the possible change in dominant processes with scale have been highlighted previously (Beven et al., 2001). However, calibrating the model to the smaller HW1 had only very subtle impacts on model parameterisation and performance. The slightly better performance of the BB parameter set on HW1 during Jul was possibly due to the greater area of peatland in HW1 and therefore a more marked non-linearity in flow response which is better captured in the BB calibration. Field observations in both HW1 and BB have shown subtle discrepancies in hydrological responses caused by differences in percentage riparian peatland (Tunaley et al., 2017), GW influence (Blumstock et al., 2015) and solar radiation (Dick et al., 2015). One difference evident from the isotope measurements was the more enriched isotope values during summer in HW1 compared to the BB, likely due to enhanced evaporation fractionation in the peatland pools. However, considering that the model failed to capture this fractionation well at either scale, this more subtle difference between the catchments was missed. Nonetheless, the dominant hydrological processes occurring in HW1 were adequately captured by the processes within the model developed for the BB, and calibrated parameters could be transferred between the catchments with similar performance.

6. Conclusion

We examined the use of models as learning tools to improve our understanding of both hydrological processes and model behaviour across non-stationary conditions. The learning framework was split into three objectives: (1) testing the model on high-frequency data; (2) testing the effect of different calibration strategies; and (3) testing model transferability downscale. The main outcomes of this study were as follows:

(1) The 6-hourly data were successfully incorporated into the model, improving the temporal resolution of the modelling, making it more relevant to the time dynamics of the isotopic and hydrometric response of the catchments. However, in the peat-influenced wet environment of the BB, the increased periodicity of isotope samples provided limited new process insights, but provided confirmatory evidence of the dominant role of mixing within the riparian area.

(2) By incorporating a calibration approach based on splitting the full data set into sub-periods, we were able to link the time variance of parameter values to different hydrological conditions, providing insights into the non-stationary nature of the dominant runoff processes. During wet periods, increased saturation results in a decrease in the storage actively involved in mixing, which causes younger water ages as more precipitation is routed laterally to the stream channel. Events with dry antecedent conditions result in a switch from groundwater domination, with associated older water ages, to a high contribution of non-linear surface water generating quick, runoff from the saturation area, resulting in young water ages. Model diagnostics on the full study period revealed poorer performance during wetting up periods, and a failure to capture the summer fractionation, due to connectivity being non-linear and spatially explicit. Hence, future work should focus on incorporating these processes into a spatially distributed model.

(3) Downscaling of the model to a < 1 km^2 catchment produced very similar parameter values and model performances, which showed the consistency of the model when applied to smaller scales and highlighted the similarity in dominant hydrological processes between the two scales.

Overall, the study highlights that by incorporating models into an integrated learning framework, with dual calibration on discharge and tracer data, we are able to extract an increased amount of information from the data and model results, and can evaluate models more rigorously than in the past.

Acknowledgements

The authors would like to thank Jonathan Dick and Audrey Innes for lab analysis and preparation of the isotope samples. In addition, we would like to thank Iain Malcolm (Marine Scotland Science) for providing AWS data. Finally, we gratefully acknowledge the European Research Council ERC (project GA 335910 VeWa) for funding the VeWa project. The data used are available from the authors. CB acknowledges support from the University of Costa Rica (project 217-B4-239 and the Isotope Network for Tropical Ecosystem Studies (ISONet)).

References

Ala-aho, P. Tetzlaff, D. Laudon, H., McNamara, J. and Soulsby, C. 2017. Using isotopes to constrain water flux and age estimates in snow-influenced catchments using the STARR (Spatially distributed Tracer-Aided Rainfall-Runoff) model. Hydrol. Earth Syst. Sci. Discuss., doi: https://doi.org/10.5194/hess-2017-106

Andréassian, V., Le Moine, N., Perrin, C., Ramos, M.H., Oudin, L., Mathevet, T., Lerat, J., Berthet, L. 2012. All that glitters is not gold: The case of calibrating hydrological models. Hydrol. Process. 26, 2206–2210. doi:10.1002/hyp.9264

Andrews, F., Croke, B., Jakeman, A. 2011. An open software environment for hydrological model assessment and development. Environ. Model. Softw. 26, 1171–1185. doi:10.1016/j. envsoft.2011.04.006

Benettin, P. Soulsby, C., Birkel, C., Tetzlaff, D., Botter, G. and Rinaldo, A. 2017. Using SAS functions and high resolution isotope data to unravel travel time distributions in headwater catchments. Water Resour. Res. 53, 1864-1878, doi: 10.1002/2016WR020117.

Berman, E.S.F., Gupta, M., Gabrielli, C., Garland, T., McDonnell, J.J. 2009. High-frequency field-deployable isotope analyzer for hydrological applications. Water Resour. Res. 45, W10201. doi:10.1029/2009WR008265

Beven, K. 2012. Rainfall-Runoff Modelling: The Primer, 2nd ed. Wiley-Blackwell, Chichester, UK.

Beven, K. 2001. How far can we go in distributed hydrological modelling? Hydrol. Earth Syst. Sci. 5, 1–12. doi:10.5194/hess-5-1-2001

Beven, K. 1993. Prophecy, reality and uncertainty in distributed hydrological modelling. Adv. Water Resour. 16, 41–51. doi:10.1016/0309-1708(93)90028-E

Birkel, C., Soulsby, C. 2016. Linking tracers, water age and conceptual models to identify dominant runoff processes in a sparsely monitored humid tropical catchment. Hydrol. Process. doi: 10.1002/hyp.10941

Birkel, C., Soulsby, C. 2015. Advancing tracer-aided rainfall-runoff modelling: A review of progress, problems and unrealised potential. Hydrol. Process. 29, 5227–5240. doi:10.1002/hyp.10594

Birkel, C., Soulsby, C., Tetzlaff, D. 2015. Conceptual modelling to assess how the interplay of hydrological connectivity, catchment storage and tracer dynamics controls nonstationary water age estimates. Hydrol. Process. 29, 2956–2969. doi:10.1002/hyp.10414

Birkel, C., Soulsby, C., Tetzlaff, D. 2014. Developing a consistent process-based conceptualization of catchment functioning using measurements of internal state variables. Water Resour. Res. 50, 3481–3501. doi:10.1002/2013WR014925

Birkel, C., Soulsby, C., Tetzlaff, D., Dunn, S., Spezia, L. 2012. High-frequency storm event isotope sampling reveals time-variant transit time distributions and influence of diurnal cycles. Hydrol. Process. 26, 308–316. doi:10.1002/hyp.8210

Birkel, C., Tetzlaff, D., Dunn, S.M., Soulsby, C. 2011. Using time domain and geographic source tracers to conceptualize streamflow generation processes in lumped rainfall-runoff models. Water Resour. Res. 47, 1–15. doi:10.1029/2010WR009547

Birkel, C., Tetzlaff, D., Dunn, S.M., Soulsby, C. 2010. Towards a simple dynamic process conceptualization in rainfall-runoff models using multi-criteria calibration and tracers in temperate, upland catchments. Hydrol. Process. 24, 260–275. doi:10.1002/hyp.7478

Bishop, K., Buffam, I., Erlandsson, M., Temnerud, J., Svartberget, K., Brook, B. 2008. Aqua Incognita : the unknown headwaters. Hydrol. Process 22, 1239–1242. doi:10.1002/hyp

Blöschl, G., Sivapalan, M. 1995. Scale Issues in Hydrological Modelling : A Review. Hydrol. Process. 9, 251–290. doi: 10.1002/hyp.3360090305

Blumstock, M., Tetzlaff, D., Malcolm, I.A., Nuetzmann, G., Soulsby, C. 2015. Baseflow dynamics: Multi-tracer surveys to assess variable groundwater contributions to montane streams under low flows. J. Hydrol. 527, 1021–1033. doi:10.1016/j.jhydrol.2015.05.019

Botter, G., Bertuzzo, E., Rinaldo, A. 2010. Transport in the hydrologic response: Travel time distributions, soil moisture dynamics, and the old water paradox. Water Resour. Res. 46, 1–18. doi:10.1029/2009WR008371

Brigode, P., Paquet, E., Bernardara, P., Gailhard, J., Garavaglia, F., Ribstein, P., Bourgin, F., Perrin, C., Andréassian, V. 2015. Dependence of model-based extreme flood

estimation on the calibration period: case study of the Kamp River (Austria). Hydrol. Sci. J. 60, 1424–1437. doi:10.1080/02626667.2015.1006632

Carey, S.K., Quinton, W.L. 2005. Evaluating runoff generation during summer using hydrometric, stable isotope and hydrochemical methods in a discontinuous permafrost alpine catchment. Hydrol. Process. 19, 95–114. doi:10.1002/hyp.5764

Chiew, F.H.S., Teng, J., Vaze, J., Post, D.A., Perraud, J.M., Kirono, D.G.C., Viney, N.R. 2009. Estimating climate change impact on runoff across southeast Australia: Method, results, and implications of the modeling method. Water Resour. Res. 45, 1–17. doi:10.1029/2008WR007338

Clark, M.P., Slater, A.G., Rupp, D.E., Woods, R. a., Vrugt, J. a., Gupta, H. V., Wagener, T., Hay, L.E. 2008. Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. Water Resour. Res. 44. doi:10.1029/2007WR006735

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., Hendrickx, F. 2012. Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. Water Resour. Res., 48. doi: 10.1029/ 2011WR011721

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. 6, 181–197.

Dick, J.J., Tetzlaff, D., Soulsby, C. 2015. Landscape influence on small-scale water temperature variations in a moorland catchment. Hydrol. Process. 29, 3098–3111. doi:10.1002/hyp.10423

Didszun, J., Uhlenbrook, S. 2008. Scaling of dominant runoff generation processes: Nested catchments approach using multiple tracers. Water Resour. Res. 44, 1–15. doi:10.1029/2006WR005242

Dooge, J.C. 1982. Parameterization of hydrologic processes, in: Land Surface Processes in Atmospheric General Circulation Models. Cambridge Univeristy Press, London.

Dunn, S.M., Mackay, R. 1995. Spatial variation in evapotranspiration and the influence of land use on catchment hydrology. J. Hydrol. 171, 49–73. doi:10.1016/0022-1694(95)02733-6

Dunn, S.M., McDonnell, J.J., Vache´, K.B. 2007. Factors influencing the residence time of catchment waters: A virtual experiment approach. Water Resour. Res. 43, 1–14. doi:10.1029/2006WR005393

Fenicia, F., Wrede, S., Kavetski, D., Pfister, L., Hoffmann, L., Savenije, H. H. G. and McDonnell, J. J. 2010. Assessing the impact of mixing assumptions on the estimation of streamwater mean residence time. Hydrol. Process., 24, 1730–1741. doi:10.1002/hyp.7595

Finger D., Vis M., Huss M. and J. Seibert, 2015. The value of multiple data set calibration versus model complexity for improving the performance of hydrological models in mountain catchments. Water Resour. Res., 51(4): 1939–1958, doi:10.1002/2014WR015712

Flügel, W. A. 1995. Delineating hydrological response units by geographical information system analyses for regional hydrological modelling using PRMS/MMS in the drainage basin of the River Bröl, Germany. Hydrol. Process. 9, 423–436. doi:10.1002/hyp.3360090313

Freeman, M.C., Pringle, C.M., Jackson, C.R. 2007. Hydrologic connectivity and the contribution of stream headwaters to ecological integrity at regional scales. J. Am. Water Resour. Assoc. 43, 5–14. doi:10.1111/j.1752-1688.2007.00002.x

Geris, J., Tetzlaff, D., McDonnell, J., Soulsby, C. 2015. The relative role of soil type and tree cover on water storage and transmission in northern headwater catchments. Hydrol. Process. 29, 1844–1860. doi:10.1002/hyp.10289

Gharari, S., Hrachowitz, M., Fenicia, F., Savenije, H.H.G. 2013. An approach to identify time consistent model parameters: Sub-period calibration. Hydrol. Earth Syst. Sci. 17, 149–161. doi:10.5194/hess-17-149-2013

Gibson, J.J. 2002. Short-term evaporation and water budget comparisons in shallow Arctic lakes using non-steady isotope mass balance. J. Hydrol. 264, 242–261. doi:10.1016/S0022-1694(02)00091-4

Gonfiantini, R. 1986. Environmental isotopes in lake studies, in: Handbook of Environmental Isotope Geochemistry, the Terrestrial Environment. Elsevier, NY, pp. 113–168.

Gupta, H. V., Kling, H., Yilmaz, K.K., Martinez, G.F. 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377, 80–91. doi:10.1016/j.jhydrol.2009.08.003

Guse, B., Pfannerstill, M., Strauch, M., Reusser, D.E., Lüdtke, S., Volk, M., Gupta, H., Fohrer, N. 2016. On characterizing the temporal dominance patterns of model parameters and processes. Hydrol. Process. 30, 2255–2270. doi:10.1002/hyp.10764

Guse, B., Reusser, D.E., Fohrer, N. 2014. How to improve the representation of hydrological processes in SWAT for a lowland catchment - temporal analysis of parameter sensitivity and model performance. Hydrol. Process. 28, 2651–2670. doi:10.1002/hyp.9777

Hannaford, J., Muchan, K., Lewis, M., Clemas, S. 2014. Hydrological Summary for the United Kingdom: August 2014. NERC Open Research Archive (NORA). Available at: http://nora.nerc.ac.uk

Hannaford, J., Barker, L., Muchan, K., Lewis, M., Clemas, S. 2014. Hydrological Summary for the United Kingdom: October 2014. NERC Open Research Archive (NORA). Available at: http://nora.nerc.ac.uk

Harman, C.J. 2015. Time-variable transit time distributions and transport: Theory and application to storage-dependent transport of chloride in a watershed. Water Resour. Res. 51, 1–30. doi:10.1002/ 2014WR015707

Herbst, M., Casper, M.C. 2008. Towards model evaluation and identification using Self-Organizing Maps. Hydrol. Earth Syst. Sci. Discuss. 4, 3953–3978. doi:10.5194/hessd-4-3953-2007

Hrachowitz, M., Savenije, H., Bogaard, T.A., Tetzlaff, D., Soulsby, C. 2013. 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

Jefferson, A.J., Bell, C.D., Clinton, S.M., Mcmillan, S.K. 2015. Application of isotope hydrograph separation to understand contributions of stormwater control measures to urban headwater streams. Hydrol. Process. 29, 5290–5306. doi:10.1002/hyp.10680

Kirchner, J.W. 2003. A double paradox in catchment hydrology and geochemistry. Hydrol. Process. 17, 871–874. doi:10.1002/hyp.5108

Kirchner, J.W., Feng, X., Neal, C., Robson, A.J. 2004. The fine structure of water-quality dynamics: the (high-frequency) wave of the future. Hydrol. Process. 18, 1353–1359. doi:10.1002/hyp.5537

Klemeš, V. 1986. Operational testing of hydrological simulation models. Hydrol. Sci. J. 31, 13–24. doi:10.1080/02626668609491024

Kling, H., Fuchs, M., Paulin, M. 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. J. Hydrol. 424-425, 264–277. doi:10.1016/j.jhydrol.2012.01.011

Kling, H., Stanzel, P., Fuchs, M., Nachtnebel, H.-P. 2015. Performance of the COSERO precipitation–runoff model under non-stationary conditions in basins with different climates. Hydrol. Sci. J. 60, 1374–1393. doi:10.1080/02626667.2014.959956

Lessels, J.S., Tetzlaff, D., Birkel, C., Dick, J., Soulsby, C. 2016. Water sources and mixing in riparian wetlands revealed by tracers and geospatial analysis. Water Resour. Res. 52, 456–470. doi:10.1002/ 2015WR017519

Lyon, S.W., Desilets, S.L.E., Troch, P.A. 2009. A tale of two isotopes: differences in hydrograph separation for a runoff event when using δD versus $\delta^{18}O$. Hydrol. Process. 23, 2095–2101. doi:10.1002/hyp.7326

Magand, C., Ducharne, A., Le Moine, N., Brigode, P. 2015. Parameter transferability under changing climate: case study with a land surface model in the Durance watershed, France. Hydrol. Sci. J.60, 1408-1423. doi:10.1080/02626667.2014.993643

McDonnell, J.J., Beven, K. 2014. Debates—The future of hydrological sciences: A (common) path forward? A call to action aimed at understanding velocities, celerities and residence time distributions of the headwater hydrograph. Water Resour. Res. 50, 5342–5350. doi:10.1002/ 2013WR015141

McGuire, K.J., McDonnell, J.J. 2006. A review and evaluation of catchment transit time modeling. J. Hydrol. 330, 543–563. doi:10.1016/j.jhydrol.2006.04.020

McGuire, K.J., McDonnell, J.J., Weiler, M., Kendall, C., McGlynn, B.L., Welker, J.M., Seibert, J. 2005. The role of topography on catchment-scale water residence time. Water Resour. Res. 41, 1–14. doi:10.1029/2004WR003657

McMillan, H., Tetzlaff, D., Clark, M., Soulsby, C. 2012. Do time-variable tracers aid the evaluation of hydrological model structure? A multimodel approach. Water Resour. Res. 48. doi:10.1029/2011WR011688

Ohlanders, N., Rodriguez, M., McPhee, J. 2013. Stable water isotope variation in a Central Andean watershed dominated by glacier and snowmelt. Hydrol. Earth Syst. Sci. 17, 1035–1050. doi:10.5194/hess-17-1035-2013

Parry, S., Muchan, K., Lewis, M., Clemas, S. 2014. Hydrological Summary for the United Kingdom: November 2014. NERC Open Research Archive (NORA). Available at: http://nora.nerc.ac.uk

Parry, S., Muchan, K., Lewis, M., Clemas, S. 2015. Hydrological Summary for the United Kingdom: July 2015. NERC Open Research Archive (NORA). Available at: http://nora.nerc.ac.uk

Peralta-Tapia A., Soulsby, C. Tetzlaff D., Sponseller R., Bishop K. and Laudon H. 2016. Hydroclimatic controls on non-stationary transit time distributions in a boreal headwater catchment. J. Hydrol. 543, 7-16. doi:10.1016/j.hydrol.2016.01.079.

Reusser, D.E., Blume, T., Schaefli, B., Zehe, E. 2009. Analysing the temporal dynamics of model performance for hydrological models. Hydrol. Earth Syst. Sci. Discuss. 13, 999–1018. doi:10.5194/hessd-5-3169-2008

Rinaldo, A., Benettin, P., Harman, C.J., Hrachowitz, M., Mcguire, K.J., van der Velde, Y., Bertuzzo, E., Botter, G. 2015. Storage selection functions: A coherent framework for quantifying how catchments store and release water and solutes 51, 4840–4847. doi:10.1002/ 2015WR017273

Rodgers, P., Soulsby, C., Waldron, S. 2005. Stable isotope tracers as diagnostic tools in upscaling flow path understanding and residence time estimates in a mountainous mesoscale catchment. Hydrol. Process. 19, 2291–2307. doi:10.1002/hyp.5677

Seibert, J. 2003. Reliability of Model Predictions Outside Calibration Conditions. Nord. Hydrol. 34, 477–492. doi:10.2166/nh.2003.028

Seibert, J., Beven, K.J. 2009. Gauging the ungauged basin : how many discharge measurements are needed? Hydrol. Earth Syst. Sci. 13, 883–892. doi:10.5194/hessd-6-2275-2009

Seibert, J., McDonnell, J.J. 2015. Gauging the Ungauged Basin: The Relative Value of Soft and Hard Data. J. Hydrol. Eng. 20, 130607193631000. doi:10.1061/(ASCE)HE.1943-5584.0000861

Seibert, J., Rodhe, A., Bishop, K. 2003. Simulating interactions between saturated and unsaturated storage in a conceptual runoff model. Hydrol. Process. 17, 379–390. doi:10.1002/hyp.1130

Seiller, G., Anctil, F., Perrin, C. 2012. Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. Hydrol. Earth Syst. Sci. 16, 1171–1189. doi:10.5194/hess-16-1171-2012

Sieber, A., Uhlenbrook, S. 2005. Sensitivity analyses of a distributed catchment model to verify the model structure. J. Hydrol. 310, 216–235. doi:10.1016/j.jhydrol.2005.01.004

Sivapalan, M. 2003. Prediction in ungauged basins: a grand challenge for theoretical hydrology. Hydrol. Process. 17, 3163–3170. doi:10.1002/hyp.5155

Soulsby, C., Birkel, C., Geris, J., Dick, J., Tunaley, C., Tetzlaff, D. 2015. Streamwater age distributions controlled by storage dynamics and nonlinear hydrologic connectivity: Modeling with high-resolution isotope data. Water Resour. Res. 51, 7759–7776. doi:10.1002/2015WR017888

Soulsby, C., Bradford, J., Dick, J., P. McNamara, J., Geris, J., Lessels, J., Blumstock, M., Tetzlaff, D. 2016. Using geophysical surveys to test tracer-based storage estimates in headwater catchments. Hydrol. Process. doi:10.1002/hyp.10889

Sprenger, M., Tetzlaff, D., Tunaley, C., Dick, J., Soulsby, C. 2017. Evaporation fractionation in a peatland drainage network affects streamwater isotope composition. Water Resour. Res. 53.851-866. doi:10.1002/2016WR019258.

Stadnyk, T.A., Delavau, C., Kouwen, N., Edwards, T.W.D. 2013. Towards hydrological model calibration and validation: Simulation of stable water isotopes using the isoWATFLOOD model. Hydrol. Process. 27, 3791–3810. doi:10.1002/hyp.9695

Tetzlaff, D., Birkel, C., Dick, J., Geris, J., Soulsby, C. 2014. Storage dynamics in hydropedological units control hillslope connectivity, runoff generation and the evolution of catchment transit time distributions. Water Resour. Res. 50, 969–985. doi:10.1002/2013WR014147

Tetzlaff, D., Soulsby, C., Waldron, S., Malcolm, I.A., Bacon, P.J., Dunn, S.M., Lilly, A., Youngson, A.F. 2007. Conceptualization of runoff processes using a geographical information system and tracers in a nested mesoscale catchment. Hydrol. Process. 21, 1289–1307. doi:10.1002/hyp.6309

Thirel, G., Andréassian, V., Perrin, C. 2015a. On the need to test hydrological models under changing conditions. Hydrol. Sci. J. 60, 1165-1173. doi:10.1080/02626667.2015.1050027

Thirel, G., Andréassian, V., Perrin, C., Audouy, J.-N., Berthet, L., Edwards, P., Folton, N., Furusho, C., Kuentz, A., Lerat, J., Lindström, G., Martin, E., Mathevet, T., Merz, R., Parajka, J., Ruelland, D., Vaze, J. 2015b. Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments. Hydrol. Sci. J. 6667, 1–16. doi:10.1080/02626667.2014.967248

Tunaley, C., Tetzlaff, D., Lessels, J., Soulsby, C. 2016. Linking high-frequency DOC dynamics to the age of connected water sources. Water Resour. Res. doi:10.1002/2015WR018419

Tunaley, C., Tetzlaff, D., Soulsby, C. 2017. Scaling effects of riparian peatlands on stable isotopes in runoff and DOC mobilization. J. Hydrol. 549, 220-235. doi: 10.1016/j.jhydrol.2017.03.056

Van Huijgevoort, M.H.J., Tetzlaff, D., Sutanudjaja, E.H., Soulsby, C. 2016. Using high resolution tracer data to constrain water storage, flux and age estimates in a spatially distributed rainfall-runoff model. Hydrol. Process. doi:10.1002/hyp.10902

Vaze, J., Post, D.A., Chiew, F.H.S., Perraud, J.M., Viney, N.R., Teng, J. 2010. Climate non-stationarity - Validity of calibrated rainfall-runoff models for use in climate change studies. J. Hydrol. 394, 447–457. doi:10.1016/j.jhydrol.2010.09.018

Wagener, T., McIntyre, N., Lees, M.J., Wheater, H.S., Gupta, H. V. 2003. Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis. Hydrol. Process. 17, 455–476. doi:10.1002/hyp.1135

Weiler, M. 2003. How does rainfall become runoff? A combined tracer and runoff transfer function approach. Water Resour. Res. 39, 1–13. doi:10.1029/2003WR002331

Wissmeier, L., Uhlenbrook, S. 2007. Distributed, high-resolution modelling of 18O signals in a meso-scale catchment. J. Hydrol. 332, 497–510. doi:10.1016/j.jhydrol.2006.08.003

Wood, E.F., Sivapalan, M., Beven, K. 1990. Similarity and scale in catchment storm response. Rev. Geophys. 28(1), 1-18. doi:10.1029/RG028i001p00001

Yu, B., Zhu, Z. 2015. A comparative assessment of AWBM and SimHyd for forested watersheds. Hydrol. Sci. J. 6667, 1–13. doi:10.1080/02626667.2014.961924

Zhou, Y., Zhang, Y., Vaze, J., Lane, P., Xu, S. 2015. Impact of bushfire and climate variability on streamflow from forested catchments in southeast Australia. Hydrol. Earth Syst. Sci. Discuss. 10, 4397–4437. doi:10.5194/hessd-10-4397-2013

Table 1: Hydrological (total rainfall 30 days prior to calibration period [P_{30}], maximum, mean and sum of precipitation [P] during the calibration period, and maximum and mean discharge [Q] recorded during calibration period) and isotope (maximum, mean, minimum, coefficient of variation [CV] of δ^2H in P and Q) characteristics of the three calibration periods.

	BB (3.2 km^2) Calibration period			HW1 (0.65 km^2) Calibration period		
	1 year	Nov	Jul	1 year	Nov	Jul
Hydrology						
P_{30} (mm)	73.6	117.2	25.8	68.8	108.8	28.2
P_{max} (mm 6h^{-1})	27.8	19	14.4	22.8	17	14.4
P_{mean} (mm 6h^{-1})	0.7	1.7	1.3	0.6	1.6	1.3
P_{sum} (mm)	957	168	95.6	913	162	93.6
Q_{max} (mm 6h^{-1})	6.24	5.66	2.84	7.07	6.55	2.54
Q_{mean} (mm 6h^{-1})	0.46	1.47	0.23	0.39	1.33	0.23
Isotopes						
δ^2H P_{max} (‰)	-13.9	-21.4	-17.4			
δ^2H P_{mean} (‰)	-57.3	-63.0	-43.5			
δ^2H P_{min} (‰)	-145.9	-141.0	-87.1			
δ^2H P CV (%)	44.7	41.7	42.1			
δ^2H Q_{max} (‰)	-51.2	-57.5	-51.2	-49.3	-58.0	-50.3
δ^2H Q_{mean} (‰)	-58.7	-62.1	-55.4	-57.7	-62.2	-54.1
δ^2H Q_{min} (‰)	-72.2	-65.2	-58.5	-70.3	-65.8	-57.2
δ^2H Q CV (%)	5.3	2.3	2.9	6.2	2.7	3.1

Table 2: Mean model parameter and performance (KGE) values and ranges, expressed as minimum and maximum values (in parentheses), in the Bruntland Burn and HW1 for the one year calibration, Nov 2014 calibration and Jul 2015 calibration

			BB			HW1		
			1 year	Nov 2014	July 2015	1 year	Nov 2014	July 2015
Parameter	Units	Initial range	Mean [min, max]	Mean [min, max]	Mean [min, max]	Mean [min, max]	Mean [min, max]	Mean [min, max]
Hydrology								
a	hr^{-1}	[0.1, 0.8]	0.35 [0.24, 0.57]	0.16 [0.10, 0.40]	0.18 [0.10, 0.25]	0.35 [0.25, 0.66]	0.23 [0.10, 0.46]	0.23 [0.10, 0.36]
b	hr^{-1}	[0.001, 0.1]	0.002 [0.001, 0.018]	0.036 [0.005, 0.100]	0.0002 [0.0001, 0.0002]	0.002 [0.001, 0.006]	0.004 [0.001, 0.006]	0.0006 [0.0001, 0.0016]
r	hr^{-1}	[0.1, 0.8]	0.54 [0.30, 0.90]	0.74 [0.24, 0.90]	0.85 [0.57, 0.90]	0.51 [0.25, 0.89]	0.41 [0.10, 0.83]	0.85 [0.55, 0.90]
k	hr^{-1}	[0.001, 0.1]	0.02 [0.01, 0.06]	0.09 [0.05, 0.10]	0.09 [0.07, 0.10]	0.02 [0.01, 0.04]	0.08 [0.02, 0.10]	0.07 [0.03, 0.10]
α	-	[0.05, 0.9]	0.61 [0.06, 0.88]	0.80 [0.42, 0.90]	0.89 [0.57, 0.90]	0.63 [0.05, 0.90]	0.70 [0.36, 0.90]	0.90 [0.89, 0.90]
KGE_Q	-	-	0.72 [0.30, 0.92]	0.64 [0.38, 0.94]	0.60 [-0.03, 0.84]	0.72 [0.30, 0.88]	0.70 [0.40, 0.93]	0.51 [0.17, 0.95]
Isotopes								
upSp	mm	[0, 500]	291 [241, 328]	432 [307, 500]	189 [16, 500]	334 [303, 499]	499 [486, 500]	259 [67, 500]
satSp	mm	[0, 1000]	84 [54, 134]	0.2 [0, 11]	97 [0, 100]	99 [93, 100]	1.5 [0, 6]	100 [99.5, 100]
lowSp	mm	[0, 1000]	990 [953, 1000]	71 [20, 135]	996 [871, 1000]	987 [839,1000]	34 [4, 78]	996[917, 1000]
KGE_δ^2H	-	-	0.75 [0.51, 0.78]	0.64 [0.21, 0.70]	0.58 [-37.25, 0.82]	0.77 [0.61, 0.81]	0.64 [0.32, 0.75]	0.56 [0.34, 0.61]

Table 3: Total storage (mm) and age (days) for both the BB and HW1 for the 1 year period (1 August 2014 – 1 August 2015) based on means of simulations from different calibration periods. Minimum and maximum values are given in parentheses, along with standard deviation.

Calibration period	Storage (mm)	Age (days)
BB		
1 year	1400 [1280, 1462] (sd = 41)	321 [6, 870] (sd = 239)
Nov 2014	512 [416, 572] (sd = 27)	150 [4,319] (sd = 69)
Jul 2015	1492 [1232, 1597] (sd = 111)	353 [1.4, 941] (sd = 263)
HW1		
1 year	1417 [1313, 1477] (sd = 35)	322.7 [7, 841] (sd = 229)
Nov 2014	552 [426, 614] (sd = 32)	143 [4, 284] (sd = 64)
Jul 2015	1472 [1277, 1553] (sd = 74)	359 [2, 938] (sd = 261)

Table 4: Cross basin test performances (KGE) between BB and HW1 for the 3 calibration periods for discharge and isotopes.

Calibration site	Period	Evaluation site			
		Discharge		Isotopes	
		BB	HW1	BB	HW1
BB	1 year	0.78	0.78	0.76	0.79
	Nov	0.57	0.56	0.66	0.65
	Jul	0.58	0.67	0.78	0.79
HW1	1 year	0.79	0.8	0.78	0.75
	Nov	0.68	0.69	0.57	0.64
	Jul	0.38	0.44	0.73	0.69

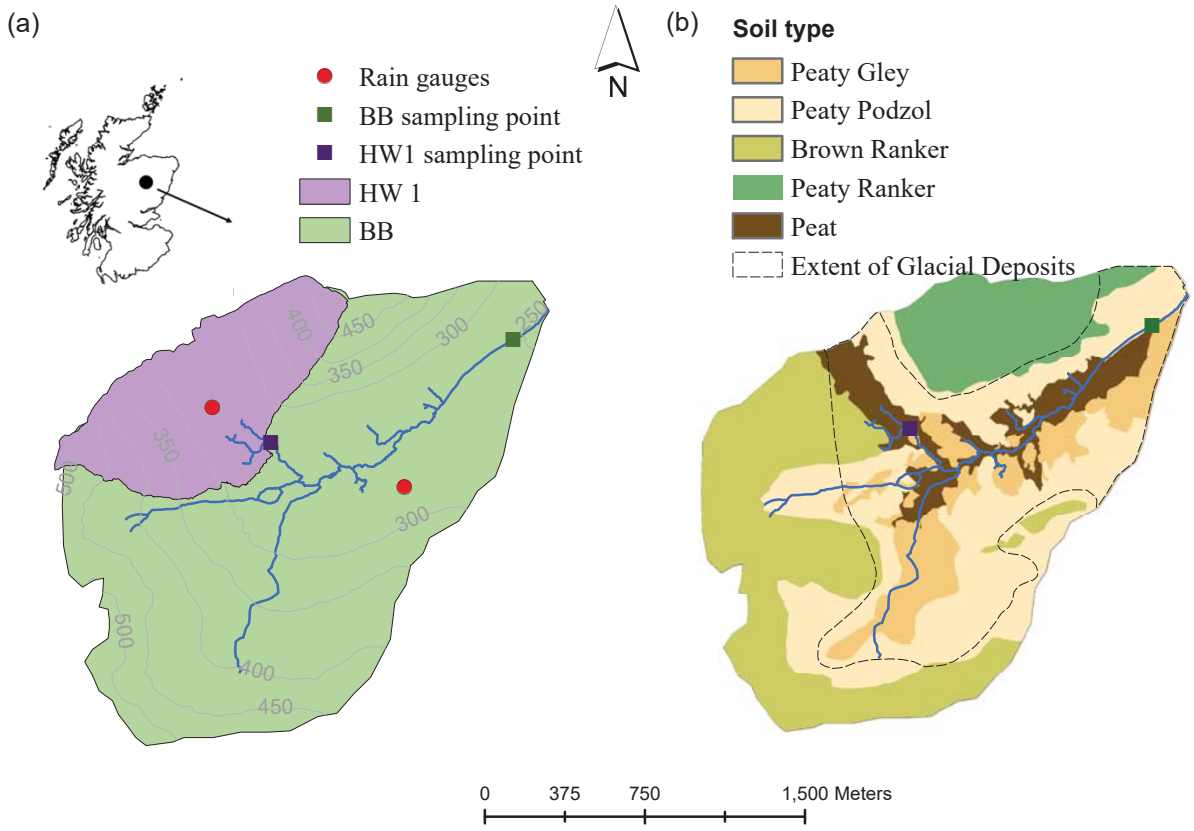

Figure 1: (a) Outlines of the nested catchments, contour lines and sampling sites; (b) dominant soil types in each of the catchments.

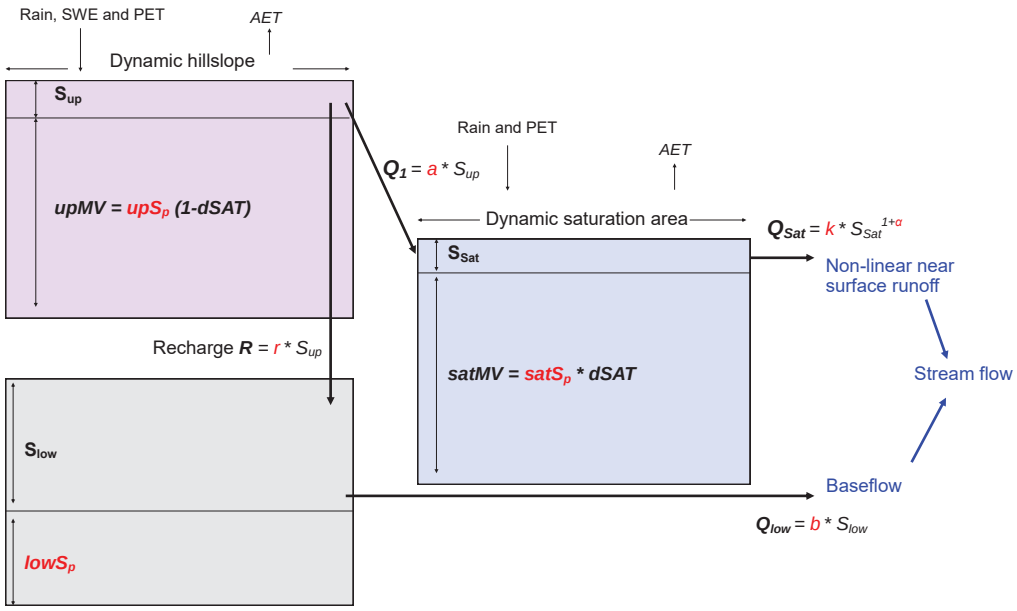

Figure 2: Schematic model structure showing the three reservoirs with associated dynamic storage (S_{up}, S_{low} and S_{sat}) and additional passive storage for time-variable mixing volumes (upS_p, $lowS_p$ and $satS_p$), which have been converted from the storage parameters according to the antecedent wetness (dSAT). The linear rate parameter a (6 hr $^{-1}$) controls the hillslope water flux to the saturated area; r (6 hr $^{-1}$) controls the groundwater recharge rate; b (6 hr $^{-1}$) controls the rate of groundwater discharge to streamflow; k (6 hr $^{-1}$) and α conceptualises saturation overland flow and controls the nonlinear runoff from the saturation area to streamflow. Calibrated parameters are displayed in red. AET and PET are actual and potential evapotranspiration, respectively.

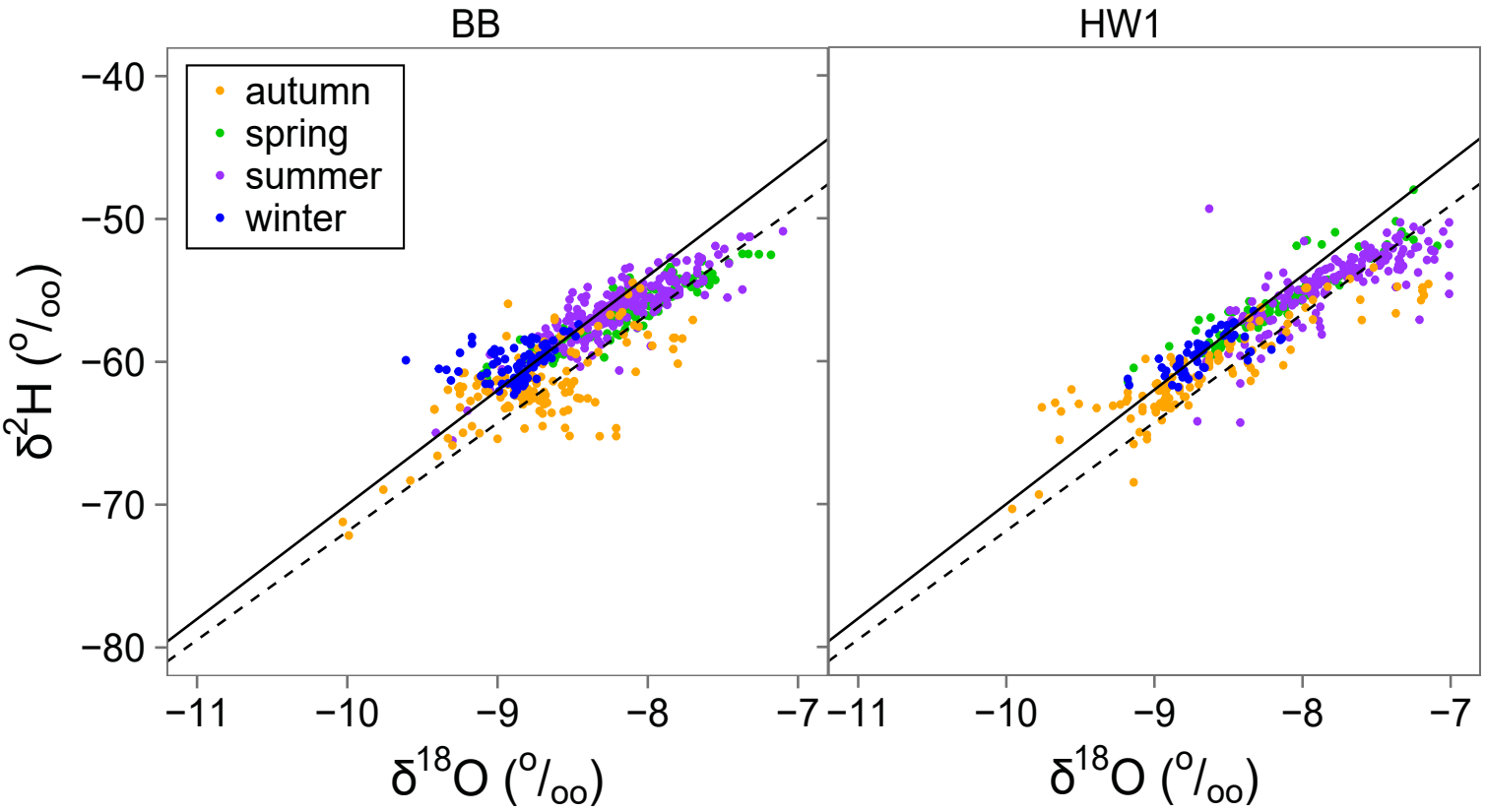

Figure 3: Streamwater stable isotopes across summer (June, July and August), autumn (September, October and November), winter (December, January and February) and spring (March, April and May) plotted along the Global Meteoric Water Line (solid line) and Local Meteoric Water Line (dashed line) for BB and HW1.

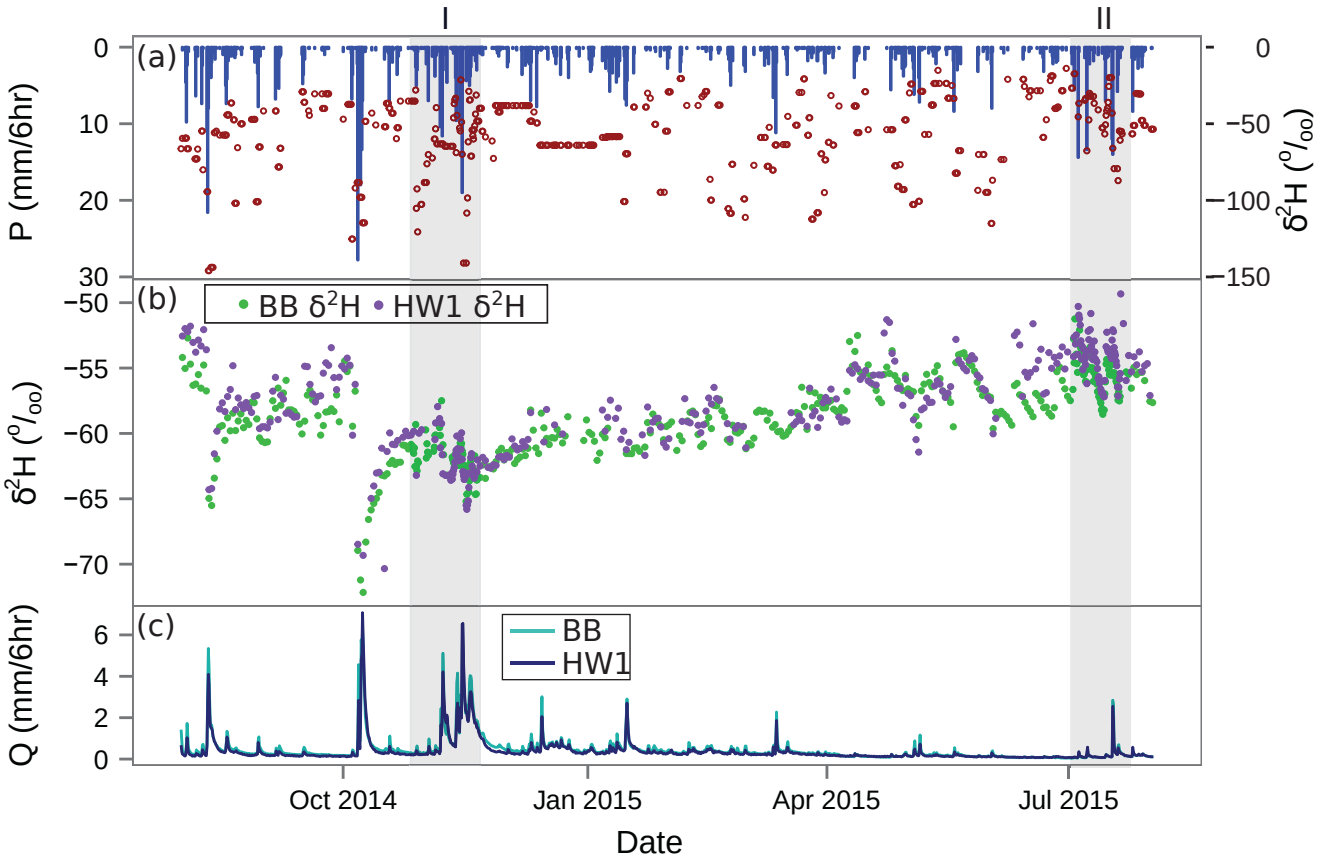

Figure 4: Time series of: (a) precipitation (blue bars) and δ^2H signatures of precipitation (red circles); (b) δ^2H of streamwater for the Bruntland Burn (BB) and HW1; (c) discharge for the BB and HW1. Shaded areas are the periods of higher frequency (6 hourly) isotope sampling (I is Nov 2014 and II is Jul 2015).

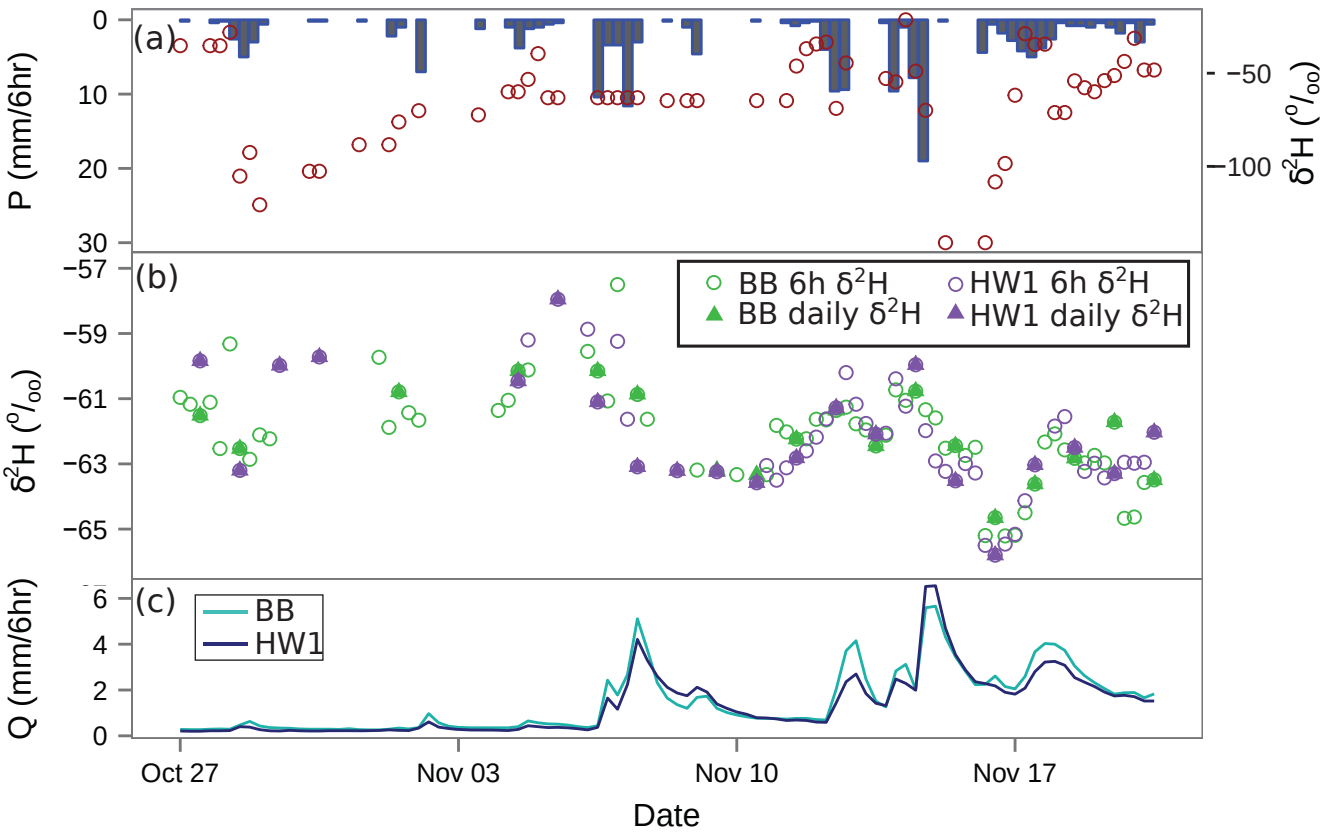

Figure 5: High frequency sampling period Nov (27 October – 20 November 2014). Time series of: (a) precipitation (blue bars) and δ^2H signatures of precipitation (red circles); (b) 6 hourly and daily δ^2H of streamwater for the Bruntland Burn (BB) and HW1; (c) discharge for the BB and HW1.

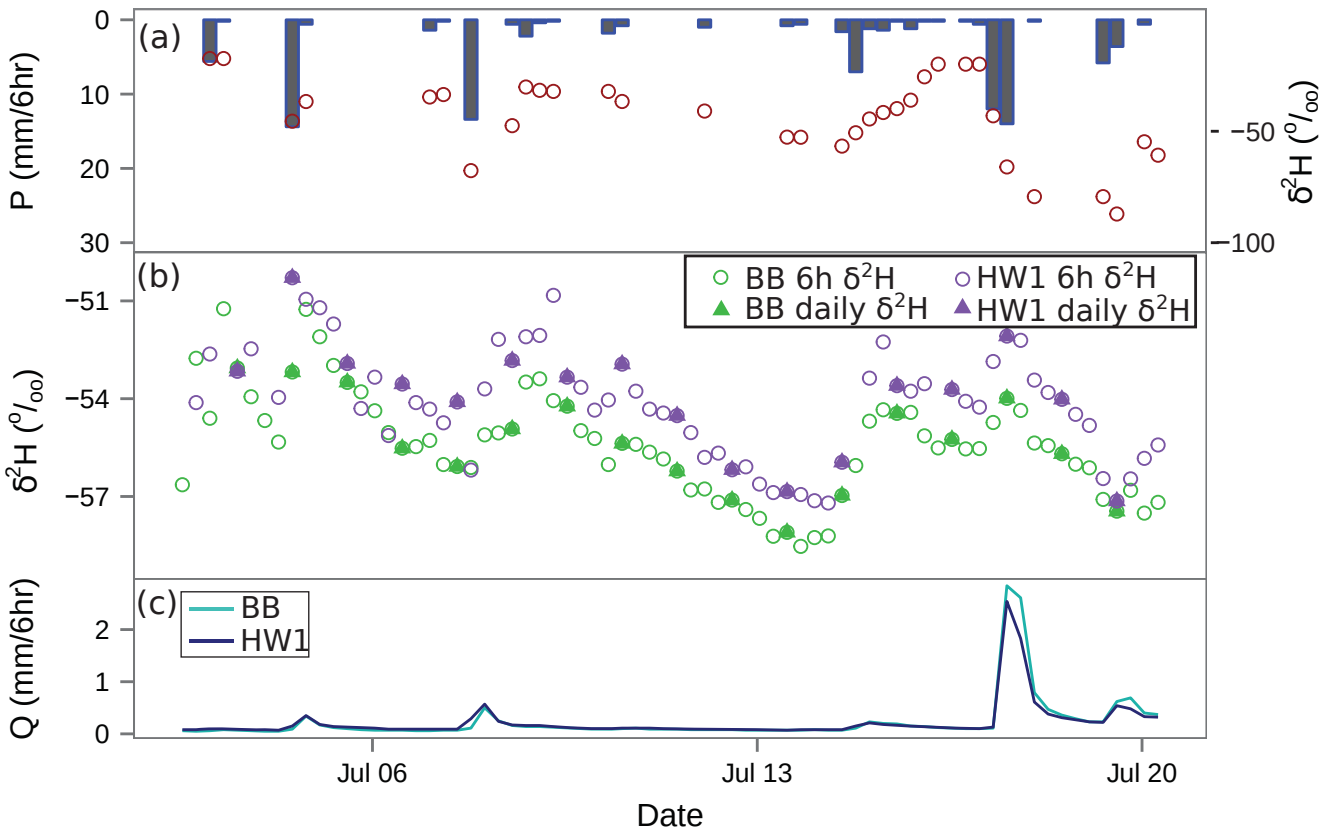

Figure 6: High frequency sampling period Jul (2 – 20 July 2015). Time series of: (a) precipitation (blue bars) and δ^2H signatures of precipitation (red circles); (b) 6 hourly and daily δ^2H of streamwater for the Bruntland Burn (BB) and HW1; (c) discharge for the BB and HW1.

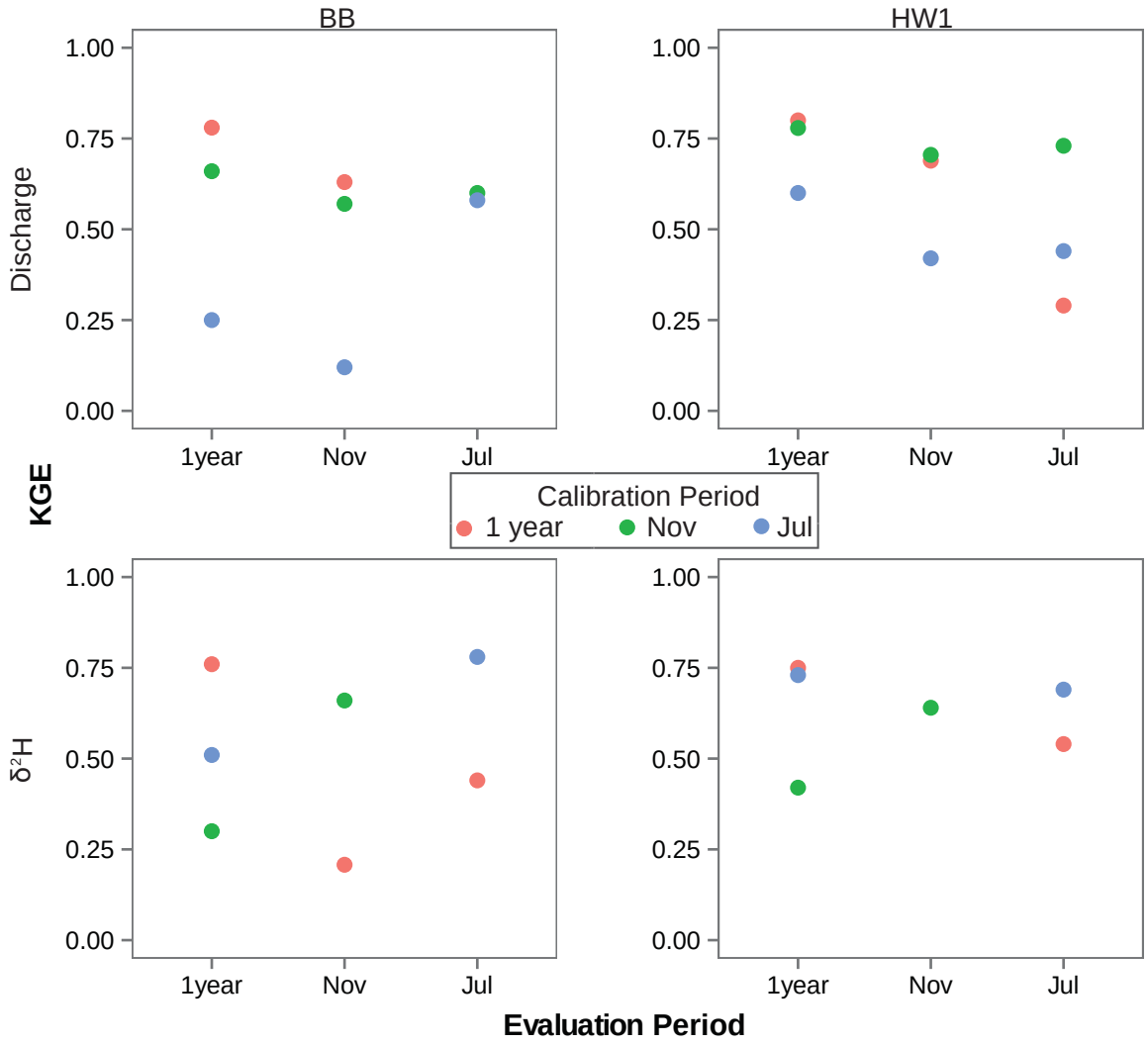

Figure 7: Comparison of model performance (KGE) between median simulations for each calibration period over different periods of evaluation. Top left: BB discharge simulations. Top right: HW1 discharge simulation. Bottom left: BB δ2H simulations. Bottom right: HW1 δ2H simulations. KGE less than 0 are not included on the plots.

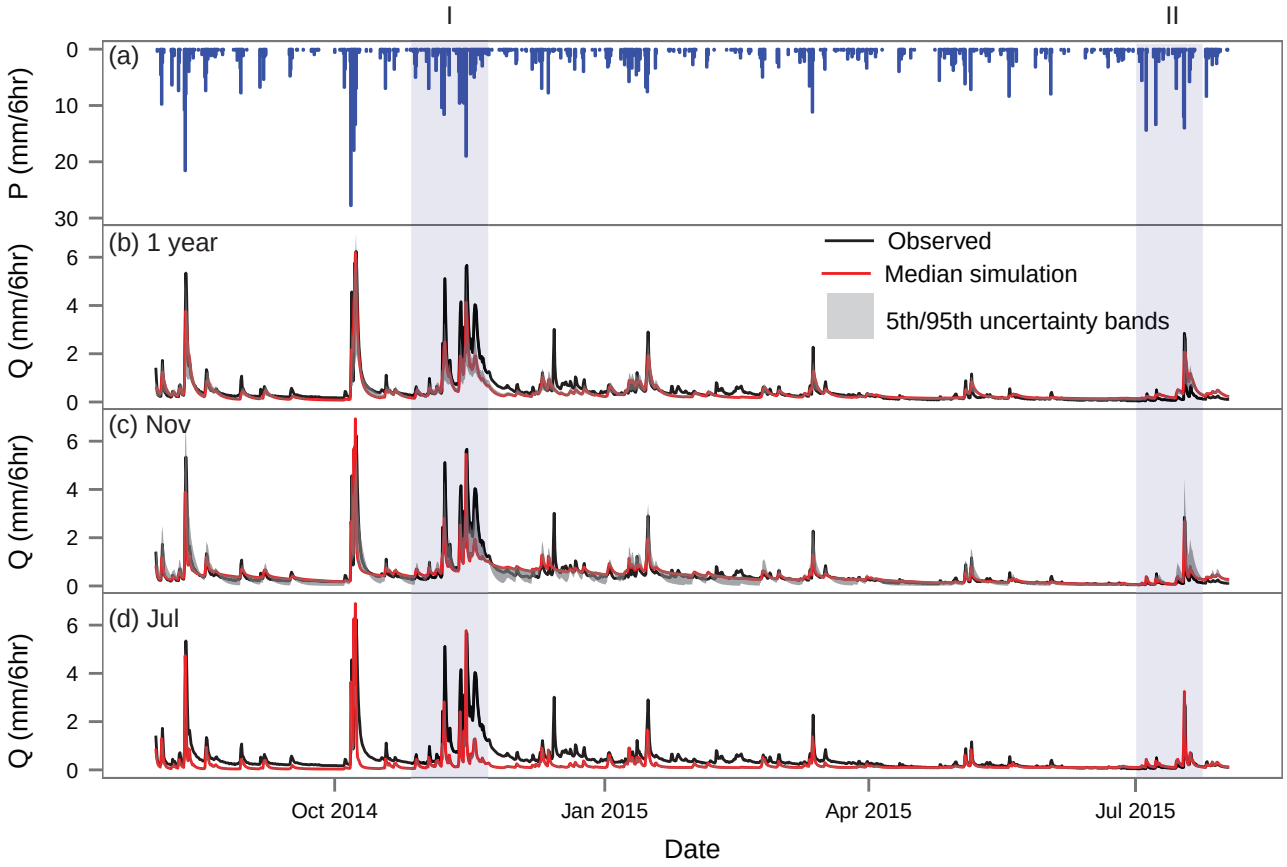

Figure 8: Comparison of the different calibration periods in the Bruntland Burn. 6- hourly data: (a) precipitation; (b) measured and simulated discharge from 1 year calibration; (c) measured and simulated discharge from Nov calibration; (d) measured and simulated discharge from Jul calibration. Uncertainty bands represent the 5th and 95th percentiles derived from the 500 best parameter sets. Shaded areas highlight the Nov (I) and Jul (II) calibration periods.

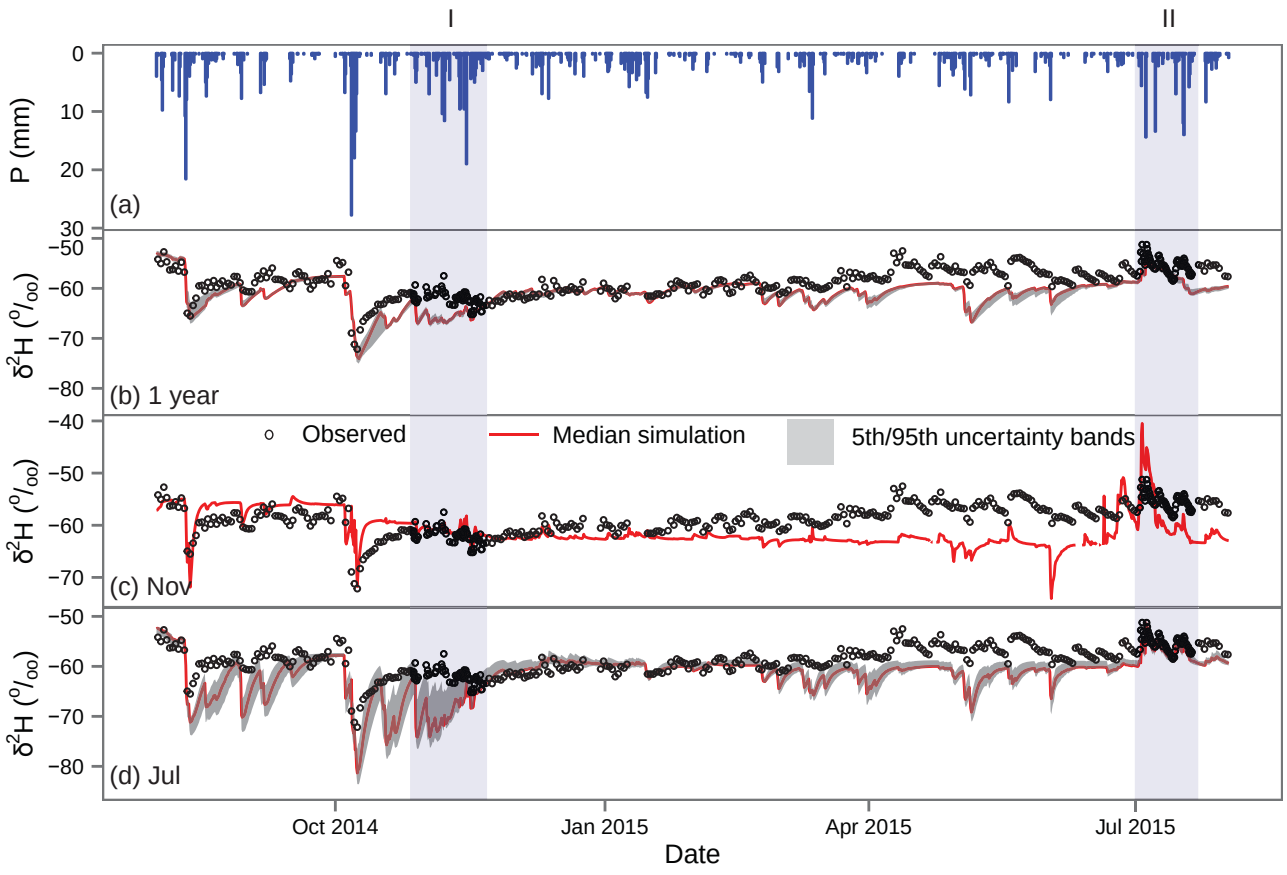

Figure 9: Comparison of the different calibration periods in the Bruntland Burn. 6 hourly data isotope data: (a) precipitation; (b) measured and simulated δ^2H from 1 year calibration; (c) measured and simulated δ^2H from the Nov calibration; (d) measured and simulated δ^2H from the Jul calibration. Uncertainty bands represent the 5^{th} and 95^{th} percentiles derived from the 500 best parameter sets. There are no uncertainty bands for the November calibration due to instability in some of the parameter sets. Shaded areas highlight the Nov (I) and Jul (II) calibration periods.

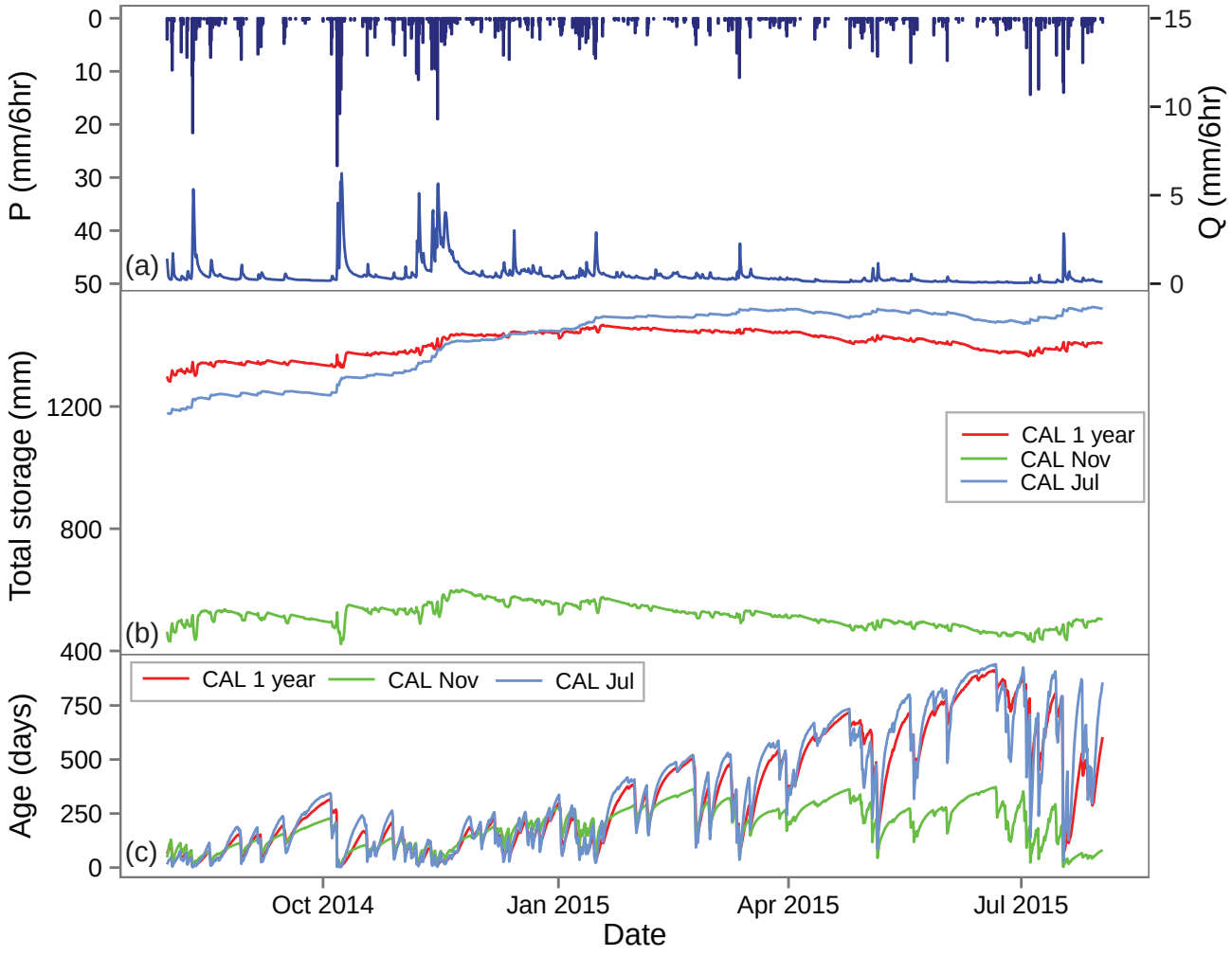

Figure 10: Comparison of the different calibration periods on total storage and age estimates in the Bruntland Burn. 6 hourly (a) precipitation and discharge; (b) total storage estimates for the 3 different calibration periods; (c) age estimates for the 3 different calibration periods. Median simulations are plotted.