What is the best time to take stream isotope samples for event-based model calibration?

Wang, Ling ; von Freyberg, Jana ; van Meerveld, Ilja ; Seibert, Jan ; Kirchner, James W

Abstract: Environmental tracer data, such as the stable water isotopic composition of streamwater and precipitation, are valuable for understanding runoff generation processes and calibrating hydrological models. Despite recent technical advancements, the collection and analysis of streamwater and precipitation samples still involve significant costs and efforts. Consequently, it is useful to study how many samples need to be collected as a basis for model calibration and what the most informative sampling times are. In previous studies, we used the Birkenes hydrological model and synthetic data to explore the value of a few stream isotope samples for event-based model calibration. Our results showed that the information from two or three selected isotope samples, and particularly a sample taken during the falling limb of the hydrograph, improved model performance significantly compared to calibration against streamflow data only. In this follow-up study, we used a unique Swiss dataset with high-frequency isotope measurements of precipitation and streamwater during six rainfall-runoff events to determine which stream isotope samples were most informative for model calibration. Several benchmarks were used to judge model validation performance. Our results showed that for 83% of the 42 possible combinations of calibration and validation events two strategically selected streamwater samples improved model performance compared to the lower benchmark when only stream stage data were used for calibration. The most informative samples also improved the calibration of a mixing-related parameter. When the model was calibrated and validated for the same event, the most informative samples for model calibration were those from the early part of the falling limb, which confirmed the results of the previous studies with synthetic data. Pre-event isotope samples were more informative when the model was calibrated for one event and then validated against other events or all events combined. Our results suggest that relatively inexpensive event-based data for model calibration can be obtained for so-far ungauged catchments by placing a water level logger in the stream and collecting precipitation isotope data and a few streamflow samples. These samples should preferably include a pre-event sample and a sample from the early falling limb. However, similar analyses for a wider range of sites and events, in combination with a selection of models with different structures and assumptions, are needed to confirm these results and to optimise sampling strategies further.

DOI: <https://doi.org/10.1016/j.jhydrol.2019.123950>

Posted at the Zurich Open Repository and Archive, University of Zurich

ZORA URL: <https://doi.org/10.5167/uzh-177631>

Journal Article

Accepted Version

The following work is licensed under a Creative Commons: Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) License.

Originally published at:

Wang, Ling; von Freyberg, Jana; van Meerveld, Ilja; Seibert, Jan; Kirchner, James W (2019). What is the best time to take stream isotope samples for event-based model calibration? *Journal of Hydrology*, 577:123950.

DOI: <https://doi.org/10.1016/j.jhydrol.2019.123950>

What is the best time to take stream isotope samples for event-based model calibration?

Ling Wang¹, Jana von Freyberg^{2,3}, Ilja van Meerveld¹, Jan Seibert^{1,4}, James W. Kirchner^{2,3}

¹Department of Geography, University of Zurich, Zurich, Switzerland

²Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland

³Swiss Federal Institute for Forest, Snow and Landscape Research (WSL), Birmensdorf, Switzerland

⁴Department of Earth Sciences, Uppsala University, Uppsala, Sweden

Correspondence: Ling Wang (ling.wang@geo.uzh.ch)

Abstract

Environmental tracer data, such as the stable water isotopic composition of streamwater and precipitation, are valuable for understanding runoff generation processes and calibrating hydrological models. Despite recent technical advancements, the collection and analysis of streamwater and precipitation samples still involve significant costs and efforts. Consequently, it is useful to study how many samples need to be collected as a basis for model calibration and what the most informative sampling times are. In previous studies, we used the Birkenes hydrological model and synthetic data to explore the value of a few stream isotope samples for event-based model calibration. Our results showed that the information from two or three selected isotope samples, and particularly a sample taken during the falling limb of the hydrograph, improved model performance

significantly compared to calibration against streamflow data only. In this follow-up study, we used a unique Swiss dataset with high-frequency isotope measurements of precipitation and streamwater during six rainfall-runoff events to determine which stream isotope samples were most informative for model calibration. Several benchmarks were used to judge model validation performance. Our results showed that two strategically selected streamwater samples improved model performance compared to calibration with only stream stage data for 83% of the 42 possible combinations of calibration and validation events. The most informative samples also improved the calibration of a mixing-related parameter. When the model was calibrated and validated for the same event, the most informative samples for model calibration were those from the early part of the falling limb, which confirmed the results of the previous studies with synthetic data. Pre-event isotope samples were more informative when the model was calibrated for one event and then validated against other events or all events combined. Our results suggest that relatively inexpensive event-based model calibration can potentially be carried out for ungauged catchments by placing a water level logger in the stream and collecting precipitation isotope data and a few streamflow samples, preferably including a pre-event sample and a sample from the early falling limb. However, similar analyses for a wider range of sites and events, in combination with a selection of models with different structures and assumptions, are needed to confirm these results and to optimise sampling strategies further.

Keywords: value of limited data, sampling time, isotope data, event-based model calibration, stream stage, ungauged catchments

1 Introduction

While it is often rather easy to achieve a good streamflow simulation with a bucket-type model, the good model performance may be achieved for the wrong reasons (Kirchner, 2006; Kirchner et al.,

1996; Seibert and McDonnell, 2002). To more rigorously test model representations of hydrological processes, the dynamics of the chemical and isotopic composition of streamwater can be used as additional calibration criteria (Birkel and Soulsby, 2015; Dunn and Bacon, 2008; Fenicia et al., 2010; Hrachowitz et al., 2015). Such tracer data can help in understanding runoff generation processes (Kendall and McDonnell, 1998), and testing and improving tracer-aided conceptual and numerical models (Birkel and Soulsby, 2015). For example, Hartmann et al. (2013) applied two models to a karst system in Southern Spain and found that hydrochemical information (e.g., $\delta^{18}\text{O}$, Cl, SO_4 and NO_3) was crucial to estimate reasonable parameter values for both models. Davies et al. (2011) used tracer breakthrough data from the hillslope-scale tracer experiment in a Swedish catchment (Nyberg et al., 1999) and streamflow data to test different hypotheses with respect to tracer injection and transport using the Multiple Interacting Pathways (MIPs) model. Although the simulated hydrographs for the different hypotheses were all similar to the measured hydrograph, the modelled tracer breakthrough curves were very different, and therefore were helpful for testing the hypotheses (Davies et al., 2011). In another study, Sprenger et al. (2015) observed soil moisture and pore water deuterium composition for three study sites at different depths, and found that the information from the soil water isotope profiles led to a more realistic parameterization of the soil hydrological models. Fenicia et al. (2008) used deuterium data for model calibration and tested three model structures for the Maimai M8 watershed to explore the relation between calibration data and model complexity. Their results suggest that isotope data can provide different information on catchment behaviour than streamflow, and that the use of more data types for calibration warrants the higher data requirements of more complex models. For most studies, when isotope data were included for model calibration, the model simulation performance of streamflow decreased, but model realism improved, compared to calibration with streamflow alone (Bergström et al., 2002; McGuire et al., 2007; Stadnyk et al., 2013; Vaché and McDonnell, 2006; Wagener, 2003). Our previous two studies (Wang et al., 2018, 2017) also showed that using isotope data for model calibration led to more constrained parameter values and a slight decrease in streamflow simulation performance compared

to calibration with streamflow data alone. As a counterexample, the results obtained by Vaché et al. (2004) can be mentioned. They found that for the San Jose basin (Chile), the parameter sets with good streamflow simulations were rejected because they did not result in a reasonable pre-event water contribution. Therefore, it is worthwhile to explore the value of isotope data for model calibration. However, the cost and effort involved in collecting and analysing a large number of streamwater samples during runoff events severely constrain the use of tracer data in most modelling studies. Particularly when hydrological models need to be applied to ungauged catchments, it is important to consider which measurements and sampling strategies are most informative and, at the same time, most cost-effective.

Decisions on the required temporal resolution of streamwater samples are usually based on the research purpose (e.g., hydrograph separation, load estimation) and practical constraints (e.g., field sampling logistics), so that the sampling frequency can vary from minutes to months per sample. For modellers, the temporal resolution (i.e., time step) of model inputs and outputs depends on both data availability and the modelling objectives (e.g., flood prediction, low flow simulations), and typically varies from an hour to a month. In this study, we focus on event-based hydrological modelling for which data of the runoff response and the isotopic composition of streamwater from one rainfall-runoff event are used to calibrate the model parameter sets and to simulate the streamflow and tracer responses for other events. Considering that a rainfall-runoff event normally lasts a few hours to days, high-frequency tracer data (e.g., hourly) might be beneficial for event-based model calibration; however, obtaining these data is expensive and may not significantly improve model calibration or validation performance if the additional samples merely "connect the dots" between lower-frequency data and thus contain little additional information. Because event-based model calibration is particularly useful when long-term data are not available, it is important to evaluate the informative value or information content of the additional event-based tracer data for model calibration.

In our previous studies (Wang et al., 2018, 2017), we used the Birkenes model (a tracer-aided bucket-type model; Hooper et al., 1988) and three design rainfall events (with limited temporal variability in rainfall intensity and isotopic composition) as model inputs to generate synthetic streamflow and streamwater isotope datasets in order to determine the value of stream isotope data for model calibration. We found that (1) model performance improved significantly when two or three isotope samples were used together with streamflow data for calibration, compared to calibration with only streamflow data, (2) the most informative samples for model calibration were more frequently taken on the falling limb of the hydrograph, and (3) systematic errors in the calibration data caused the times of the most informative samples to be more clustered and earlier than for error-free data (Wang et al., 2018, 2017). To test these findings in real-world catchments, field-based datasets can be used in a similar framework. In this follow-up study, we used recently published unique high-frequency isotope and stream stage data for a small stream near Birmensdorf, Switzerland (von Freyberg et al., 2017). Streamflow data were not available for this catchment, but previous studies have demonstrated the value of water level time series for model calibration, particularly in humid and temperate climates (Seibert and Vis, 2016). Thus, the stream stage and isotope data were used to calibrate and evaluate the Birkenes model and to assess whether real-world data lead to similar results as the synthetic data in our earlier work. More specifically, we address the following research questions:

1. What is the value of one or two isotope samples for event-based model calibration compared to calibration with stream stage data only?
2. What part of the hydrograph contains the most informative samples for event-based model calibration (pre-event, rising limb, near-peak, early falling limb or late falling limb) and does this vary from event to event?

The results of this study would be particularly useful for hydrological modelling studies in ungauged catchments since they would benefit considerably from a more targeted planning of fieldwork. While it might not be feasible to measure streamflow and obtain high-frequency tracer data in such

catchments, the installation of a water level logger, rain gauge and the collection of a few streamwater and precipitation samples during a few events might be sufficient for model calibration.

2 Methods

2.1 General approach

To assess the value of stream isotope data for model calibration, we used a similar approach to Seibert and Beven (2009) and Hartmann et al. (2017). We compared the validation performance of a model that was calibrated with a subset of the stream isotope data to 1) the validation performance of the same model calibrated with the complete isotope dataset and 2) the validation performance for the same model calibrated without any isotope data. As in our previous studies (Wang et al., 2018, 2017), we used the Birkenes model (Christophersen et al., 1982; Hooper et al., 1988; section 2.3). However, in contrast to the previous studies that were based on synthetic data, we based the present analysis on hourly measurements of precipitation and stream stage and the isotopic composition of precipitation and streamwater for six rainfall-runoff events (Figure 1, section 2.2). More specifically, our modelling approach included three main steps: model calibration, model validation and model evaluation. Figure 2 shows the workflow for model calibration with isotope data from one stream isotope sample ($n = 1$).

- *Step 1 Model calibration:* We selected one of the six events for model calibration (E_c , where $E_{c=1}$ means event E1 was selected as the calibration event). We then calibrated the model based on all stream stage data and one stream isotope sample from this event using the SCE-UA algorithm (Duan et al., 1992, 1993). Because the Birkenes model predicts streamflow and not stream stage, we used the Spearman rank correlation coefficient as the objective function to compare the simulated streamflow to measured stream stage (Seibert and Vis, 2016), along with the relative error of the simulated and measured isotopic composition of

streamflow for the chosen sample time. The Spearman rank correlation coefficient and the relative error were weighted in the combined objective function for calibration (see section 2.4). To find the optimal parameter sets, we used 100 independent calibration trials with different initial parameter values, which led to 100 calibrated parameter sets. We then selected the five parameter sets with the best calibration performance for use in the further analyses.

- *Step 2 Model validation:* The five parameter sets obtained from Step 1 were used to simulate the streamflow and stream isotopic composition for all six events. These simulations were compared to the observed stream stage and isotope data from each event to calculate the model performance based on the same objective function (i.e., the Spearman rank correlation coefficient for streamflow and the average relative error for the isotopic composition of streamflow; see Step 1 and see Section 2.4). The median validation performance for the five chosen parameter sets was used as the representative validation performance for the calibration with this particular stream isotope sample. This validation procedure was carried out for each of the six events individually ($E_{v=1-6}$; where $E_{v=1}$ means event E1 was chosen as the validation event), and for all six events together based on the average validation performance for the six events ($E_{v=all}$). Thus, there were three different validation scenarios: validation for the event that was also used for model calibration (*same-event*), validation for one of the five other events (*other-event*), and validation for all six events together (*all-event*; Figure 2). Steps 1 and 2 were repeated for each isotope sample for each calibration event (E_c).
- *Step 3 Model evaluation:* All samples during an event used for model calibration were ranked based on the validation performance for each of the seven validation options for an event (one same-event validation, five other-event validations, and one all-event validation). We then selected the three samples with the best validation performance (the *top-3 samples*) for each of the seven validation options (i.e., $E_{v=1-6}$ and $E_{v=all}$) for each calibration event (i.e., $E_{c=1-6}$)

and determined the timing of these samples relative to the hydrograph (pre-event, rising limb, near-peak, early falling limb and late falling limb; see section 2.5).

For the calibration with two samples ($n = 2$), we used the same workflow (Figure 2), except that we used each potential isotope sample pair during an event (i.e., the possible combinations of two samples) for calibration and used the average relative error of the simulated and measured isotopic composition of streamflow for the calibration objective function (instead of the relative error).

Note that for the same-event validation and all-event validation, the calibration data (including one or two streamwater samples) and the validation data (including all samples for the event or all events) are not completely independent. However, we still use the term validation for model testing purposes.

2.2 High-frequency isotope dataset

For model calibration, we used streamwater and precipitation isotope measurements from a small catchment (0.44 km^2) located near the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) in Birmensdorf, Switzerland (von Freyberg et al., 2017).

In brief, the catchment elevation ranges between 540 and 700 m above sea level (asl). Half of the catchment is covered with open grassland and grain fields, and the other half is a suburban residential area. The soil consists mainly of colluvial soil and partly brown soil. Mean annual precipitation for the area is around 1134 mm y^{-1} (measurement station Zürich/Fluntern, 1981-2010, Swiss Federal Office of Meteorology and Climatology (MeteoSwiss)).

The dataset includes 30-minute to hourly measurements of stable water isotopes in streamwater and precipitation, 10-minute data on stream stage, and daily precipitation sums from 13 February until 11 March 2016. Streamwater and precipitation isotopic composition were measured on-site using Picarro's Continuous Water Sampler module (CWS) coupled to a wavelength-scanned cavity ring-down spectrometer (model L2130-i; Picarro Inc., Santa Clara, CA, USA). If no precipitation occurred,

streamwater was analysed every 30 minutes. For every 30-minute interval that the precipitation collector was full, only precipitation samples were analysed, resulting in an approximately hourly sampling frequency for streamwater or precipitation during events (depending on the precipitation volume during each 30-minute interval). Stream stage was measured with a pressure transducer (model DL/N70; STS Sensor Technik Sirnach AG, Sirnach, Switzerland), which was corrected for changes in air pressure. For detailed information on data collection and study site characteristics, we refer to von Freyberg et al. (2017).

Hourly precipitation data were obtained from the CombiPrecip dataset from MeteoSwiss (Erdis et al., 2012; Sideris et al., 2014). This 1 km² gridded dataset combines gauge (SwissMetNet) and radar (MeteoSwiss) precipitation measurements. The measured daily precipitation at the field site and the daily sums of the CombiPrecip dataset were linearly correlated ($P_{\text{field}} = 0.8 * P_{\text{CombiPrecip}} + 0.55$ (all in mm); $R^2 = 0.86$; von Freyberg et al. (2017)). The hourly CombiPrecip data were used for this study because they allowed us to run the model at an hourly time step. However, they may include errors and bias in the amount and distribution of the areal average precipitation.

Six rainfall events (E) during snow-free conditions were selected from the isotope dataset for model calibration (Figure 1 and Table 1). Each event was defined to start at the beginning of the rainfall event and to end when the next rainfall event started or when the stream stage and isotopic composition of streamflow had returned to pre-event values. This selection resulted in event durations between 13h and 38h (Table 1). Rainfall amounts for the six events ranged from 5.5 to 13.3 mm. Events E1, E2, E3 and E5 caused a rapid stream stage response with a steep rising limb, whereas the stream stage responded more slowly during the larger E4 and E6 events. Event E4 was the longest event (38 h) and was characterised by two peaks and complex changes in the isotopic composition of precipitation and streamwater. In general, the streamwater isotopic composition lagged the response of stream stage (Figure 1). For the relatively short E5 event (13 h), no isotope data were available during peak flow. Event E5 was still used for this study because in many real

situations we may not have peak flow samples (as it is difficult to know when peak flow occurs and the duration of peak flow is very short).

All datasets, except the precipitation data, had to be converted to hourly time steps. The hourly stream stage data included the stream stage measurement at each hour. The precipitation isotope data were aggregated by taking the mean value of the precipitation isotope samples measured within each hour. Since these model input data were required at each time step, when no samples were available within the hour (e.g., the precipitation collector was not full because not enough rain fell), the isotopic composition of the neighbouring sample (mostly the later sample unless no samples were taken afterwards) was used. The streamwater isotope data were similarly aggregated by taking the mean value of the streamwater isotope samples measured within that hour. Unlike the precipitation data, the streamwater isotope data were not required for each time step (neither for calibration nor validation); thus values of NaN (i.e., Not a Number) were assigned to time steps for which no streamwater samples were available. This aggregation procedure resulted in 68 precipitation and 140 stream isotope data points for the studied events.

Unfortunately, no rating curve was available for the site, and thus model calibration was only possible against stream stage (i.e., stream level). However, it has previously been shown that in humid catchments the Spearman rank correlation coefficient can be effective for model calibration on stream stage observations (Seibert and Vis, 2016).

2.3 The Birkenes model

The Birkenes model is a lumped bucket-type coupled flow and tracer model. It was developed to simultaneously simulate streamflow and its isotopic composition (Christophersen et al., 1982; de Grosbois et al., 1988; Hooper et al., 1988) and has been applied to many regions of the world (Christophersen and Wright, 1981; Grip et al., 1985; Lundquist et al., 1990; Neal et al., 1988) for different modelling purposes. We chose the Birkenes model (in the version described by de Grosbois

et al. (1988) and Hooper et al. (1988) for this study because (1) it is a simple model with low data requirements (i.e., it only needs data for precipitation and its isotopic composition as input, and streamflow and its isotopic composition for calibration), and (2) its structure and mixing assumptions are similar to those of more recent coupled flow and tracer models, such as the SoftModel (Fenicia et al., 2008), the CIM model (Birkel et al., 2010) and the conceptual model used by Kirchner (2019, 2016).

The Birkenes model consists of two linear reservoirs (A and B) that represent a fast and a slow response. It has seven parameters that need to be calibrated: two flow rate parameters (AK , BK) that describe the relation between storage and flow from the fast and slow reservoir, respectively, three storage parameters ($AMIN$, $BMIN$, $BSIZE$) that define thresholds for flow generation (Q_A , Q_B and Q_{OVER}), one parameter ($AKSMX$) that describes the proportion of flow from the fast reservoir to the slow reservoir and a constant baseflow parameter ($QBASE$). Similar to many other coupled flow and tracer models, the model assumes complete mixing in both reservoirs. The two mixing-related parameters ($AMIN$ and $BMIN$) determine the mixing; the other parameters are flow-related parameters. Evapotranspiration from reservoir A (ET_A) was set to 0.03 mm h^{-1} based on the typical evaporation rates for the area in winter; it was assumed that there was no evaporation from reservoir B. For additional information on model structure and parameters, we refer to de Grosbois et al. (1988).

2.4 Model calibration

To evaluate the informative value of individual stream isotope samples for event-based model calibration, we assumed that stream stage data, plus isotope data for only one or two streamwater samples ($n = 1$ or 2) would be available for model calibration. We also assumed that the complete precipitation isotope dataset was available. We used a 60-week warm-up period before the simulation period with the hourly precipitation data extracted from the CombiPrecip dataset. The

isotopic composition of precipitation for the warm-up period was set to -11.04 ‰, based on the relatively constant isotopic composition of pre-event streamwater (-11.04 ± 0.21 ‰, von Freyberg et al., 2017). After the warm-up period, the hourly isotopic composition of precipitation (aggregated from the original dataset, see section 2.2) was used for the model simulations. We calibrated the model to the event stream stage data and the subset of stream isotopic data (either one or two samples per event) by minimizing the value of the combined objective function (F). This was done for each calibration event ($E_{c=1}$ to $E_{c=6}$) and every subset of isotopic data (i.e., every available sample and every possible sample pair). The combined objective function (F ; Eq. 1) consists of two parts: the goodness of fit for the simulation of the streamflow (F_Q) and for the simulation of the isotopic composition of streamwater (F_C):

$$F = \sqrt{\frac{F_Q^2 + F_C^2}{2}} \quad (\text{Eq. 1})$$

The Birkenes model predicts streamflow but only stream stage data were available for our study. Therefore, we used, as suggested by Seibert and Vis (2016), the Spearman rank correlation coefficient R_S (Spearman, 1904) between the simulated streamflow (Q_{sim}) and the measured stream stage as the basis for the objective function for the streamflow dynamics (F_Q):

$$F_Q = 1 - R_S \quad (\text{Eq. 2})$$

The objective function for the isotopic composition of the streamwater (F_C) was based on the average relative error:

$$F_C = \frac{1}{n} \sum \frac{|C_{Qobs}(i) - C_{Qsim}(i)|}{\text{Max}(C_{Qobs}) - \text{Min}(C_{Qobs})} = \frac{1}{n} \sum \frac{|C_{Qobs}(i) - C_{Qsim}(i)|}{2 \text{ ‰}} \quad (\text{Eq. 3})$$

where $C_{Qobs}(i)$ is the observed isotopic composition of streamwater at time i , $C_{Qsim}(i)$ is the simulated isotopic composition of streamwater at time i , and n is the number of streamwater samples used for calibration or validation. The range of the isotopic composition of streamwater was set to 2 ‰ for all events based on the average range in the isotopic composition of streamwater for the six events (1.8 ‰). We chose a constant value for the range in the isotopic composition because in reality, we

cannot predict the change in isotopic composition of streamflow for individual events beforehand, and because this allowed us to compare the validation performance for different events. The weights of the two parts of the combined objective function can affect the model calibration (Hooper et al., 1988). We, therefore, tested different weights for F_Q and F_C in the combined objective function (F) by adjusting the denominator in the calculation of F_C (i.e., using smaller and larger values for the range in the isotopic composition of streamwater; Eq. 3). The effect of the different weights on the model calibration results and on the derived most informative sampling times (based on the ranking of the model validation performance) was relatively small. Larger weight for F_C did not change the median value of F_Q (i.e., the streamflow was simulated well regardless of the weight for F_Q in the combined objective function), although the values of F_Q varied slightly more. We, therefore, chose an objective function that put a larger weight on the isotopic composition of streamwater than the streamflow to obtain good fits for both the isotopic composition of the streamwater and the streamflow (for more discussion on the effect of the weights of F_Q and F_C in F , see discussion section 4.4).

As in our previous studies (Wang et al., 2018, 2017), the SCE-UA method (Duan et al., 1992, 1993) was chosen for automatic parameter optimization because the genetic algorithm used in this method is considered robust and efficient (Francés et al., 2007; Yapo et al., 1996) and is available as an open source Matlab code. The algorithm results in one optimal parameter set for each initial selection of parameter values (seed). To account for the influence of the initial selection of the seed parameter values from which the optimization starts, 100 different independent seeds were used for each calibration, resulting in 100 optimised parameter sets for each calibration. The five parameter sets with the best calibration performance were selected for further analyses (*Calibration*, Figure 2). The initial parameter ranges were set to 0.1 to 10 times the default parameter values based on model calibration to field data from the Birkenes catchment in Norway (AMIN = 13 mm, BMIN = 40 mm, BSIZE = 40 mm, AK = $3.33 \cdot 10^{-2} \text{ h}^{-1}$, BK = $1.90 \cdot 10^{-3} \text{ h}^{-1}$, AKSMX = 0.75; de Grosbois et al., 1988), except for AK and BK, for which the minimum values were set to 0. QBASE is generally set to the minimum observed flow (de Grosbois et al., 1988). In this study, the range of QBASE was set between

0 mm h⁻¹ to 0.3 mm h⁻¹ because only stream stage was measured and the minimum streamflow was unknown.

Although the SCE-UA method is generally considered an efficient parameter optimization algorithm, we found that the initial selection of parameter values (seed) considerably influenced the calibration performance. To improve model calibration efficiency, we used the pragmatic approach and calibrated the model using the SCE-UA method with a large number of seeds (100), and afterwards selected the five best parameter sets based on the calibration performance.

2.5 Model validation and determination of the best sampling times

For each of the five selected parameter sets obtained from the calibration with the stream stage data and the isotope data from one sample (or sample pair) for an event, the model was validated using the combined objective function (F , Eq.1), the stream stage data, and the complete isotope dataset. The model performance was evaluated by the objective function for the validation period using the parameter values from the calibration period. There were three different validation scenarios (Figure 2) depending on which event was used for validation (E_v): (1) validation for the event to which the model was calibrated (same-event); (2) validation for each of the five other events (other-event); (3) the validation for all six events combined (all-event). For each validation scenario, the median value of the objective function from the five different parameter sets was defined as the representative validation performance, named as $F(E_v)$ (*Validation*, Figure 2). The overall validation performance for all six events ($E_{v=all}$) was based on the average value of $F(E_v)$ for all six events.

For each combination of calibration and validation events (i.e., each combination of E_c and E_v), all samples (or sample pairs) used for calibration were ranked based on the value of the objective function for the validation ($F(E_v)$), and the three samples (or sample pairs) with the best validation performance were selected as the top-3 samples (or top-3 sample pairs).

To investigate the timing of these samples, each event was divided into five sub-periods (I to V)

based on the relative stream stage change (i.e., the difference between the stream stage and the minimum stream stage, divided by the difference between the maximum and minimum stream stage during the event; Figure 3a).

- *Pre-event*: from the start of the event until the relative stream stage change reached 5%;
- *Rising limb*: when the relative stream stage change increased from 5% to 95%;
- *Near-peak*: when the relative stream stage change was larger than 95%;
- *Early falling limb*: when the relative stream stage change dropped from 95 to 50%;
- *Late falling limb*: when the relative stream stage change dropped below 50%.

Each event was also divided into three sub-periods based on the isotopic composition of streamwater (C_Q): (1) *pre-change*, (2) *change*, (3) *post-change* (Figure 3b). During the *pre-change* and *post-change* periods, the isotopic composition ($\delta^{18}O$) of the streamwater samples was close to the isotopic composition of baseflow (i.e., C_Q was within -11.04 ± 0.50 ‰; Figure 3b). Please note that -11.04 ± 0.50 ‰ is an arbitrary threshold based on the field observation that the isotopic composition of pre-event streamwater was relatively constant, averaging -11.04 ± 0.21 ‰, even though the isotopic composition of streamwater could change by up to 4.80 ‰ (E6) during the event (von Freyberg et al., 2017). While it is generally not possible to determine these sub-periods in real time, they are still useful to assess the informative value of data collected during different flow situations.

To determine the informative value of stream samples taken during each of these periods for model calibration, we calculated how often the top-3 samples occurred in each sub-period and then divided this number by the number of samples for each sub-period if they were selected randomly. Thus, a value larger than 1 means that a top-3 sample occurred more frequently during this period than expected based on the actual sample distribution (i.e., random chance). For example, for the calibration and validation for event E1 with one sample ($E_{c=1}$, same-event, $n = 1$), one of the three top-3 samples occurred during the early falling limb ($1/3 = 0.33$) and two occurred during the late falling limb ($2/3 = 0.67$), whereas three of the 18 samples were taken during the early falling limb

sub-period ($3/18=0.17$) and 13 were taken during the late falling limb ($13/18=0.72$), so that the relative frequencies with which early falling limb samples and late falling limb samples were chosen as the top-3 samples were 2.00 ($0.33/0.17$) and 0.92 ($0.67/0.72$) respectively. We repeated this calculation for each calibration event and the three validation scenarios (same-event, other-event and all-event) separately and then calculated the mean and standard error of the relative frequencies for all six calibration events.

2.6 Benchmarks

When field data are used for model calibration, it is difficult to evaluate the model performance based on the value of the objective function alone because there may be errors in both the data and the model structure that make it impossible to obtain a perfect model fit. Introducing benchmarks can facilitate comparisons of model performance (Seibert et al., 2018), and, therefore, we included four different benchmark calibrations in our analysis:

1. Event lower benchmark (*ELB*): the model was calibrated only with stream stage data from the calibration event E_c .
2. Event upper benchmark (*EUB*): the model was calibrated with stream stage data and all streamwater isotope data from the calibration event E_c .
3. Overall lower benchmark (*OLB*): the model was calibrated with stream stage data for all six events by optimizing the average value of the objective function for the six events.
4. Overall upper benchmark (*OUB*): the model was calibrated with stream stage data and all available streamwater isotope data for all six events by optimizing the average value of the objective function for the six events.

For each of the benchmark calibrations, 100 optimised parameter sets were derived, and the five parameter sets with the best calibration performance were used for the validation. The median of

the validation performances ($F(E_v)$) for these five parameter sets was defined as the benchmark validation performance, named ELB, EUB, OLB and OUB.

ELB and EUB were used to compare the performance of the model calibrated with subsets of stream isotope samples against the performance of the model calibrated with no (for ELB) or all (for EUB) isotope data for the same calibration event E_c . For instance, if the validation performance of the model calibrated with a subset of isotope samples from an event is better than the ELB, then this subset of isotope data has an added value for model calibration compared to calibration with stream stage data alone; if the validation performance of the model calibrated with this subset of data is close to EUB, then the value of this subset of isotope data for model calibration is as good (or bad) as that of the complete (hourly) isotope dataset used for calibration (and additional samples thus have little added value). Similarly, the OLB and OUB were used to compare the performance of the model calibrated with a subset of data from different events to the overall performance of the model calibrated with the data from all six events.

For the evaluation of the parameter values, the calibrated parameter values obtained by using a subset of samples for calibration were compared to the calibrated parameter values for the ELB, EUB, OLB and OUB as well.

3 Results

3.1 Improvement in overall model validation performance by including streamwater isotope data in model calibration

Figure 4 shows the validation performance for the models calibrated with one or two samples (i.e., $n = 1$, $n = 2$) for the six different calibration events (E_c), as well as the four benchmarks (i.e., ELB, EUB, OLB and OUB). The validation performance differed for the different calibration and validation event combinations, and was, unsurprisingly, generally best when the model was calibrated and validated

for the same event (same-event). The performance of the event lower benchmark (ELB) and overall lower benchmarks (OLB) was already relatively good, indicating that calibrating the model using only stream stage data resulted already in good model fits. The difference between the overall lower benchmark (OLB) and the overall upper benchmark (OUB) was small, while the differences between the event upper benchmark (EUB) and event lower benchmark (ELB) was generally larger. This indicates that for individual events the isotope data did improve the model performance, but when stage data were used for the calibration for all six events, the improvement in model performance for each event due to the isotope data was small (this is further elaborated on in discussion section 4.2).

In total, there were 42 event calibration and validation combinations (i.e., six calibration events ($E_{c=1-6}$) times seven different validations per calibration event (i.e., $E_{v=1-6}$ plus $E_{v=all}$)). The median validation performance ($F(E_v)$) for all calibrations with one sample (or one sample pair), which is considered to represent a situation where samples would be taken randomly (R), was better than the performance of the model calibrated with only the stream stage data (ELB) for only half of the cases (i.e., 20 and 21 out of 42 cases for $n = 1$ and $n = 2$ respectively; Table 3). However, when one of the top-3 samples or top-3 sample pairs was used for calibration, the model validation performance was better than for the calibration with only stream stage data (i.e., better than the ELB) for the majority of the cases (32 out of 42 cases = 76% for the calibration with one sample, and 35 out of 42 = 83% for calibration with two samples; Table 3). The performance of the model calibrated with one of the top-3 sample pairs was as good as (or even better than) the event upper benchmark (EUB; calibration with stream stage and all available isotope data for that event) for 39 out of 42 cases (= 93%), and was as good as or better than the overall upper benchmark (OUB; calibration with all stream stage and all available isotope data for all six events) for 21 out of the 42 cases (50%; Table 3).

3.2 Effect of sampling time on model validation performance

3.2.1 Effect of isotope data on the simulations of the streamwater isotopic composition

As shown in section 3.1, including isotope data in model calibration improved the model validation performance for some of the cases. As an example, Figure 5 shows the simulated hydrograph for event E1 using the parameter sets obtained from calibration with no isotope data ($n = 0$) and all available isotope data for this event ($n = 18$), as well as for the calibration with each of the top-3 samples and the worst sample ($n = 1$) and each of the top-3 sample pairs and worst sample pair ($n = 2$), together with the observed stream stage and isotopic composition of streamwater. With the data from one of the top-3 samples included in the calibration, the simulated isotopic composition of the streamwater (C_Q) improved considerably compared to calibration with no isotope data. The addition of the second sample from the top-3 sample pairs improved the simulations further (particularly the simulated C_Q during the late falling limb) and resulted in a performance close to that for the calibration with all available stream isotope data. There was a clear difference in the performance of the model calibration based on the top-3 samples (or sample pairs) and the worst sample (or sample pair), which means that the sampling time matters when only one or two samples are available for model calibration (the effects of the sampling time on model validation performance are discussed further in the following sections). These improvements in the simulated isotopic composition of the streamwater by adding only one or two top-3 samples were observed for all six same-event validation cases and some of the other-event validation cases (e.g., when the model was calibrated using data from event E1 and validated for event E6; $E_{c=1}$ & $E_{v=6}$). Therefore, the information contained in one or two selected streamwater isotope samples can improve the simulations of the streamwater isotopic composition and sometimes the simulations were as good as the simulations when all samples were used for calibration.

3.2.2 One sample ($n = 1$)

The model validation performance depended on when a sample was taken during an event (visualized by the colours of the circles for each subplot in Figure 6) and differed for the different validation events (i.e., the rank in model validation performance for a given sampling time differed for the different validation events, E_v).

- When the model was calibrated for event E1 ($E_{c=1}$) and validated for events E1 or E6 ($E_{v=1}$ or $E_{v=6}$), the early falling limb samples led to the best model performance and were, thus, considered most informative. However, when the model was validated for events E2, E3, E4 or all events, the pre-event samples led to better model performance. When the model was validated for event E5, both pre-event and early falling limb samples were most informative.
- When the model was calibrated for event E2 ($E_{c=2}$) and validated for events E2 or E3, the early falling limb samples led to the best model performance. However, when the model was validated for events E1 or E6, the pre-event samples resulted in the best model performance. The late falling limb samples were informative for all validation events.
- When the model was calibrated for event E3 ($E_{c=3}$), the early falling limb samples led to the best model performance, except for validation for event E1, for which the pre-event and near-peak samples led to better model validation performance.
- When the model was calibrated for event E4 ($E_{c=4}$), the early falling limb and late falling limb samples led to a better model performance than the other samples, except when the model was validated for event E4, for which the pre-event and rising limb samples led to better model performance.
- When the model was calibrated for event E5 ($E_{c=5}$), the late falling limb samples led to the best model performance when the model was validated for events E1 and E6, but the rising limb samples led to the best model performance for the other validation events. However, it should be noted that no samples were taken during the near-peak and early falling limb sub-periods for E5.

- When the model was calibrated for event E6 ($E_{c=6}$), the pre-event samples led to good model performance, except when the model was validated for event E6. The early falling limb samples also led to good model performance, except when the model was validated for event E2 for which the late falling limb led to better model performance.

In general, when the model was calibrated with the stream stage data and one stream isotope sample and validated for the same event (same-event), samples taken during the period that the stream water isotopic composition changed (change sub-period, Figure 3b) led to the best model performance. When the model was calibrated for one event and then validated for another event (other-event validation) or all six events (all-event validation), samples taken before the isotopic composition of the streamwater changed (pre-change sub-period; Figure 3b) led to better model performance.

The use of one sample for model calibration helped to improve the calibration of the mixing-related parameter AMIN. The calibrated values for parameter AMIN varied greatly between the 100 calibration trials (Figure 7). The spread of the calibrated parameter values was smaller (i.e., the colour of the bins is darker in Figure 7) when using one sample for model calibration ($n = 1$) compared to the event lower benchmark (ELB) and overall lower benchmarks (OLB). The parameter values obtained when using one of the top-3 samples was closer to the parameter value obtained for the event upper benchmark (EUB) than when using one of the other samples for calibration. For E5, the parameter identifiability was better for the later falling limb samples (i.e., more than half of the calibrated AMIN values in the lowest bin) but calibration with one of the top-3 samples led to a parameter distribution that was closer to that for the EUB. However, it should also be noted that the parameter distribution of AMIN for the six event upper benchmarks (EUB) is very different, indicating model related errors (see discussion in section 4.4).

3.2.3 Two samples ($n = 2$)

Similar to the results for the calibration with one sample, the model validation performance for the sample pairs varied and depended on which event was used for validation. The samples taken during the period that the isotopic composition of streamflow changed (change sub-period) led to better model performance when the model was calibrated and validated for the same event, while the pre-change samples led to the best model performance when the model was validated for another event or all six events.

3.2.4 Best sampling sub-period relative to the number of samples in the dataset

Figure 8 shows the relative frequency of the top-3 samples and top-3 sample pairs in a sub-period, compared to the relative sampling frequency during that sub-period for the three types of validation (same-event, other-event and all-event validation). We can see that for all three validation scenarios the early falling limb samples were more frequently selected as top-3 samples than expected based on random sampling (i.e., they have a relative frequency larger than 1) and the near-peak period and late falling limb period were less likely to be chosen as top-3 samples. For the all-event validation scenario, the relative frequency of the near-peak samples was larger than one, but this was caused by the results from only one of the six events (all top-3 samples and top-3 sample pairs during the near-peak sub-period were for calibration event E1; none of the other five calibration events had a top-3 sample or top-3 sample-pair during near-peak flow conditions). The large variation in the relative frequency of the top-3 samples during a certain period (i.e., large error bars in Figure 8) was caused by the differences in the timing of the top-3 samples or sample pairs (Figure 6), as well as the small number of top-3 samples and top-3 sample pairs and the small number of samples during the pre-event, rising limb and near-peak sub-periods in the dataset.

When the model was validated for the same event as for which it was calibrated (same-event), the early falling limb samples were more frequently selected as the top-3 samples than would be

expected from the sampling frequency during this sub-period. The pre-event samples were selected as the top-3 samples less frequently than expected based on random sampling (pre-event samples were selected as top-3 samples only for event E4, Figure 6). Even though the relative frequency of the top 3-samples during the rising limb period was larger than 1, nearly half of them occurred during the pre-change sub-period when the isotopic composition was similar to that of baseflow (3 out of 6 for $n = 1$ and 5 out of 9 for $n = 2$, Table 4). In addition, all of these pre-change sub-period samples were from calibration for event E5, for which there were no near-peak and early falling limb samples available (Figure 6).

When the model was validated for a different event than the one for which it was calibrated, or when it was validated for all events, the pre-event (or more precisely pre-change) sub-period samples were more frequently chosen as one of the top-3 samples, followed by the early falling limb samples. Similar to the same-event results, the majority (68-100%) of the top-3 samples during the rising limb period occurred during the pre-change sub-period (number in brackets, Table 4). These results suggest that for validation on a different event or all events, pre-change samples were more likely to be chosen as the top-3 samples than early falling limb samples and are thus considered more informative.

4 Discussion

4.1 Synthetic data vs. real data

The results from the present study with real-world data confirmed the previous results based on synthetic-data (Wang et al., 2018, 2017), in that (1) the model performance improved significantly when two strategically selected samples were used for calibration (Figure 4), and (2) falling limb samples were often the most informative samples when the model was calibrated and validated for the same event (Figures 6 & 8). For many rainfall events, the change in the isotopic composition was

largest 2.5 ± 1.0 h after peak flow (von Freyberg et al., 2017), which may explain why the near-peak samples (based on stream stage) were less informative than the early falling limb samples.

However, when the model was validated using data from another event than the one for which it was calibrated, or for all six events together, the pre-event and rising limb samples (particularly those taken during the pre-change sub-period when the isotopic composition of the streamwater was close to that of baseflow) were the most informative samples for model calibration. This indicates that pre-event isotope samples are valuable for model calibration. We attribute this to the clear anchoring effect by the baseflow samples for the other-event and all-event validations, such that when we get the isotopic composition of baseflow right, the model performs relatively well for the entire event.

In our previous synthetic data studies, all parameter values were well identified through model calibration. In the present study, we observed that the flow-related parameters (e.g., BK and QBASE) were well constrained and consistent for the different calibration and validation event combinations, while the parameters related to mixing processes (AMIN and BMIN) were less constrained and their optimum values were differed among the different calibration events. This is consistent with other studies using the Birkenes model for real-world situations, which have shown that the mixing-related parameters were poorly identifiable with real-world data, indicating that the Birkenes model is over-parameterized despite its simple structure (Hooper et al., 1988; Lundquist et al., 1990). For the same-event validation (for both synthetic and real-world data), inclusion of isotope data in the calibration always improved the model performance. However, for the other-event validation with real-world data, this was not always the case and the additional isotope data may, in fact, cause overfitting for that data point and a reduction in overall model performance (Figure 4 and Table 3). This overfitting likely explains why model performance was not improved by including data from a randomly selected isotope sample for nearly half of the calibration and validation event combinations, compared to the model calibrated with stream stage data only (ELB) (Table 3).

4.2 Is more data always better?

The overall lower and upper benchmarks (OLB and OUB) and event lower and upper benchmarks (ELB and EUB) were used to facilitate the comparison of the validation performance for the different model calibrations. When the model was calibrated and validated for the same event (same-event validation), the event upper benchmark (EUB) was always better than the event lower benchmark (ELB). For instance, when the model was calibrated and validated for E1 ($E_{v=1}$ & $E_{c=1}$), the EUB of E1 was better than the ELB of E1 (Figure 4). However, when the model was calibrated and validated for different events (other-event validation), the EUB was not always better than the ELB, suggesting that the inclusion of isotope information from one event may not improve the validation performance for another event (Figure 4). For example, when the model was validated for E1, the EUB of E4 was worse than the ELB of E4 ($E_{v=1}$ & $E_{c=4}$ in Figure 4). This might be explained by the overfitting to the data and differences in the dynamics of the isotopic composition for the individual events, which can result in a poor fit for certain events when the model is calibrated to another event, or because the Birkenes model is not able to simulate different events consistently due to limitations in the model structure and assumptions (e.g., the assumption of complete mixing). However, our example for validation event E1 ($E_{v=1}$) in Figure 4 shows that each of the top-3 samples or top-3 sample pairs from calibration event E4 ($E_{c=4}$) led to a better model performance than the event lower and upper benchmarks (ELB and EUB) for event E4 ($E_{v=1}$ & $E_{c=4}$ in Figure 4), indicating that the use of one or two selected informative samples can be more beneficial than using all available samples in the calibration. We also observed that although most of the samples during the calibration event (e.g., E1 and E6; Figure 4) did not improve the model performance for the validation for the other events (e.g., E2-5; Figure 4), the pre-event samples improved the validation performance for the other events considerably (Figure 6), suggesting that these samples are particularly valuable for model calibration.

The small difference between the overall lower benchmark (OLB) and the overall upper benchmark (OUB) indicates that calibration against stream stage data alone was almost as good as the

calibration against both stream stage and isotope data when data from multiple events were used. The parameter values were not constrained further by using all available stream isotope samples in model calibration. This suggests that (1) stream stage data from all multiple (in this case all six) events were relatively good for model calibration and (2) calibration on a diverse set of events with different dynamics in the isotopic composition of stream water diluted the informative value of each individual streamwater isotope sample for model calibration. The simplicity of the Birkenes model may also explain why more complex dynamics cannot be simulated.

4.3 Implications for ungauged catchments

Conventionally, catchment-scale hydrological models are calibrated and validated using the Nash–Sutcliffe model efficiency or another objective function that compares the simulated and observed streamflow volumes. The approach of combining stream stage data and the Spearman rank correlation as objective function was proposed and tested for more than 600 gauged basins in the USA by Seibert and Vis (2016). Their results indicate that stream stage data alone can provide good model simulation results in humid catchments. This approach was also used in the present study because only stream stage data were available. The use of stream stage data for model calibration has two advantages over the use of streamflow data: (1) the Spearman rank correlation puts more similar weight on different flow conditions and it (2) it helps to reduce the effect of observation uncertainties on model calibration, because stream stage observations have smaller uncertainties than stream discharge measurements. The disadvantage is that this calibration approach only focuses on ranking and therefore the absolute streamflow volume is not constrained.

Model-based studies that explore the value of limited data can provide guidance on hydrological data collection or experimental design. This is particularly helpful for data-scarce catchments or short-term projects with a small budget, where decisions need to be made on which measurements to take. For example, Sprenger et al. (2015) recommended that only one or two sampling campaigns

are necessary to collect soil isotopic samples to estimate flow and transport parameters in the unsaturated zone. Similarly McIntyre and Wheeler (2004) showed that decreasing the total number of in-river phosphorus samples in a two-month calibration period from 62 (daily) to four (event-based) samples led to only a slight decrease in model calibration performance, especially when there were errors in the data and model structure. Seibert and McDonnell (2015) found that ten streamflow observations during high flow provide almost as much information for model calibration as three months of continuously measured streamflow. Juston et al. (2009) similarly suggested that streamflow and groundwater measurements taken during a large event are more valuable for model calibration than fixed-interval measurements. Our results suggest that event-based model calibration for ungauged catchments will benefit from the installation of a water level logger at the catchment outlet and a few streamwater isotope samples collected before storm events and during the falling limb of the hydrograph (together with precipitation and precipitation isotope data). This approach is particularly efficient because it avoids the need to collect extensive stream stage and discharge measurements during a range of flow conditions to obtain a rating curve, as well as the logistically challenging collection of streamwater samples during the rising limb or peak flow of the hydrograph. If the most informative sampling times are known, field hydrologists can try to obtain samples during these preferred periods. For example, we can combine forecasts of precipitation with hydrological modelling to estimate the maximum and minimum stream stages of the target events, together with the real-time stream stage measurements, and collect streamwater samples during the chosen sub-periods manually or by automatic samplers.

4.4 Limitations of this study

To explore the value of a few streamwater isotope samples for model calibration, we used the average hourly isotopic composition of precipitation as input data, even though it is unlikely that hourly isotope data are available for ungauged study sites. Lower-frequency isotope data for precipitation (e.g., a bulk sample) will affect model performance (Birkel et al., 2010); longer time

series of rainfall and isotopic composition are crucial for model calibration (Sprenger et al., 2015). Our previous study has shown that errors in the precipitation data (which directly influence the average isotopic composition of precipitation for the catchment) can have a large effect on the model calibration and lead to early sampling times (rising limb) being more informative compared to the error-free cases (Wang et al., 2018).

The Birkenes model is a simple hydrological model. We assumed that the model is capable of simulating the catchment behaviour properly. However, the model might not be suitable for simulating all types of events. Even though the simulation of streamflow was relatively good and consistent, the calibrations for the different events led to significantly different values for the mixing-related parameters (e.g., AMIN, Figure 7). This suggests that there are indeed errors in the model assumptions and structure and that the model does not represent the catchment behaviour well during all events. It would be valuable to test a large selection of different events and models to explore the interactions between the data and model, and their effects on the informative content of streamwater isotope samples for model calibration.

The objective function F (Eq. 1) included two parts: F_Q and F_C . Previous studies using the Birkenes model have shown that the weights of the two objective functions affect model calibration (Hooper et al., 1988). Therefore, we tested different weights for F_Q and F_C by adjusting the denominator in the calculation of F_C (Eq. 3). For 19 of the 42 (=45%) event calibration and validation combinations (i.e., six calibration events ($E_{C=1-6}$) times seven different validations per calibration event (i.e., $E_{V=1-6}$ plus $E_{V=all}$)), the value of F was dominated by the value of F_C because the value of F_Q and the variations in F_Q were very small (i.e., the simulation of streamflow was relatively good based on the stream stage data alone and not affected by the inclusion of streamwater isotope data in the calibration). The value of F was dominated by F_Q for only five cases, due to the poor validation performance of streamflow for event E4. For the other 18 cases (43% of all cases), the combined objective function was influenced by both F_Q and F_C . Considering that the simulation of the streamflow dynamics was

already good based on the stream stage data (except for event E4), we selected an objective function that put more weight on F_c to also get a good simulation for the stream isotopic composition.

5 Conclusions

We tested the informative value of one or two streamwater isotope samples for event-based model calibration for a small study catchment. Our results show that there is value in using stream isotope data for model calibration (compared to calibration with stream stage data only) but that high-frequency isotope data for one or a few events may not necessarily significantly improve model calibration compared to a few strategically selected streamwater samples. The timing of the streamwater isotope samples used for model calibration affected the model performance. The inclusion of one or two randomly selected streamwater isotope samples in model calibration improved model validation performance, compared to calibration on stream stage data alone, for only half of the cases. However, when the most informative samples (i.e., one of the top-3 sample pairs) were used in model calibration, model performance improved compared to calibration on stream stage data alone for 83% of the cases. The samples taken during the early falling limb of the hydrograph were most informative for model calibration when the model was validated for the same event. However, when the model was validated for another event or all six events together, samples taken prior to the event or during the rising limb of the hydrograph were most informative. Samples taken near the flow peak were less informative for model calibration than other samples. These results are consistent with the results from previous studies using synthetic data, suggesting that the high informative value of the isotope samples taken on the early falling limb for model calibration might be real, and that studies with synthetic data can provide useful information and insights. Above all, this study demonstrates the value of hydrological modelling as a tool to inform sampling strategies. However, the results might be affected by the simplicity of the Birkenes model (e.g., a model with higher complicity might be able to simulate streamwater dynamics better, so that higher-

frequency data could improve model performance more). Therefore, similar studies using different models based on high-frequency isotope data, spanning several catchments and a larger range of events, are needed to ensure the generality of the results.

Acknowledgments

We thank Sergio Maffioletti and Marc Vis for IT support related to the use of the ScienceCloud at the University of Zurich, which enabled us to run the computational-intensive simulations on virtual machines. We also thank the three anonymous reviewers for their helpful comments. This work was funded by the Swiss National Science Foundation (Project-143995).

References

- Bergström, S., Lindström, G., Pettersson, A., 2002. Multi-variable parameter estimation to increase confidence in hydrological modelling. *Hydrol. Process.* 16, 413–421.
<https://doi.org/10.1002/hyp.332>
- Birkel, C., Dunn, S.M., Tetzlaff, D., Soulsby, C., 2010. Assessing the value of high-resolution isotope tracer data in the stepwise development of a lumped conceptual rainfall-runoff model. *Hydrol. Process.* 24, 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. *Hydrol. Process.* 29, 5227–5240.
<https://doi.org/10.1002/hyp.10594>
- Christophersen, N., Seip, H.M., Wright, R.F., 1982. A model for streamwater chemistry at Birkenes, Norway. *Water Resour. Res.* 18, 977–996. <https://doi.org/10.1029/WR018i004p00977>
- Christophersen, N., Wright, R.F., 1981. Sulfate budget and a model for sulfate concentrations in stream water at Birkenes, a Small forested catchment in southernmost Norway. *Water Resour. Res.* 17, 377–389. <https://doi.org/10.1029/WR017i002p00377>

744 Davies, J., Beven, K., Nyberg, L., Rodhe, A., 2011. A discrete particle representation of hillslope
 745 hydrology: hypothesis testing in reproducing a tracer experiment at Gårdsjön, Sweden. *Hydrol.*
 746 *Process.* 25, 3602–3612. <https://doi.org/10.1002/hyp.8085>

747 de Grosbois, E., Hooper, R.P., Christophersen, N., 1988. A multisignal automatic calibration
 748 methodology for hydrochemical models: A case study of the Birkenes Model. *Water Resour.*
 749 *Res.* 24, 1299–1307. <https://doi.org/10.1029/WR024i008p01299>

750 Duan, Q., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual
 751 rainfall-runoff models. *Water Resour. Res.* 28, 1015–1031. <https://doi.org/10.1029/91WR02985>

752 Duan, Q.Y., Gupta, V.K., Sorooshian, S., 1993. Shuffled complex evolution approach for effective and
 753 efficient global minimization. *J. Optim. Theory Appl.* 76, 501–521.
 754 <https://doi.org/10.1007/BF00939380>

755 Dunn, S.M., Bacon, J.R., 2008. Assessing the value of Cl – and $\delta^{18}\text{O}$ data in modelling the
 756 hydrological behaviour of a small upland catchment in northeast Scotland. *Hydrol. Res.* 39, 337.
 757 <https://doi.org/10.2166/nh.2008.134>

758 Erdin, R., Frei, C., Künsch, H.R., 2012. Data Transformation and Uncertainty in Geostatistical
 759 Combination of Radar and Rain Gauges. *J. Hydrometeorol.* 13, 1332–1346.
 760 <https://doi.org/10.1175/JHM-D-11-096.1>

761 Fenicia, F., McDonnell, J.J., Savenije, H.H.G., 2008. Learning from model improvement: On the
 762 contribution of complementary data to process understanding. *Water Resour. Res.* 44, 1–13.
 763 <https://doi.org/10.1029/2007WR006386>

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

767 Francés, F., Vélez, J.I., Vélez, J.J., 2007. Split-parameter structure for the automatic calibration of

768 distributed hydrological models. *J. Hydrol.* 332, 226–240.
 769 <https://doi.org/10.1016/j.jhydrol.2006.06.032>

770 Grip, H., Jansson, P.-E., Johnsson, H., Nilsson, S.I., 1985. Application of the “Birkenes” Model to Two
 771 Forested Catchments on the Swedish West Coast. *Ecol. Bull.* 37, 176–192.

772 Hartmann, A., Barberá, J.A., Andreo, B., 2017. On the value of water quality data and informative
 773 flow states in karst modelling. *Hydrol. Earth Syst. Sci.* 21, 5971–5985.
 774 <https://doi.org/10.5194/hess-21-5971-2017>

775 Hooper, R.P., Stone, A., Christophersen, N., de Grosbois, E., Seip, H.M., 1988. Assessing the Birkenes
 776 Model of stream acidification using a multisignal calibration methodology. *Water Resour. Res.*
 777 24, 1308–1316. <https://doi.org/10.1029/WR024i008p01308>

778 Hrachowitz, M., Fovet, O., Ruiz, L., Savenije, H.H.G., 2015. Transit time distributions, legacy
 779 contamination and variability in biogeochemical 1/f α scaling: how are hydrological response
 780 dynamics linked to water quality at the catchment scale? *Hydrol. Process.* 29, 5241–5256.
 781 <https://doi.org/10.1002/hyp.10546>

782 Juston, J., Seibert, J., Johansson, P.-O., 2009. Temporal sampling strategies and uncertainty in
 783 calibrating a conceptual hydrological model for a small boreal catchment. *Hydrol. Process.* 23,
 784 3093–3109. <https://doi.org/10.1002/hyp.7421>

785 Kendall, C., McDonnell, J.J., 1998. *Isotope Tracers in Catchment Hydrology*, 1st ed. Elsevier, New
 786 York.

787 Kirchner, J.W., 2019. Quantifying new water fractions and transit time distributions using ensemble
 788 hydrograph separation: theory and benchmark tests. *Hydrol. Earth Syst. Sci.* 23, 303–349.
 789 <https://doi.org/10.5194/hess-23-303-2019>

790 Kirchner, J.W., 2016. Aggregation in environmental systems – Part 2: Catchment
 791 mean transit times and young water fractions under hydrologic nonstationarity. *Hydrol. Earth*

792 Syst. Sci. 20, 299–328. <https://doi.org/10.5194/hess-20-299-2016>

793 Kirchner, J.W., 2006. Getting the right answers for the right reasons: Linking measurements,
 794 analyses, and models to advance the science of hydrology. *Water Resour. Res.* 42, 1–5.
 795 <https://doi.org/10.1029/2005WR004362>

796 Kirchner, J.W., Hooper, R.P., Kendall, C., Neal, C., Leavesley, G., 1996. Testing and validating
 797 environmental models. *Sci. Total Environ.* 183, 33–47. [https://doi.org/10.1016/0048-](https://doi.org/10.1016/0048-9697(95)04971-1)
 798 [9697\(95\)04971-1](https://doi.org/10.1016/0048-9697(95)04971-1)

799 Lundquist, D., Christophersen, N., Neal, C., 1990. Towards developing a new short-term model for
 800 the Birkenes catchment—lessons learned. *J. Hydrol.* 116, 391–401.

801 McGuire, K.J., Weiler, M., McDonnell, J.J., 2007. Integrating tracer experiments with modeling to
 802 assess runoff processes and water transit times. *Adv. Water Resour.* 30, 824–837.
 803 <https://doi.org/10.1016/j.advwatres.2006.07.004>

804 McIntyre, N.R., Wheeler, H.S., 2004. Calibration of an in-river phosphorus model: prior evaluation of
 805 data needs and model uncertainty. *J. Hydrol.* 290, 100–116.
 806 <https://doi.org/10.1016/j.jhydrol.2003.12.003>

807 Neal, C., Christophersen, N., Neale, R., Smith, C.J., Whitehead, P.G., Reynolds, B., 1988. Chloride in
 808 precipitation and streamwater for the upland catchment of river severn, mid-wales; some
 809 consequences for hydrochemical models. *Hydrol. Process.* 2, 155–165.
 810 <https://doi.org/10.1002/hyp.3360020206>

811 Nyberg, L., Rodhe, A., Bishop, K., 1999. Water transit times and flow paths from two line injections of
 812 ³H and ³⁶Cl in a microcatchment at Gårdsjön, Sweden. *Hydrol. Process.* 13, 1557–1575.
 813 [https://doi.org/10.1002/\(SICI\)1099-1085\(19990815\)13:11<1557::AID-HYP835>3.0.CO;2-S](https://doi.org/10.1002/(SICI)1099-1085(19990815)13:11<1557::AID-HYP835>3.0.CO;2-S)

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

816 Seibert, J., McDonnell, J.J., 2015. Gauging the Ungauged Basin: Relative Value of Soft and Hard Data.
817 J. Hydrol. Eng. 20, A4014004. [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0000861](https://doi.org/10.1061/(ASCE)HE.1943-5584.0000861)

818 Seibert, J., McDonnell, J.J., 2002. On the dialog between experimentalist and modeler in catchment
819 hydrology: Use of soft data for multicriteria model calibration. Water Resour. Res. 38, 23-1-23–
820 14. <https://doi.org/10.1029/2001WR000978>

821 Seibert, J., Vis, M.J.P., 2016. How informative are stream level observations in different geographic
822 regions? Hydrol. Process. 30, 2498–2508. <https://doi.org/10.1002/hyp.10887>

823 Seibert, J., Vis, M.J.P., Lewis, E., van Meerveld, H.J., 2018. Upper and lower benchmarks in
824 hydrological modelling. Hydrol. Process. 32, 1120–1125. <https://doi.org/10.1002/hyp.11476>

825 Sideris, I. V., Gabella, M., Erdin, R., Germann, U., 2014. Real-time radar-rain-gauge merging using
826 spatio-temporal co-kriging with external drift in the alpine terrain of Switzerland. Q. J. R.
827 Meteorol. Soc. 140, 1097–1111. <https://doi.org/10.1002/qj.2188>

828 Spearman, C., 1904. The Proof and Measurement of Association between Two Things. Am. J. Psychol.
829 15, 72. <https://doi.org/10.2307/1412159>

830 Sprenger, M., Volkmann, T.H.M., Blume, T., Weiler, M., 2015. Estimating flow and transport
831 parameters in the unsaturated zone with pore water stable isotopes. Hydrol. Earth Syst. Sci. 19,
832 2617–2635. <https://doi.org/10.5194/hess-19-2617-2015>

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

836 Vaché, K.B., McDonnell, J.J., 2006. A process-based rejectionist framework for evaluating catchment
837 runoff model structure. Water Resour. Res. 42. <https://doi.org/10.1029/2005WR004247>

838 Vaché, K.B., McDonnell, J.J., Bolte, J., 2004. On the use of multiple criteria for a posteriori model
839 rejection: Soft data to characterize model performance. Geophys. Res. Lett. 31, n/a-n/a.

840 <https://doi.org/10.1029/2004GL021577>

841 von Freyberg, J., Studer, B., Kirchner, J.W., 2017. A lab in the field: high-frequency analysis of water
842 quality and stable isotopes in stream water and precipitation. *Hydrol. Earth Syst. Sci.* 21, 1721–
843 1739. <https://doi.org/10.5194/hess-21-1721-2017>

844 Wagener, T., 2003. Evaluation of catchment models. *Hydrol. Process.* 17, 3375–3378.
845 <https://doi.org/10.1002/hyp.5158>

846 Wang, L., van Meerveld, H., Seibert, J., 2018. Effect of Observation Errors on the Timing of the Most
847 Informative Isotope Samples for Event-Based Model Calibration. *Hydrology* 5, 4.
848 <https://doi.org/10.3390/hydrology5010004>

849 Wang, L., van Meerveld, H.J., Seibert, J., 2017. When should stream water be sampled to be most
850 informative for event-based, multi-criteria model calibration? *Hydrol. Res.* 48, 1566–1584.
851 <https://doi.org/10.2166/nh.2017.197>

852 Yapo, P.O., Gupta, H.V., Sorooshian, S., 1996. Automatic calibration of conceptual rainfall-runoff
853 models: sensitivity to calibration data. *J. Hydrol.* 181, 23–48. [https://doi.org/10.1016/0022-](https://doi.org/10.1016/0022-1694(95)02918-4)
854 1694(95)02918-4

855

857 *Table 1: Characteristics of the six events used for model calibration and validation.*

Event	Stage at start of the event (mm)	Maximum stream stage (mm)	Total precip. (mm)	Precip. duration (h)	Maximum precip. intensity (mm/h)	12 h antecedent precip. (mm)	Event duration (h)	No. of isotope data points (-)
E1	442	482	6	8	2	0	18	18
E2	446	475	9	11	2	3	32	30
E3	357	461	6	8	2	1	23	22
E4	446	466	13	18	2	2	38	34
E5	456	492	8	10	4	2	13	9
E6	451	481	11	19	1	1	31	27

858

859

Table 2: Total number of model calibrations for each calibration event and the overall benchmarks.

The event lower benchmark (ELB) describes the calibration based on only stream stage data for that event ($n = 0$). The event upper benchmark (EUB) describes the calibration with all available stream stage and isotope data for that event (n varies between 9-30 depending on the event; see Table 1).

The overall lower benchmark (OLB) describes the calibration with the stream stage data for all six events but no isotope data ($n = 0$). The overall upper benchmark (OUB) is the calibration with all available stream stage and streamwater isotope data for the six events ($n = 140$).

Calibration event (E_c)	Event lower benchmark (ELB; $n = 0$)	Event upper benchmark (EUB)		Sum
		$n = 1$	$n = 2$	
$E_{c=1}$	1	18	153	173
$E_{c=2}$	1	30	435	467
$E_{c=3}$	1	22	231	255
$E_{c=4}$	1	34	561	597
$E_{c=5}$	1	9	36	47
$E_{c=6}$	1	27	351	380
$E_{c=all} OLB$				1
$E_{c=all} OUB$				1
Total				1921

869 *Table 3: Positive ('+') or negative ('-') change in model validation performance ($F(E_v)$) when the model*
870 *was calibrated with one or two random isotope samples ('R'), or one of the top-3 samples or top-3*
871 *sample pairs ('T3'), compared to the performance for the calibration with stream stage data only*
872 *(ELB; i.e. total number of plusses) . Also shown are the number of combinations of E_c and E_v for which*
873 *the model validation performance was the same as or better than the event upper benchmark (EUB;*
874 *calibration using all stream stage and streamwater isotope data available for that event) and the*
875 *overall upper benchmark (OUB; calibration using all stream stage and streamwater isotope data*
876 *available for all six events).*

	$E_{c=1}$				$E_{c=2}$				$E_{c=3}$				$E_{c=4}$				$E_{c=5}$				$E_{c=6}$			
	n = 1		n = 2		n = 1		n = 2		n = 1		n = 2		n = 1		n = 2		n = 1		n = 2		n = 1		n = 2	
	R	T3	R	T3	R	T3	R	T3	R	T3	R	T3	R	T3	R	T3	R	T3	R	T3	R	T3	R	T3
$E_{v=1}$	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	-	+	-	+	+	+	+	+
$E_{v=2}$	-	-	-	-	+	+	+	+	-	-	-	+	-	+	-	+	-	+	-	+	-	-	-	-
$E_{v=3}$	-	-	-	-	+	+	+	+	+	+	+	+	-	+	-	+	-	+	-	+	-	-	-	-
$E_{v=4}$	-	-	-	-	+	+	+	+	-	+	-	+	+	+	+	+	+	+	+	+	-	-	-	+
$E_{v=5}$	-	-	-	-	+	+	+	+	-	+	-	+	+	+	+	+	+	+	+	+	-	+	+	+
$E_{v=6}$	+	+	+	+	+	+	+	+	-	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+
$E_{v=all}$	-	-	-	-	+	+	+	+	-	+	-	+	-	+	-	+	+	+	+	+	-	-	-	+
$\geq$ ELB (No. of +)	2	2	2	2	7	7	7	7	2	6	2	7	3	7	3	7	4	7	4	7	2	3	3	5
$\geq$ EUB	3	6	5	6	2	6	2	7	2	6	1	6	1	7	1	7	2	4	2	7	1	5	1	6
$\geq$ OUB	2	2	2	3	3	3	3	6	2	2	2	3	1	2	1	4	2	3	2	3	2	2	2	2

877

878

879 *Table 4: Number of top-3 samples (n = 1) or top-3 sample pairs (n = 2) that belong to each sub-period*
880 *based on the relative change in stream stage for the three validation scenarios (same-event, SE;*
881 *other-event, OE; and all-event, AE). The numbers in brackets show the number of samples taken*
882 *during both the rising limb (based on stream stage) and the pre-change period (based on C_0) when*
883 *the isotopic composition was close to that of baseflow. See Figure 3 for the definition of the different*
884 *sub-periods.*

Sub-period	No. of samples in dataset	n = 1			n = 2		
		SE	OE	AE	SE	OE	AE
Pre-event	15	1	10	2	2	41	6
Rising limb (pre-change)	20 (7)	6(3)	14(12)	4(4)	9(5)	22(15)	7(6)
Near-peak	9	1	4	1	1	9	2
Early falling limb	40	7	31	7	10	54	13
Late falling limb	56	3	31	4	14	54	8
Sum	140	18	90	18	36	180	36

885

886

Figure 1: Time series of precipitation (P), stream stage, and the aggregated isotopic composition ($\delta^{18}\text{O}$) of precipitation (C_P) and streamflow (C_Q) for the six events used for the model calibration and validation. Note that these are the values after aggregation and gap-filling with neighbouring samples for isotopic composition ($\delta^{18}\text{O}$) of precipitation (C_P) or assignment of values of NaN (i.e., Not a Number) to time steps for which no streamwater samples were available; they therefore differ from the original single-sample isotope values of von Freyberg et al. (2017).

Figure 2: Workflow for model calibration and validation with one sample ($n = 1$). The three coloured boxes highlight the three different validation scenarios (same-event, other-event and all-event validation) based on the combination of calibration event (E_c) and validation event (E_v). The value of the model validation objective function was calculated five times (for the five parameter sets), marked as X 5.

Figure 3: Separation of the event into sub-periods (a) based on the relative change in stream stage (I: pre-event; II: rising limb; III: near-peak; IV early falling limb; V: late falling limb) and (b) based on the isotopic composition of streamwater (1: pre-change; 2: change; 3: post-change).

905

906

907

908

909

Figure 4: Validation performance ($F(E_v)$) for all subsets of sample combinations and the benchmarks. Note the inverse y-axis implying that results are better towards the top; also the scale is different for each validation event (E_v) to better show the model performance differences for each calibration event (E_c , marked with different colours). For each E_c , the results shown from left to right are: ELB ($n = 0$), $n = 1$, $n = 2$, EUB ($n =$ all available samples for the event). The same-event validation, $E_c = E_v$, is indicated with bold font in the x-axis text.

910

911 Figure 5: Simulated hydrographs for event E1 when calibrated with a different number of samples ($n = 0, 1, 2, 18$). The circles mark the times of the top-3
 912 samples (large circle: top 1, middle circle: top 2, small circle: top 3) and the triangles mark the worst sample used for calibration based on the validation
 913 performance for E1 ($E_{C=1}$, $E_{V=1}$).

Figure 6: The observed stream stage (black line) and stream isotopic composition (circles) with the rank of the stream isotope samples based on the validation performance in colour for each calibration event (E_c ; columns) and validation event (E_v ; rows) combination. The Top-3 samples for each calibration and validation combination are indicated with red triangles.

918

919

920

921

922

Figure 7: Distribution of 100 calibrated values for parameter AMIN. For each event, the model was calibrated using event stream stage and the isotope data from one sample ($n=1$), no sample (ELB), and all event samples (ELB), as well as with stream stage data for all six events without and with all 140 samples (OLB and OUB, respectively). The x-axis shows the time for each event in hours; the y-axis shows the range of the values for the parameter AMIN, which is divided to 20 bins. A darker colour in each bin indicates a higher identifiability of AMIN (i.e., more of the seeds result in the same parameter value). For a uniform

923 *distribution the number in each bin would be 5, therefore values ≤ 5 are shown in white. Please note that all 100 parameter sets were considered here, while*
924 *only the five parameter sets with the best calibration performance were selected for the further analyses. Calibration with one of the top-3 samples highlighted*
925 *in red led to the best same-event model validation performance.*

Figure 8: The relative frequency of the top-3 samples and top-3 sample pairs in each sub-period (a: $n = 1$; b: $n = 2$) for the three different validation scenarios: validation for the same event (same-event in orange); validation for one of the five other events (other-event in blue), and validation for all six events (all-event in gray). The error bar shows the standard error, which was calculated from the relative frequency of the top-3 samples for the six individual calibration events. A relative frequency larger than 1 indicates that more of the top-3 samples or sample pairs occurred during this period than would be expected based on random chance. See figure 3 for the description of the different sub-periods. Note that the number of possible validation options varied for the different validation scenarios and that the error bars are thus not directly comparable.