Received: 23 October 2020 | Revised: 5 May 2021 | Accepted: 5 May 2021
DOI: 10.1002/hyp.14223

UNCERTAINTIES IN MEASUREMENTS

WILEY

Effects of streamflow isotope sampling strategies on the calibration of a tracer-aided rainfall-runoff model

Jamie Lee Stevenson[1] | Christian Birkel[2] | Aaron J. Neill[1] | Doerthe Tetzlaff[3,4,1] | Chris Soulsby[1]

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

[2] Department of Geography and Water and Global Change Observatory, University of Costa Rica, San José, Costa Rica

[3] Geographisches Institut, Humboldt University Berlin, Berlin, Germany

[4] IGB Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany

Correspondence
Dr. Jamie Lee Stevenson, Northern Rivers Institute, School of Geosciences, University of Aberdeen, Aberdeen, UK.
Email: r01js19@abdn.ac.uk

Funding information
European Research Council, Grant/Award Number: VeWa GA335910; Leverhulme Trust, Grant/Award Number: RPG-2018-375

Abstract

Isotopes are increasingly used in rainfall-runoff models to constrain conceptualisations of internal catchment functioning and reduce model uncertainty. However, there is little guidance on how much tracer data is required to adequately do this, and different studies use data from different sampling strategies. Here, we used a 7-year time series of daily stable water isotope samples of precipitation and streamflow to derive a range of typical stream sampling regimes and investigate how this impacts calibration of a semi-distributed tracer-aided model in terms of flow, deuterium and flux age simulations. Over the 7 years weekly sampling facilitated an almost identical model performance as daily, and there were only slight deteriorations in performance for fortnightly sampling. Monthly sampling resulted in poorer deuterium simulations and greater uncertainty in the derived parameter sets ability to accurately represent catchment functioning, evidenced by unrealistic reductions in the volumes of water available for mixing in the saturation area causing simulated water age decreases. Reducing sampling effort and restricting data collection to 3 years caused reductions in the accuracy of deuterium simulation, though the deterioration did not occur if sampling continued for 5 years. Analysis was also undertaken to consider the effects of reduced sampling effort over the driest and wettest hydrological years to evaluate effects of more extreme conditions. This showed that the model was particularly sensitive to changes in sampling during dry conditions, when the catchment hydrological response is most non-linear. Across all dataset durations, sampling in relation to flow conditions, rather than time, revealed that samples collected at flows >Q50 could provide calibration results comparable to daily sampling. Targeting only extreme high flows resulted in poor deuterium and low flow simulations. This study suggests sufficient characterization of catchment functioning can be obtained through reduced sampling effort over longer timescales and the targeting of flows >Q50.

KEYWORDS
model calibration, sampling strategies, stable isotopes, tracer-aided modelling, uncertainty, water age

1 | INTRODUCTION

Stable water isotopes which occur naturally in precipitation can show strong inter-annual and intra-annual variability, with the damping and lagging of this variability in catchment discharge able to indicate the hydrological functioning of a catchment in terms of storage, transit times and response to changing antecedent conditions (Bowen, 2008; Bowen et al., 2019; Birkel et al., 2014; Birkel & Soulsby, 2015). Therefore, their use within hydrological models can help to better represent internal catchment functioning, improve process realism and reduce uncertainty (Ala-aho et al., 2017; van Huijgevoort et al., 2016; Weiler, 2003). Such improvement is achieved through correctly capturing a catchment's storage-flux interactions which determines how fast a perturbation in the system, for example, precipitation input, is transferred to the hydrograph (celerity) and the speed (velocity) a water particle takes to move through the system (McDonnell & Beven, 2014). Therefore, dual calibration of rainfall-runoff models to both the flow hydrograph and streamflow tracer composition can increase confidence that a calibrated model gives "the right answers for the right reasons" and potentially reduce model uncertainty (Kirchner, 2006).

However, to integrate tracers into rainfall-runoff models requires increasing model complexity through conceptualisation of storage mixing volumes, and associated parameters, needed to dampen and lag the isotopic input signal (Birkel et al., 2014). This needs to be done carefully to avoid overparameterization that may make parameter identification more problematic and actually increase uncertainty (Kirchner, 2006). Additionally, streamflow simulation performance is often compromised through dual calibration with isotopes (Wang et al., 2019). In turn this can, in certain situations, lead to scepticism about the usefulness of tracers (Seibert et al., 2003). Despite such concerns, better conceptualisation of flux-storage interactions (e.g., Birkel et al., 2011) has led to a steady increase in the use of tracer aided models (Birkel & Soulsby, 2015), and efforts to acquire datasets with sufficient temporal extent to capture the variability of a catchment's water travel time and residence time distributions (McDonnell & Beven, 2014; Remondi et al., 2018; Stadnyk & Holmes, 2020). Only by acquiring such datasets can researchers reduce uncertainty around whether models can sufficiently reproduce the rainfall-runoff response and tracer input–output transformations in a way that adequately captures a variety of events, antecedent conditions and catchment states.

However, there is surprisingly little guidance on how much tracer data is needed to calibrate a rainfall-runoff model in terms of the longevity and frequency of sampling. Long-term isotope time series are still relatively rare as most datasets are collected during short-term projects. Furthermore, most studies tend to conduct weekly sampling, though high frequency sampling through a series of events is becoming more common (e.g., Knapp et al., 2019; Zhang et al., 2019). However, financial and logistical costs of high frequency data collection can be restrictive, especially when research is focussed on understanding longer-term water balance dynamics rather than short-term extreme events such as floods. Consequently, Sprenger et al. (2019) called for more research to better understand how often to take samples to optimize resource efficiency. Such resource efficiency must be balanced with collecting sufficient data to ensure model calibration reduces overall uncertainty in the combined simulation of flows and tracers. The question of adequate data collection and reduction of model uncertainty versus resource efficiency is not new to the field of hydrological modelling. McIntyre and Wheater (2004) showed that limited collection of phosphorus data caused increased model uncertainty and, hence, was of limited value.

Defining the exact quantity of data needed to calibrate tracer-aided models is indeed difficult and may well be site specific and depend on the responsiveness of a catchment in terms of event-based, seasonal and inter-annual variability (e.g., Hrachowitz et al., 2011). These factors, in turn, depend on both internal catchment properties and hydroclimatic drivers (Hrachowitz et al., 2010).

Previous studies have touched upon the issue, and though most have been constrained by relatively limited data availability, they have drawn informative insights. For example, Birkel, Dunn et al. (2010a) concluded that weekly sampling during a 13-month period was inadequate to characterize the temporal variability of precipitation and streamflow isotopes of a small agricultural catchment. Birkel et al. (2012) extended this work showing that even daily sampling could mask true isotope dynamics during storm events. When investigating the impact to mean transit times, calculated through the use of lumped models and a ~23-month dataset, Timbe et al. (2015) found that model parameters were highly sensitive to changes in isotope sampling resolution. Similarly, using an 18-month dataset, transit time distributions were considerably altered when the resolution from weekly isotopic sampling was increased to, for example, daily sampling (Stockinger et al., 2016).

Conversely, increasing the data resolution has also been shown to be less valuable in some instances, as found by Tunaley et al. (2017) who concluded, using a 15-month dataset, that sub-daily sampling did not reveal new process insights but rather confirmed those derived from daily sampling. It has further been shown that when considering event-based model calibration, using both synthetic and observed data, relatively few isotopic samples can provide sufficient information to characterize event dynamics (Wang et al., 2017, 2019). This is similar to conclusions of earlier work by Seibert and Beven (2009) who found relatively few flow gauging's were needed to calibrate a hydrological model, provided that a sufficient range of the flows across the hydrograph were captured. Pool et al. (2017) also concluded that a small number of strategically selected runoff measurements can be adequate when using a bucket-type model in virtually ungauged catchments.

Here we build on previous research by utilizing a 7-year time series of stable isotopes in daily samples of precipitation and streamflow to investigate how a reduced or targeted sampling effort for streamflow isotope samples impacts the calibration of a semi-distributed tracer-aided model. Such datasets are rare (von Freyberg et al., 2017) and are difficult to obtain but provide a valuable opportunity to improve our understanding of the information content and potential redundancy of high intensity, long term isotope datasets. More specifically, we addressed the following research objectives:

1. To evaluate how sampling streamflow isotopes according to differing temporal frequencies or flow percentiles impacts the calibration of a tracer-aided rainfall runoff model.
2. To identify the role of dataset longevity in the findings of Objective 1 by comparing 3, 5 and 7 year duration time series.
3. To assess if the impacts of a change in streamflow isotope sampling regime on calibration are exacerbated during more hydrologically extreme years.

2 | STUDY SITE DESCRIPTION

The study was undertaken using data collected in the 3.2 km^2 Bruntland Burn (BB) catchment, which has been described in detail previously (e.g., Birkel et al., 2011; Tetzlaff et al., 2014). Briefly, the BB is a subcatchment of the 31 km^2 Girnock Burn and consists of steep valley sides dominated by freely draining shallow podzol and ranker soils that facilitate groundwater recharge (Figure 1). The geology is predominantly metamorphic in the southern part of the catchment and granite bedrock in the north. During wetter conditions, the hillslopes can, via lateral flow, become hydrologically connected to the wide valley bottom that is predominated by deposits of glacial till overlain with poorly draining peats. These soils facilitate saturation overland flow and remain close to saturation throughout the year. The spatial extent of this saturated area can range between 4% and 69% of the catchment depending on antecedent wetness and precipitation event characteristics (Birkel et al., 2011). The climate is temperate/ boreal oceanic and most precipitation events are of low intensity, with 50% of the mean annual total of ~1000 mm falling in events <10 mm. Snow accounts for <10% of inputs, with mean annual potential evapotranspiration of ~450 mm and mean winter and summer temperatures ranging from ~0 to ~12°C respectively.

3 | DATA AND METHODS

3.1 | Hydrological and isotopic data

Streamflow samples were collected for stable isotope analysis at the catchment outlet (Figure 1) daily at 14:00 h between 01/10/2011 and 30/09/2018 using an ISCO 3700 autosampler. The same method was used to collect precipitation (P) samples for isotopic analysis at the same location, though samples were cumulated over a 24-hour period

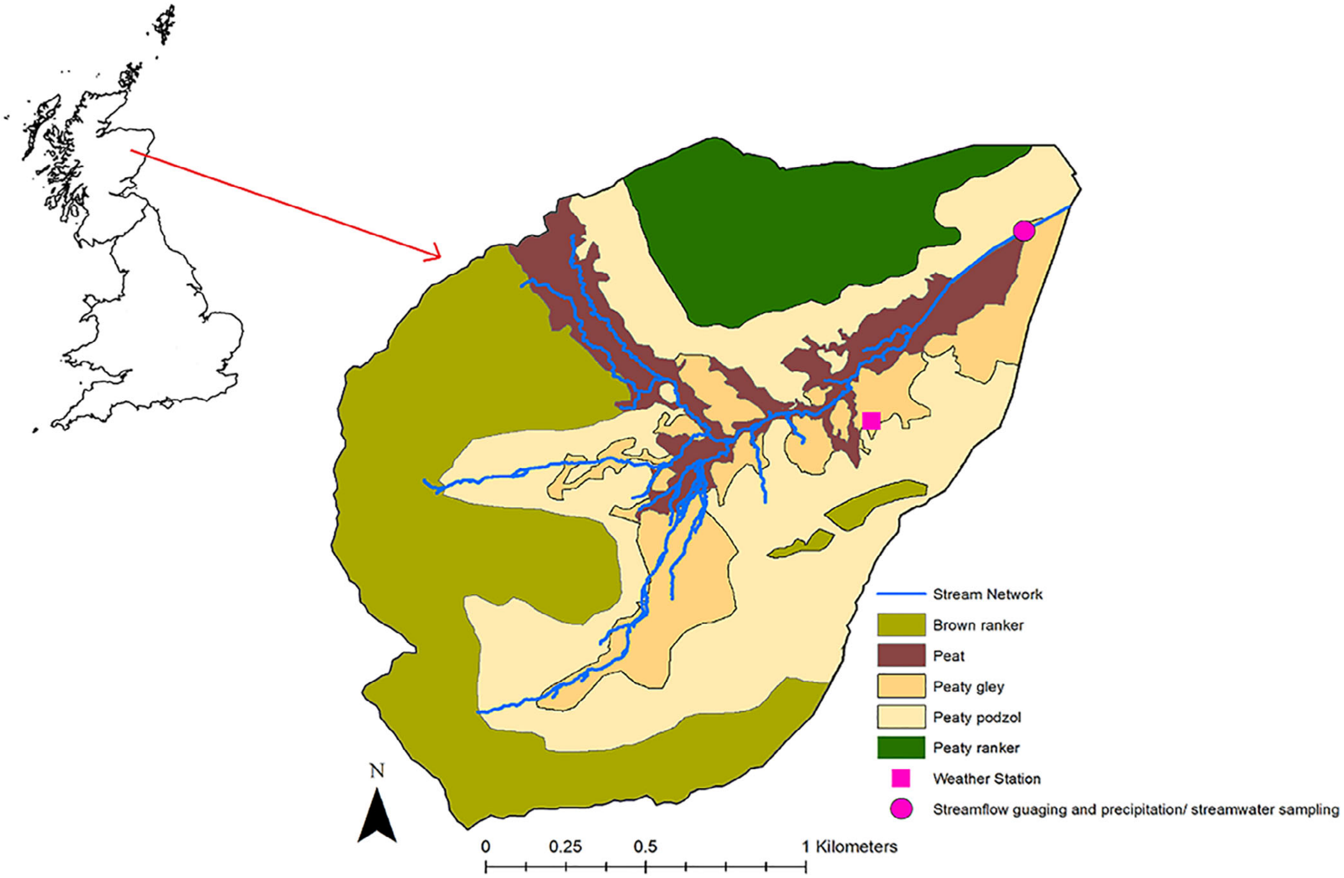

FIGURE 1 Data collection points, stream network and soil types of the Bruntland Burn with geographical reference

starting at 00:00 h. To prevent fractionation, paraffin was added to storage bottles prior to collection. All samples were analysed for deuterium (δ^2H) and oxygen-18 (δ^{18}O) isotopes at the University of Aberdeen using a Los Gatos DLT-100 laser isotope analyser (precision of ±0.4‰ for δ^2H and ±0.1‰ for δ^{18}O). Due to the higher relative precision, we used δ^2H in the study. To infill 603 missing precipitation isotope samples, a spline (a series of polynomial equations which collectively create a curve that must pass through a set of control points) was fitted to the un-weighted monthly P isotope averages. The spline predicted value for missing data was then combined with noise randomly from a uniform distribution between ±1 standard deviation of the observed P isotope for the month in question. This method proved superior, R^2 of 0.14 between available observed and predicted data, to other climate-based multiple linear regression models that were trialled. Discharge was obtained from stage height measurements in a rated section at the catchment outlet at 15-minute intervals and aggregated to an hourly and daily resolution. Precipitation input data was collected within the catchment at 15-minute intervals and subsequently aggregated to daily values. Daily PET was calculated using the Penman–Monteith equation from data collected at an automated weather station in the catchment.

Figure 2 displays the precipitation, streamflow and associated isotope time series. The close relationship between precipitation and streamflow is evident. Winters are generally wetter, with higher streamflow, though rainfall is fairly evenly distributed through the year, and high flows can occur in summer. The winters of 2013–14 and 2015–16 were noticeably wet, with the summer of 2018 being notably dry. There was a strong day-to-day variability in the isotopic composition of precipitation, reflecting different air mass sources, though seasonality is evident as well with colder winter precipitation generally being more depleted than summer precipitation. Streamflow isotopes reflect this seasonality more clearly, though are highly damped in comparison to precipitation as a result of mixing with the large volume of stored water in the catchment (Birkel et al., 2011). More direct stream isotope responses to precipitation are evident during wetter periods (e.g., winter 2013–14) whilst variation is more compressed during dry periods (e.g., the 2017–18 hydrological year).

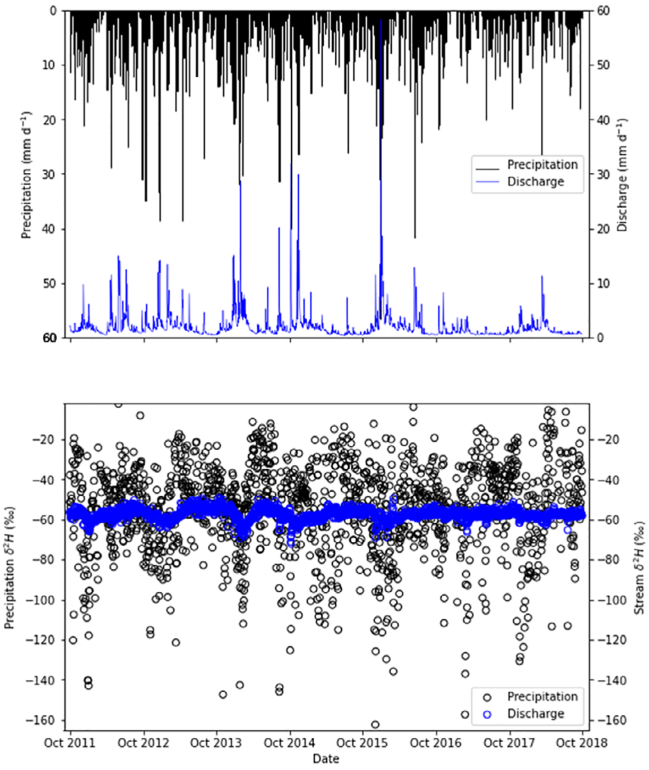

FIGURE 2 Time series of observed daily precipitation and discharge rates (top) and respective isotopic signatures (bottom)

3.2 | Model description

The coupled, semi-distributed, dynamic saturation area flow-tracer model (D-Sat) used here was developed iteratively and is described in a series of papers (e.g., Birkel et al., 2011, 2014, 2015; Birkel, Tetzlaff, Dunn, & Soulsby, 2010b; Soulsby et al., 2015). Figure 3 shows the basic model structure which conceptualizes the catchment as three interacting compartments; with hillslope, dynamic saturation area and groundwater reservoirs. Each has an associated dynamic storage (S_{up}, S_{sat} and S_{low} respectively), with transfer of water from S_{up} to S_{sat} and S_{low} controlled by calibrated linear rate parameters a and r, respectively. Streamflow is generated via groundwater using the linear rate parameter b, and saturation overland flow is controlled by the rate parameter k and a non-linearity parameter α. Each compartment is also characterized by a calibrated mixing volume parameter (upS_p, $satS_p$ and $lowS_p$) to represent the storage needed to dampen the isotope outputs in relation to precipitation inputs, but these volumes do not affect the dynamic water storage fluxes.

A key aspect of the model structure is the non-linear expansion and contraction of the saturation area, which is calculated at each timestep using an antecedent precipitation index-type algorithm (see Birkel et al., 2010b). Both the time-varying distribution of precipitation between the hillslope and saturation area and the compartments' mixing volumes are calculated using this index.

Previous use of the model (e.g., Tunaley et al., 2017) has shown the need to account for the non-conservative behaviour of isotopes in the hillslope and saturation areas due to evaporative fractionation in the catchments peaty soils (Sprenger et al., 2017). In this study, an adapted version of the equation used by Benettin et al. (2017) was employed to account for this, though terms relating to the preferential selection of old or new water were removed given the well-mixed approach to water storage components employed in D-sat. Benettin et al. (2017) also used a uniform approach to fractionation across the catchment, however, exploratory investigations suggested that model results were improved if fractionation was calculated separately for the hillslope and saturation area using two calibrated parameters; β_{Up} and β_{sat}. For the saturation area the isotopic composition was updated to reflect evaporative fractionation over the course of a timestep via the following equation:

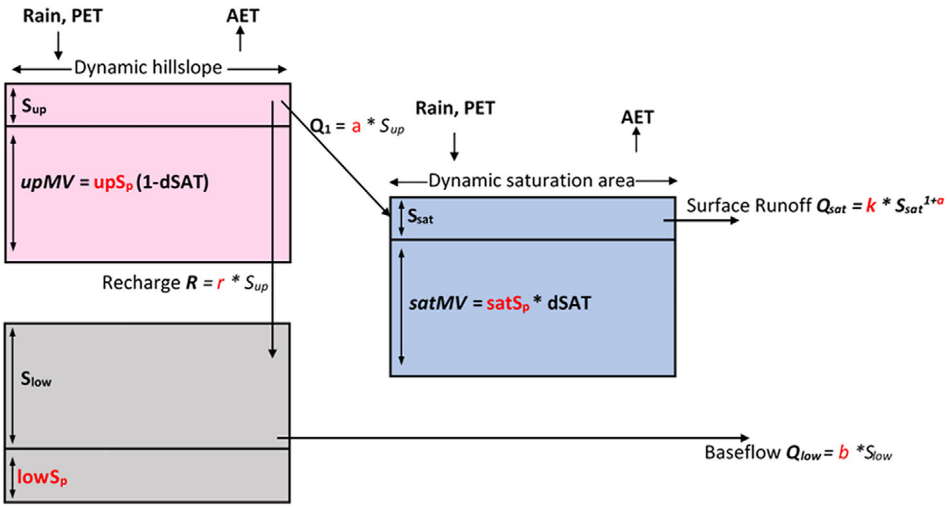
Rain, PET
AET
Dynamic hillslope
S_{up}
$upMV$ = upS_p (1-dSAT)
$Q_1 = a * S_{up}$
Rain, PET
AET
Dynamic saturation area
Surface Runoff $Q_{sat} = k * S_{sat}^{1+\alpha}$
S_{sat}
$satMV$ = $satS_p$ * dSAT
Recharge $R = r * S_{up}$
S_{low}
Baseflow $Q_{low} = b * S_{low}$
$lowS_p$

FIGURE 3 Schematic model structure, with basic equations, displaying the three dynamic reservoirs controlling streamflow (S_{up}, S_{sat} and S_{low}) and associated passive storages regulating the isotopic composition of landscape units. Hillslope and saturation areas are time-variable, calculated according to the catchments antecedent wetness. Isotope fractionation in the hillslope and saturation area controlled by calibrated parameters β_{Up} and β_{Sat}, respectively; Equation (1). Other calibrated parameters are shown in red

$$\frac{d\left[\delta^2 H_{sat}\right]}{dt} = (1 - \beta_{Sat}) \frac{AET_{sat(t)}}{ST_{sat(t)}} \left(\delta^2 H_{sat(t)} + 1000\right), \qquad (1)$$

where $\delta^2 H_{sat}$ is the isotopic composition prior to fractionation, AET_{sat} is actual evapotranspiration and ST_{sat} is total storage (dynamic storage + passive mixing volume) in the saturation area prior to the loss of AET_{sat} at time t. An equivalent equation was used to update the isotopic composition of the hillslope store. The model structure also facilitates the tracking of water as it transits from precipitation to storage and streamflow, with this time-stamping approach used to estimate the age of water being routed to the stream from the groundwater reservoir and saturation area at each timestep. The age of water transiting from the saturation area to the stream is hereafter referred to as Age_Q_{sat}.

3.3 | Modelling experiments

3.3.1 | Subsampling of the δ^2H dataset

To investigate how the δ^2H sampling frequency in streamflow during model calibration impacts simulated daily outputs, the 7-year daily dataset was subsampled (Figure 4 and Table 1) to mimic various temporal and flow percentile-based sampling regimes. Every other day, every third day, weekly, fortnightly and monthly were selected as temporally focussed sampling strategies, with weekly and fortnightly being consistently Monday and the first day of the month chosen for the monthly resolution. Because high flows are essential to characterize the hydrological functioning of a catchment (Seibert & Beven, 2009), we also used a sampling strategy based on flow percentiles. Flows >Q_{50}, >Q_{25}, >Q_{10} and >Q_{05} were selected for flow percentile subsampling to ascertain if such sampling provided sufficient information content for model calibration. Specific targeting of flows <Q_{50} is not included as analysis of the discharge rate distribution showed little variation in isotope values as they were essentially baseflows derived from groundwater.

To assess the impact of dataset longevity and represent studies utilizing shorter datasets, the full dataset was shortened to 3 and 5 years, through removal of the most recent data, and subsampled according to the above protocol. Finally, the driest (01/10/17–30/09/18) and wettest (01/10/13–30/09/14) hydrological years, in terms of total precipitation, were also subsampled as above. Repeating experiments during these individual years would allow us to understand if more hydrologically extreme periods were more sensitive to a reduction in sampling effort. Whilst comparison to an "average" 1 year period was considered, such a time period was elusive – potentially due to recent increases in climatic variation. Therefore, only the most extreme years were included in analysis to avoid comparison to an "average" year not truly being so and causing a bias within result discussions.

3.3.2 | Model calibration

We used a daily time step for the modelling. This is consistent with the response time of the catchment, where previous research has shown that even during events, the sub-daily variation in isotopic composition of stream water is limited (Tunaley et al., 2017). All required model inputs other than δ^2H (Figure 3) were prepared at this daily resolution to allow results to be directly attributable to changes in streamflow isotope sampling. These inputs spanned the exact temporal extent as the respective δ^2H calibration dataset (e.g., when calibrating to 1, 3, 5 or 7 years of isotope data then 1, 3, 5 or 7 years of, for example, streamflow data would be included respectively). Prior to each model run, a standard 4 year period (01/10/2011–30/09/2015) was used for spin-up as preliminary work had shown this period is optimal for initialising storages. Leaving the period consistent between sampling resolutions and datasets ensured that changes in model performance were due to calibration effects of the reduced δ^2H sampling alone, as opposed to a change in the spin-up period.

For the model calibration we used a non-dominated genetic sorting algorithm (NSGA2 by Deb et al., 2002) which simultaneously optimized, with equal weighting, the modified Kling-Gupta efficiency

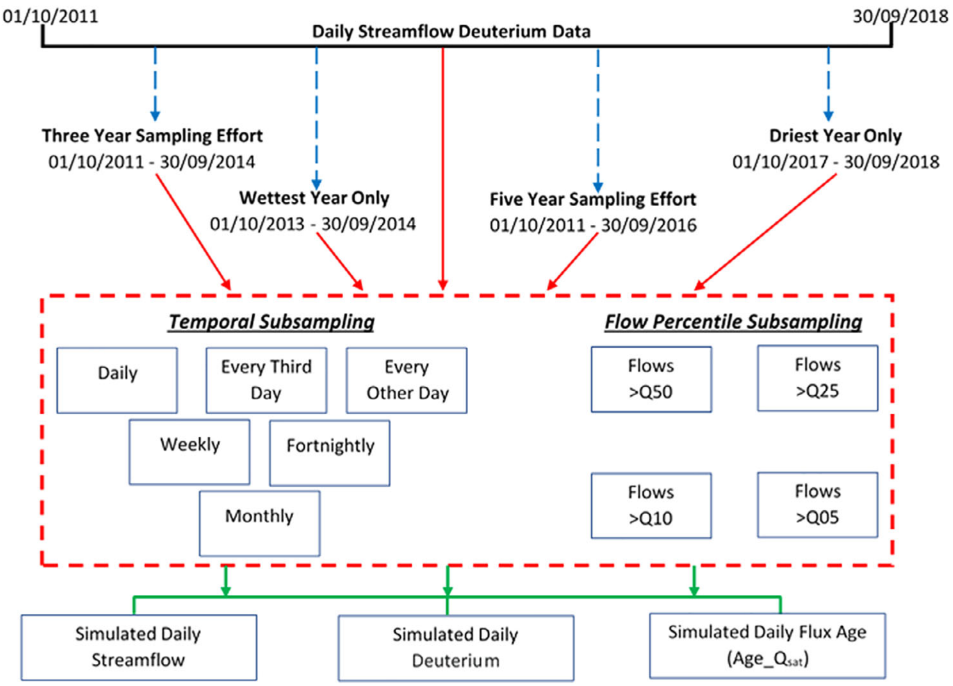

FIGURE 4 Schematic diagram of modelling experiments. The full 7-year dataset was subsampled according to various temporal and flow percentile specifications, as were four shorter datasets, to assess the impact of such sampling on the simulated daily outputs

TABLE 1 Number of δ^2H samples within each dataset

Sampling frequency *Dataset duration*	Daily	Every other day	Every third day	Weekly	Fortnightly	Monthly	>Q50	>Q25	>Q10	>Q05
01/10/2011–30/09/2018	2557	1279	853	366	183	84	1279	640	256	128
01/10/2011–30/09/2014	1096	548	366	157	79	36	548	274	110	55
01/10/2011–30/09/2016	1827	914	609	261	131	60	914	457	183	92
01/10/2013–30/09/2014	365	183	122	53	27	12	183	92	37	19
01/10/2017–30/09/2018	365	183	122	53	27	12	183	92	37	19

(Kling et al., 2012), hereafter KGE, for both flow and δ^2H within predefined parameter ranges (Table 2). In line with previous use of the model (e.g., Soulsby et al., 2015), 500 parameter sets were constrained over 100 iterations (resulting in 50 000 different parameter combinations tested) to provide a final optimal parameter population of 500 which, in absence of formal uncertainty analysis, provided confidence intervals around average simulated values. Performance metrics and graphical representations of simulated outputs were calculated and created using the median simulation from the 500 parameter sets at each timestep.

3.3.3 | Model evaluation

Throughout all modelling experiments, simulated flow, δ^2H and Age_Q_{sat} were evaluated for change as simulated flow and δ^2H represent the core model focus, whilst Age_Q_{sat} is indicative of any change in process realism (i.e., if flow performance remained relatively unchanged between model runs but Age_Q_{sat} was substantially altered, it would indicate the same flows were being simulated from differing mechanisms). Age_Q_{sat} was chosen over the total age of discharge as contributions of water from the groundwater store result in older water (i.e., >3 years) which are much less identifiable (Benettin et al., 2017). In contrast, the fluxes from the saturation area are responsible for contributing younger water (≤30 days old) to flow, so allowing the percentage of daily discharge comprised of young water to be calculated. Model evaluation, unless stated otherwise, was always based upon the daily simulated value versus the daily observed value, spanning the time period of the δ^2H calibration dataset, regardless of the δ^2H calibration datasets resolution. Thus, for example, the model calibrated with 3 years of monthly isotope data was evaluated based on 3 years of daily isotope data (given that simulated outputs were at daily resolution, regardless of model input resolution).

4 | RESULTS

4.1 | Effects of sampling frequency using the full 7-year dataset

Both streamflow and isotopes could, in terms of model performance metrics and visual appearance, be adequately and almost identically

TABLE 2 Parameter space explored to simultaneously optimize Kling-Gupta efficiency of Q and δ^2H

Parameter	a	b	r	k	α	upS_p	$satS_p$	$lowS_p$	β_{Up}	β_{Sat}
Unit	per day	per day	per day	per day	-	mm	mm	mm	-	-
Lower Bound	0.2	0.0001	0.2	0.001	0.1	0	0	0	0.95	0.95
Upper Bound	0.8	0.1	0.9	0.1	0.9	500	1000	1000	1	1

simulated, by daily, every 2 days, every 3 days, weekly and fortnightly sampling (Figure 5 and Table 3). Only when sampling was reduced to monthly intervals was there a substantial deterioration in the reliability of flow simulations, accuracy of isotope simulations and the associated uncertainty around process representation (Figure 5 and Table 3). Calibration using monthly samples also resulted in much younger water ages being simulated (Figure 6), as a result of a substantially decreased (~2.7 times) mean volume of water being available for mixing in the saturation area passive store (*$satS_p$*; Table 4). Parameter identifiability was also notably poorer as reflected through the increase in the standard deviation of *$satS_p$* within the retained parameter values, with both of these factors again translating into a higher uncertainty in the accuracy of process representation. Though simulated ages were much younger, analysis (not shown) of the simulated time series evidenced age dynamics, for example, general patterns of age fluctuations across time, were similarly captured to daily derived simulations.

The change in mixing volume also contributed to much poorer δ^2H simulations for monthly sampling (Figure 5 and Table 3), with the dynamics being poorly reproduced. Conversely, monthly sampling did result in an improvement in the flow performance due to high flows being better captured, though at the cost of process representation change in the flow simulation component of the model. Here, the mean value for the *b* parameter controlling groundwater flow was greatly increased (Table 4), causing the groundwater store to empty faster. The calibration process compensated for this by reducing the mean recharge rate into the reservoir (*r* parameter) to ensure that sufficient water was available to capture higher flows sourced from the hillslope and saturation reservoirs. These changes, coupled with poor δ^2H simulations, indicated increased uncertainty in the ability of the monthly dataset to represent hydrological processes when used for model calibration.

Sampling according to flow percentiles demonstrated there was little impact on δ^2H simulations when targeting flows >Q_{50} (Figure 5 and Table 3). Sampling only flows >Q_{10} led to substantive deteriorations in δ^2H simulations, which was unsurprisingly worsened when using only flows >Q_{05} (Figure 5). A substantive under-representation of low flows, as shown by reductions in logNSE values (Table 3) and graphically (Figure 5), further reduced certainty in the >Q_{10} and >Q_{05} derived parameter sets providing realistic representations of catchment function. At such extreme percentile targeting, the simulated Age_Q_{sat} remained resilient, being comparable to those simulations derived from daily sampling due to the mean *$satS_p$* parameter remaining relatively consistent. In contrast, the mean *$lowS_p$* parameter was ~17 times greater, causing the poorer δ^2H simulations.

4.2 | Influence of dataset longevity

Comparing modelled outputs for the time period that is common (01/10/2011–30/09/2014) to the three datasets of varying longevity (3, 5 and 7 years) evidenced differences in results. Slightly improved flow efficiency statistics occurred with a reduction in sampling longevity, particularly apparent for low flow simulations with (log NSE's of 0.44, 0.48 and 0.59 for 7, 5 and 3 years of weekly sampling, respectively). Shortening the sampling period to 3 years also resulted in a slight decrease in the number of days, within the common time period, where simulated saturated overland flow was ≤30 days old in comparison to 5 or 7 years of weekly data (69%, 77% and 77% of days, respectively). The slight reduction in the number of days with young water was caused by a modest increase in the mean *$satS_p$* parameter when using the 3-year dataset (Table 4 and Table S1).

KGE values calculated for δ^2H simulations during the common time period only were slightly impacted through a reduction in dataset duration (0.72, 0.77 and 0.76 for 3, 5 and 7 years of weekly data, respectively). However, visually, (Figure 7) simulations derived from 3 years of data were substantially poorer than suggested by the KGE, with overly depleted predictions for summer apparent. Simulations from 5 years of calibration data were visually more comparable to those with the 7-year dataset because the mean and standard deviation of *upS_p* and *$satS_p$* parameter values were similar to those derived from 7 years of data (Table S2). Thus, the reduction in sampling length to 5 years did not increase uncertainty around the accuracy of process representation in comparison to the 7 year dataset. As with the 7 year dataset (Section 4.1) patterns described for weekly data were comparable to sub-weekly resolutions.

When sampling according to flow percentiles, a reduction in dataset duration led to an improvement in Q simulations for the common time period, though 5 years proved better for low flows (log NSE's of 0.52, 0.57 and 0.50 for 3, 5 and 7 years of >Q_{50} sampling, respectively). δ^2H simulations (not shown) became poorer through a reduction in sampling longevity; with >Q_{50} sampling resulting in KGE's of 0.68, 0.74 and 0.76 for 3, 5 and 7 years of data, respectively. The reduction in KGE when sampling for 3 years indicated that the datasets information content for model calibration was reduced in comparison to the longer datasets, so meaning relative certainty in the derived parameter sets ability to accurately represent catchment functioning and associated δ^2H was also reduced. Age_Q_{sat} was little impacted (not shown), with no discernible change to either the dynamics or simulated ages with a reduced data longevity at >Q_{50}. This stability resulted from the *$satS_p$* parameter values remaining similar, whereas the other parameter (*$lowS_p$*) strongly influencing δ^2H

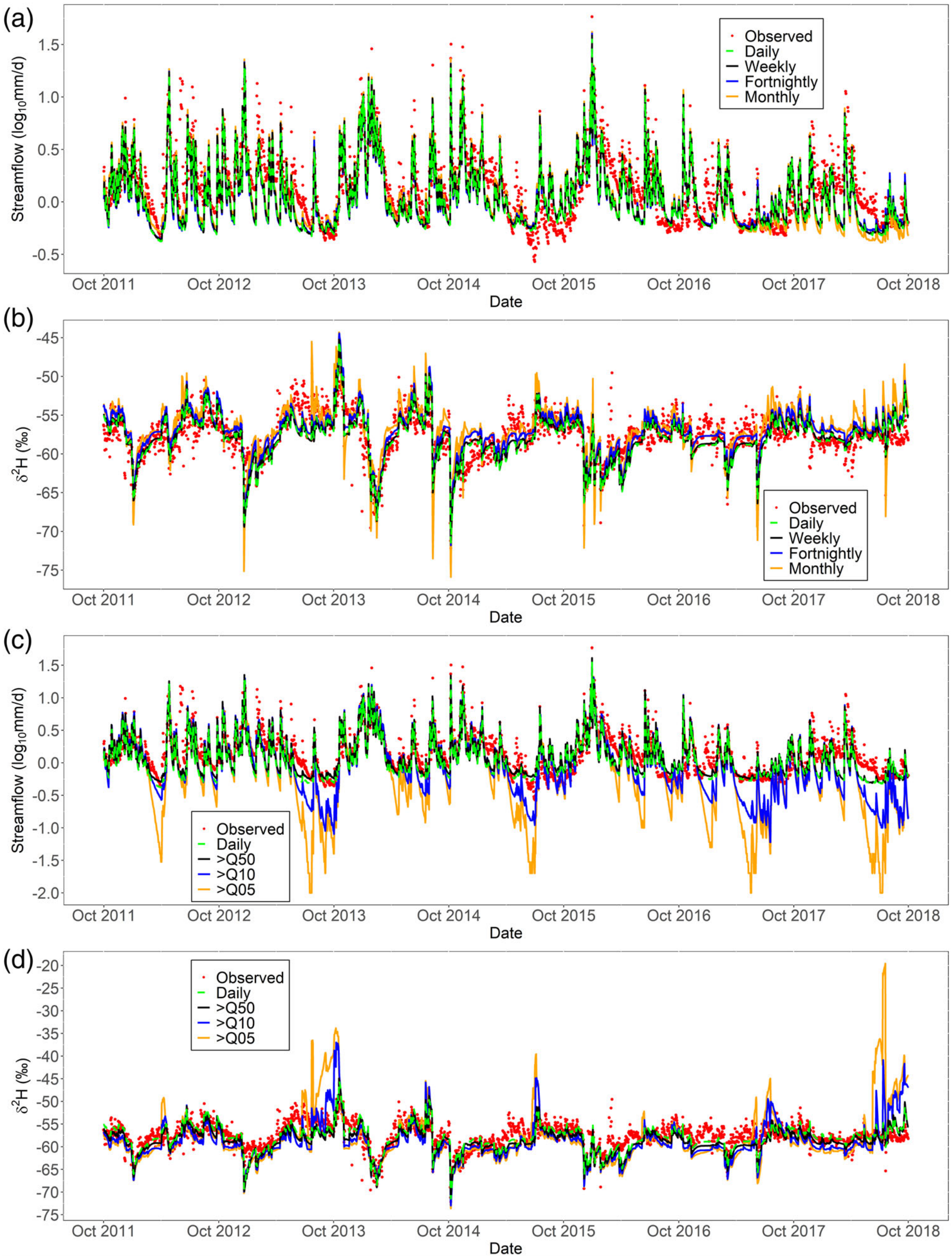

FIGURE 5 Time series of simulated daily streamflow and δ^2H from selected temporal (a,b) and flow percentile subsampling (c,d) across the 7-year dataset. Plotted values calculated using the median value at each timestep of the 500 retained parameter sets

simulations was much more sensitive to reducing the longevity of sampling to 3 years (Table 4 and Tables S1 and S2), explaining reductions in δ^2H performance but still reasonable Age_Q_{sat}.

When considering model outputs across the full extent of both the 3 and 5 year datasets (not the common time period) low flow simulations were worse, as observed in the 7 year time series, by targeting flows >Q_{10}, though this was less severe for the 3 year dataset (Table 3). Monthly resolution again caused reductions in the mean *$satS_p$* parameter when sampling for both 3 and 5 years, therefore decreasing simulated water ages and providing an unrealistic process representation. Identifiability of the *$satS_p$* parameter was also poorer as evidenced by the standard deviation of retained parameter sets (Supplementary Material Tables 2 and 3); with these statistics also being indicative of increased model uncertainty. Regarding δ^2H simulations, the 3 year dataset was less sensitive to both the more extreme temporal and percentile subsampling, whereas the 5 year dataset responded similarly to the 7 year dataset with simulations much worsened by monthly and >Q_{10} sampling (Table 3).

4.3 | Effects of sampling in more hydrologically extreme years

Flow simulations during the wet year (01/10/2013–30/09/2014) were the best of all datasets particularly with regard to simulation of low flows (log NSE's Table 3). The weekly resolution data however did

TABLE 3 Model performance statistics. Statistics calculated by comparing the median simulated daily value (and in parentheses the 5th and 95th percentiles) obtained from the 500 retained parameter sets derived from varying δ^{2}H resolution calibration datasets against the daily observed value for the extent of the time period indicated

KLING-GUPTA EFFICIENCY STATISTICS										
Isotopes (δ^{2}H)										
Sampling frequency										
Dataset duration	**Daily**	**Every other day**	**Every third day**	**Weekly**	**Fortnightly**	**Monthly**	**>Q50**	**>Q25**	**>Q10**	**>Q05**
Full 7 years (01/10/2011–30/09/2018)	0.72 (0.72, 0.72)	0.72 (0.72, 0.73)	0.73 (0.72, 0.73)	0.72 (0.72, 0.72)	0.72 (0.72, 0.72)	0.65 (0.65, 0.64)	0.71 (0.71, 0.70)	0.68 (0.69, 0.67)	0.37 (0.60, −0.31)	−0.30 (0.67, −0.44)
3 years (01/10/2011–30/09/2014)	0.73 (0.69, 0.69)	0.73 (0.70, 0.66)	0.70 (0.66, 0.74)	0.72 (0.67, 0.68)	0.71 (0.66, 0.70)	0.74 (0.74, 0.59)	0.68 (0.70, 0.55)	0.68 (0.67, 0.64)	0.63 0.65, 0.39)	0.65 (0.64, 0.41)
5 years (01/10/2011–30/09/2016)	0.76 (0.76, 0.72)	0.75 (0.75, 0.71)	0.75 (0.75, 0.74)	0.76 (0.76, 0.72)	0.74 (0.75, 0.71)	0.65 (0.71, 0.62)	0.73 (0.74, 0.69)	0.71 (0.70, 0.58)	0.29 (0.56, 0.16)	0.15 (0.59, 0.05)
Extreme wet year (01/10/2013–30/09/2014)	0.84 (0.83, 0.85)	0.85 (0.83, 0.85)	0.83 (0.83, 0.83)	0.84 (0.83, 0.85)	0.83 (0.83, 0.83)	0.64 (0.46, 0.67)	0.80 (0.82, 0.78)	0.76 (0.78, 0.70)	0.72 (0.77, 0.67)	0.25 (0.60, 0.23)
Extreme dry year (01/10/2017–30/09/2018)	0.63 (0.62, 0.60)	0.61 (0.62, 0.60)	0.61 (0.58, 0.58)	0.62 (0.62, 0.59)	0.60 (0.59, 0.59)	0.04 (0.03, −0.14)	0.46 (0.44, 0.43)	−1.07 (−4.06, −3.79)	−0.15 (−0.14, −0.18)	−0.34 (0.0, −0.2)
Hydrology (Q)										
Full 7 years (01/10/2011–30/09/2018)	0.73 (0.61, 0.78)	0.73 (0.59, 0.78)	0.71 (0.58, 0.78)	0.73 (0.61, 0.78)	0.72 (0.61, 0.77)	0.76 (0.71, 0.78)	0.75 (0.68, 0.78)	0.75 (0.71, 0.78)	0.75 (0.65, 0.81)	0.73 (0.65, 0.80)
3 years (01/10/2011–30/09/2014)	0.70 (0.68, 0.71)	0.70 (0.68, 0.71)	0.70 (0.68, 0.71)	0.70 (0.68, 0.71)	0.70 (0.68, 0.71)	0.70 (0.65, 0.71)	0.71 (0.68, 0.72)	0.71 (0.67, 0.73)	0.71 (0.64, 0.72)	0.71 (0.68, 0.71)
5 years (01/10/2011–30/09/2016)	0.77 (0.66, 0.80)	0.78 (0.66, 0.80)	0.76 (0.64, 0.80)	0.77 (0.68, 0.80)	0.78 (0.74, 0.80)	0.78 (0.71, 0.80)	0.72 (0.62, 0.79)	0.78 (0.70, 0.80)	0.78 (0.72, 0.81)	0.76 (0.58, 0.78)
Extreme wet year (01/10/2013–30/09/2014)	0.78 (0.71, 0.76)	0.72 (0.49, 0.67)	0.80 (0.73, 0.75)	0.72 (0.55, 0.68)	0.69 (0.47, 0.77)	0.82 (0.71, 0.79)	0.81 (0.81, 0.77)	0.82 (0.81, 0.77)	0.82 (0.81, 0.79)	0.69 (0.68, 0.64)
Extreme dry year (01/10/2017–30/09/2018)	0.49 (0.43, 0.55)	0.5 (0.45, 0.54)	0.49 (0.41, 0.55)	0.48 (0.4, −0.54)	0.45 (0.38, 0.55)	0.53 (0.49, 0.54)	0.48 (0.38, 0.55)	0.49 (0.40, 0.56)	0.50 (0.42, 0.55)	0.49 (0.39, 0.58)
NASH-SUTCLIFFE STATISTICS (COMPUTED ON NATURAL LOGARITHM CONVERTED VALUES)										
Hydrology (Q)										
Full 7 years (01/10/2011–30/09/2018)	0.52 (0.43, 0.48)	0.52 (0.43, 0.59)	0.53 (0.42, 0.58)	0.52 (0.44, 0.58)	0.51 (0.39, 0.57)	0.53 (0.48, 0.58)	0.57 (0.51, 0.60)	0.57 (0.51, 0.60)	0.01 (0.48, −2.36)	−1.59 (0.49, −2.49)
3 years (01/10/2011–30/09/2014)	0.60 (0.42, 0.63)	0.60 (0.46, 0.64)	0.60 (0.43, 0.63)	0.59 (0.43, 0.63)	0.60 (0.45, 0.63)	0.43 (−0.41, 0.56)	0.50 (0.01, 0.55)	0.52 (0.06, 0.56)	0.39 (0.47, −1.08)	0.46 (0.53, −0.23)
5 years (01/10/2011–30/09/2016)	0.56 (0.36, 0.64)	0.56 (0.34, 0.65)	0.57 (0.43, 0.64)	0.57 (0.39, 0.64)	0.56 (0.35, 0.61)	0.52 (0.37, 0.65)	0.61 (0.40, 0.64)	0.66 (0.36, 0.67)	−0.39 (0.35, −1.53)	−1.13 (0.31, −1.74)
Extreme wet year (01/10/2013–30/09/2014)	0.72 (0.48, 0.73)	0.77 (0.50, 0.68)	0.64 (0.46, 0.71)	0.77 (0.49, 0.71)	0.75 (0.50, 0.75)	0.46 (0.12, 0.62)	0.20 (−0.29, 0.60)	0.41 (−0.10, 0.57)	0.40 (0.57, 0.18)	0.36 (0.47, −0.11)
Extreme dry year (01/10/2017–30/09/2018)	−0.1 (−0.33, 0.09)	−0.08 (−0.27, 0.08)	−0.16 (−0.43, 0.14)	−0.09 (−0.29, 0.08)	0.06 (−0.34, 0.38)	−0.35 (−0.81, −0.20)	−0.32 (−0.7, −0.07)	−1.12 (−3.16, −0.07)	−0.4 (0.01, −2.21)	−0.84 (0.06, −2.59)

cause a volumetric under-estimation of higher flows, something not encountered in the multi-year datasets. This resulted in a lowered KGE in comparison to daily resolution data during the same year (Table 3). This decrease was principally driven by a reduction in the mean non-linearity α parameter, which was not fully compensated for by an increase in the mean *k* parameter (Table S3). Conversely, δ^2H simulations during the wet year were similar to temporal subsampling as the 7-year dataset, being little impacted until only monthly resolution data was used (Table 3). Here a substantial increase in the *lowS*$_p$ and decrease in the *satS*$_p$ mixing volumes caused a more substantive reductions in performance (Table S3 and Table 3). The alteration in the *satS*$_p$ parameter again caused a reduction in simulated water ages and more unrealistic process representation.

Regarding percentile subsampling in the wet year, flow simulations responded similarly to those when subsampling the full dataset (Section 4.1), apart from targeting flows >Q$_{05}$, which worsened simulations across the entire hydrograph (KGE, Table 3). In contrast to the 7-year dataset, sampling flows >Q$_{10}$ in the wet year had a less severe effect on δ^2H simulations despite the notable alteration in the mean *lowS*$_p$ parameter in both instances. Such a parameter alteration was less influential during the wet year as flows were less dependent on groundwater and dominated by saturation area fluxes, with the parameter important in controlling the isotopic signature of this store (*satS*$_p$) remaining consistent when targeting flows >Q$_{10}$. Again, this stability in mean *satS*$_p$ values meant that Age_Q$_{sat}$ simulations (not shown) were little impacted when considering any of the flow percentile sampling strategies.

Model performance for calibration using samples taken in the dry year (01/10/17–30/09/18) was the worst of all datasets with streamflow KGE's of only ~0.5 for all sampling strategies, whilst log NSE values indicated very poor simulations (Table 3) regardless of data resolution. Such poor simulations suggest reduced confidence in the accuracy of process representation. As with other datasets, calibration to monthly data again led to a reduction in the mean *satS*$_p$ parameter causing an increase in the number of days with an Age_Q$_{sat}$ ≤30 days old, whilst a ~4 times increase of the mean *k* parameter value (Table S4) facilitated better simulation of high flows. Simulations of δ^2H were also poor in comparison to other datasets, and particularly sensitive to monthly subsampling, with this being especially evident towards the end of the simulation period (Figure 8).

Sampling according to flow percentiles proved particularly impactful during the dry year, with targeting of flows >Q$_{25}$, >Q$_{10}$, and >Q$_{05}$ all causing substantial deteriorations in the δ^2H simulations (Table 3), due to systematic changes in the calibrated *satS*$_p$ and *lowS*$_p$ parameter values (Table S4). This had a severe impact on Age_Q$_{sat}$ simulations (Figure 9) with 3% and 2%, respectively, of days having an Age_Q$_{sat}$ ≤30 days as the volume of water available for mixing (*satS*$_p$) was greatly increased, in comparison to 50% of days having an Age_Q$_{sat}$ value of ≤30 days old with daily sampling during the dry year. Clearly, the lack of information in data collected during dry years results in the poorest simulations and largest uncertainties around the calibration datasets ability to accurately represent hydrological processes.

5 | DISCUSSION

5.1 | Impact of temporal frequency and longevity of data on model performance

In this study we took a 7 year time series of daily stable isotopes samples of precipitation and streamflow to investigate how sampling frequency and longevity impacts the calibration of a semi-distributed tracer-aided model in terms of the accuracy, process representation and uncertainty of flow, deuterium and saturation area flux age simulations for a catchment in the Scottish Highlands. To do so we resampled the streamflow isotope dataset to reflect typical streamflow sampling regimes.

When using 7 years of data for calibration, weekly sampling gave almost identical simulations to daily sampling and similar uncertainty bounds for model performance (Table 3). Indeed, only a slight deterioration was evident when fortnightly samples were used. This goes against the seemingly logical argument that higher frequency data provides greater information content for model calibration (an argument disproved at the event scale by Wang et al., 2017, 2019), as well as the findings of previous studies (e.g., Birkel et al., 2010a; Birkel et al., 2012; Stockinger et al., 2016; Timbe et al., 2015). However, differences in findings must be contextualized given that the modelling approaches used by previous studies were, apart from Birkel et al. (2010a), fundamentally different to the D-sat model employed in our study. Therefore, underlying assumptions and model structures could mean nuances in each model's sensitivity to data input change, and so results from different studies will not always be directly comparable. Moreover, catchment characteristics will play an integral role in the sensitivity of a model to reduced sampling. For instance, the study using a broadly comparable model structure (Birkel et al., 2010a) was based in a catchment subject to intensive agricultural activities (including drainage and soil compaction) in direct contrast to the BB land-use, and thus hydrological regime. In the BB streamflow δ^2H variations is particularly damped in comparison to the isotopic signature of precipitation. This is related to the hydrological importance of the valley bottom wetlands which store large quantities of water that mix with hillslope drainage and act as "isostats", largely setting the isotopic composition of the stream (see Tetzlaff et al., 2014). Consequently, weekly sampling in this catchment captures a large proportion of the isotopic variation, whereas in the small agricultural catchment used in Birkel et al. (2010a) and Birkel et al. (2012) weekly sampling was inadequate in capturing the distribution tails of the isotopic composition. The authors statistically demonstrated this by showing how a move from daily sampling to weekly sampling increased kurtosis (the weight of a distributions tails in comparison to the distribution centre) by 46%. For our 7-year dataset this only caused an 11% increase in kurtosis. The catchment characteristics and statistical properties described here also help to explain why weekly sampling was further shown to be unimpactful to δ^2H simulations in the other datasets of varying temporal longevity and hydrological conditions (Table 3).

However, decreasing the longevity of weekly data from 7 to 3 years (Section 4.2 and Figure 7) caused a reduction in δ^2H

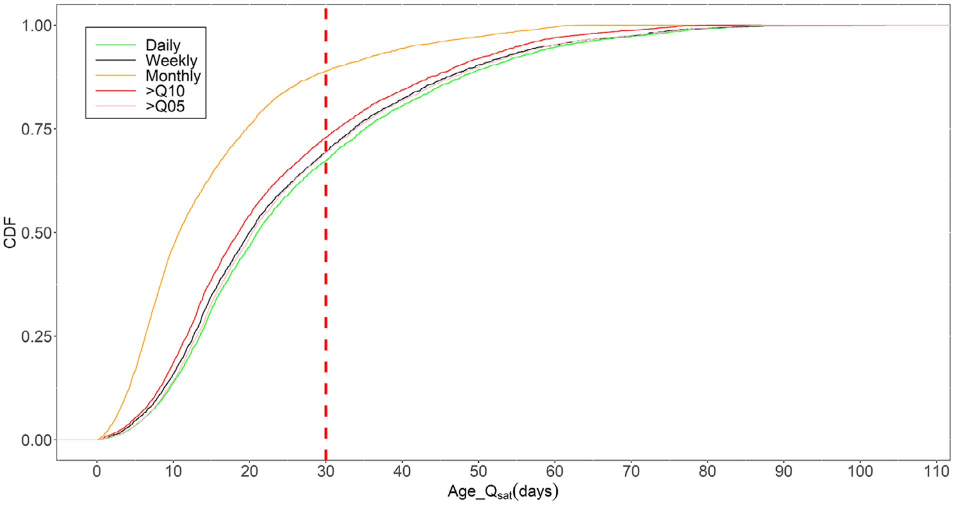

FIGURE 6 CDF's of daily Age_Q$_{sat}$, derived from simulation across the full 7-year time period (01/10/11–30/09/18) with different sampling resolutions. Plotted values calculated using the median value at each timestep of the 500 retained parameter sets

simulation performance for the common time period. As results from weekly sampling were directly comparable to results from daily sampling for the 3-year dataset, this would indicate that extending the sampling period, not the temporal frequency, is more beneficial for calibration of the D-Sat model in the BB. Such an approach helps reduce uncertainty around the datasets ability to produce parameter sets which more accurately represent catchment functioning. This finding may be attributable to the longer time period having a greater variation of hydrological conditions, allowing the calibration process to achieve a more robust parameter set that relates to a wider range of catchment responses. However, extending sampling beyond the 5-year period provided limited improvement in δ^2H simulations. In extending the sampling period from 5 to 7 years, the most extreme dry year since 2002–2003 was included in the calibration data, capturing the effects of a fundamental shift in catchment functioning. Under such conditions, the predictability of streamflow response to precipitation inputs is lower as a result of more non-linear interactions between input fluxes and storage dynamics. Previous work has shown that the semi-distributed approach of the D-sat model has limitations in representing and simulating the re-wetting of saturation areas under drier conditions as spatially heterogeneous areas re-wet to saturation and contribute to runoff at varying rates (Soulsby et al., 2015). Therefore, inclusion of this extra data does not necessarily aid the characterization of more common, predictable, hydrological patterns experienced within the timeframe under consideration and a more spatially distributed approach to fully exploit the additional information in calibration (e.g., Kuppel, Tetzlaff, Maneta, & Soulsby, 2018) is required.

Indeed, when calibrating with daily data during the dry year, model performance was substantially poorer for both flows and isotopes than for any other dataset, with this visually evident for δ^2H simulations in Figure 8. These results are likely due to reasons discussed above, which may have been further impacted by the spin-up period not containing data from the dry year. Such poor performance was in direct contrast to the temporal subsampling of the wet year which had the best initial performance statistics and was also no more sensitive to temporal subsampling than the full 7-year dataset. These findings are likely a product of the catchment exhibiting a more linear, predictable and immediate response to precipitation inputs when catchment storage is high due to high antecedent wetness and shallow flow paths. This is because the BB has been shown to respond in a manner consistent with recent theoretical work on tracer-aided modelling which has shown an “inverse storage” effect with stream water ages becoming lower when storage is high, as more organized lateral flow paths occur more frequently and non-linear interactions with sub-surface storage is reduced (Harman, 2015; Soulsby et al., 2015).

Temporal subsampling also showed the importance of including sufficient isotopic data for model calibration to help reduce uncertainty around if good model simulations, indicated by objective function statistics, were a result of accurate process representation. For example, flow calibration statistics (KGE) derived from monthly resolution sampling were better, or in one case equal to, those when using daily data; a finding in line with other studies where inclusion of tracer data causes the flow simulations to worsen (e.g., Bergström et al., 2002; McGuire et al., 2007; Stadnyk et al., 2013). However, there was a fundamental shift in the turnover and age of water in the model domain using monthly data, as shown by the increase in the number of days where Age_Q$_{sat}$ was ≤30 days old. This means that despite good flow simulations, the storage-flux-age interactions were known to be unrealistic, based on other studies in the catchment (e.g., Benettin et al., 2017; Kuppel et al., 2018). Assessment of parameter changes showed that this change stemmed from a reduction in the $satS_p$ parameter, so reducing the mixing volume available in the saturation area. Thus, despite reasonable simulations of flow, calibration with monthly data is insufficient to obtain parameter sets that represent hydrological processes in models where water age and storage dynamics are important, such as biogeochemical models, limiting their usefulness for environmental change assessment (e.g., Dick et al., 2015). Increasing the temporal longevity of monthly resolution sampling did not resolve this issue as the same problems occurred in the longer datasets.

TABLE 4 Summary statistics of the 500 parameter sets retained from NSGA2 calibration when subsampling the full dataset (01/10/2011–30/09/2018)

Parameter and ***Resolution***	**Min**	**Max**	**Mean**	**Stan. Dev.**	**Parameter and** ***Dilution***	**Min**	**Max**	**Mean**	**Stan. Dev.**
a *Daily*	0.29	0.53	0.45	0.06	**b** *Daily*	0.0004	0.0020	0.0008	0.0004
a *Weekly*	0.30	0.51	0.46	0.04	**b** *Weekly*	0.0004	0.0020	0.0008	0.0004
a *Monthly*	0.24	0.43	0.35	0.06	**b** *Monthly*	0.0008	0.0023	0.0015	0.0004
a *>Q50*	0.36	0.55	0.50	0.04	**b** *>Q50*	0.0007	0.0021	0.0012	0.0004
a *>Q10*	0.32	0.46	0.4	0.04	**b** *>Q10*	0.0019	0.0782	0.0312	0.0226
a *>Q05*	0.32	0.51	0.38	0.06	**b** *>Q05*	0.0012	0.0879	0.0532	0.0299
R *Daily*	0.54	0.80	0.62	0.05	**k** *Daily*	0.0113	0.0192	0.0163	0.0021
R *Weekly*	0.54	0.90	0.72	0.14	**k** *Weekly*	0.0127	0.0215	0.0180	0.0022
R *Monthly*	0.31	0.51	0.41	0.04	**k** *Monthly*	0.0230	0.0575	0.0372	0.0123
R *>Q50*	0.58	0.90	0.79	0.10	**k** *>Q50*	0.0200	0.0200	0.0200	<0.100
R *>Q10*	0.32	0.58	0.45	0.07	**k** *>Q10*	0.0250	0.058	0.0499	0.0090
R *>Q05*	0.30	0.51	0.43	0.05	**k** *>Q05*	0.0227	0.0527	0.034	0.0046
α *Daily*	0.89	0.90	0.90	0.00	**upS**$_{\mathbf{p}}$ *Daily*	496.95	500.00	499.75	0.37
α *Weekly*	0.85	0.90	0.90	0.01	**upS**$_{\mathbf{p}}$ *Weekly*	497.89	500.00	499.67	0.41
α *Monthly*	0.74	0.9	0.83	0.06	**upS**$_{\mathbf{p}}$ *Monthly*	451.65	499.99	494.82	7.01
α *>Q50*	0.87	0.90	0.90	<0.10	**upS**$_{\mathbf{p}}$ *>Q50*	497.50	500.00	499.72	0.31
α *>Q10*	0.60	0.88	0.67	0.07	**upS**$_{\mathbf{p}}$ *>Q10*	489.41	500.00	499.02	1.78
α *>Q05*	0.56	0.87	0.74	0.05	**upS**$_{\mathbf{p}}$ *>Q05*	403.30	500.00	452.67	30.21
satS$_{\mathbf{p}}$ *Daily*	70.95	88.94	81.93	3.70	**lowS**$_{\mathbf{p}}$ *Daily*	0.04	340.59	57.28	75.56
satS$_{\mathbf{p}}$ *Weekly*	64.01	110.50	77.30	4.14	**lowS**$_{\mathbf{p}}$ *Weekly*	0.01	458.85	55.54	82.44
satS$_{\mathbf{p}}$ *Monthly*	1.290	70.90	29.92	20.23	**lowS**$_{\mathbf{p}}$ *Monthly*	0.05	283.14	135.05	65.89
satS$_{\mathbf{p}}$ *>Q50*	75.87	175.28	81.12	7.44	**lowS**$_{\mathbf{p}}$ *>Q50*	11.61	899.52	269.45	141.78
satS$_{\mathbf{p}}$ *>Q10*	94.03	108.41	98.8	2.52	**lowS**$_{\mathbf{p}}$ *>Q10*	819.73	999.99	992.15	20.93
satS$_{\mathbf{p}}$ *>Q05*	88.68	173.86	95.81	8.41	**lowS**$_{\mathbf{p}}$ *>Q05*	943.36	999.99	994.73	9.80
β_{Up} *Daily*	0.98	0.99	0.98	<0.1	β_{Sat} *Daily*	0.96	0.97	0.96	<0.10
β_{Up} *Weekly*	0.98	0.98	0.98	<0.1	β_{Sat} *Weekly*	0.96	0.97	0.97	<0.10
β_{Up} *Monthly*	0.97	0.99	0.98	<0.1	β_{Sat} *Monthly*	0.95	0.97	0.96	<0.10
β_{Up} *>Q50*	0.98	0.99	0.99	<0.1	β_{Sat} *>Q50*	0.95	0.96	0.95	<0.10
β_{Up} *>Q10*	0.99	1.00	0.99	<0.1	β_{Sat} *>Q10*	0.95	0.95	0.95	<0.10
β_{Up} *>Q05*	0.99	1.00	1.00	<0.1	β_{Sat} *>Q10*	0.95	0.95	0.95	<0.10

5.2 | Impact of targeting specific sections of the hydrograph on model performance

Sampling only the highest flows, >Q_{10} and >Q_{05}, caused poor simulation of low flows, as would be expected given that the isotope data would not contain information pertaining to such conditions. However, as other parts of the hydrograph were simulated quite well, the KGE calibration metric did not reflect such poor low flow simulations and actually improved (Table 3) when sampling only flows >Q10. As with monthly resolution data, the streamflow KGE belies an increase in uncertainty around if derived parameter sets are able to accurately characterize hydrological processes given such extreme flow targeting caused severe decreases in δ^2H simulations. However, as targeting flows >Q_{50} had generally little impact on streamflow and δ^2H simulations, apart from the low flows of the hydrologically extreme years, it appears that the isotopic composition of the lowest flows did not add substantive additional information for calibration. The likely explanation is that sampling only higher than median flows in the BB still captures the range of δ^2H variation, including data from baseflows, allowing the calibration procedure to explore switches between baseflow dominance to stormflow and changing antecedent conditions. Furthermore, low flows are dominated by groundwater for which isotopic composition variability is small as such flows are not representing the variability of different areas within a catchment that can contribute to streamflow, as shown in Swiss catchments by Florianic et al. (2019). Moreover, previous work in the BB has shown that even when using daily data, the lack of variation in the isotopic composition of groundwater results in poorer simulations of low flow

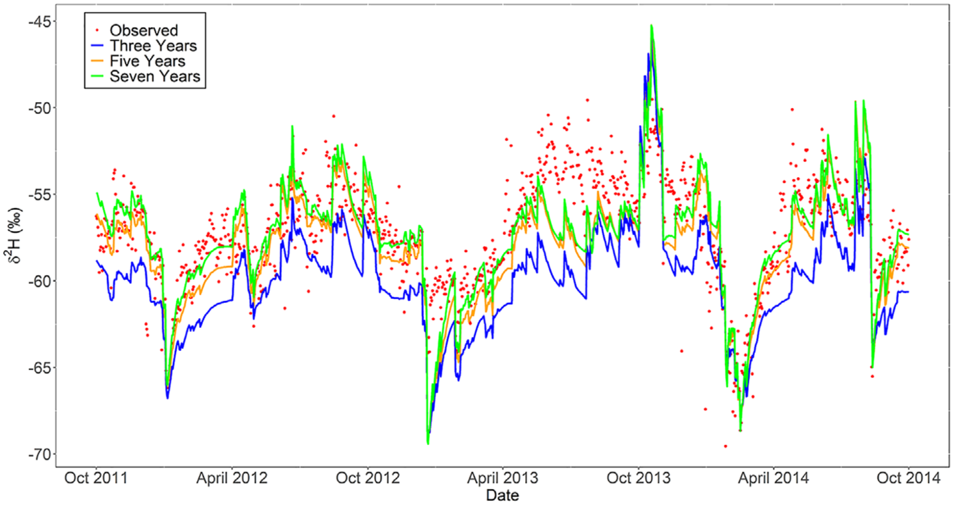

FIGURE 7 Time series of simulated daily δ^2H for the time period common to the three datasets of varying length when sampling at a weekly resolution. Plotted values calculated using the median value at each timestep of the 500 retained parameter sets

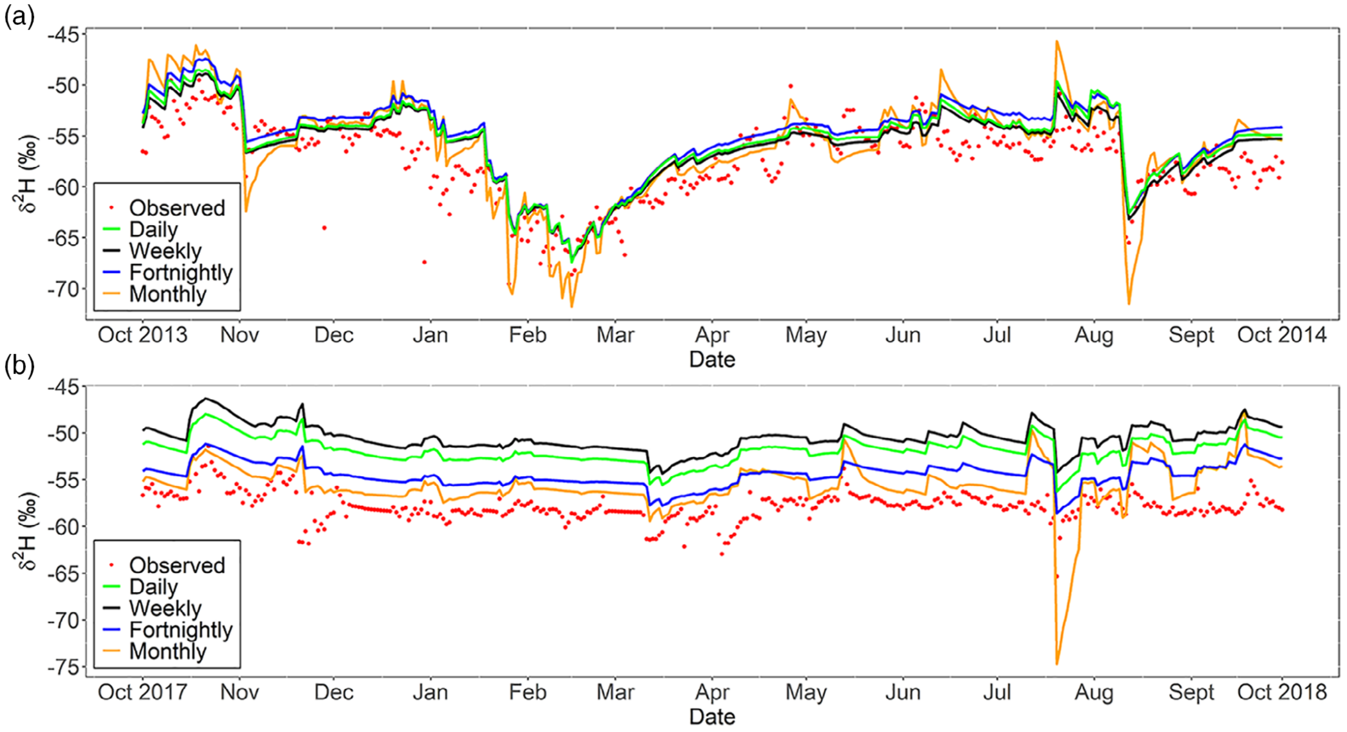

FIGURE 8 The impact on daily δ^2H simulations of select temporal subsampling during (a) the wet year and (b) the dry year. Calculated using the median value at each timestep of the 500 retained parameter sets

δ^2H variability (Soulsby et al., 2015). The dominance of groundwater contributions during lower flows also explains why the Age_Q$_{sat}$ simulations were, in all but the dry year, little impacted by flow percentile subsampling. The retained isotopic samples contain information specifically from time periods when the saturation area will be most volumetrically, and therefore age, variable; meaning the calibration process is still able to provide a correct representation of the parameter integral for Age_Q$_{sat}$ calculation (*satS$_p$*), despite reductions in isotope data.

Such results from flow percentile subsampling link with those resulting from temporally subsampling the dataset. Therefore, we conclude that instead of increasing the frequency of collection the increase in sampling longevity is of greater importance (Section 4.2). Also consistent with findings from temporal subsampling, the results based on calibration for the dry year of 2017–18 was extremely sensitive to reductions in data. This was clearly shown in the Age_Q$_{sat}$ statistics (Figure 9) where much older water ages were simulated as the calibrated mean *satS$_p$* value came close to the maximum permitted value (Table S4). This, coupled with the poorer daily resolution simulations and sensitivity of flow and δ^2H simulations to percentile targeted sampling during this period, means reduced sampling during drier periods can give misleading insights into catchment functioning.

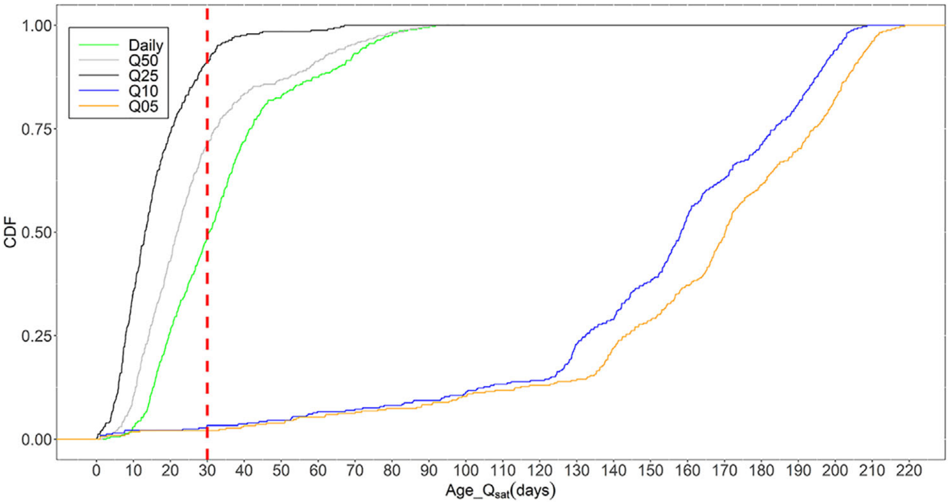

FIGURE 9 CDF's of daily Age_Q_{sat} simulations when sampling according to flow percentile thresholds during the dry year (01/10/17–30/09/18). Calculated using the median value at each timestep of the 500 retained parameter sets

5.3 | How can sampling strategies be designed to maximize resource use and information content for model calibration?

Data collection is the primary mechanism by which understanding of a catchments hydrological processes can be gained, and in tracer hydrology this invariably involves sample collection. However, there is often uncertainty regarding a datasets representativeness in term of sampling the full range of hydrological conditions that characterize catchment function (Birkel et al., 2010a). Consequently, such uncertainty also affects modelling when such datasets are being used in calibration. Data collection is invariably limited by time and financial resources, but the resultant datasets must be sufficiently robust to be used in model calibration and consequently environmental change studies. Ideally, researchers should know how much data, and from which part of the hydrograph, is required to understand and capture sufficient variations in catchment hydrological behaviour (Sprenger et al., 2019). Of course, this is very challenging in ungauged catchments (e.g., Tetzlaff et al., 2013). Consequently, knowledge on the information content of differing sampling strategies at well monitored long-term sites like the BB can help minimize the resource burden of data collection in new tracer studies, enabling an ever-wider range of catchments to be monitored, helping to globally reduce uncertainty in modelling landscape scale hydrological processes.

Our findings show that for this upland catchment weekly sampling of streamflow isotopes provides comparable information for tracer-aided model calibration as daily sampling during the same time period. Therefore, the resources for data collection can be reduced substantially, whilst levels of uncertainty in the datasets representation of the hydrological system is not increased. Indeed, our study shows that for the BB catchment, resources would be better focussed on extending the sampling period, rather than the sample frequency, to allow a greater range of hydrological conditions to be experienced, providing greater information content for calibration. Targeting only flows >Q_{50} also provides adequate information content in the BB whilst halving the sampling effort; however, there are two caveats. Firstly, a slight decrease in δ^2H simulation according to the calibration target metric became clear (KGE; Table 3) and whilst sampling was reduced by half, which though beneficial, is of less benefit than reducing sampling to weekly intervals. Secondly, the practicality of sampling such flows may be problematic because the flow distribution is inherently unknown. Here, findings should be viewed as demonstrating how flows <Q_{50} have less information content for calibration than those >Q_{50}. Therefore, where sampling effort is not regimented, as with weekly sampling, targeting medium and higher flows is the best use of resources. Such flows can be predicted using a researcher's own judgement and with simpler antecedent precipitation index-type algorithms (e.g., Birkel et al., 2010b) to identify when a catchment is likely to experience higher saturation rates and corresponding flow rates. More realistically, however, if autosamplers were used, only a sub-set of the collected samples would need to be analysed based on the measured flows.

If the objective of a sampling regime is to specifically characterize the Age_Q_{sat} dynamics then our results show that sampling may be reduced further (e.g., target only flows >Q_{10}) given the relative stability of the $satS_p$ parameter regardless of the percentile data used for calibration. However, caution must be employed given that other parameter values are substantially affected, and thus the uncertainty around characterization of the hydrological system. Such changes could feasibly impact the simulated volume and age of water entering and exiting the saturation area as hydrological conditions change between years and beyond those conditions tested here. This assertion is highlighted when considering how Age_Q_{sat} simulations were especially sensitive to flow percentile targeting during the driest year (Figure 9).

It is important to stress that these findings are derived for a northern upland catchment and should be viewed in that context. Nevertheless, similar approaches could be used in the increasingly large number of catchments where frequent tracer data are being collected for multiple years. It is likely that the nature of hydrological response times and degree of storage and mixing will be the key determinant for the required sampling frequency needed for model calibration and to determine travel time distributions. For example, in urban or humid tropical catchments (e.g., Correa et al., 2020) sub-daily sampling is likely essential to characterize isotope dynamics during storm peaks, whilst more seasonal flow regimes in groundwater dominated catchments can be adequately characterized by monthly sampling (Smith et al., 2021).

6 | CONCLUSIONS

We used a tracer-aided conceptual rainfall-runoff model to investigate how model calibration, process representation uncertainty, and flow, isotope and flux age simulations were impacted through reduced sampling of streamflow isotopes. Weekly sampling provided simulations and calibrated parameter value combinations, comparable to those derived from daily sampling and so uncertainty in the datasets ability to represent the hydrological processes was not increased. Similarly, samples taken at flows above the median (>Q_{50}) contained sufficient information for model calibration. Instead of increasing sampling resolution, sampling longevity should be extended where possible to allow a greater range of hydroclimatic conditions to be captured. However, the results did show that benefits plateaued after 5 years, though this could be due to the inclusion of a more hydrologically extreme year that the model could not simulate well. Importantly, we also demonstrated how infrequent sampling (e.g., monthly or only during extreme high flows) is inadequate to capture hydrological process variability. During drought conditions reducing sampling has a relatively greater impact as the calibration could not fully characterize the system dynamics. In both of these situations, the use of a model calibrated on this data as a baseline for water quality or environmental change studies would have an associated uncertainty increase in the quality and accuracy of that baseline. Furthermore, we also demonstrated how water ages can be used as an additional metric of process representation change beyond the simulation of isotopic values.

This research has enhanced understanding of both the amount of data required to provide sufficient information on a catchments function and where within the flow hydrograph this information is most informative. The study also casts doubt, in catchments such as the BB, on the assertion that higher frequency data is necessarily better. These conclusions can help inform optimized resource use for hydrologically similar catchments, enabling researchers to best target their sampling. Future work should replicate our study using other types of hydrological models and in catchments where hydrological processes are fundamentally different to the montane, wetland influenced catchment considered here.

ACKNOWLEDGEMENTS

We are grateful to the Leverhulme Trust ISOLAND project (RPG-2018-375) and the European Research Council VeWa project (VeWa project GA 335910) for funding and the lead author wishes to thank G.M. Stevenson for support leading up to, and throughout, the project.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the PURE data repository of the University of Aberdeen: https://abdn.pure.elsevier.com/en/datasets/water-isotopes-at-bruntland-burn-catchment (Kuppel, Tetzlaff, & Soulsby, 2018).

ORCID

Jamie Lee Stevenson https://orcid.org/0000-0003-2042-9130
Christian Birkel https://orcid.org/0000-0002-6792-852X
Doerthe Tetzlaff https://orcid.org/0000-0002-7183-8674
Chris Soulsby https://orcid.org/0000-0001-6910-2118

REFERENCES

Ala-aho, P., Tetzlaff, D., Laudon, H., McNamara, J., & 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. *Hydrology and Earth System Sciences*, *21*(10), 5089–5110. https://doi.org/10.5194/hess-2017-106

Benettin, P., Soulsby, C., Birkel, C., Tetzlaff, D., Botter, G., & Rinaldo, A. (2017). Using SAS functions and high-resolution isotope data to unravel travel time distributions in headwater catchments. *Water Resources Research*, *53*(3), 1864–1878. https://doi.org/10.1002/2016WR020117

Bergström, S., Lindström, G., & Petterson, A. (2002). Multi-variable parameter estimation to increase confidence in hydrological modelling. *Hydrological Processes*, *16*(2), 413–421. https://doi.org/10.1002/hyp.332

Birkel, C., Dunn, S. M., Tetzlaff, D., & Soulsby, C. (2010a). Assessing the value of high-resolution isotope tracer data in the stepwise development of a lumped conceptual rainfall-runoff model. *Hydrological Processes*, *24*(16), 2335–2348. https://doi.org/10.1002/hyp.7763

Birkel, C., & Soulsby, C. (2015). Advancing tracer-aided rainfall-runoff modelling: A review of progress, problems and unrealised potential. *Hydrological Processes*, *29*(25), 5227–5240. https://doi.org/10.1002/hyp.10594

Birkel, C., Soulsby, C., & Tetzlaff, D. (2014). Developing a consistent process-based conceptualization of catchment functioning using measurements of internal state variables. *Water Resources Research*, *50*(4), 3481–3501. https://doi.org/10.1002/2013WR014925

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. *Hydrological Processes*, *29*, 2956–2969. https://doi.org/10.1002/hyp.10414

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. *Hydrological Processes*, *26*(2), 308–316. https://doi.org/10.1002/hyp.8210

Birkel, C., Tetzlaff, D., Dunn, S. M., & Soulsby, C. (2010b). Towards a simple dynamic process conceptualization in rainfall-runoff models using multi-criteria calibration and tracers in temperate, upland catchments. *Hydrological Processes*, *24*, 260–275. https://doi.org/10.1002/hyp.7478

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 Resources Research*, *47*, 1–15. https://doi.org/10.1029/2010WR009547

Bowen, G. J. (2008). Spatial analysis of the intra-annual variation of precipitation isotope ratios and its climatological corollaries. *Journal of Geophysical Research*, *113*(D5), D05113. https://doi.org/10.1029/2007JD009295

Bowen, G. J., Cai, Z., Fiorella, R. P., & Putman, A. L. (2019). Isotopes in the water cycle: Regional- to global-scale patterns and applications. *Annual Review of Earth and Planetary Sciences*, *47*, 453–479. https://doi.org/10.1146/annurev-earth-053018-060220

Correa, A., Gutierres, J., Dehaspe, J., Quesada, A. M. D., Sánchez, R., Soulsby, C., & Birkel, C. (2020). A spatially-distributed assessment of non-stationary green and blue water ages in a pristine tropical rainforest. *Geophysical Research Abstracts*, *21*, 01–01.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multi-objective genetic algorithm: NSGA2- II. *IEEE Transactions on Evolutionary Computation*, *6*(2), 182–197. https://doi.org/10.1109/4235.996017

Dick, J. J., Tetzlaff, D., Birkel, C., & Soulsby, C. (2015). Modelling landscape controls on dissolved organic carbon sources and fluxes to streams. *Biogeochemistry*, *122*(2), 361–374. https://doi.org/10.1007/s10533-014-0046-3

Florianic, M. G., Fischer, B. M., Molnar, P., Kirchner, J. W., & van Meerveld, H. J. (2019). Spatial variability in specific discharge and streamwater chemistry during low flows: Results from snapshot sampling campaigns in eleven Swiss catchments. *Hydrological Processes*, *33*(22), 2847–2866. https://doi.org/10.1002/hyp.13532

Harman, C. J. (2015). Time-variable transit time distributions and transport: Theory and application to storage-dependent transport of chloride in a watershed. *Water Resources Research*, *51*, 1–30. https://doi.org/10.1002/2014WR015707

Hrachowitz, M., Soulsby, C., Tetzlaff, D., & Malcolm, I. A. (2011). Sensitivity of mean transit time estimates to model conditioning and data availability. *Hydrological Processes*, *25*(6), 980–990. https://doi.org/10.1002/hyp.7922

Hrachowitz, M., Soulsby, C., Tetzlaff, D., Malcolm, I. A., & Schoups, G. (2010). Gamma distribution models for transit time estimation in catchments: Physical interpretation of parameters and implications for time-variant transit time assessment. *Water Resources Research*, *46*(10), W10536. https://doi.org/10.1029/2010WR009148

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*(3), W03S04. https://doi.org/10.1029/2005WR004362

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, *424*, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011

Knapp, J. L. A., Neal, C., Schlumpf, A., Neal, M., & Kirchner, J. W. (2019). New water fractions and transit time distributions at Plynlimon, Wales, estimated from stable water isotopes in precipitation and streamflow. *Hydrology and Earth System Sciences*, *23*, 4367–4388. https://doi.org/10.5194/hess-23-4367-2019

Kuppel, S., Tetzlaff, D., Maneta, M. P., & Soulsby, C. (2018). EcH$_2$O-iso 1.0: Water isotopes and age tracking in a process-based, distributed ecohydrological model. *Geoscientific Model Development*, *11*, 3045–3069. https://doi.org/10.5194/gmd-11-3045-2018

Kuppel, S., Tetzlaff, D., & Soulsby, C. (2018). *Water isotopes at Bruntland Burn catchment*. Retrieved from https://abdn.pure.elsevier.com/en/datasets/water-isotopes-at-bruntland-burn-catchment

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 Resources Research*, *50*(6), 5342–5350. https://doi.org/10.1002/2013WR015141

McGuire, K. J., Weiler, M., & McDonell, J. J. (2007). Integrating tracer experiments with modelling to assess runoff processes and water transit times. *Advances in Water Resources*, *30*(4), 824–837. https://doi.org/10.1016/j.advwatres.2006.07.004

McIntyre, N. R., & Wheater, H. S. (2004). Calibration of an in-river phosphorus model: Prior evaluation of data needs and model uncertainty. *Journal of Hydrology*, *290*(1-2), 100–116. https://doi.org/10.1016/j.jhydrol.2003.12.003

Pool, S., Viviroli, D., & Seibert, J. (2017). Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration? *Journal of Hydrology*, *554*, 613–622. https://doi.org/10.1016/j.jhydrol.2017.09.037

Remondi, F., Kirchner, J. W., Burlando, P., & Fatichi, S. (2018). Water flux tracking with a distributed hydrological model to quantify controls on the spatio-temporal variability of transit time distributions. *Water Resources Research*, *54*(4), 3081–3099. https://doi.org/10.1002/2017WR021689

Seibert, J., & Beven, K. J. (2009). Gauging the ungauged basin: How many discharge measurements are needed? *Hydrology and Earth System Sciences*, *13*(6), 883–892.

Seibert, J., Rodhe, A., & Bishop, K. (2003). Simulating interactions between saturated and unsaturated storage in a conceptual runoff model. *Hydrological Processes*, *17*(2), 379–390. https://doi.org/10.1002/hyp.1130

Smith, A. A., Tetzlaff, D., Kleine, L., Maneta, M., & Soulsby, C. (2021). Upscaling land-use effects on water partitioning and water ages using tracer-aided ecohydrological models. *Hydrology and Earth System Sciences*, (Preprint). https://doi.org/10.5194/hess-2020-539

Soulsby, C., Birkel, C., Geris, J., Dick, J., Tunaley, C., & Tetzlaff, D. (2015). Streamwater age distributions controlled by storage dynamics and nonlinear hydrologic connectivity: Modelling with high-resolution isotope data. *Water Resources Research*, *51*, 7759–7776. https://doi.org/10.1002/2015WR017888

Sprenger, M., Stumpp, C., Weiler, M., Aeschbach, W., Allen, S. T., Benettin, P., Dubbert, M., Hartmann, A., Hrachowitz, M., Kirchner, J. W., McDonnel, J. J., Orlowski, N., Penna, D., Pfahl, S., Rinderer, M., Rodriguez, N., & Werner, C. (2019). The demographics of water: A review of water ages in the critical zone. *Reviews of Geophysics*, *57*(3), 800–834. https://doi.org/10.1029/2018RG000633

Sprenger, M., Tetzlaff, D., Tunaley, C., Dick, J., & Soulsby, C. (2017). Evaporation fractionation in a peatland drainage network affects stream water isotope composition. *Water Resources Research*, *53*(1), 851–866. https://doi.org/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. *Hydrological Processes*, *27*(25), 3791–3810. https://doi.org/10.1002/hyp.9695

Stadnyk, T. A., & Holmes, T. L. (2020). On the value of isotope-enabled hydrological model calibration. *Hydrological Sciences Journal*, *65*(9), 1525–1538. https://doi.org/10.1080/02626667.2020.1751847

Stockinger, M. P., Bogena, H. R., Lücke, A., Diekkrüger, B., Cornelissen, T., & Vereecken, H. (2016). Tracer sampling frequency influences estimates of young water fraction and Streamwater transit time distribution. *Journal of Hydrology*, *541*, 952–964. https://doi.org/10.1016/j.jhydrol.2016.08.007

Tetzlaff, D., Al-Rawas, G., Blöschl, G., Carey, S. K., Fan, Y., Hrachowitz, M., Kirnbauer, R., Jewitt, G., Laudon, H., McGuire, K. J., Sayama, T., Soulsby, C., Zehe, E., & Wagener, T. (2013). Process realism: Flow paths and storage. In *Runoff prediction in ungauged basins – Synthesis across processes, places and scales* (pp. 53–70). Cambridge University Press.

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 Resources Research*, *50*(2), 969–985.

Timbe, E., Windhorst, D., Celleri, R., Timbe, L., Crespo, P., Frede, H. G., Feyen, J., & Breuer, L. (2015). Sampling frequency trade-offs in the assessment of mean transit times of tropical montane catchment waters under semi-steady-state conditions. *Hydrology and Earth*

System Sciences, *19*(3), 1153–1168. https://doi.org/10.5194/hess-19-1153-2015

Tunaley, C., Tetzlaff, D., Birkel, D., & Souslby, C. (2017). Using high-resolution isotope data and alternative calibration strategies for a tracer-aided runoff model in a nested catchment. *Hydrological Processes*, *31*(22), 3962–3978. https://doi.org/10.1002/hyp.11313

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. *Hydrological Processes*, *30*(25), 4761–4778. https://doi.org/10.1002/hyp.10902

von Freyberg, J., Studer, B., & Kirchner, J. W. (2017). A lab in the field: High-frequency analysis of water quality and stable isotopes in stream water and precipitation. *Hydrology and Earth System Sciences*, *21*(3), 1721–1739. https://doi.org/10.5194/hess-21-1721-2017

Wang, L., van Meerveld, H. J., & Seibert, J. (2017). When should stream water be sampled to be most informative for event-based, multi-criteria model calibration? *Hydrology Research*, *48*(6), 1566–1584. https://doi.org/10.2166/nh.2017.197

Wang, L., von Freyberg, J., van Meerveld, I., Seibert, J., & Kirchner, J. (2019). What is the best time to take stream isotope samples for event-based model calibration? *Journal of Hydrology*, *577*, 123950. https://doi.org/10.1016/j.jhydrol.2019.123950

Weiler, M. (2003). How does rainfall become runoff? A combined tracer and runoff transfer function approach. *Water Resources Research*, *39*(11). https://doi.org/10.1029/2003WR002331

Zhang, Z., Chen, X., Cheng, Q., & Soulsby, C. (2019). Storage dynamics, hydrological connectivity and flux ages in a karst catchment: Conceptual modelling using stable isotopes. *Hydrology and Earth System Sciences*, *23*, 51–71. https://doi.org/10.5194/hess-23-51-2019

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Stevenson, J. L., Birkel, C., Neill, A. J., Tetzlaff, D., & Soulsby, C. (2021). Effects of streamflow isotope sampling strategies on the calibration of a tracer-aided rainfall-runoff model. *Hydrological Processes*, *35*(6), e14223. https://doi.org/10.1002/hyp.14223