Journal of Hydrology 621 (2023) 129604

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Guidance on large scale hydrologic model calibration with isotope tracers

Tegan L. Holmes [a,b,*], Tricia A. Stadnyk [b,a], Masoud Asadzadeh [a], John J. Gibson [c,d]

[a] University of Manitoba, Civil Engineering, Winnipeg, MB R3T 5V6, Canada
[b] University of Calgary, Geography, Calgary AB T2N 1N4, Canada
[c] InnoTech Alberta, 3-4476 Markham Street, Victoria BC V8Z 7X8, Canada
[d] University of Victoria, Geography, Victoria BC V8W 2Y2, Canada

A B S T R A C T

Standard hydrologic model evaluation and calibration approaches focus on the accurate simulation of streamflow, disregarding internal process simulations. Stable isotope tracers can provide additional information on water sources, and process flux and storage, which can be used to inform model calibration. This study assesses the added value of isotope data in comparison to current best-practice flow-only calibration methods and evaluates the merits and limitations of isotope simulation performance metrics for the purposes of hydrological model calibration. Following several years of regular isotope sampling and measurement, an isotope-enabled process-based hydrologic model was tested on a large watershed in western Canada (Athabasca River), which allowed model calibration using global sensitivity analyses, Monte Carlo simulations, and multi-objective optimizations. Isotope tracer data were found to improve both process and streamflow component identifiability and produced some minor improvement in individual parameter value identifiability. Calibrating to optimize both flow and isotope simulation performance produced better flow simulation ensembles, with improved observation capture and validation performance, relative to calibrating to optimize flow simulations alone. Using an isotope simulation performance metric which includes timing error as a secondary optimization objective led to more robust streamflow modeling, even in mesoscale watersheds with limited isotope observation datasets.

1. Introduction

Hydrologic models are essential tools for hydrologists, used to predict runoff or streamflow in both short-term forecasting and long-term climate change projection applications. However, the purpose or application of hydrologic models vary, from overall basin water balance estimates, to predicting flood volumes, or for the investigation of flow generation processes, such as baseflow in regional groundwater impact studies, among other applications (Clark et al., 2017). Likewise, hydrologic models vary in structure and complexity, from fully physically-based models to conceptual models, and in scale, from hillslope to global simulation of water fluxes (Guse et al., 2021; Refsgaard et al., 2022). To reliably predict streamflow under climatic conditions different than those in limited observation records or estimate flows in ungauged locations, hydrologic models must accurately represent the basic physical processes most influential in generating streamflow at a specific scale of application (Duethmann et al., 2020).

One of the major hurdles facing large-scale process-based models is the limited quantity of information available to define or inform runoff generation processes and water movement (Fatichi et al., 2016; Kirchner, 2006; Stevenson et al., 2021). Only in the most intensively monitored research sites, i.e., in small scale catchments, are process variables and flowpaths, e.g., transpiration loss or wetland flux rate, defined at the daily timescale. Streamflow records, with potentially some low-frequency water tracer or point process observations, are often all that can be hoped for in model training at the *meso*-scale (Coulibaly et al., 2013; Gibson et al., 2020). This lack of data poses a major challenge for the reliability of process-based models: how can the modeler be assured of accuracy at the process level when the only available observations are the final, cumulative, streamflow? Can any confidence be placed in the model when innumerable combinations of hydrologic process simulations add up to the same total streamflow, and only the total streamflow is verified by the modeler?

Part of the solution is increasing the information available to assess model accuracy, both at the internal process level (i.e., hydrologic compartments) and summative (i.e., streamflow) (Kirchner, 2006). Previous research has shown that stable isotope tracer data (i.e., ratios of water molecules containing ^{18}O or ^{2}H to standard water) can provide additional information for the evaluation of hydrologic models (Ala-aho et al., 2018, 2017; He et al., 2019; Holmes et al., 2020; Stadnyk and Holmes, 2020; Tunaley et al., 2017). Stable isotope tracers are particularly useful in remote or inaccessible watersheds as they are non-reactive and naturally occurring (reducing field work requirements), while their variable concentration in precipitation and evaporating

* Corresponding author at: University of Calgary, 2500 University Dr NW, Department of Geography, Calgary, AB T2N 1N4, Canada.
E-mail address: tegan.holmes@ucalgary.ca (T.L. Holmes).

https://doi.org/10.1016/j.jhydrol.2023.129604

Available online 29 April 2023

water bodies can provide additional information on influential water sources and hydrological processes (Brooks et al., 2018; Oshun et al., 2016; Peralta-Tapia et al., 2015).

Observational data, whether flow or tracer observations, can feed into the model development through parameter calibration. In process-based models, which aim to emulate real-world hydrologic storage and mass flux, some parameter values can be estimated using field measurements or remotely sensed data; however, there are generally other parameters (with limited physical basis) for which good or reasonable values are unmeasurable or unknown (e.g., regional average soil conductivity) (Acero Triana et al., 2019; Fatichi et al., 2016). These unknown parameter values necessitate calibration, where parameter values are selected or adjusted to achieve an acceptable model performance. There are two approaches commonly used in the literature: trying vast numbers of parameter value combinations (generated either at random or using a sampled distribution) and using those with the best model performance or using a search algorithm to locate optimal parameter value combinations (Beven and Binley, 1992; Efstratiadis and Koutsoyiannis, 2010; Pechlivanidis et al., 2011). There are benefits to both approaches: generating large numbers of independent solutions allows for much more complex statistical analyses and can ensure coverage of the entire parameter space, while search algorithms are much better at identifying good quality solutions (i.e., solutions which match observed data sufficiently), at a lower computational cost (Acero Triana et al., 2019; Pechlivanidis et al., 2011).

Regardless of the approach chosen, the calibration depends on quantifying the quality of the simulated model output using some sort of performance metric or set of metrics (Bennett et al., 2013). The quantified model performance can be used to identify 'best' solutions, set an acceptable model performance threshold, or be used by a search algorithm as an objective. Performance metrics are not limited to simulated streamflow. For example, if the hydrologic model is capable of simulating both flow and isotope tracer composition, both data types can be used to quantify the model performance quality (He et al., 2019; Nan et al., 2021; Stadnyk and Holmes, 2020; Tunaley et al., 2017). However, to date, there are no universal guidelines on the best performance metrics for isotope tracer-aided model calibration and metric selection has largely been ad-hoc or 'best guess' in the tracer-aided calibration literature (Holmes et al., 2020). Tracers in large-scale watersheds are generally irregularly or infrequently sampled which can influence metric selection; different metrics vary in their sensitivity to different properties of the data timeseries (such as variability or maxima); therefore, by using a mixture of multiple metrics (to evaluate different aspects of the flow data, or to simultaneously evaluate both flow and tracer data), model evaluation or calibration can be rendered more reliable or comprehensive (Bennett et al., 2013; Mizukami et al., 2019).

For the internal process representation, the precision of a simulation depends only on the identifiability of the parameter values (provided the model structure and equations are static)(Guse et al., 2021). If the calibration process results in a narrow range of values for a given parameter, that parameter is well-identified, and if all parameters influencing a modeled process are well-identified, the range of flow contributions from the process will be narrow. However, identifiable parameters do not necessarily lead to parameter or process accuracy, they may still not be representative of the real-world system the model is intended to emulate (Guse et al., 2020). Due to issues such as model structure errors or forcing data uncertainty, parameters which consistently produce optimal performance metric values may also be consistently wrong (e.g., over-estimating evaporation to compensate for biased precipitation inputs or not including sublimation in the model).

Adding tracer data to the calibration process can improve the accuracy of the internal process representation, particularly for soil water storages, and improve the identifiability of some parameters (He et al., 2019; Holmes et al., 2020). However, parameter identifiability may also decrease when calibrating with isotope tracer data, as parameter values producing optimal flow simulation results can be contradicted by the second calibration target, the tracer simulation (Holmes et al., 2020). Effects on flow simulation accuracy and precision are likewise unclear: including isotope data in the calibration can decrease flow simulation performance during the calibration period, and the flow simulation uncertainty may increase. On the other hand, flow simulation performance during the validation period may improve when tracer data are included in calibration. The value added by isotope tracer data to calibration remains elusive and ill-defined, as previous studies have either used only mixed flow and isotope data, used different calibration methods for flow-only and mixed flow and isotope calibration (limiting comparability) or used a single isotope tracer performance metric with a fixed weight in calibration (limiting generalizability).

This study aims to address two remaining gaps in the isotope-aided modeling literature at the large-scale: quantifying the added value of isotope data relative to current best-practice flow-only calibration methods, and a comparison of the merits and limitations of different isotope simulation performance metrics for the purposes of model calibration. In particular, we address the following questions:

- Does use of stable isotope tracer data in calibration alter parameter identifiability, and if so, is this change an improvement?
- Does use of stable isotope tracer data in calibration lead to different flow simulation results, and if so, is this change an improvement?
- Can a tracer performance metric be recommended to maximize the benefits of simulating isotope tracers in model calibrations?

Our study is conducted in the Oil Sands region of Alberta, within the Athabasca River basin. This region was chosen because there exists one of the longest large river isotope sampling programs in Canada, the watershed contains several hydrometric gauges for calibration of both tributaries and mainstem, and because it is a high-latitude drainage system representative of cold regions seasonal hydrology (permitting an evaluation of snow and glacier melt).

2. Methods

2.1. Hydrologic model and study area

2.1.1. The CHARM/isoWATFLOOD model

Assessing parameter identifiability and change in flow simulation performance resulting from the inclusion of stable isotope tracers in model calibration requires an isotope-enabled hydrologic model; this study uses the CHARM/WATFLOOD hydrologic model and isoWATFLOOD, the isotope simulation module. Both CHARM and isoWATFLOOD are open-source models that use a combination of conceptual and physics-based hydrologic process representations for relatively computational efficient distributed modeling of meso- and large-scale watersheds (Kouwen, 2018). The isoWATFLOOD dual-isotope tracer model simulates oxygen-18 and deuterium compositions for all water storages and fluxes in the base CHARM model, with the assumptions that all non-evaporative fluxes have the same concentration as the originating simulated storage and that hydrologic storages are completely mixed through depth (Holmes, 2016; Stadnyk et al., 2013; Stadnyk and Holmes, 2020). The linked hydrologic and isotope tracer simulations both output daily results by default, but the internal simulation runs at an hourly time-step.

The CHARM/WATFLOOD model structure divides the watershed domain into grid cells with defined drainage directions, and then subdivides each cell into grouped response units or GRU (with the area of each GRU generally determined from landcover data), as shown in Fig. 1. The default structure of a GRU (used for most types of landcover, such as forests, grass or shrubland) has a vertical soil column divided into two soil layers, the upper and lower soil zones, both of which contribute lateral flow to the channel network. Surface water from snowmelt or rain can either infiltrate vertically or runoff directly to the channel network. Evapotranspiration is only modeled for wetlands and

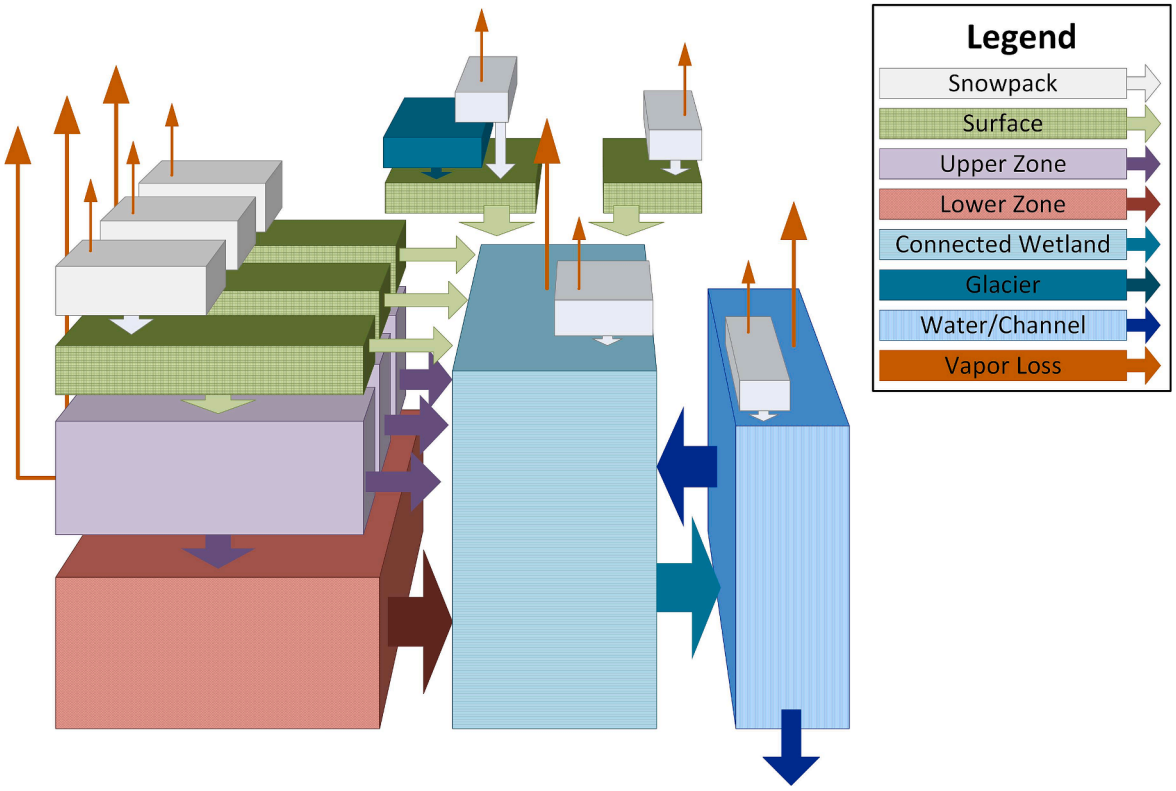

Fig. 1. Schematic of grouped response units (GRU) representing water storage (blocks) and fluxes between storages (arrows) within a single headwater grid cell, with glaciers, connected wetlands and three soil-based GRU.

water in upper soil storages; the evaporation component is fractionating, and the isotopic composition of the evaporate is modeled using equations from Gibson et al. (2016a, 2008). Snowpacks melt and accumulate independently for each GRU and soils freeze and thaw based on air temperature, with reduced conductivity in frozen soil (permafrost is not modeled). There are four special GRU classes without soil layers: glacier, impervious, connected wetland and water or channel. In the glacier and impervious classes all rain and snowmelt becomes direct runoff, and glacier GRU also generate glacier melt flows. Connected wetlands (fens or riparian zones), where present, accumulate all lateral outflows from the land and soil, and have a bi-directional connectivity with stream channels. Wetland areas disconnected from the channel network are treated as a default GRU. The water or channel GRU is the end point for all flows generated within the grid cell, and it is the only GRU connected to other cells, with inflows from upstream cells and an outflow downstream. More detailed descriptions of process representation and equations can be found in Holmes, 2016 and Holmes et al. (2022).

All modeled processes have parameters controlling the simulated flux and storage, with the values of these parameters being either consistent across all GRU, or separate GRU having individual parameter values representing landcover or soil specific hydrologic response. There is a very large number of parameters (over 300 for the Athabasca basin model) which can be set for a CHARM/isoWATFLOOD model, but only a minority truly benefit from calibration. The values of most parameters are best determined from previously published studies, and simulations are largely insensitive to their value. This study will calibrate only the most sensitive, highest priority parameters, based on model developer recommendations and previous isoWATFLOOD calibration studies, including Holmes et al. (2020) and Stadnyk and Holmes (2020), listed in Table 1.

Table 1
Calibrated parameters, listing the process affected by the parameter, the parameter names, and the GRU types affected by the parameter value. Parameters with unique values for different GRU classes are italicized.

Parameter	Process	Parameter name	Internal name	Applicable Grouped Response Units
Surface soil conductivity	Infiltration	k F (surf)	ak	All soil-based
Horizontal upper soil zone conductivity	Interflow	k F (horz)	rec	All soil-based
Open water PET to AET factor	Evaporation	PET F	fpet	Water, connected wetland
Snowmelt rate factor	*Snowmelt*	*melt rate*	*fm*	*All*
Upper soil zone soil water retention cap	*Soil storage and ET*	*soil ret*	*retn*	*All soil-based*
Vertical upper soil zone conductivity	Recharge	k F (vert)	ak2	All soil-based
Baseflow equation constant	*Baseflow*	*C*	*flz*	*All soil-based*
Baseflow equation power	Baseflow	pwr	pwr	All soil-based
Channel roughness factor	*Channel velocity*	*n*	*r2n*	*Water*
Wetland porosity	Wetland storage	θ (wet)	theta	Connected wetland
Wetland conductivity	Wetland velocity	k (wet)	kcond	Connected wetland
Glacier melt factor	Glacier melt	glac F	gladjust	Glacier

2.1.2. Athabasca watershed

The Athabasca River runs north-east from the Rocky Mountains toward the Peace-Athabasca Delta and Lake Athabasca and has a total watershed area of 156,000 km^2 in the Canadian provinces of Alberta and Saskatchewan, on Treaty 6 and 8 land (Fig. 2; hydrometric gauge details provided in Table S1).

Soils in the Athabasca basin are predominately loam, with more clay in the middle of the basin and sandier soil in and around the Athabasca Oil Sands region in the downstream (Alberta Geological Survey, 2013; Shangguan et al., 2014). Regional groundwater flows contribute 5% or less of the flow of the Athabasca River and its tributaries; there are still small areas of actively degrading permafrost (Gibson et al., 2016b; Vitt et al., 2000). The largest tributaries in the south of the Athabasca basin are the Pembina, McLeod and Lesser Slave Rivers; the Clearwater, Firebag and MacKay Rivers are the largest tributaries in the downstream portion of the basin. Temperatures in the Athabasca basin vary widely, with a monthly average temperature difference of 36 °C over the course of a year, and average temperatures below freezing for 5 months of the year (Environment and Climate Change Canada, 2020). The basin receives 270 mm of rainfall and 180 cm of snowfall in an average year, but there is substantial spatial and interannual variability in precipitation (Environment and Climate Change Canada, 2020). The flow of the Athabasca River is not regulated, either for flood control or hydroelectric generation, but water is diverted for agricultural use and industrial uses in the oil sands (Rosa et al., 2017).

The CHARM model used in this study discretizes the Athabasca River basin area firstly with a 0.4° longitude by 0.2° latitude grid, and the resulting cells are further divided into 10 different grouped response unit types, with areas determined from land cover data from the ESA (European Space Agency, 2017). The overall prevalence of the GRU types in the watershed model and their modeled structures (as described in 2.1.1) are listed in Table 2.

Table 2
Grouped response units for the Athabasca watershed model with landcover types and prevalence.

GRU Name	GRU type	Prevalence (%)	Landcover types
Grass	soil	8.1	herbaceous, cropland, grassland
Disconnected wetland	soil	8.8	wetlands: all vegetation heights
Connected wetland	connected wetland	2.2	wetlands: all vegetation heights
Mixed forest	soil	15.1	mixed and deciduous forest
Coniferous forest	soil	54.1	coniferous needle leafed forest
Shrub	soil	6.2	shrubland and sparse trees
Impervious	impervious	0.03	urban, consolidated bare ground
Barren	soil	1.4	sparse short vegetation, bare ground
Water	water	3.8	open water
Glacier	glacier	0.2	permanent ice and snow

2.1.3. Forcing and evaluation data

2.1.3.1. Meteorological data. The coupled isotope-hydrologic models were run using forcings based on observations at 56 Environment and Climate Change Canada (ECCC) weather gauges in the watershed and surounding area (Environment and Climate Change Canada, 2020). Forcing data in individual grid cells were interpolated at an hourly time step for humidity and air temperature, and at a daily time step for accumulated daily precipitation. All gauges with observation data for the relevant time interval were included in the inverse distance squared

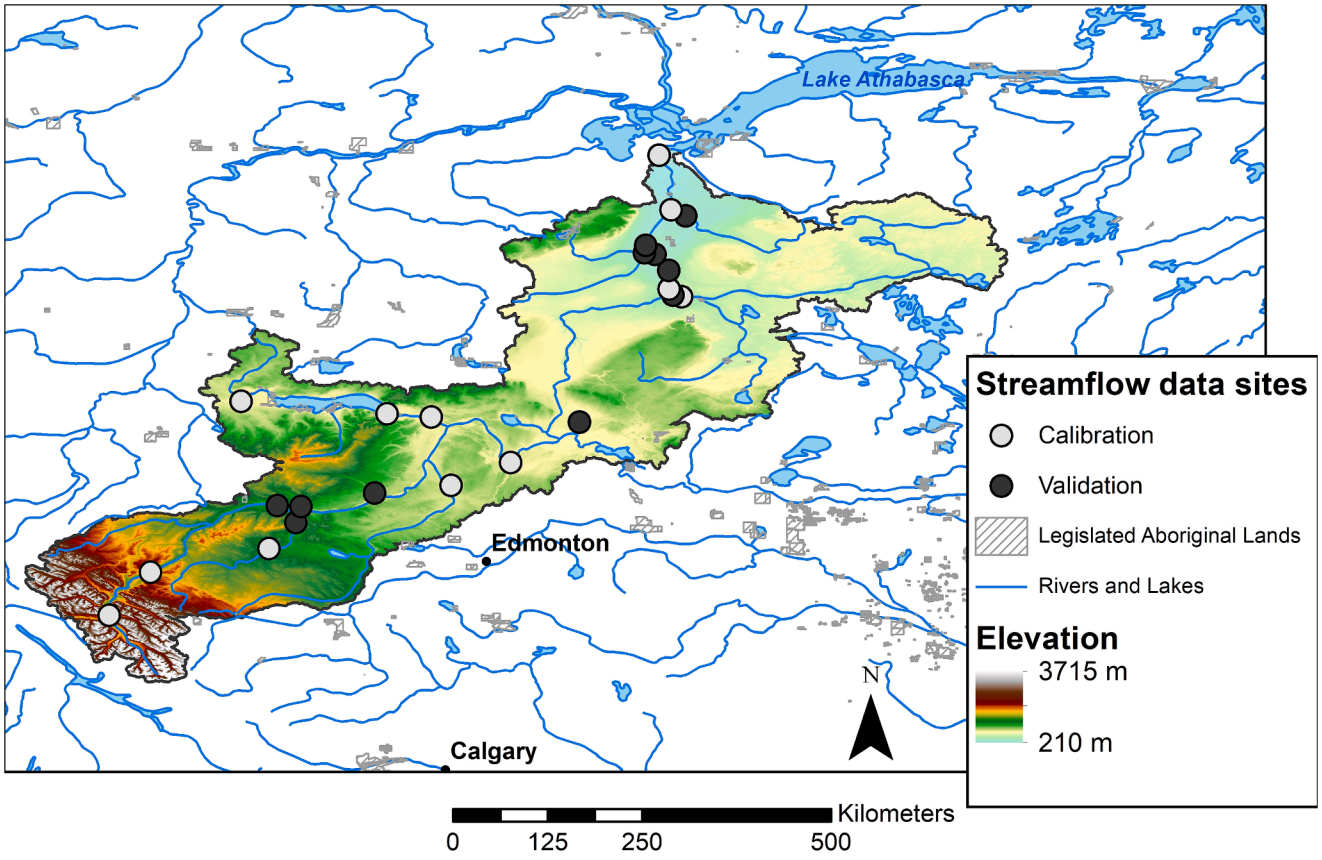

Fig. 2. Map of calibration and validation data locations in the Athabasca River basin, with elevation data from Holmes et al. (2022).

weighting estimate, and lapse rates of 0.2 mm/km and −5 °C/km were applied for precipitation and temperature estimates respectively (Kouwen, 2018; Minder et al., 2010). The isotope tracer model also requires isotopic compositions of precipitation to run; due to the scarcity and consistency of observations for isotopes in precipitation, these compositions were generated from an empirical model. For this study, average monthly isotopic compositions of precipitation were estimated from local climate and geographic data, based on geostatistical regressed equations originating from Delavau et al. (2015), and adapted by Holmes, 2016 to incorporate deuterium. The Delavau et al. (2015) model has been applied successfully in several previous studies (Gibson et al., 2021; Holmes et al., 2020, 2022; Stadnyk and Holmes, 2020), with model input uncertainties (derived from the empirically modelled isotopes in precipitation signatures) generally being smaller than the reported uncertainty bounds from model structure and parameter selection. It is noted that the use of an empirical model, particularly one at a monthly timestep, comes with limitations and assumptions. The impact of model forcing on isoWATFLOOD isotope-streamflow smulation, including the isoP model, was investigated by Delavau et al. (2017), and reports well-constrained input (forcing) uncertainty in all seasons except for the spring freshet, where flow (source) variability was the highest (i.e., snowmelt and large rain events) and isotopic compositions are more variable within a monthly timestep.

2.1.3.2. Flow and isotope data. Simulated model outputs were compared to historical hydrometric data from the Water Survey of Canada (Environment and Climate Change Canada, 2018). From a total of 20 continuous or seasonal (i.e., continuous only during the open water season) flow gauges with daily data between 2002 and 2015, the 10 gauges with the highest quality (i.e., most complete) data series were used to calibrate the model, and the remaining 10 were used for validation (see Fig. 2 for spatial distribution and the supplement for detailed gauge information). Gauged areas for the Athabasca and its tributaries range between 1000 and 137,500 km^2. Streamflow data have an estimated uncertainty of approximately ± 10%, with higher uncertainty expected during peak flow and ice-on periods (Kiang et al., 2018; Westerberg et al., 2020).

In addition to the WSC flow data, the Alberta Environmental Monitoring, Evaluation and Reporting Agency's Long-Term River Network monitoring program collected stable isotope tracer data at hydrometric gauges in the Athabasca watershed between 2002 and 2014 (Gibson et al., 2016b). Water samples for isotope tracer analyses were analyzed using either a Micromass IsoPrime Dual Inlet/Gas Chromatograph (University of Waterloo Environmental Isotope Laboratory, pre-2009) or a Thermo Scientific Delta V Advangage Dual Inlet/HDevice system (Alberta Innovates Technology Futures, Victoria, 2009 and later) (Gibson et al., 2016b). Both machines have estimated analytical uncertainties of ± 0.1‰ for oxygen-18 and ± 1‰ for deuterium, and all results are reported relative to VSMOW (Gibson et al., 2016b). To prevent post-collection isotopic fractionation, water samples were sealed in 30 mL high-density polyethylene bottles and analyzed within one year of sample collection (Gibson et al., 2019; Spangenberg, 2012).

2.2. Model calibration methods

Assessing the value added by isotope data to model calibration relative to current best-practice flow-only calibration requires the application of a multi-stage calibration workflow, completed both with and without stable isotope tracer data. The overall methodology is outlined in the flow chart in Fig. 3, and details are covered in the

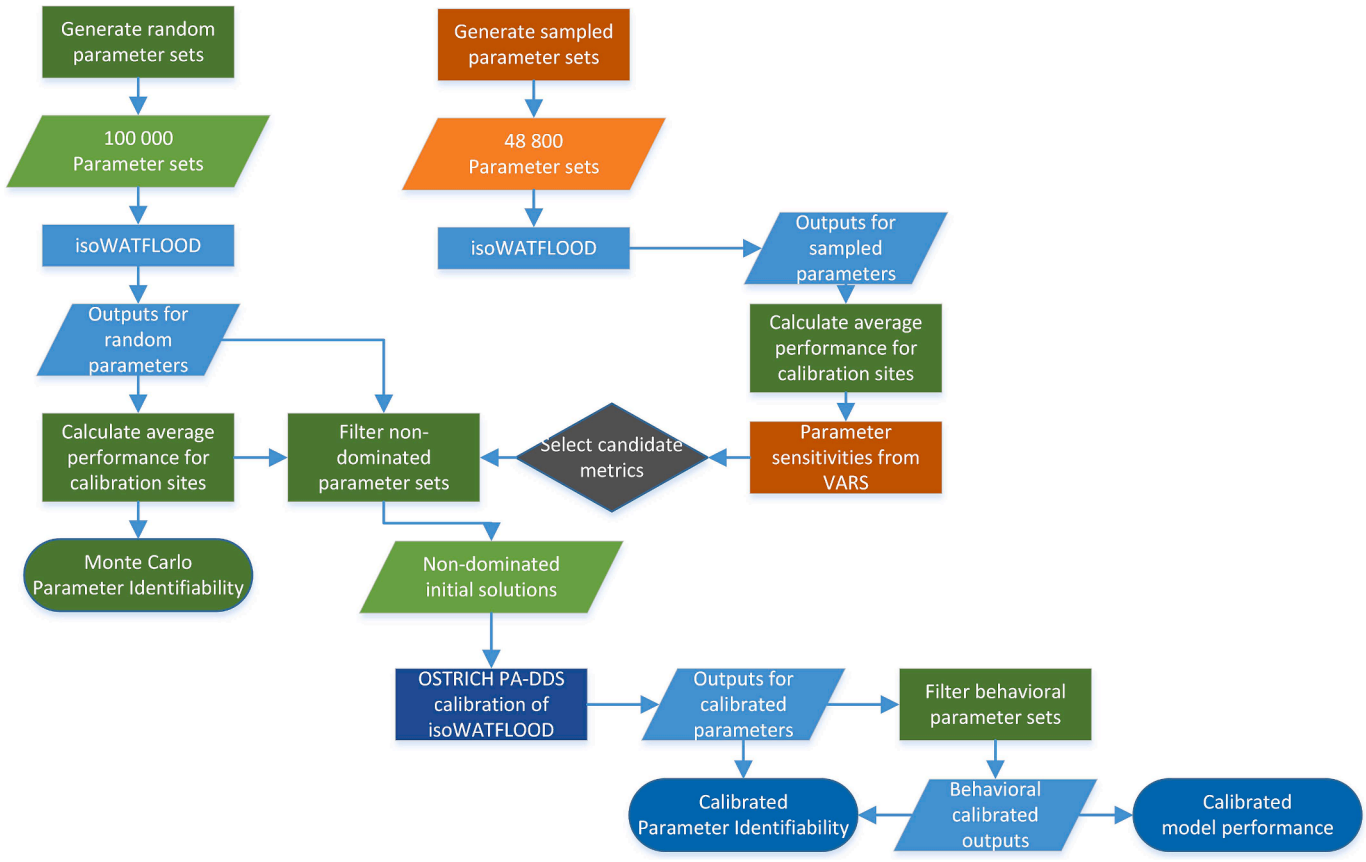

Fig. 3. Flow chart of the study calibration methodology, with isoWATFLOOD processes and outputs in blue, VARS in orange, and independent scripts in green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

following sections.

Two separate assessments of parameter identifiability were performed. The first was a Monte Carlo style evaluation, based on a large number of randomly generated parameter sets, which has the advantages of independent sampling and limited user decisions (sections 3.1 and 3.2). The second was a set of optimizations, using current recommended methods (assessing parameter sensitivities, selecting useful metrics, running multi-objective optimizations and selecting behavioral parameter sets), that produces higher quality solutions, but also depends on a series of user choices (sections 3.3 and 3.4).

2.2.1. Model performance metrics

A variety of metrics were evaluated as potential calibration objectives, selected based on the most commonly applied metrics from the literature. Metrics vary in their responsiveness to different types of simulation error (such as timing or volume bias), and likewise in their parameter sensitivities (Holmes et al., 2022). In calculating all performance metrics, only simulated data on days that have flow or isotope observations are considered.

The normalized root mean square error metric, a simple residual error metric primarily focused on timing and bias error, was evaluated as a calibration objective for both the isotope and flow simulations:

$$NRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{s,i} - x_{o,i}\right)^2}/\overline{x}_o \tag{1}$$

Where $x_{o,i}$ is observation i, $x_{s,i}$ is the corresponding simulated value, $\overline{x}_o$ is the mean value of all observations and n is the number of observations. The Kling-Gupta efficiency (KGE) metric, balanced between bias, variability and timing error, was likewise evaluated as a calibration objective for both the isotope and flow simulations, as were its constituent components, the bias (β), the relative variability (α) and the correlation (r) (Gupta et al., 2009).

$$r = \frac{\sum_{i=1}^{n}\left(x_{o,i} - \overline{x}_o\right)\left(x_{s,i} - \overline{x}_s\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{o,i} - \overline{x}_o\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_{s,i} - \overline{x}_s\right)^2}} \tag{2}$$

$$\alpha = \frac{\sigma_s}{\sigma_o} \tag{3}$$

$$\beta = \frac{\overline{x}_s}{\overline{x}_o} \tag{4}$$

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{5}$$

Where $\overline{x}_s$ is the mean value of all simulated values with corresponding observations and σ_o and σ_s are the standard deviations of the observed and simulated data.

A selection of flow signature metrics (which do not include simulating timing error) were considered as potential calibration objectives for the isotope and flow simulations: the error in the slope of the flow duration curve (SFDC) (for flows) or the slope of the duration curve (SDC) (for isotopes), and the high- and low-flow signatures, at the 5% and 95% exceedance probabilities (Viglione et al., 2013):

$$SFDC\ or\ SDC = 100\left(\frac{x_{s,30} - x_{s,70}}{40\overline{x}_s} - \frac{x_{o,30} - x_{o,70}}{40\overline{x}_o}\right) \tag{6}$$

$$Q_5 = \frac{x_{o,5} - x_{s,5}}{x_{o,5}} \tag{7}$$

$$Q_{95} = \frac{x_{o,95} - x_{s,95}}{x_{o,95}} \tag{8}$$

Where x_5, x_{30}, x_{70} and x_{95} are the data with exceedance probabilities of 5%, 30%, 70% and 95%. The previously described relative variability and bias metrics are also flow signatures (Shafii and Tolson, 2015).

An additional signature metric for the isotope simulation exclusively was included in the comparison of potential calibration objectives: the error in the simulated slope of the local mixing line (LML). The LML slope error, which relies on the simulation of both isotopes and focuses on simulation variability error, measures the mismatch between the simulated LML slope and the observed LML slope and like the above flow signature metrics does not include timing error (Stadnyk and Holmes, 2020):

$$LMLmE = \frac{\sum_{i=1}^{n}\left(O_{s,i} - \overline{O}_s\right)\left(D_{s,i} - \overline{D}_s\right)}{\sum_{i=1}^{n}\left(O_{s,i} - \overline{O}_s\right)^2} - \frac{\sum_{i=1}^{n}\left(O_{o,i} - \overline{O}_o\right)\left(D_{o,i} - \overline{D}_o\right)}{\sum_{i=1}^{n}\left(O_{o,i} - \overline{O}_o\right)^2} \tag{9}$$

Where $O_{o,i}$ is oxygen-18 observation, and $O_{s,i}$ is the corresponding simulated oxygen-18 value, and $D_{o,i}$ and $D_{s,i}$ are the observed and simulated values for deuterium.

Finally, two metrics were evaluated as potential calibration objectives for only the flow simulation: the Nash-Sutcliffe efficiency (NSE), and the log transform version of NSE were calculated exclusively for the flow simulation (Nash and Sutcliffe, 1970). These traditional metrics are included due to their long-standing use in hydrologic model evaluations; NSE is highly affected by large residuals during high flows, while logNSE is relatively affected by smaller residuals during low flows.

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(x_{s,i} - x_{o,i}\right)^2}{\sum_{i=1}^{n}\left(x_{o,i} - \overline{x}_o\right)^2} \tag{10}$$

$$logNSE = 1 - \frac{\sum_{i=1}^{n}\left(log\left(x_{s,i}\right) - log\left(x_{o,i}\right)\right)^2}{\sum_{i=1}^{n}\left(log\left(x_{o,i}\right) - \overline{log\left(x_o\right)}\right)^2} \tag{11}$$

In addition to individual simulation metrics, which can potentially be used as calibration objectives, there are also ensemble performance metrics quantifying the performance of groups of simulations (e.g., the outcome of multi-objective calibration). In this study ensemble performance is assessed using two indices to quantify the two key components of ensemble quality: the containment ratio, and the relative band width. The containment ratio quantifies the ensemble accuracy and is simply the ratio of observations contained within the ensemble upper and lower bounds to the total number observations (Xiong et al., 2009). The average relative band width quantifies ensemble precision and is calculated as (Xiong et al., 2009):

$$RB = \frac{\sum_{i=1}^{n}\left(x_{s,i}^{u} - x_{s,i}^{l}\right)/x_{o,i}}{n} \tag{12}$$

Where n is number of observations, $x_{o,i}$ is observation i, $x_{s,i}^{u}$ is the maximum corresponding simulated value from the ensemble (i.e., the upper simulated bound) and $x_{s,i}^{l}$ is the minimum corresponding simulated value from the ensemble (i.e., the lower simulated bound).

2.2.2. Parameter sensitivity and metric selection

Parameter sensitivities for the various performance metrics are evaluated to design informed calibration strategies utilizing isotope tracer data. The best method to analyze sensitivity to this end is a global analysis (rather than local), which evaluates parameter sensitivity over the entire specified parameter space (Song et al., 2015). This study uses VARS, a global sensitivity approach which uses variograms to quantify global parameter sensitivity (Razavi and Gupta, 2016). VARS, which is implemented in the VARS-TOOL software package, was selected for the sensitivity analysis over older alternative approaches due its relative computational efficiency, which is a key consideration in applying a global analysis to highly parameterized models with longer run-times (Razavi et al., 2019). The VARS-TOOL software (V2 MATLAB version) was used to generate 48,800 sampled parameter sets and the recommended $IVARS_{50}$ sensitivity index was used, along with bootstrapped 90% confidence intervals (sampling specifics can be found in Table S2 in the supplement) (Razavi and Gupta, 2016).

The parameter sensitivity results presented in this study have been normalized using the total sensitivity of a given metric, to simplify comparisons of potential calibration objectives. The averaged performance (for all 10 flow and 17 isotope simulation performance metrics described in 2.2.1) at all 10 calibration sites was used as the response variable in the parameter sensitivity analysis. The resulting parameter sensitivity values are used to select possible metrics for the multi-objective optimization (Section 2.2.4).

In calibrating a hydrologic model, the aim is to identify 'good' values for all calibrated parameters. In order to identify a 'good' value, it is necessary that the calibration objective changes in response to changes in the value of that parameter. If the objective metric is unresponsive to changes in the value of a parameter, an optimization algorithm cannot distinguish between 'good' and 'bad' values for that parameter, and it will remain uncalibrated. Parameter sensitivity data are quantifications of this responsiveness; insensitive parameters will not be successfully calibrated. To maximize the number of parameters which might potentially be calibrated, metric combinations minimizing insensitive parameters are preferred. The following rules were applied to quantify and rank the sensitivity coverage of performance metrics, to select candidate metrics for this study in a more objective and reproducible manner:

1. Metrics (for flow and isotope data separately) are ranked from most types of error covered (e.g., timing, variability or bias errors, as categorized in Holmes et al., (2022)), to least
2. Parameter sensitivities are classified as highly sensitive (over 10% of total sensitivity, bootstrapped sensitivity uncertainty less than 10% of total sensitivity), likely sensitive (over 3% of total sensitivity, sensitivity uncertainty less than 25%), possibly sensitive (over 1% of total sensitivity), and insensitive (i.e., no detected sensitivity) (less than 1% of total sensitivity)
3. The metric with the most broadly distributed sensitivity (i.e., highest number of parameters which are at least possibly sensitive) is selected as the primary candidate metric (when using mixed data types, the flow metric is primary)
4. When searching for second candidate metric, the metric increasing the sensitivity (i.e., higher classification based on step 2) for the most parameters, from the primary candidate base is selected
5. In cases of a tie (i.e., equal numbers of sensitive parameters) the metric covering the most types of error (from step 1) is chosen

Seven different combinations of data types are used to select candidate metric pairs: flow-only, isotope-only, oxygen-18 only, deuterium only, flow and oxygen-18, flow and deuterium, and flow and both isotope tracers.

2.2.3. Monte Carlo analysis

Exploring the relationship between parameter values and model performance requires sampling the full range of all parameters and running the hydrologic simulation with each parameter set to generate simulation results. For this study, a Monte Carlo analysis is performed, where all 27 calibrated parameters are sampled across their entire calibration range using a uniform random distribution. A total of 100,000 random parameter sets are generated and run in the isoWATFLOOD model of the Athabasca basin. The simulation performance for the resulting flow and tracer outputs was quantified using performance metrics (Section 2.2.1) and averaged for calibration sites. The Monte Carlo results are used to assess the parameter identifiability for each performance metric independently, by selecting the best 0.1% solutions for each metric and finding the distribution (i.e., range, median and interquartile range) of parameter values for those 100 simulations. Parameter values are classified as very well identified if the total range of calibrated parameter values is less than 5% of the calibration range and the interquartile range (IQR) is less than 2%, well identified if the range is less than 10% and IQR less than 5%, identified if either the range is less than 20% or the IQR than 10%, and somewhat identified either the range is less than 50% or the IQR is less than 25%.

The key benefit of the Monte Carlo analysis is that parameter sets, and therefore simulation results, are completely independent of each other, permitting robust statistical treatment of the results. Monte Carlo simulations have also been widely used to calibrate models with mixed flow and isotope data in previous studies (Delavau et al., 2017; Neill et al., 2019; Piovano et al., 2020). However, using randomly generated parameter sets is computationally inefficient (requiring vast numbers of independent simulations), and the best performing simulations have a very low likelihood of matching the quality of solutions produced by an optimization algorithm (Efstratiadis and Koutsoyiannis, 2010).

2.2.4. Multi-Objective optimization

The current best practice for hydrologic model calibration is the use of optimization algorithms, particularly multi-objective automated search algorithms. An optimization algorithm may start from a random point in the parameter space (or from a specified initial solution), but rather than continuing to generate random solutions, it will search the adjacent region of the parameter space for better solutions. The algorithm judges solutions as better or worse based on the value of the objective function (or objective functions for multi-objective algorithms); the selection of an objective is therefore a key decision in model calibration that warrants further investigation.

This study uses the PA-DDS (Pareto archiving dynamically dimensioned search) algorithm for multi-objective model calibration, following the results in Holmes et al., (2020). Based on the DDS algorithm (Tolson and Shoemaker, 2007), the PA-DDS algorithm is a computationally efficient search method where the search space is gradually constrained as the algorithm completes iterations, but with the additional ability to retain an archive of solutions which are not dominated (i.e., all solutions which are not categorically outperformed, considering multiple objectives) (Asadzadeh et al., 2014; Asadzadeh and Tolson, 2013). Model calibrations were performed using the OSTRICH program (v17.12.19), a model-independent calibration tool that implements the PA-DDS algorithm among others (Matott, 2017). For each set of candidate objectives, five separate calibration trials were run, with different random seeds (the method for choosing performance metrics as possible candidates is covered in 2.2.2). Each trial ran for 1000 iterations, and all non-dominated solutions from each of the five trials were retained in the initial results analyses (section 3.3). The five trials for each candidate performance metric pair shared initial solutions: non-dominated solutions for the candidate metrics were selected from the Monte Carlo results, as better-quality initial solutions can reduce the computational budget required for optimization. While PA-DDS is not limited to two objectives, this study is limited to paired objectives, as increasing the number of objectives generally increases the computational requirements (Asadzadeh et al., 2014). Calibrations using an optimization algorithm generally have better simulation performance than those using random parameter sets in a Monte Carlo analysis and using multi-objective optimization generates a substantial number of parameter sets producing good flow or tracer simulations (Efstratiadis and Koutsoyiannis, 2010). However, the solutions from a particular trial are not fully independent of each other (as the algorithm generates new solutions from progressive modifications of previous ones), which limits or complicates the statistical treatment of the results.

3. Results

3.1. Monte Carlo parameter identifiability

The best 100 solutions from the 100 000 randomly generated parameter sets run in the Monte Carlo analysis are selected for each flow and isotope performance metric independently. Box-whisker plots of all parameters (normalized to the calibration range) and all assessed metrics are shown in Fig. 4 (mean, range and IQR values for all parameters

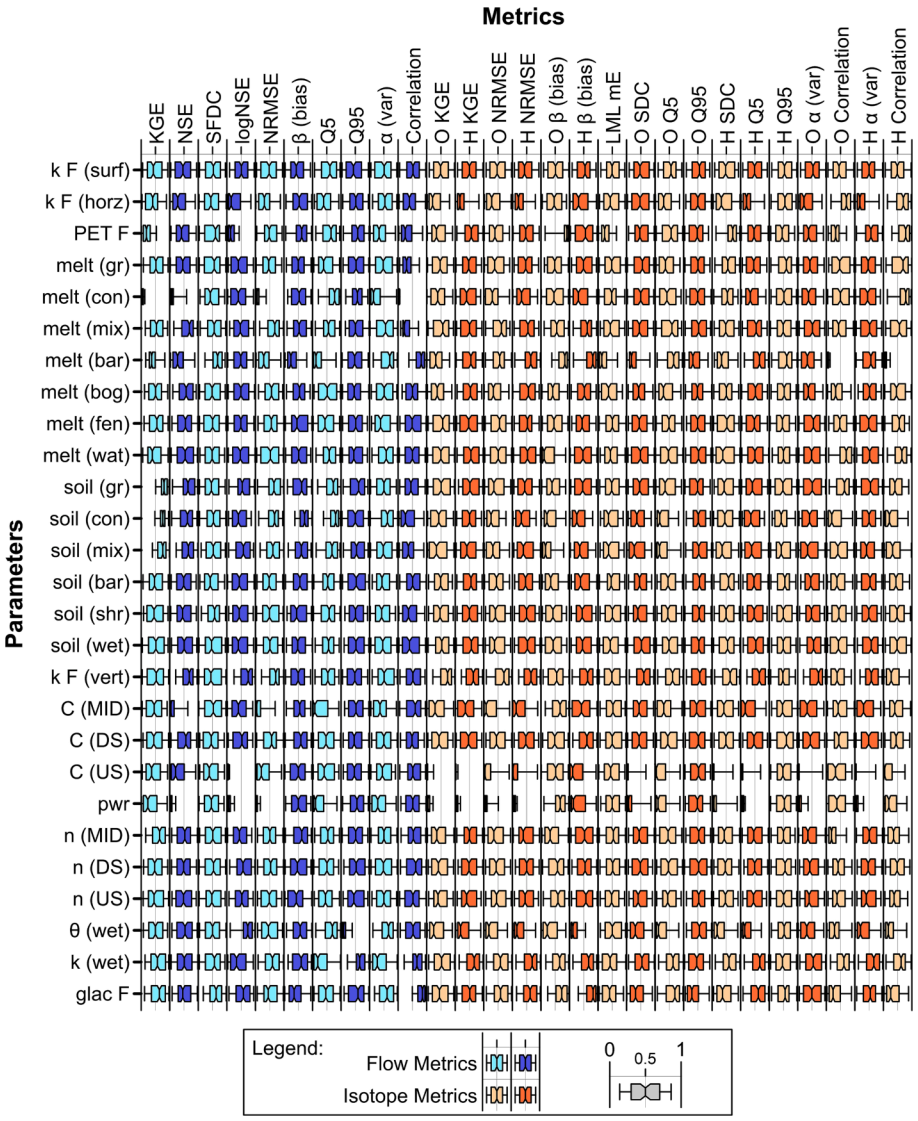

Fig. 4. Box-whisker plots of normalized parameter values for the best 100 solutions from the Monte Carlo analysis, based on each performance metric independently. Whiskers extend to the 5/95 percentile, and flow and isotope metrics are shown in blue and orange respectively (lighter and darker shades are used to differentiate columns). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and metrics are shown in Tables S6 to S8).

Overall, there is limited identifiability for most parameters; most boxplots exceed 50% of the calibration range (i.e., parameter values are unidentifiable; categorized parameter identifiability from the Monte Carlo analysis is shown in Table S9). Parameter values for melt rates for coniferous forest and barren ground, the baseflow power, and the baseflow coefficient in the mountains are identifiable when calibrating to some metrics. Most parameters have some relationship with at least one performance metric, where the parameter values for the 100 selected solutions are either limited to or skewed towards only part of the total calibration range. For example, the distribution of values for the horizontal conductivity factor (k F (horz)) is skewed for the majority of performance metrics although it is only somewhat identifiable for a single metric, KGE H, and the parameter value range is somewhat identified for the wetland porosity parameter for Q correlation and H bias. Only the surface soil conductivity parameter and the downstream channel roughness (k F (surf) and n (DS)) are completely unidentifiable, maintaining a uniform distribution of parameter values across the entire possible parameter range. Generally, flow and isotope performance metrics are aligned in either identifying or not identifying parameter values, but not in all cases. For example, the flow and oxygen-18 performance metrics identify opposite ends of the possible parameter value range for the soil water retention capacities for coniferous and mix forest. While intriguing, the Monte Carlo results do not identify parameter values well and are inconclusive on the utility of isotope data for improving parameter identifiability. The parameter space, with 27 interacting calibrated parameters, is both large and complex and even the best simulations produced were of mediocre quality (e.g., best flow KGE range between 0.48 and 0.52).

3.2. *Parameter sensitivity and candidate objective metric selection*

A VARS global sensitivity analysis was performed for all evaluated performance metrics independently, using average performance at calibration sites as the response variable. The normalized sensitivities and sensitivity reliabilities are shown in Fig. 5.

A few parameters are consistently insensitive across all evaluated metrics, namely the surface soil conductivity (i.e., kF (surf)) and connected wetland snowmelt rate (i.e., melt(fen)), but generally parameters are sensitive for some metrics and not others. Both flow and isotope metrics are sensitive to some snowmelt and baseflow parameters. Isotope performance metrics are more sensitive to upper zone soil

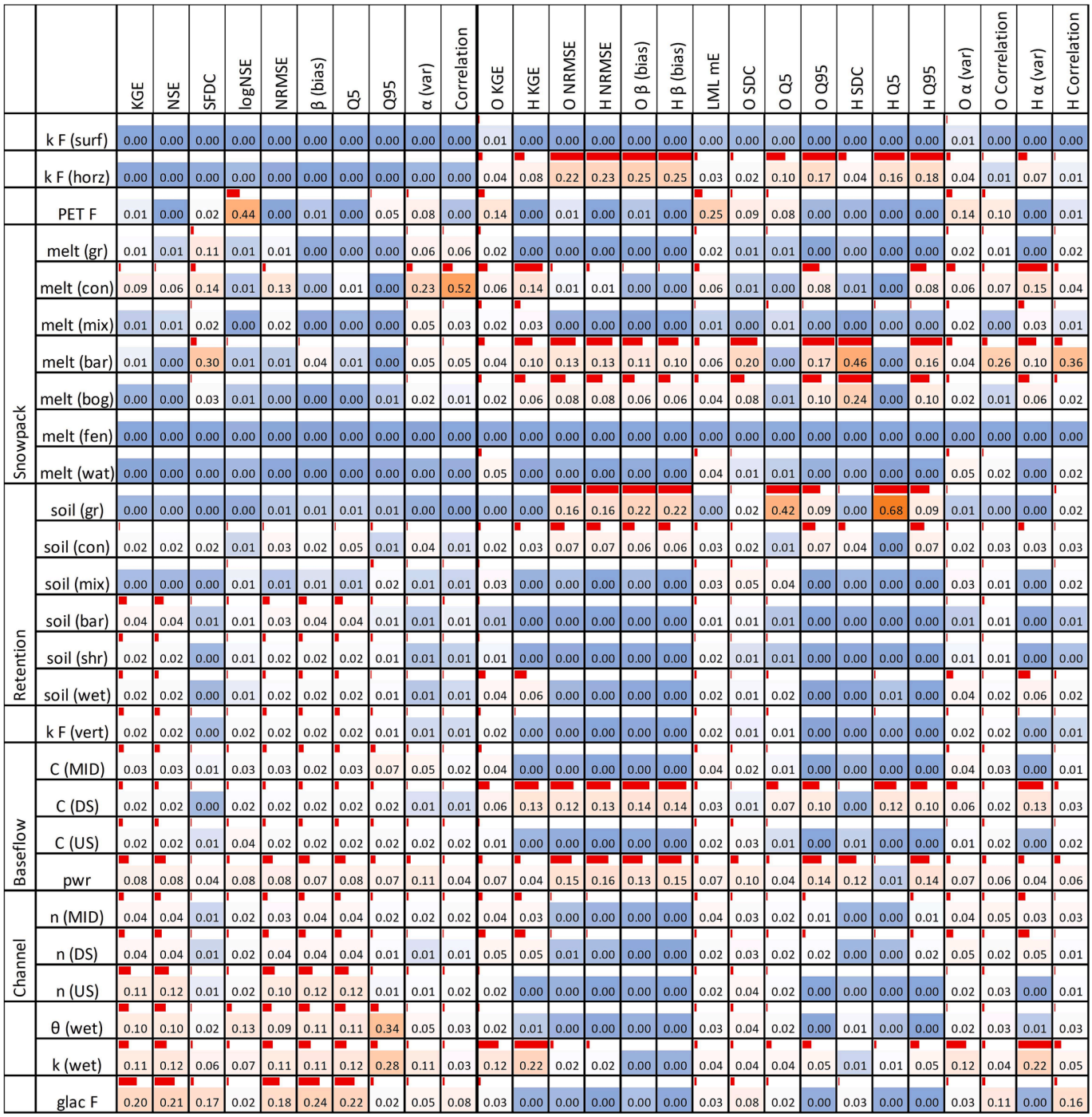

Fig. 5. Parameter sensitivity for averaged calibration site performance from the VARS analysis, with the reliability of the sensitivity results indicated by the length of the red bars. Highly sensitive and insensitive parameters are highlighted in orange and blue respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

retention parameters in dominant land classes, relative to the flow performance metrics, but these sensitivities are not generally reliable. Likewise, flow metrics are more sensitive to wetland and glacier parameters but not always reliably. The LML slope error, and the oxygen-18 KGE and SDC standout for their balanced sensitivities, with nearly equal shares of the total sensitivity for each individual parameter.

Sensitive parameters are likely to be at least somewhat identifiable (based on the Monte Carlo identifiability results in Fig. 4; Table S3 provides correlation metrics between Monte Carlo IQR and VARS sensitivity), but with too many anomalies to make either analysis type a strong predictor of the other. The melt rate for coniferous forest is both sensitive and identifiable using flow performance metrics, but glacier melt is sensitive and generally poorly identified with the same metrics. Unidentifiable parameters match with consistently to the insensitive ones, yet isotope metrics are better at identifying baseflow coefficients in the upstream and worse at identifying the same coefficient in the downstream (C (DS)) than would be predicted based on the parameter sensitivities.

The data presented in Fig. 5 are the basis of the metric candidate identification process, where performance metric pairs providing the widest parameter sensitivity coverage (following the quantitative criteria in 2.2.2) are selected for use in multi-objective calibration. The

categorized parameter sensitivity coverage for all individual metrics and the candidate paired performance metrics is shown in Tables S4 and S5 respectively, in the supplementary data.

Candidate pairs were selected when considering only flow metrics, flow and isotope metrics and only isotope metrics (see section 2.2.2 for the selection rules). One additional candidate pair was added, as the NRMSE for isotope simulation performance is too commonly used to be overlooked in this analysis. Overall, there is limited differences in the categorized parameter sensitivity coverage between the eight different candidate calibration metric pairs. By maximizing the number of sensitive parameters within the data constraints, the resulting coverage is similar, with most parameters being somewhat sensitive (i.e., not insensitive) to at least one of the paired metrics. The lower zone, wetland and channel parameters have similar sensitivity for all candidate pairs, and there is more variation between candidates for upper zone, melt and evaporation parameter sensitivities.

3.3. Multi-Objective optimization

Each candidate pair of performance metrics were used as objectives in five independent multi-objective calibration trials, using the non-dominated solutions (for that candidate pair) from the Monte Carlo analysis as initial solutions. The outcomes of these 40 calibration trials are shown in Fig. 6, along with the initial solutions.

Simulation performance after calibration is substantially improved compared to the initial random solutions (e.g., most solutions with KGE Q as an objective now have KGE Q exceeding 0.5). Fig. 6 is not strictly a trade-off front, as neither KGE Q nor NRMSE ^{18}O are objectives for all calibrations, but it still clearly demonstrates the cost of calibrating with an isotope tracer objective. The best flow simulation performance (x-axis) does not achieve the best isotope tracer (y-axis) simulation performance. If equivalent weight is given to isotope and flow performance, the flow simulation performance in calibration will be lower than if flow performance is considered alone. However, considering the uncertainty in performance metric values, a decrease in calibration KGE Q of 0.03 (the difference between maximizing flow performance and giving equal weight to flow and isotope performance) is not particularly serious. The two isotope tracers are similar but not identical, e.g., when KGE ^{2}H is an objective, the calibration performance does not quite overlap with solutions found with KGE ^{18}O as an objective. On the other hand, using NRMSE ^{18}O or KGE ^{18}O as objectives produces equivalent calibration performances. There is a clear segregation of objective pair types: the flow-only pair has the best flow performance, the isotope-only pairs have the best isotope simulation performance, and the mixed flow-isotope pairs fall between the other two, with some overlap. The difference between the best flow performance for the flow-only and mixed flow-isotope calibrations is negligible (0.586 vs 0.583 KGE Q).

The isotope-only calibrations do not generally produce good flow simulations, with a mean and median flow KGE of 0.28, flow simulations are of similar quality to those from random parameter sets in the Monte Carlo analysis (mean and median KGE 0.20 and 0.21). Isotope-only calibrations only outperformed random solutions in avoiding very poor flow simulations, with a minimum flow KGE of 0.05, compared to −0.27 from the Monte Carlo analysis.

To assess the effect of using isotope tracer data in calibration on identified parameter values, the relationships between flow and isotope simulations performance and calibrated parameter values for several key parameters are illustrated in Fig. 7, with other parameters in the

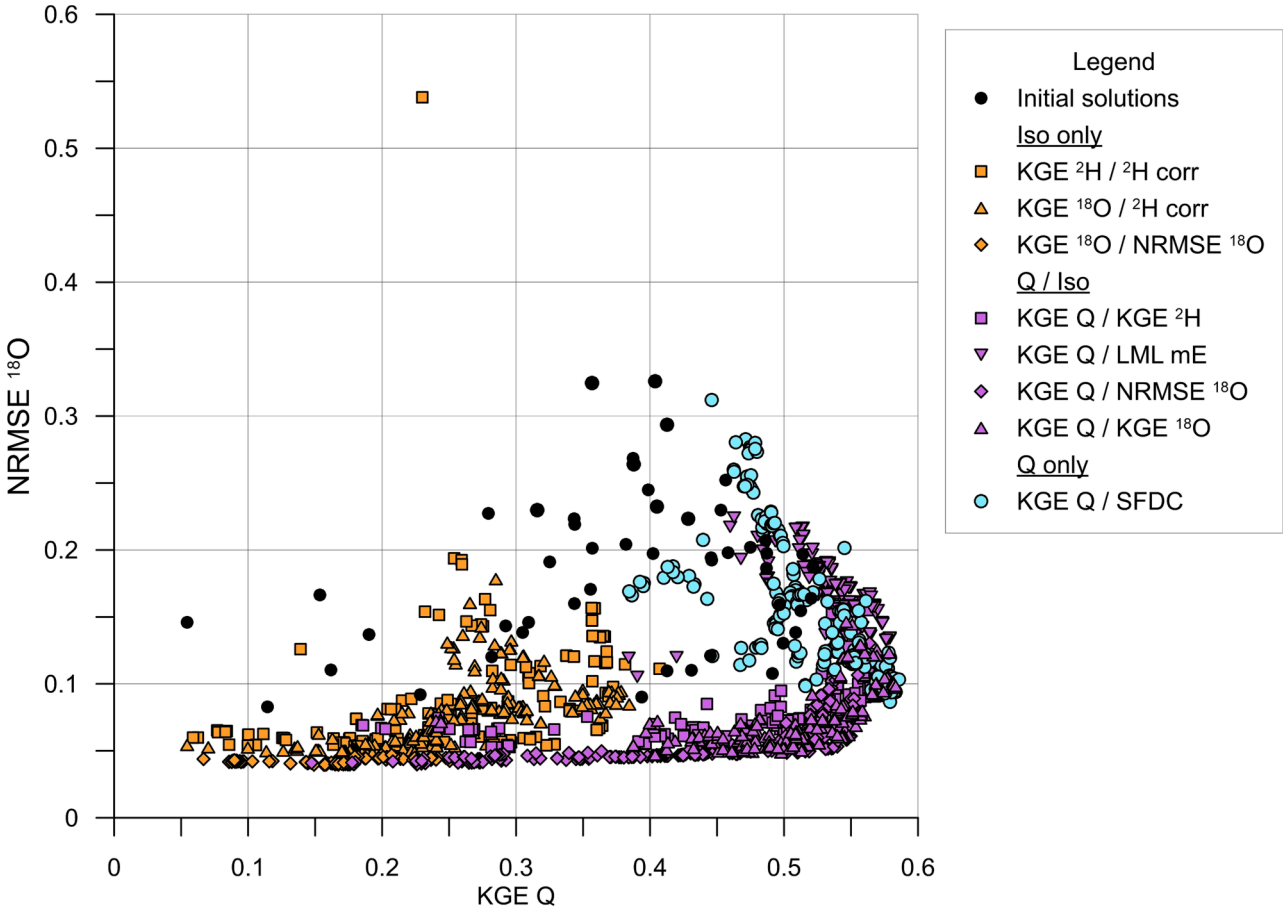

Fig. 6. Calibration performance for the PA-DDS solutions for all candidate metric combinations with initial solutions taken from the Monte Carlo analysis (black dots).

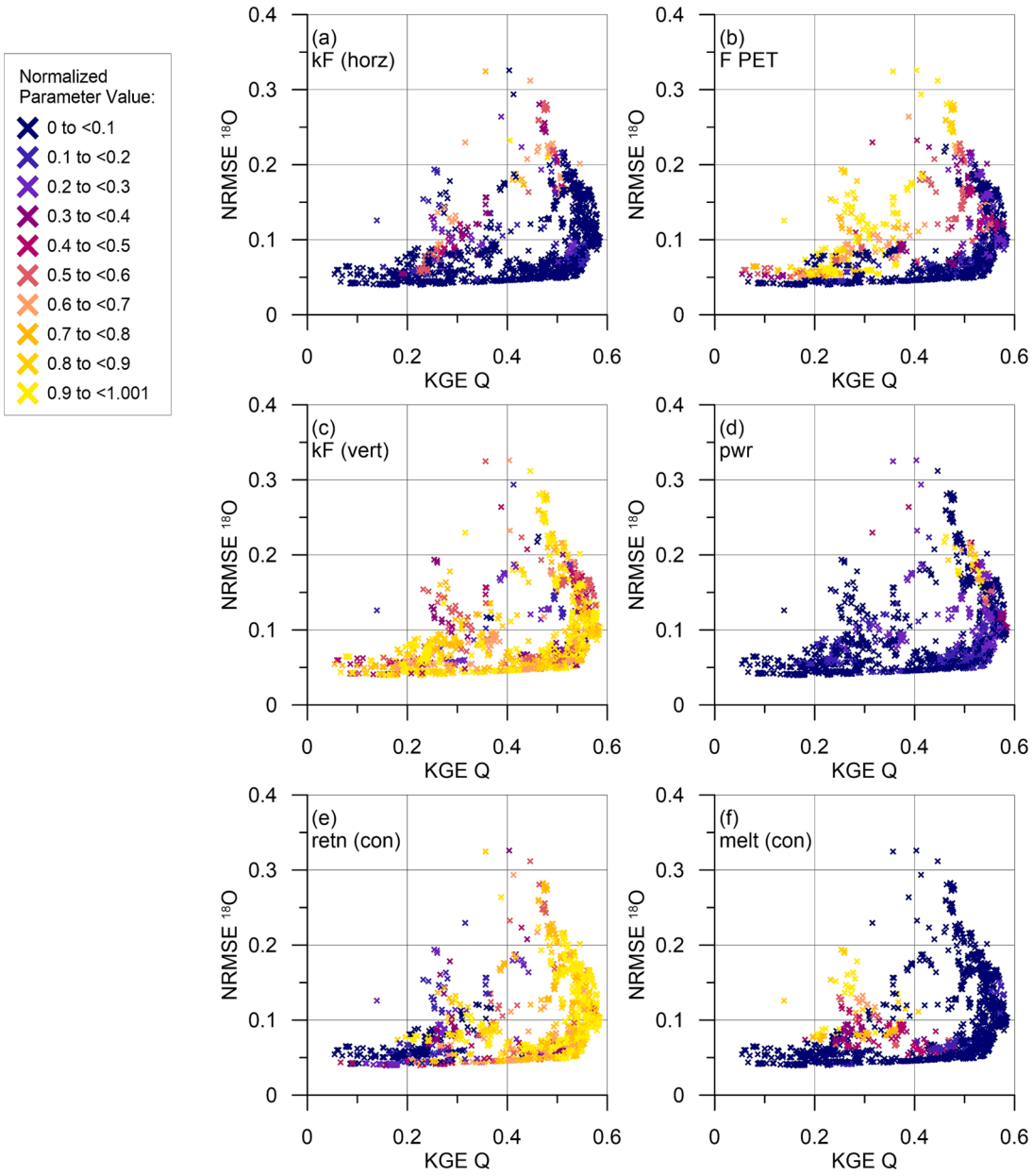

Fig. 7. Calibration performance for all PA-DDS solutions, with point color indicating normalized parameter value.

supplementary material (S1-S4); three outlier solutions with NRMSE ^{18}O greater than 0.5 are not shown due to space constraints.

These highly influential parameters generally have identified parameter values or correlated simulation performance and parameter values. Some are identified by both flow and isotope results (i.e., kF, the horizontal and vertical soil conductivity), but others are better identified by either the flow (i.e., coniferous forest soil retention) or the isotope performance (i.e., the evaporation adjustment factor) with parameter values only consistent for the best performing simulations of one type.

The final step in the calibration process is the selection of acceptable solutions from the total output of the 40 non-dominated (Pareto) solution produced by the multi-objective optimization algorithm. As the flow simulation is the primary model output of interest for a hydrologic model, only solutions with calibration KGE Q greater than 0.5 will be considered behavioral. This threshold entirely removes isotope-only calibrations from further consideration, as no behavioral solutions were produced by calibrating to isotope tracer performance alone. The average flow simulation at validation sites for behavioral solutions is shown in Fig. 8, along with the calibration performance.

As is typical for hydrologic models, the validation performance decreases relative to calibration performance, with average KGE Q exceeding 0.5 for calibration sites, and ranging between 0.26 and 0.42 for validation sites. The best validation flow simulation performance for behavioral solutions is best predicted by the isotope simulation performance in calibration, not the flow simulation performance in calibration; the best validation KGE Q are for solutions with low NRMSE ^{18}O, and the worst are for solutions with higher NRMSE ^{18}O.

Behavioral solutions from the calibrated parameter sets have better parameter identifiability than the best solutions from random parameter sets; however, solutions generated from the same PA-DDS multi-objective calibration trial are not truly independent which limits the comparability to the Monte Carlo results (the five trials per candidate objective pair are independent, rather than all parameter sets). To evaluate the effect of using isotope tracer data in calibration on parameter identifiability for MOO, the normalized parameter values from all behavioral solutions, for all parameters, are shown in the

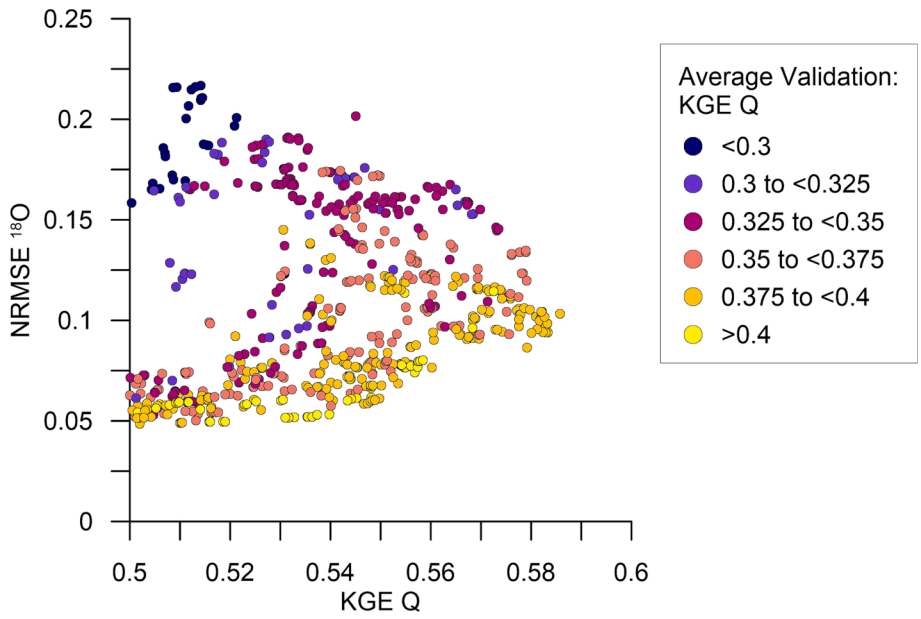

Fig. 8. Calibration and validation performance for PA-DDS solutions with behavioral flow (KGE Q $\geq$ 0.5) calibration performance.

histograms in Fig. 9 (mean, range and IQR values for these distributions are shown in Tables S10to S12).

The majority of parameters have at least somewhat identified values, and a few are well-identified (e.g., coniferous forest snowmelt rate and the baseflow coefficient in the upstream region). Soil retention, soil conductivity, wetland parameters and the evaporation adjustment factor are identified or somewhat identified for most objective pairs. All five candidate objective pairs have broadly similar identifiable and unidentifiable parameters, and usually agree on approximately the same parameter values. The two oxygen-18 calibrations (using KGE ^{18}O and NRMSE ^{18}O as secondary objectives) have extremely similar parameter values and parameter identifiability, and the calibration using KGE ^{2}H as the second objective is similar to both oxygen-18 calibrations, except for the glacier melt factor. The flow-only calibration and the calibration using the LML slope error as the second objective diverge from the other calibrations for some parameters. As an example, the three pairs using an isotope metric including timing error all converge on a low value of wetland porosity, but LML slope error calibration tends toward a high porosity value, and the flow-only calibration does not identify a value for porosity. It is noted, however, that using isotope data in calibration does not consistently improve parameter identifiability (e.g., the evaporation factor and wetland porosity are better identified, but mixed wood upper zone soil water retention has a less identifiable parameter value).

3.4. Calibrated ensemble performance

The final stage of evaluating calibration effectiveness and changes in flow simulation results from the use of stable isotope tracer data in calibration is the assessment of the ensemble performance. All behavioral solutions from the five PA-DDS calibration trials for each candidate objective pair were included, with equal weight, in an ensemble of parameter sets, such that simulation uncertainty resulting from parameter selection can be identified (total numbers of members are listed in Table 3). The average annual hydrographs for the three most divergent ensembles of behavioral solutions are shown in Fig. 10, at the Athabasca River at Fort McMurray, which covers most of the modeled area (other hydrographs shown in Figure S5 in the supplementary material).

Parameter identifiability does not decisively improve when isotope data are included in calibration, but the resulting ensembles produce distinctive component contributions to streamflow. Mixed flow-isotope calibrations have similar uncertainty for total flow to a flow-only calibration, with slightly improved observation containment (at this downstream gauge). On the other hand, the isotope-enabled calibration ensembles have substantially higher confidence in the component contributions generating the total flow (e.g., the range of contributions from the upper soil zone for isotope-enabled calibrations is less than half that of the flow-only calibrations). The percent contributions from the upper and lower soil zones are listed in Table 3, along with the range in percent contributions, the ensemble containment ratios, and relative bounds.

Average contributions from the two soil zones are consistent for all calibrated ensembles, but the ranges in those contributions are substantially different. All mixed flow-isotope calibrated ensembles have much lower maximum upper zone contributions, and much higher minimum lower zone contributions than the flow-only ensemble. This improvement in flow component identifiability is linked to improvements in process flux identifiability for soils, (see Figure S6 in the supplement). Median and inter-quartile ranges for soil fluxes are similar for all calibration objective pairs, but the flow-only calibration has substantially longer tails (i.e., higher uncertainty) than mixed isotope-flow calibrations.

Similarity in the total flow uncertainty and the improvement in the ensemble observation containment is also supported quantitatively across all gauge sites. The three mixed flow-isotope calibrations using an isotope performance metric that includes timing errors all have better containment ratios than the flow-only calibration at both calibration and validation gauges, and these improvements in the mean containment ratio are statistically significant. There is no significant change in the relative band width for the ensembles calibrated with flow and oxygen-18 data. The ensemble calibrated with KGE Q and KGE ^{2}H has wider relative bounds. The final flow-isotope calibrated ensemble, calibrated to maximize KGE Q and minimize LML slope error, has significantly narrower relative band width, but a lower containment ratio than the flow-only calibration.

4. Discussion

4.1. Parameter identifiability

Overall, parameter values were not well-identified, regardless of methodology, metric or data type, with a few exceptions. The small number of well-identified parameter values is unsurprising given the

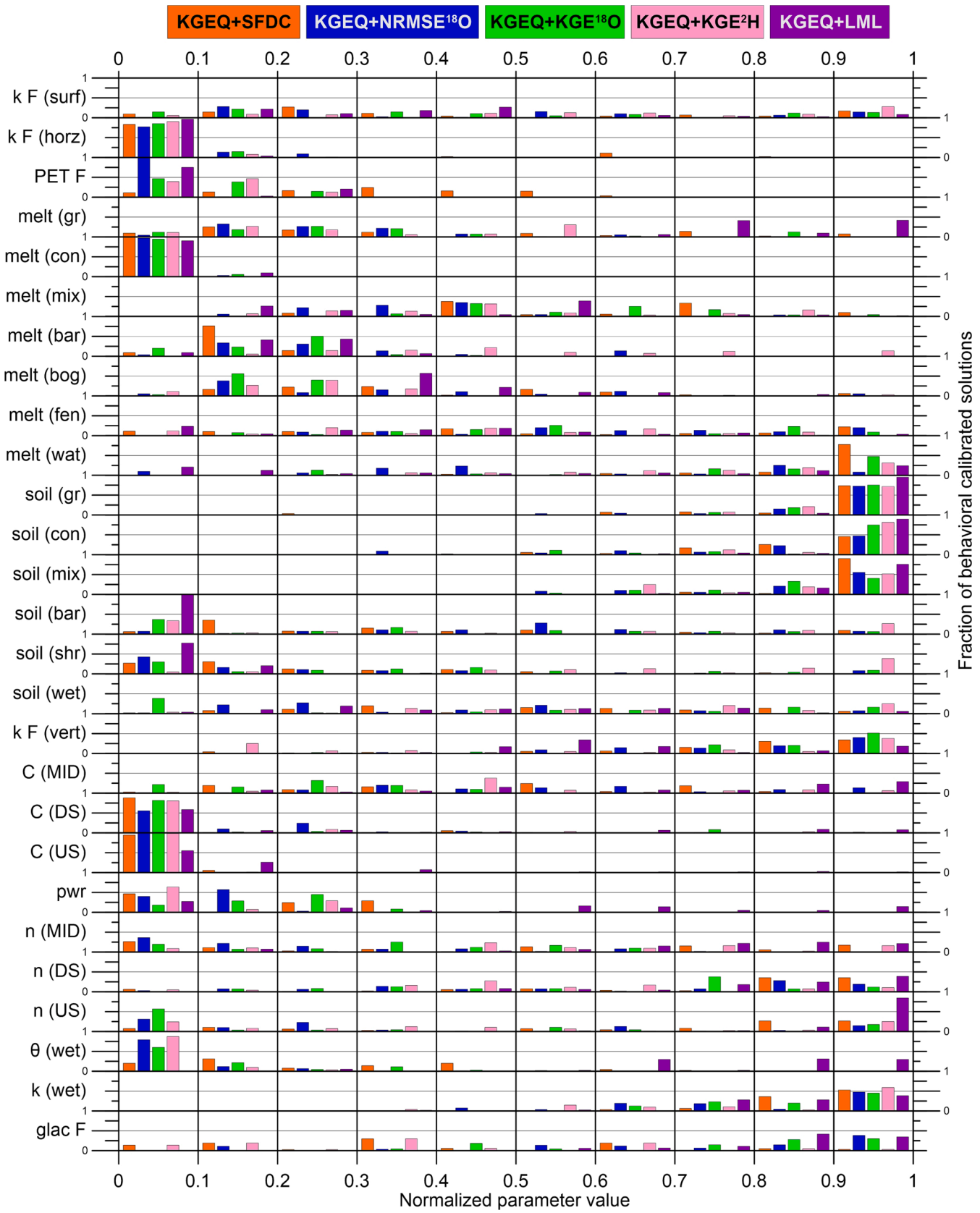

Fig. 9. Parameter identifiability for behavioral calibrated solutions. All 27 normalized parameters' histograms are shown for the 5 candidate objective metric combinations.

Table 3
Average containment ratios and relative band width for all behavioral calibrated ensembles, with percent contributions from the upper and lower zones to the total streamflow.

Objectives:		KGEQ/ SFDC	KGEQ/ NRMSEO	KGEQ/ KGEO	KGEQ/ KGEH	KGEQ/ LML
Behavioral solutions		144	110	143	124	165
Containment Ratio	Calibration	0.34	0.44	0.45	0.42	0.28
	Validation	0.29	0.33	0.31	0.33	0.23
	Difference from flow-only (p-value)	–	0.001	0.026	0.001	0.005
Relative Band-Width	Calibration	0.86	0.92	0.87	0.96	0.81
	Validation	1.28	1.24	1.01	1.39	0.81
	Difference from flow-only (p-value)	–	0.809	0.065	0.019	0.025
Lower zone (% contribution)	Average	62	61.4	63.6	59.7	63.4
	Range (max/min)	26.3 (69/43)	19.9 (71/51)	15.9 (71/55)	14.4 (67/53)	16.2 (71/55)
Upper zone (% contribution)	Average	7.6	6.8	5.4	6.4	5.6
	Range (max/min)	25.8 (29/3)	12.1 (15/3)	9.9 (13/3)	10.7 (13/3)	11.5 (14/3)

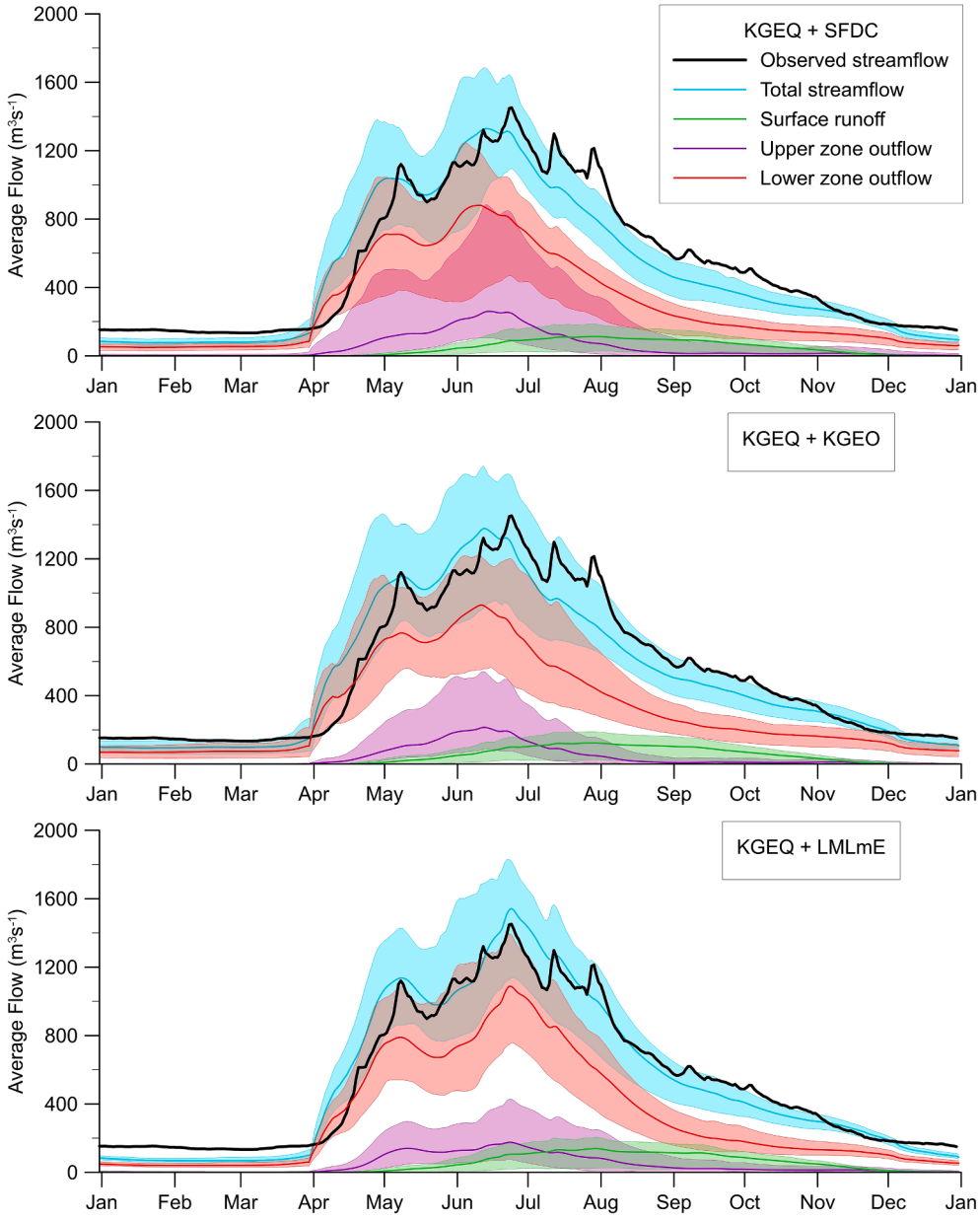

Fig. 10. Average annual hydrographs for the simulation period (2001 to 2015), with behavioral calibrated ensembles shown, along with surface, upper and lower zone virtual tracer flows at Fort McMurray 07DA001 (mean and absolute range are included).

model complexity, number of calibrated parameters, and large and heterogeneous study area. The best identified parameters for both flow and isotope data were the baseflow parameters, which had generally constrained values regardless of the assessment method. That baseflow is the best identified process is entirely unsurprising, as the majority of total streamflow is routed through the lower soil zone as baseflow, and the process is controlled by a limited number of parameters. The snow melt rate parameter for coniferous forest (the dominant land class) was well-identified by most flow metrics, including KGE Q, but was not well-identified by isotope tracer metrics. On the other hand, the PET factor is substantially better identified by isotope tracer metrics than flow metrics alone. In some cases, parameter values may be best identified by a combination of flow and isotope tracer simulation performance. The values of horizontal soil conductivity and wetland porosity are most constrained when both flow and isotopes are optimized. Isotope and flow data only identified different values for the same parameter for upper zone soil water retentions, with the best flow performance associated with very high retention, and the best isotope performance associated with low retention. It is entirely possible to have isotope simulation performance of approximately similar quality with either low or high retention but if flow simulation performance is not considered in parameter value selection, lower retention values are identified. Altogether, isotope tracer data, whether δ^2H or $\delta^{18}O$, does not lead to dramatic alterations in parameter identifiability, but the limited changes do tend to be improvements. With so many interacting parameters within the isoWATFLOOD model, individual parameter identifiability is weak, even with additional tracer data. A simpler, more conceptual model than isoWATFLOOD would have more identifiable parameter values. It is also important to note that the choice of parameter value limits for calibration can impact parameter identifiability when identifiability is defined in terms of the potential parameter value range, as was done in this study. Excessively wide limits can artificially increase identifiability, while overly restricted limits on parameter values either mask identifiability (limits overlap optimal parameter values) or overestimate identifiability (optimal values fall outside limits). This study has relied on developer recommendations for limiting parameter values in calibration and the shortage of identifiability from the Monte Carlo results suggest that parameter limits were not excessively broad, but the tendency for well-identified parameter values to approach either the upper or lower bound in the MOO results suggests that some optimal parameter values lie outside the specified limit. These parameters may be less identifiable than categorized based on range and IQR relative to total potential parameter range, however, if the optimization can push parameter values to the limit, the parameter is not unidentifiable.

The addition of isotope performance to the calibration was more adept at constraining the overall process contributions to streamflow (Fig. 10) than identifying parameter values. The use of isotope tracer data as a calibration objective avoided many outlying flux simulations and flow component combinations that were included by calibrations using only flow data. It is not unexpected that fluxes are better constrained than parameter values, in models with large numbers of interacting processes and parameters. Given parameter identifiability is typically desired as a proxy for process identifiability, it is preferable to have better process identification without better parameter value identification than the reverse.

4.2. Simulation performance

Both flow-only and mixed isotope-flow multi-objective calibrations resulted in numerous acceptable flow simulations (i.e., KGE Q greater than 0.5 in calibration), while calibrating to isotope data alone did not. For predictions in ungauged basins (PUB), should a basin have sparse isotope data, using these as a lone calibration objective may still be a better alternative than purely random parameter value selection within general recommended ranges, as isotope tracers may rule out some of the least accurate flow simulations. That said, caution should be used when interpreting any long-term model simulations as this study demonstrates that isotopes alone are clearly not a reliable substitute for hydrometric data. PA-DDS calibrated solutions using KGE Q as an objective revealed flow simulations that outperform any randomly generated parameter set, while the isotope-only calibrations result in flow simulation performance similar to the random solutions from the Monte Carlo analysis. The Monte Carlo analysis was an inefficient method of identifying behavioral parameter sets for both flow and isotope tracer simulations; numerous higher quality solutions were identified using multi-objective calibration using 1% of the computational budget of the Monte Carlo analysis.

The mixed isotope-flow multi-objective calibrations resulted in solutions with similar flow simulation performance, and better isotope tracer simulation performance compared to the flow-only calibration. The differences between isotope enabled and flow-only calibrations in flow simulation performance for calibration gauges was negligible, and validation performance improved when isotope simulation performance was included as a calibration objective. Considered as ensembles, the mixed isotope-flow calibrations generally outperform the flow-only calibration, with better observation containment for calibrations when an isotope performance metric which include timing error is used as the second objective, and similar relative bounds on isotope-enabled calibrated ensembles. The sub-component ensembles benefited most from the inclusion of isotope tracer data in the calibration, with substantially narrower (more precise) ranges. The addition of either oxygen-18 or deuterium data excludes process-contribution combinations that lead to acceptable total streamflow simulations. The increased confidence in the flow sub-component contributions may be a contributing factor in the positive correlation between isotope simulation performance in calibration and flow simulation validation performance. This study has found a similar interrelationship between isotope tracer performance and flow simulation performance in validation to previous work with the isoWATFLOOD model, but in a different watershed, and using a different calibration methodology (Stadnyk and Holmes, 2020). Overall, isotope tracer data are beneficial to the flow simulation; adding isotope tracer data to the calibration leads to flow simulations comparable or better than simulations calibrated with flow data alone.

4.3. Calibration recommendations

Isotope tracer simulations and observation data have demonstrable value as a supplement to hydrometric data in process-based hydrologic model calibration. If possible, calibration using an isotope simulation performance metric as a secondary objective in a multi-objective optimization is recommended. Results of this study indicate that mixed isotope-flow calibrations result in better streamflow simulation, and improved identifiability of streamflow sub-component simulation than flow-only calibrations.

Monte Carlo simulations are not well-suited to identifying parameter values in large-scale process-based models or for tracer-aided calibrations in such models. Process-based models typically have large numbers of parameters to calibrate and the interactions between these parameters have complex effects on simulation outcomes. Exploring the parameter space of such models (as a Monte Carlo analysis aims to) is computationally onerous, particularly for large-scale models with significant run times, and does not produce better calibration outcomes than multi-objective optimization.

It is important that the isotope performance metric includes timing error for the isotope tracer simulation. Neither KGE nor NRMSE has a decisive advantage over the other as a calibration objective for tracer simulation. This study found no reason not to use NRMSE to compare isotope simulation quality, as it was equivalent to using KGE for improving isotope and flow model performance. Using KGE to quantify tracer simulation performance has the advantage of aligning with a common flow performance metric, but the metric is more vulnerable to sampling gaps or biases than NRMSE. The variability component of the

KGE metric is particularly unreliable when sampling programs do not include seasons (e.g., no winter sample collection) or have disproportionate numbers of samples in some periods (e.g., intensive summer sample collection for a student research project). Previous research found no benefit from calibrating KGE for both isotope tracers simultaneously, as the oxygen-18 and deuterium simulations are correlated and contain roughly the same information for calibration (Holmes et al., 2020). The results of this study show no decisive differences between calibrations using oxygen-18 and calibrations using deuterium: both isotopes improved ensemble and validation performance, increased confidence in flow component simulations, and neither lead to substantial change in parameter identifiability. The choice between stable isotope tracers (i.e., oxygen-18 or deuterium) is dependant on the model or the data available; in this study, oxygen-18 is preferrable as it can be simulated independently (the current isoWATFLOOD model code is limited to either oxygen-18 or both isotopes simultaneously), which may not be the case for other modeling packages.

In summary:

1. Use an isotope tracer performance metric as a second objective in multi-objective optimizations of process-based hydrologic models
2. Choose an isotope tracer simulation performance metric that includes timing error
   - o Consider KGE for consistently sampled isotope data sets
   - o Use NRMSE for irregularly sampled isotope data sets

5. Conclusions

Isotope tracer simulations, and specifically calibrations including tracer model performance optimization, are beneficial for both process representation and streamflow simulations. These benefits, including better ensemble observation containment, higher validation performance scores for flow simulations, and improved process contribution identification, can be realized even in mesoscale watersheds, and with limited isotope observation datasets. Flow data should remain the primary focus of hydrologic model calibration but using an isotope simulation performance metric which includes timing error as a secondary calibration objective leads to more robust streamflow modeling.

CRediT authorship contribution statement

Tegan L. Holmes: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Visualization. **Tricia A. Stadnyk:** Conceptualization, Methodology, Resources, Writing – review & editing, Funding acquisition. **Masoud Asadzadeh:** Methodology, Writing – review & editing. **John J. Gibson:** Resources, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Holmes, Tegan (2023), "Research data for 'Guidance on large scale hydrologic model calibration with isotope tracers'", Mendeley Data, V1, https://data.mendeley.com/datasets/h5p6mgych7/1.

Acknowledgments

The authors acknowledge this study occurred within and about Treaty 8 and 6 regions, lands which are, or have historically been, home to no less than nine Indigenous peoples of Canada: the Dane-zaa, Sekani, Secwepemc (Shuswap), Salish, Ktunaxa, Nakoda/Stoney, Woodland Cree, Chipewyan (Denesoline), and Métis. The authors gratefully acknowledge those who have contributed to data collection required to conduct this study, including the Water Survey of Canada, Environment and Climate Change Canada and Innotech Alberta. This research was supported by the Natural Sciences and Engineering Research Council of Canada [CRD 462584-2013], and Global Water Futures [NSERC CFREF-GWF 418474].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2023.129604.

References

Acero Triana, J.S., Chu, M.L., Guzman, J.A., Moriasi, D.N., Steiner, J.L., 2019. Beyond model metrics: The perils of calibrating hydrologic models. J Hydrol (Amst) 578, 124032.

Ala-aho, P., Tetzlaff, D., McNamara, J.P., Laudon, H., Soulsby, C., 2017. Using isotopes to constrain water flux and age estimates in snow-influenced catchments using the STARR (Spatially distributed Tracer-Aided Rainfall–Runoff) model. Hydrol. Earth Syst. Sci. 21, 5089–5110. https://doi.org/10.5194/hess-21-5089-2017.

Ala-aho, P., Soulsby, C., Pokrovsky, O.S., Kirpotin, S.N., Karlsson, J., Serikova, S., Vorobyev, S.N., Manasypov, R.M., Loiko, S., Tetzlaff, D., 2018. Using stable isotopes to assess surface water source dynamics and hydrological connectivity in a high-latitude wetland and permafrost influenced landscape. J Hydrol (Amst) 556, 279–293. https://doi.org/10.1016/j.jhydrol.2017.11.024.

Alberta Geological Survey, 2013. Bedrock Geology of Alberta [WWW Document]. accessed 9.20.21. https://open.canada.ca/data/en/dataset/5155d48c-ce34-4493-b4f6-fb4eb94fb348.

Asadzadeh, M., Tolson, B., 2013. Pareto archived dynamically dimensioned search with hypervolume-based selection for multi-objective optimization. Engineering Optimization 45 (12), 1489–1509.

Asadzadeh, M., Tolson, B.A., Burn, D.H., 2014. A new selection metric for multiobjective hydrologic model calibration. Water Resour Res 50, 7082–7099. https://doi.org/10.1002/2013WR014970.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environmental Modelling and Software 40, 1–20. https://doi.org/10.1016/j.envsoft.2012.09.011.

Beven, K., Binley, A., 1992. The future of distributed models: Model calibration and uncertainty prediction. Hydrol Process 6, 279–298. https://doi.org/10.1002/hyp.3360060305.

Brooks, J.R., Mushet, D.M., Vanderhoof, M.K., Leibowitz, S.G., Christensen, J.R., Neff, B. P., Rosenberry, D.O., Rugh, W.D., Alexander, L.C., 2018. Estimating Wetland Connectivity to Streams in the Prairie Pothole Region: An Isotopic and Remote Sensing Approach. Water Resour Res 54, 955–977. https://doi.org/10.1002/2017WR021016.

Clark, M.P., Bierkens, M.F.P., Samaniego, L., Woods, R.A., Uijlenhoet, R., Bennett, K.E., Pauwels, V.R.N., Cai, X., Wood, A.W., Peters-Lidard, C.D., 2017. The evolution of process-based hydrologic models: Historical challenges and the collective quest for physical realism. Hydrol Earth Syst Sci 21, 3427–3440. https://doi.org/10.5194/hess-21-3427-2017.

Coulibaly, P., Samuel, J., Pietroniro, A., Harvey, D., 2013. Evaluation of Canadian national hydrometric network density based on WMO 2008 standards. Canadian Water Resources Journal 38, 159–167. https://doi.org/10.1080/07011784.2013.787181.

Delavau, C., Chun, K.P., Stadnyk, T., Birks, S.J., Welker, J.M., 2015. North American precipitation isotope (δ18O) zones revealed in time series modeling across Canada and northern United States. Water Resour Res 51, 1284–1299. https://doi.org/10.1002/2014WR015687.

Delavau, C., Stadnyk, T., Holmes, T., 2017. Examining the impacts of precipitation isotope input (δ18O ppt) on distributed, tracer-aided hydrological modelling. Hydrol Earth Syst Sci 21, 2595–2614. https://doi.org/10.5194/hess-21-2595-2017.

Duethmann, D., Blöschl, G., Parajka, J., 2020. Why does a conceptual hydrological model fail to correctly predict discharge changes in response to climate change? Hydrol Earth Syst Sci 24, 3493–3511. https://doi.org/10.5194/hess-24-3493-2020.

Efstratiadis, A., Koutsoyiannis, D., 2010. One decade of multi-objective calibration approaches in hydrological modelling: a review. Hydrological Sciences Journal-Journal Des Sciences Hydrologiques 55 (1), 58–78.

Environment and Climate Change Canada, 2018. Water Survey of Canada: Historical hydrometric data [WWW Document]. accessed 5.26.21. https://wateroffice.ec.gc.ca.

Environment and Climate Change Canada, 2020. Historical climate data [WWW Document]. accessed 5.26.21. https://climate.weather.gc.ca/historical_data/search_historic_data_e.html.

Fatichi, S., Vivoni, E.R., Ogden, F.L., Ivanov, V.Y., Mirus, B., Gochis, D., Downer, C.W., Camporese, M., Davison, J.H., Ebel, B., Jones, N., Kim, J., Mascaro, G., Niswonger, R., Restrepo, P., Rigon, R., Shen, C., Sulis, M., Tarboton, D., 2016. An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. J Hydrol (Amst) 537, 45–60. https://doi.org/10.1016/j.jhydrol.2016.03.026.

Gibson, J.J., Birks, S.J., Yi, Y., 2016a. Stable isotope mass balance of lakes: A contemporary perspective. Quat Sci Rev 131, 316–328. https://doi.org/10.1016/j.quascirev.2015.04.013.
Gibson, J.J., Birks, S.J., Edwards, T.W.D., 2008. Global prediction of δA and δ2H- δ18O evaporation slopes for lakes and soil water accounting for seasonality. Global Biogeochem Cycles 22. https://doi.org/10.1029/2007GB002997.
Gibson, J.J., Yi, Y., Birks, S.J., 2016b. Isotope-based partitioning of streamflow in the oil sands region, northern Alberta: Towards a monitoring strategy for assessing flow sources and water quality controls. J Hydrol Reg Stud 5, 131–148. https://doi.org/10.1016/j.ejrh.2015.12.062.
Gibson, J.J., Yi, Y., Birks, S.J., 2019. Isotopic tracing of hydrologic drivers including permafrost thaw status for lakes across Northeastern Alberta, Canada: A 16-year, 50-lake assessment. J Hydrol Reg Stud 26, 100643.
Gibson, J.J., Holmes, T., Stadnyk, T.A., Birks, S.J., Eby, P., Pietroniro, A., 2020. 18O and 2H in streamflow across Canada. J Hydrol Reg Stud 32, 100754.
Gibson, J.J., Holmes, T., Stadnyk, T.A., Birks, S.J., Eby, P., Pietroniro, A., 2021. Isotopic constraints on water balance and evapotranspiration partitioning in gauged watersheds across Canada. J Hydrol Reg Stud 37, 100878.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J Hydrol (Amst) 377 (1-2), 80–91.
Guse, B., Kiesel, J., Pfannerstill, M., Fohrer, N., 2020. Assessing parameter identifiability for multiple performance criteria to constrain model parameters. Hydrological Sciences Journal 65, 1158–1172. https://doi.org/10.1080/02626667.2020.1734204.
Guse, B., Fatichi, S., Gharari, S., Melsen, L.A., 2021. Advancing Process Representation in Hydrological Models: Integrating New Concepts, Knowledge, and Data. Water Resour Res 57. https://doi.org/10.1029/2021WR030661.
He, Z., Unger-Shayesteh, K., Vorogushyn, S., Weise, S.M., Kalashnikova, O., Gafurov, A., Duethmann, D., Barandun, M., Merz, B., 2019. Constraining hydrological model parameters using water isotopic compositions in a glacierized basin, Central Asia. J Hydrol (Amst) 571, 332–348. https://doi.org/10.1016/j.jhydrol.2019.01.048.
Holmes, T., 2016. Assessing the value of stable water isotopes in hydrologic modeling: a dual isotope approach (MSc). University of Manitoba, Winnipeg.
Holmes, T., Stadnyk, T.A., Kim, S.J., Asadzadeh, M., 2020. Regional Calibration With Isotope Tracers Using a Spatially Distributed Model: A Comparison of Methods. Water Resour Res 56. https://doi.org/10.1029/2020WR027447.
Holmes, T.L., Stadnyk, T.A., Asadzadeh, M., Gibson, J.J., 2022. Variability in flow and tracer-based performance metric sensitivities reveal regional differences in dominant hydrological processes across the Athabasca River basin. J Hydrol Reg Stud 41, 101088. https://doi.org/10.1016/j.ejrh.2022.101088.
Kiang, J.E., Gazoorian, C., McMillan, H., Coxon, G., Le Coz, J., Westerberg, I.K., Belleville, A., Sevrez, D., Sikorska, A.E., Petersen-Øverleir, A., Reitan, T., Freer, J., Renard, B., Mansanarez, V., Mason, R., 2018. A Comparison of Methods for Streamflow Uncertainty Estimation. Water Resour Res 54, 7149–7176. https://doi.org/10.1029/2018WR022708.
Kirchner, J.W., 2006. Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. Water Resour Res 42. https://doi.org/10.1029/2005WR004362.
Kouwen, N., 2018. WATFLOOD/CHARM Canadian Hydrological And Routing Model . Waterloo.
Matott, L.S., 2017. OSTRICH: an Optimization Software Tool. Documentation and User's Guide, Version 17 (12), 19.
Minder, J.R., Mote, P.W., Lundquist, J.D., 2010. Surface temperature lapse rates over complex terrain: Lessons from the Cascade Mountains. J Geophys Res 115, D14122. https://doi.org/10.1029/2009JD013493.
Mizukami, N., Rakovec, O., Newman, A.J., Clark, M.P., Wood, A.W., Gupta, H., v., Kumar, R., 2019. On the choice of calibration metrics for "high-flow" estimation using hydrologic models. Hydrol Earth Syst Sci 23, 2601–2614. https://doi.org/10.5194/hess-23-2601-2019.
Nan, Y., Tian, L., He, Z., Tian, F., Shao, L., 2021. The value of water isotope data on improving process understanding in a glacierized catchment on the Tibetan Plateau. Hydrol Earth Syst Sci 25, 3653–3673. https://doi.org/10.5194/hess-25-3653-2021.
Nash, J.E., Sutcliffe, J., v, 1970. River flow forecasting through conceptual models part I—A discussion of principles. J Hydrol (Amst) 10, 282–290.
Neill, A.J., Tetzlaff, D., Strachan, N.J.C., Soulsby, C., 2019. To what extent does hydrological connectivity control dynamics of faecal indicator organisms in streams? Initial hypothesis testing using a tracer-aided model. J Hydrol (Amst) 570, 423–435. https://doi.org/10.1016/j.jhydrol.2018.12.066.
Oshun, J., Dietrich, W.E., Dawson, T.E., Fung, I., 2016. Dynamic, structured heterogeneity of water isotopes inside hillslopes. Water Resour Res 52, 164–189. https://doi.org/10.1002/2015WR017485.
Pechlivanidis, I.G., Jackson, B.M., Mcintyre, N.R., Wheater, H.S., 2011. Catchment scale hydrological modelling: A review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications. Global NEST journal 13, 193–214.
Peralta-Tapia, A., Sponseller, R.A., Tetzlaff, D., Soulsby, C., Laudon, H., 2015. Connecting precipitation inputs and soil flow pathways to stream water in contrasting boreal catchments. Hydrol Process 29, 3546–3555. https://doi.org/10.1002/hyp.10300.
Piovano, T.I., Tetzlaff, D., Maneta, M., Buttle, J.M., Carey, S.K., Laudon, H., McNamara, J., Soulsby, C., 2020. Contrasting storage-flux-age interactions revealed by catchment inter-comparison using a tracer-aided runoff model. J Hydrol (Amst) 590, 125226. https://doi.org/10.1016/j.jhydrol.2020.125226.
Razavi, S., Gupta, H., v., 2016. A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. Theory. Water Resour Res 52, 423–439. https://doi.org/10.1002/2015WR017558.
Razavi, S., Sheikholeslami, R., Gupta, H., v, Haghnegahdar, A., 2019. VARS-TOOL: A toolbox for comprehensive, efficient, and robust sensitivity and uncertainty analysis. Environmental Modelling & Software 112, 95–107. https://doi.org/10.1016/j.envsoft.2018.10.005.
Refsgaard, J.C., Stisen, S., Koch, J., 2022. Hydrological process knowledge in catchment modelling – Lessons and perspectives from 60 years development. Hydrol Process 36. https://doi.org/10.1002/hyp.14463.
Rosa, L., Davis, K.F., Rulli, M.C., D'Odorico, P., 2017. Environmental consequences of oil production from oil sands. Earths Future 5, 158–170. https://doi.org/10.1002/2016EF000484.
Shafii, M., Tolson, B.A., 2015. Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. Water Resour Res 51, 3796–3814. https://doi.org/10.1002/2014WR016520.
Shangguan, W., Dai, Y., Duan, Q., Liu, B., Yuan, H., 2014. A global soil data set for earth system modeling. J Adv Model Earth Syst 6 (1), 249–263.
Song, X., Zhang, J., Zhan, C., Xuan, Y., Ye, M., Xu, C., 2015. Global sensitivity analysis in hydrological modeling: Review of concepts, methods, theoretical framework, and applications. J Hydrol (Amst) 523, 739–757.
Spangenberg, J.E., 2012. Caution on the storage of waters and aqueous solutions in plastic containers for hydrogen and oxygen stable isotope analysis. Rapid Communications in Mass Spectrometry 26 (22), 2627–2636.
Stadnyk, T.A., Delavau, C., Kouwen, N., Edwards, T.W.D., 2013. Towards hydrological model calibration and validation: simulation of stable water isotopes using the isoWATFLOOD model. Hydrol Process 27, 3791–3810. 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, 1525–1538. https://doi.org/10.1080/02626667.2020.1751847.
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. Hydrol Process 35. https://doi.org/10.1002/hyp.14223.
Tolson, B.A., Shoemaker, C.A., 2007. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. Water Resour Res 43. https://doi.org/10.1029/2005WR004723.
Tunaley, C., Tetzlaff, D., Birkel, C., Soulsby, C., 2017. Using high-resolution isotope data and alternative calibration strategies for a tracer-aided runoff model in a nested catchment. Hydrol Process 31, 3962–3978. https://doi.org/10.1002/hyp.11313.
Viglione, A., Parajka, J., Rogger, M., Salinas, J.L., Laaha, G., Sivapalan, M., Blöschl, G., 2013. Comparative assessment of predictions in ungauged basins - Part 3: Runoff signatures in Austria. Hydrol Earth Syst Sci 17, 2263–2279. https://doi.org/10.5194/hess-17-2263-2013.
Vitt, D.H., Halsey, L.A., Zoltai, S.C., 2000. The changing landscape of Canada's western boreal forest: the current dynamics of permafrost. Canadian Journal of Forest Research 30 (2), 283–287.
Xiong, L., Wan, M., Wei, X., O'Connor, K.M., 2009. Indices for assessing the prediction bounds of hydrological models and application by generalised likelihood uncertainty estimation. Hydrological Sciences Journal 54, 852–871. https://doi.org/10.1623/hysj.54.5.852.