Journal of Hydrology: Regional Studies 41 (2022) 101088

Contents lists available at ScienceDirect

Journal of Hydrology: Regional Studies

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

Variability in flow and tracer-based performance metric sensitivities reveal regional differences in dominant hydrological processes across the Athabasca River basin

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 3R4, Canada

ARTICLE INFO

Keywords:
Hydrologic modeling
Parameter sensitivity
Process simulation
Performance metrics
Isotope tracers

ABSTRACT

Study region: Athabasca River basin, Alberta, Canada (156,000 km^2).
Study focus: Hydrology often relies upon hydrologic models in data-sparse regions; however, it is unclear if such models are reliably accurate, or if internal process simulations are reasonable representations of watershed function. Standard model evaluation and calibration approaches often prioritize accurate reproduction of recorded streamflow, ignoring process simulation fidelity, regardless of the intended model application. This study evaluates whether combined use of streamflow and isotope tracer performance metrics can improve representation of simulated streamflow-generating processes within a large river basin, the Athabasca watershed, to inform calibration of a process-based, distributed hydrologic model.
New hydrological insights for the region: Flow-based performance metrics were found to be sensitive to processes influencing streamflow volume and timing, but insensitive to internal flow paths and storage volumes. Although somewhat less reliable than flow metrics, isotope tracer performance metrics are found to be most sensitive to processes influencing mixing and water age, and appreciably responsive to many other processes. We demonstrate that process-based hydrologic models for rivers such as the Athabasca River cannot be optimally calibrated using streamflow metrics alone, as such optimizations cannot tune parameters or process representations to which the objective function is insensitive. Importantly, isotope tracers have demonstrable value for informing process-based hydrologic model optimization by providing a window into the subsurface black box within complex regional-scale simulations.

1. Introduction

Hydrologic models are broadly used to simulate flow generating processes in watersheds, typically with the goal of producing runoff and streamflow assessments. The flow timing and water volumes from these assessments affect predictions of ecosystem function and resilience, water supply and hydroelectric generation, and the extent of damage resulting from flooding and drought (Carlisle et al., 2011; Wan et al., 2021; Buttle et al., 2016). An accurate prediction of streamflow in both the short-term and long-term

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

https://doi.org/10.1016/j.ejrh.2022.101088
Received 31 May 2021; Received in revised form 5 April 2022; Accepted 14 April 2022

(for climate change projections) is therefore of key importance for water resources operations and planning. Standard methods of evaluating models tend to judge simulation quality based on the accurate reproduction of historical streamflow, while the fidelity of process simulation is often ignored, regardless of the intended use of the simulation results (Clark et al., 2011).

Increasing the physical basis of a model to improve process simulation and decrease model dependence on historical data is intuitively attractive, but also necessitates higher resolution input data and increases the computational demand for running the model (Clark et al., 2017; Peters-Lidard et al., 2017). Accurately modeling streamflow is particularly challenging in remote regions as data availability is generally low: streamflow and weather gauges are rare and have limited record lengths (Coulibaly et al., 2013). Data limitations increase the uncertainty, but also the need, for hydrologic modeling in mid- to high latitude regions with limited accessibility. To reliably simulate flows in ungauged basins, or to predict flows under non-stationary climatic conditions, it is critical for hydrologic models to accurately represent the physical processes generating streamflow (Duethmann et al., 2020). Information on individual hydrologic processes is even more rare than weather or hydrometric data, and limited accessibility in remote regions stalls the expansion of data networks.

Additional observations, such as stable isotope tracer data (i.e. ratios of water molecules containing ^{18}O or ^{2}H to standard water), have been used to add information on hydrologic processes. Stable isotope tracers are naturally occurring, non-reactive tracers of water source and processes resulting from their variable occurrence in precipitation and evaporating water bodies (Birkel and Soulsby, 2015; Bowen et al., 2019). A hydrologic model capable of simulating both flow and isotope tracer composition can therefore be more robustly evaluated against additional observed data. A few models have already combined isotope and flow simulations, such as the isoWATFLOOD model (Stadnyk et al., 2013), the IWBMIso model (Belachew et al., 2016), or the STARR model (van Huijgevoort et al., 2016). Linked hydrologic-tracer simulations can be compared to both flow and tracer observations, further expanding the options for model evaluation, simulation performance metric choice, and the identification of more physically meaningful model parameter values during model calibration (Holmes et al., 2020; Stadnyk and Holmes, 2020; Tunaley et al., 2017; Yamanaka and Ma, 2017).

Model performance metrics are used to evaluate model accuracy, or as objective functions in automated model calibration algorithms. They are broadly categorized as either residual error metrics, based on aggregation of differences between simulated and observed data point pairs; or data set comparison metrics, based on differences between a population property of the observed and the simulated data sets (Bennett et al., 2013). Despite the wealth of literature on performance metrics, there remains no general consensus on the best performance metric(s) to evaluate either flow or isotope tracer simulations, and moreover, these two types of data typically differ in temporal resolution and consistency of sampling, which can significantly impact metric accuracy and therefore selection (Bennett et al., 2013; Mizukami et al., 2019). The Kling-Gupta (KGE) and Nash-Sutcliffe efficiency (NSE) metrics are frequently used in the literature, but flow signature metrics are also shown to improve the evaluation of specific flow simulation characteristics and identify hydrologically consistent parameter sets (Shafii and Tolson, 2015; Knoben et al., 2019; Sahraei et al., 2020). Isotope simulations are most frequently evaluated using some variant of a residual error metric, but other metrics, such as the Kling-Gupta efficiency have also been applied (He et al., 2019; Tunaley et al., 2017). These metrics all evaluate the model performance at local points (either individually or as an averaged performance) using integrated (or cumulative) data, as both flow and flow tracers are the final summation of a multitude of hydrologic processes across a watershed.

A process-based model can potentially be verified along individual flow paths, as such processes are both simulated and intended to match real-world fluxes. It cannot be assumed that metrics designed to evaluate cumulative model performance will also be capable of evaluating or identifying the individual processes contributing to flow (Oreskes et al., 1994). Previous research on process-based model calibration identifies that streamflow performance only informs a sub-set of simulated processes (Acero Triana et al., 2019; Newman et al., 2017), and that process sensitivity to model performance varies seasonally (Bajracharya et al., 2020; Pfannerstill et al., 2015; Wagener et al., 2003). The literature is, however, short in analyzing the processes tracer performance metric are sensitive to. The actual capacities of metrics (for both streamflow and tracer simulations) to react to changes in the internal simulation of critical hydrologic processes (on both inter- and intra-annual time scales) would be of considerable utility in designing calibration strategies for process-based models (Mizukami et al., 2019). Sensitivity analyses are well-adapted to address this point, and relative sensitivities of model parameters have previously been used to inform model calibration (Razavi and Gupta, 2015; Song et al., 2015; Haghnegahdar et al., 2017). Sensitivity analyses are distinct from model calibration, as they do not identify optimal or even necessarily good parameter values, but rather they identify linkages between parameters and metrics.

This study will evaluate whether performance metrics respond to changes in simulated streamflow-generating processes for the purposes of guiding hydrologic modeling choices. To this end, global sensitivity analyses are utilized to answer the following questions:

1. Which processes are flow simulation performance metrics sensitive to, and is there any temporal or spatial variability in this sensitivity; and
2. Are there processes which isotope tracer metrics are sensitive to that streamflow metrics are insensitive to, and vice-versa.

These results will be used to assess the value of various metrics or datasets in adequately informing the calibration of a process-based hydrologic model, rather than comparing calibration outputs as has been done previously. Our aim is to provide guidance for the selection of performance metrics in tracer-aided calibration, and an awareness of inherent trade-off between traditional flow-based calibration and tracer-aided calibration. Our study focuses on the Canadian Oil Sands region in Alberta, Canada, where it is critical to assess the reliability of water supply forecasts given this region is undergoing significant future change resulting from anthropogenic development and climate change, including glacial retreat, permafrost thaw and increased forest fires (Gibson et al., 2019a, 2019b; Nenzén et al., 2020; Stahl et al., 2008). Understanding the hydrology and water supply of this region is a key goal of the Alberta Oil Sands Monitoring strategy (Government of Canada, 2021).

2. Methods

2.1. Athabasca River basin

The Athabasca River runs north-east from the Rocky Mountains to Lake Athabasca and the Peace-Athabasca Delta. It is the most southerly part of the Mackenzie River basin (Fig. 1). The Athabasca River watershed is located in the north of the Canadian provinces of Alberta and Saskatchewan, on Treaty 6 and 8 territory. The total watershed area is 156,000 km^2; elevations and land use vary widely from upstream to down. The upper reaches of the Athabasca are alpine or foothills regions, with steep slopes and some glaciers, most notably the Athabasca Glacier in the Columbia Icefield (Intsiful and Ambinakudige, 2021). The lower reaches, which coincide with the Athabasca Oil Sands region, have subdued relief and abundant wetlands. The soils in the basin are primarily loam, with higher clay prevalence in the mid-reach, some sandy or coarse soil in the downstream region and some exposed rock or shallow soil in the upstream areas (Shangguan et al., 2014). The rock underlying the Athabasca basin is predominately sedimentary stone from the Cretaceous and Paleogene, with some older rocks exposed in the Rocky Mountains and a small area with Precambrian granite in the north (Alberta Geological Survey, 2013). Substantial agricultural activity occurs between the upper and lower reaches; boreal coniferous forests are prevalent throughout the basin. Minimal quantities of water are diverted for agriculture, while approximately 1% of annual flow is used for activities in the oil sands (Rosa et al., 2017). There is some sporadic permafrost in the region which is actively degrading; deeper bedrock formations contribute small amounts of flow to the Athabasca River and its tributaries (3–5% of annual flow) (Gibson et al., 2016; Vitt et al., 2000). The landscape features of the Athabasca River basin, and its upstream, middle and downstream regions are summarized in Table 1.

The climate of the Athabasca watershed is highly seasonal; mean monthly temperatures (averaged across the entire basin) range from − 19 °C to + 17 °C, with a mean annual temperature of 0 °C over the study period (2002–2015). The long-term average annual precipitation is 450 mm; in the downstream reaches, approximately 60% of precipitation falls as rain, but the upstream reaches are colder than the basin average and a larger fraction falls as snow (Environment and Climate Change Canada, 2020).

The Athabasca River basin is, in many respects, an ideal watershed case study for the utility of isotope tracers in large-scale hydrologic modeling. The basin contains a wide range of elevations and land cover, from the glacial headwaters, through mixed-use grasslands and ending in wetland-dominated boreal forest, within a moderately-sized basin. The rivers in the Athabasca watershed are not regulated by any major reservoirs or hydro-electric developments. The Athabasca River basin is also relatively accessible, compared to many mid- to high-latitude watersheds, and oil sands developments have led to expanded research and longer-term monitoring in the region.

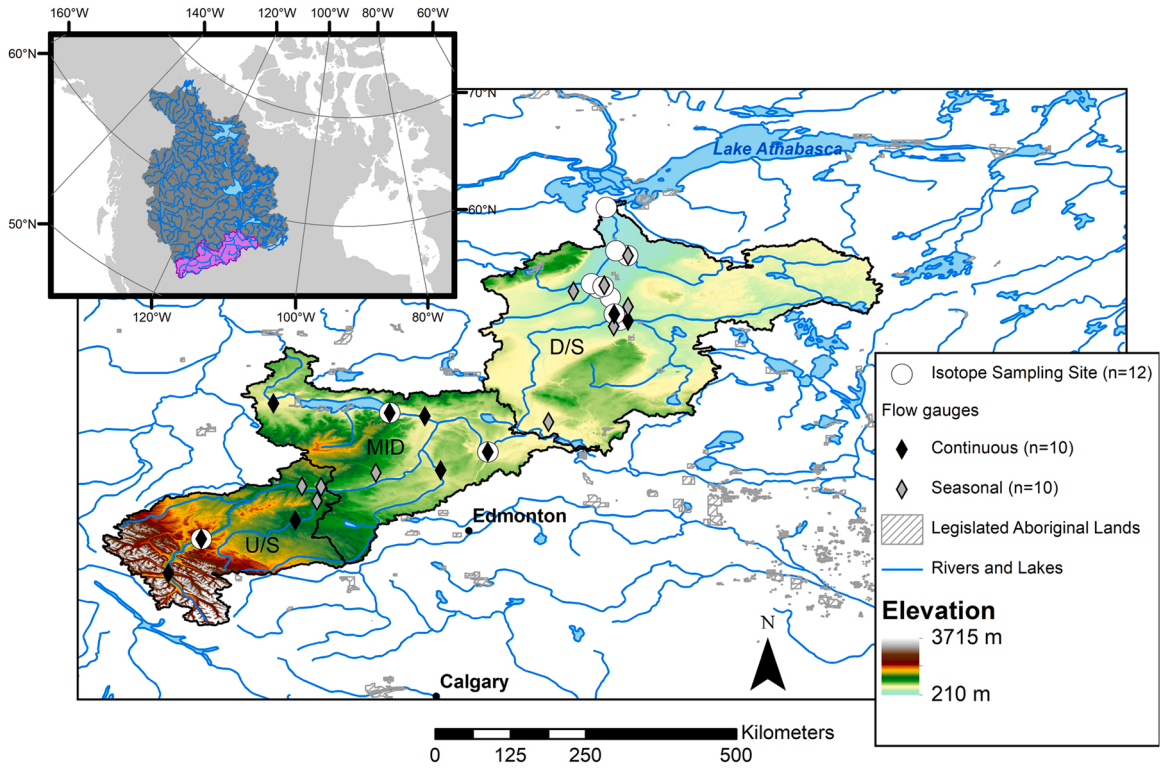

Fig. 1. The Athabasca watershed with Water Survey of Canada flow gauges and sampling sites for isotope compositions of streamflow. The Mackenzie River basin with the Athabasca River watershed highlighted is shown in the inset.

Table 1
Topographic, soil and land cover data summary for the Athabasca River basin and its upstream (U/S), mid-reach (MID), and downstream (D/S) regions as in this study (see Fig. 1 for region boundaries).

		All	U/S	MID	D/S
Area (km^2)		156,000	32,200	46,000	77,700
Slope (%)	Average	0.35	1.21	0.28	0.15
Elevation (m)	Maximum	3715	3715	1379	866
	Minimum	211	689	494	211
	Mean	659	1375	725	522
Soil (%)	Sand/coarse	9.4	0.0	0.0	19.0
	Loam (low clay)	62.0	85.0	44.9	62.5
	Loam (with clay)	25.1	14.7	49.0	15.2
	Clay/clay mix	3.5	0.2	6.1	3.4
Land cover (%)	Grass	8.1	5.3	23.0	0.4
	Wetland	11.0	4.0	12.3	13.2
	Mixed Wood	15.1	10.8	30.7	7.5
	Coniferous	54.1	65.2	25.0	66.8
	Shrub	6.2	7.1	4.1	7.1
	Impervious	0.0	0.1	0.0	0.0
	Barren	1.4	5.9	0.0	0.4
	Water	3.8	0.8	4.7	4.6
	Glacier	0.2	0.9	0.0	0.0

2.2. Hydrologic model setup & parameterization

2.2.1. Model

The Athabasca River watershed was modeled using CHARM/WATFLOOD and its associated dual-isotope simulation model, iso-WATFLOOD. CHARM is an open source, distributed hydrologic model with a mixture of physically based and conceptual process representation (Kouwen, 2018). The isotope tracer models for CHARM simulate the isotopic concentrations of oxygen-18 and deuterium in all of the storages and fluxes used in the original hydrologic model; individual hydrologic storages are assumed to be completely mixed through depth, and fluxes generally have the same concentration as the source storage, except evaporative fluxes, which are subject to isotopic fractionation (Stadnyk and Holmes, 2020). Both the hydrologic and tracer simulations run on an hourly time-step, with daily simulated model output.

The Athabasca River basin model divides the watershed area into 320 grid cells, with a nominal cell size of 0.4° longitude by 0.2°

Table 2
List of potential significant parameters included in the sensitivity analysis, including the parameter names, the process affected by the parameter and the affected GRU with a list of the decoupled land classes to which coupled parameters are applied.

Parameter description	Parameter name	Internal name	Hydrologic Process	Applicable GRU	Decoupled classes
Surface soil conductivity	k F (surf)	ak	Infiltration	All soil-based	–
Horizontal upper soil zone conductivity	k F (horz)	rec	Interflow	All soil-based	–
PET to AET factor	PET F	fpet	Evaporation	Water, connected wetland	–
Snowmelt rate factor	melt rate	fm	Snowmelt	All	low vegetation (grass+shrub), coniferous, mixed, bare (barren+impervious), disconnected wetland, connected wetland, water
Upper soil zone soil water retention cap	soil ret	retn	Soil storage and ET	All soil-based	grass, coniferous, mixed, barren, shrub, wetland
Vertical upper soil zone conductivity	k F (vert)	ak2	Recharge	All soil-based	–
Baseflow equation constant	C	flz	Baseflow	All soil-based	Upstream, mid-basin, downstream
Baseflow equation power	pwr	pwr	Baseflow	All soil-based	–
Channel roughness factor	n	r2n	Channel velocity	Water	Upstream, mid-basin, downstream
Wetland porosity	θ (wet)	theta	Wetland storage	Connected wetland	–
Wetland conductivity	k (wet)	kcond	Wetland velocity	Connected wetland	–
Glacier melt factor	glac F	gladjust	Glacier melt	Glacier	–

latitude (actual cell sizes are adjusted based on drainage area); each cell is subdivided into 10 grouped response unit (GRU) types, based on land cover data from the ESA (European Space Agency, 2017). The majority of these GRU types are modeled with soil layers, namely the grass (8.1%), coniferous (54.1%) and mixed forest (15.1%), shrub (6.2%), disconnected wetland (8.8%), and barren (1.4%) classes. The glacier (0.2%) and impervious (0.03%) classes have no modeled soil storages (all rain and snowmelt becomes direct runoff), and glacier GRU also generate glacier melt flows. Open water (3.8%) and wetlands connected to the stream network (2.2%) also have no soil storages, but rain or snowmelt is added directly to the wetland, channel or lake rather than running off.

2.2.2. Process representation

The CHARM/WATFLOOD model has two soil layers, both of which can generate sub-surface flows to the channel network. Rain or snowmelt can either infiltrate to the upper soil zone or runoff. Water in the upper soil zone may recharge the lower soil zone, flow out to the channel network or connected wetlands, or evapotranspire. Water in the lower soil zone may only flow out to the channel network or connected wetlands. All types of GRU have potential snowpack storages. Snow and glacier melt rates are calculated as a function of air temperature and melting snowpacks cover fractional areas of each GRU. The upper soil zone under a snowpack is considered frozen, and all soil fluxes (infiltration, recharge and interflow) have substantially reduced rates in frozen soils, but permafrost is not included in the model. Connected wetlands, where present, collect outflows from all GRU with soil layers, and have bi-directional flow with the channel network, with direction determined by the relative water levels. More detailed descriptions and full equations can be found in Holmes (2016).

All processes listed above have parameters controlling the simulated flux; parameter values can be consistent across GRU, or separate GRU can have different parameter values specific to that class. As a mixed physically based and conceptual model, there are a very large number of parameters which can be altered in setting up and calibrating a watershed model (over 250 for the Athabasca model). However, the vast majority have minimal impact on the simulation (e.g. overland flow roughness factors), or can be estimated from the literature (surface depression storage caps). This study will focus on potentially significant parameters identified by previous studies, including Holmes et al. (2020), and model developer recommendations, with a minimum of one parameter per simulated process, as listed in Table 2.

2.3. Meteorological data

The hydrologic and isotope tracer models were run using four meteorological forcings: hourly air temperature and humidity, daily total precipitation and monthly average isotopic compositions of precipitation. The precipitation, temperature and humidity forcings were based on observations at Environment and Climate Change Canada (ECCC) weather stations (Environment and Climate Change Canada, 2020). Forcing data for each grid cell at each time step were estimated using inverse distance squared weighting, with a temperature lapse rate of − 5 °C/km and a precipitation lapse rate of 0.2 mm/km; 56 weather stations were included in the calculation, provided there was observation data for that time interval (Minder et al., 2010; Kouwen, 2018). The isotopic compositions of precipitation were estimated from the empirical model developed by Delavau et al. (2015), which uses a geospatial interpolation and a

Table 3
Hydrometric gauges and isotope sampling sites in the Athabasca River basin.

		Latitude (°)	Longitude (°)	Isotope Samples	Drainage Area (km^2)	Operation Schedule
07AA002	ATHABASCA RIVER NEAR JASPER	52.91	-118.06		3,870	Continuous
07AD002	ATHABASCA RIVER AT HINTON	53.42	-117.57	159	9,760	Continuous
07AE001	ATHABASCA RIVER NEAR WINDFALL	54.21	-116.06		19,600	Seasonal
07AG004	MCLEOD RIVER NEAR WHITECOURT	53.99	-115.84		9,110	Seasonal
07AG007	MCLEOD RIVER NEAR ROSEVEAR	53.70	-116.16		7,140	Continuous
07AH001	FREEMAN RIVER NEAR FORT ASSINIBOINE	54.41	-114.96		1,660	Seasonal
07AH003	SAKWATAMAU RIVER NEAR WHITECOURT	54.20	-115.78		1,150	Seasonal
07BC002	PEMBINA RIVER AT JARVIE	54.45	-113.99		13,100	Continuous
07BE001	ATHABASCA RIVER AT ATHABASCA	54.72	-113.29	146	74,600	Continuous
07BF002	WEST PRAIRIE RIVER NEAR HIGH PRAIRIE	55.45	-116.49		1,150	Continuous
07BK001	LESSER SLAVE RIVER AT SLAVE LAKE	55.31	-114.76	17	13,600	Continuous
07BK007	DRIFTWOOD RIVER NEAR THE MOUTH	55.26	-114.23		2,100	Continuous
07CA006	WANDERING RIVER NEAR WANDERING RIVER	55.17	-112.39		1,120	Seasonal
07CD001	CLEARWATER RIVER AT DRAPER	56.68	-111.20	44	30,800	Continuous
07CD004	HANGINGSTONE RIVER AT FORT MCMURRAY	56.60	-111.41		960	Seasonal
07DA001	ATHABASCA RIVER BELOW MCMURRAY	56.78	-111.40	126	133,000	Continuous
07DA006	STEEPBANK RIVER NEAR FORT MCMURRAY	56.89	-111.20	37	1,320	Seasonal
07DA008	MUSKEG RIVER NEAR FORT MACKAY	57.21	-111.55	70	1,460	Seasonal
07DB001	MACKAY RIVER NEAR FORT MACKAY	57.12	-112.01	26	5,570	Seasonal
07DC001	FIREBAG RIVER NEAR THE MOUTH	57.65	-111.20	44	6,500	Seasonal
07DD011	ATHABASCA RIVER AT OLD FORT	58.37	-111.52	120	160,000	–
AB07DA0750	ELLS RIVER	57.30	-111.68	36	2,500	–
AB07DA0980	ATHABASCA RIVER U/S FIREBAG	57.72	-111.38	68	154,400	–

multiple linear regression of geographic and climatic indicators. No field measurements of isotopes in precipitation within the Athabasca basin boundaries were used in the development or validation of the geospatial isotope model, but meteoric water samples from both immediately south and north of the watershed were included (Delavau et al., 2011). The climate zone models covering the Athabasca basin had modeled precipitation residual IQR of 3.6 and 4.7‰ (for $\delta^{18}O$) in validation and the model adequately captured the seasonality of isotopes in precipitation (i.e. highly depleted precipitation in winter and annual variation of 15‰ for $\delta^{18}O$) (Delavau et al., 2015).

2.3.1. Flow and isotope data

Historical hydrometric data from the Water Survey of Canada were used to calculate model performance metrics (Environment and Climate Change Canada, 2018). A total of 20 continuous or seasonal (i.e. continuous only during the open water season) hydrometric stations with daily data (m^3s^{-1}) between 2002 and 2015 were used in the analysis, listed in Table 3 (see Fig. 1 for spatial distribution, and Appendix A for average annual hydrographs and isotope sample data). Gauged areas ranged between 960 and 133,000 km^2. The uncertainty in the streamflow data are approximately ± 10% on average, with higher uncertainty during peak flow and ice-on periods (Kiang et al., 2018; Westerberg et al., 2020).

A monthly water isotope sampling campaign on the Athabasca River and several tributaries was conducted for the Alberta Environmental Monitoring, Evaluation and Reporting Agency's Long-Term River Network monitoring program (Gibson et al., 2016). Sampling at hydrometric gauges in the Athabasca basin began in 2002, and continued through 2014, with variable sampling frequency; some years, sampling occurred approximately monthly, while some gauges have data gaps longer than one year. All water samples were sealed in 30 mL high-density polyethylene bottles and analyzed at either the University of Waterloo Environmental Isotope Laboratory or at Alberta Innovates Technology Futures, Victoria (Gibson et al., 2016). High-density polyethylene bottles have been shown to be effective at preventing isotopic fractionation, and all samples were sealed and analyzed within 1 year of sample collection (Gibson et al., 2019a, 2019b; Spangenberg, 2012). Water samples were analyzed using a Micromass IsoPrime Dual Inlet/Gas Chromatograph pre-2009, and from 2009 on, using a Thermo Scientific Delta V Advantage Dual Inlet/HDevice system, with an estimated analytical uncertainty of ± 0.1‰ for oxygen-18 and ± 1‰ for deuterium for both periods (Gibson et al., 2016). Isotope results are reported in δ notation in permil (‰), relative to V-SMOW.

2.4. Performance metrics

A variety of metrics were selected to quantify simulation performance, based on the most commonly applied metrics from the literature. As noted throughout this section, metrics are sensitive to various characteristics of the distribution of the error residuals, and therefore including multiple metrics in model evaluations is expected to expand the number of sensitive parameters. Only simulated data on days that have flow or isotope observations are considered for calculating these performance metrics. Firstly, the normalized root mean square error was used for both the flow and isotope simulations, calculated as:

$$NRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{s,i} - x_{o,i}\right)^2 \Big/ \overline{x_0}} \tag{1}$$

Where n is number of observations, $x_{o,i}$ is observation i, $x_{s,i}$ is the corresponding simulated value and $\overline{x}_o$ is the observation mean. The traditional Nash-Sutcliffe efficiency (NSE), another residual error metric, and the log transform of NSE were calculated exclusively for the flow simulation (Nash and Sutcliffe, 1970). While *NRMSE* and *NSE* are highly sensitive to large residuals that often happen due to mis-timed simulations in high-flow periods, *logNSE* is more sensitive to small residuals that happen during low-flow periods.

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

$$logNSE = 1 - \frac{\sum_{i=1}^{n}\left(log\left(x_{s,i}\right) - log\left(x_{o,i}\right)\right)^2}{\sum_{i=1}^{n}\left(log\left(x_{o,i}\right) - \overline{log(x_o)}\right)^2} \tag{3}$$

The Kling-Gupta efficiency (*KGE*) metric, and all three of its constituent components (i.e. the correlation r, the relative variability α and the bias β) were used for both the isotope and flow simulations. *KGE* and *NSE* share the same components r,α, and β, but *KGE* gives them the same weight as opposed to *NSE* that relatively undermines variability (Gupta et al., 2009).

$$r = \frac{\sum_{i=1}^{n}\left(x_{o,i} - \overline{x}_o\right)\left(x_{s,i} - \overline{x}_s\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{o,i} - \overline{x}_o\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_{s,i} - \overline{x}_s\right)^2}} \tag{4}$$

$$\alpha = \frac{\sigma_s}{\sigma_o} \tag{5}$$

$$\beta = \frac{\overline{x}_s}{\overline{x}_o} \tag{6}$$

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \tag{7}$$

Where σ_s and σ_o are the standard deviations of the simulated and observed data.

Three popular flow signature metrics were also applied to both the isotope and flow simulations. Eq. 8 was used to calculate the slope of the flow duration curve (*SFDC*) (for flows) and the slope of the duration curve (*SDC*) (for isotopes) (Viglione et al., 2013):

$$SDC = \quad 100\left(\frac{x_{s,30} - \quad x_{s,70}}{40\overline{x}_s} - \frac{x_{o,30} - \quad x_{o,70}}{40\overline{x}_o}\right) \tag{8}$$

Where x_{30} and x_{70} are the data with exceedance probabilities of 30% and 70%. The high- and low-flow signatures, at the 5% and 95% exceedance probabilities were calculated using Eqs. 9 and 10:

$$Q_5 = \quad \frac{x_{o,5} - \quad x_{s,5}}{x_{o,5}} \tag{9}$$

$$Q_{95} = \quad \frac{x_{o,95} - \quad x_{s,95}}{x_{o,95}} \tag{10}$$

Where x_5 and x_{95} are the data with exceedance probabilities of 5% and 95%.

There are multiple other possible permutations of the SDC, and high- and low-flow signatures, which include slopes in log scale, alternate exceedance probabilities, and mean high and low flow comparisons, but these are all fundamentally variants evaluating the same parts of the hydrograph (Bajracharya et al., 2020; Shafii and Tolson, 2015; Yilmaz et al., 2008). The relative variability and bias components of KGE can also be flow signatures based on discharge statistics (Shafii and Tolson, 2015).

Finally, three metrics quantifying the representativeness of the simulated slope of the isotope-derived local mixing line (LML), which uses both isotope simulations in combination, were included in the analysis: the LML slope error, the LML intercept error and the LML fit error (Stadnyk and Holmes, 2020). The LML slope error is the difference between the best fit slope for the simulated river isotope compositions and the best fit slope for the observed river isotope compositions:

$$LML \quad mE = \frac{\sum_{i=1}^{n}\left(O_{s,i} - \overline{O}_s\right)\left(D_{s,i} - \overline{D}_s\right)}{\sum_{i=1}^{n}\left(O_{s,i} - \overline{O}_s\right)^2} - \frac{\sum_{i=1}^{n}\left(O_{o,i} - \overline{O}_o\right)\left(D_{o,i} - \overline{D}_o\right)}{\sum_{i=1}^{n}\left(O_{o,i} - \overline{O}_o\right)^2} \tag{11}$$

Where $O_{o,i}$ is oxygen-18 observation i, and $O_{s,i}$ is the simulated oxygen-18 value for the time observation i was taken, and $D_{o,i}$ and $D_{s,i}$ are likewise the observation and simulated value for the deuterium data. Similarly, the LML intercept error is the difference between the best fit line intercept for the simulated river isotope compositions and the best fit line intercept for the observed river isotope compositions:

Table 4
Summary of the 29 performance metrics considered, listing which simulation types each metric was applied to, and a qualitative assessment of the simulation error types the metric responds to (filled circles indicate strong responses and empty circles indicate some response).

	Simulation			Error type				
	Flow	Isotope		Timing	Bias	Variability	Upper quantile	Lower quantile
		^{18}O	^{2}H					
NRMSE	X	X	X	●	●	○	○	○
NSE	X			●	●	○	○	○
logNSE	X			●	○			●
KGE	X	X	X	●	●	●	○	○
β (bias)	X	X	X		●			
α (var)	X	X	X			●	○	○
Correlation	X	X	X	●				
SDC	X	X	X		○	●		
Q5	X	X	X				●	
Q95	X	X	X					●
LML mE			X		○	●	○	○
LML bE			X		●	○	○	○
LML RE			X			●		

$$LML \quad bE = \overline{D}_s - \frac{\overline{O}_s \sum_{i=1}^{n} (O_{s,i} - \overline{O}_s)(D_{s,i} - \overline{D}_s)}{\sum_{i=1}^{n} (O_{s,i} - \overline{O}_s)^2} - \overline{D}_o + \frac{\overline{O}_o \sum_{i=1}^{n} (O_{o,i} - \overline{O}_o)(D_{o,i} - \overline{D}_o)}{\sum_{i=1}^{n} (O_{o,i} - \overline{O}_o)^2} \tag{12}$$

The LML fit error is simply the difference between the R^2 values for the best fit lines through the simulated and observed river isotope compositions:

$$LML \quad RE = R_s^2 - R_o^2 \tag{13}$$

The performance metrics used in this study are summarize in Table 4, which includes a qualitative assessment of which error types each metric responds to (as, according to the literature previously referenced here, the sensitivity of metrics to various error types differ). Error types considered are simulation timing, simulation bias, the simulation variability, and errors in high quantiles (e.g. peak flows, enriched isotope concentrations) and low quantiles (e.g. low flow periods, snowmelt freshet isotope signatures).

2.5. Parameter sensitivity and visualization

Parameter sensitivity analyses quantify the variation in a response variable - either a model performance metric or simply a model output variable - to changes in parameter values, either locally (around a particular parameter value) or globally (across a wide range of possible parameter values). A global sensitivity analysis (GSA) can illuminate the relative importance of different hydrologic processes within a particular watershed (Razavi and Gupta, 2015; Song et al., 2015), but the response is often aggregated across broad areas and longer time periods, such that results represent an "average" or aggregate level of sensitivity. This can be misleading for directing further research and in identifying the most significant unknowns (Bajracharya et al., 2020).

This study uses variogram-based GSA, an approach which aims to both improve the characterization of sensitivity and computational efficiency using the variogram (measuring variance of differences in the response surface over the parameter space) and quantifying global parameter sensitivity by integrating the variogram across multiple scales (Razavi and Gupta, 2016a). This method has been implemented in the "Variogram Analysis of Response Surfaces" (VARS) framework, the basis of the VARS-TOOL software (Razavi et al., 2019). This tool was selected for the GSA in this study due to the relative efficiency of the method, which was required to apply GSA to a large-scale process-based hydrologic model. The methodology for the application of VARS to the isoWATFLOOD model is illustrated in Fig. 2.

The VARS tool was used to generate parameter sets using star sampling (a sampling methodology for computationally efficient coverage of the full parameter space): from 200 star centers, a sampling resolution of 0.1 of the total parameter range used to generate cross sections of uniformly spaced parameter cross sections, a total of 48,800 parameter sets were generated (specifics of the sampling space are listed in Table B.1 in the appendix) (Razavi and Gupta, 2016b). According to Razavi and Gupta (2016b), this sample size was deemed sufficient for VARS to calculate sensitivity reliably for our study. In VARS, the variograms are integrated for multiple perturbation scales, but the $IVARS_{50}$ (integrated variogram across a range of scales, from 0% to 50% of the scale range) index is the most comprehensive index for global, rather than local, sensitivity and is therefore used exclusively in our results (Razavi and Gupta,

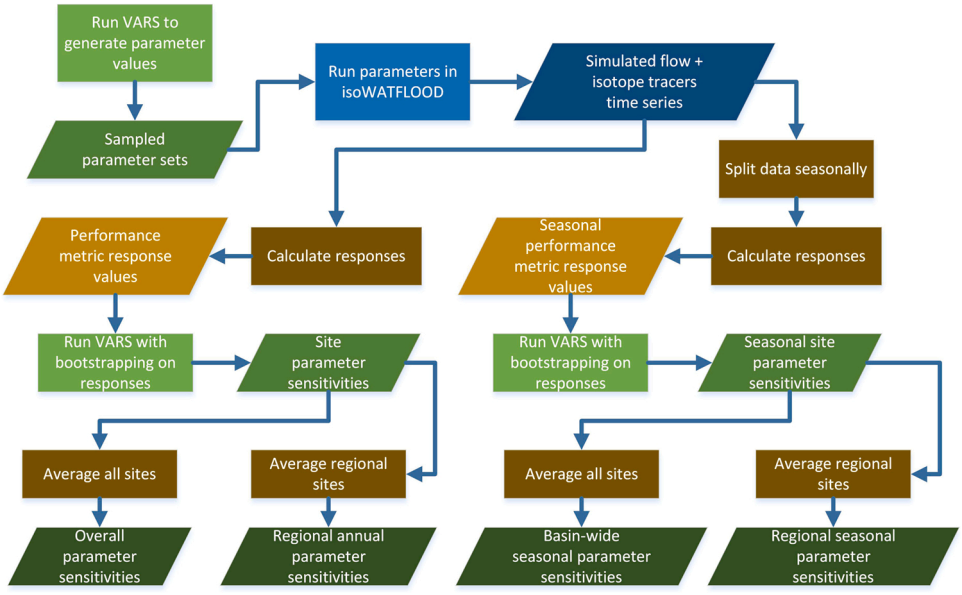

Fig. 2. Flow chart of the methodology for isoWATFLOOD simulations and generating parameter sensitivities from the VARS analyses. Processes are indicated with rectangles and data with parallelograms; VARS is shaded in green, isoWATFLOOD in blue and external scripts in brown.

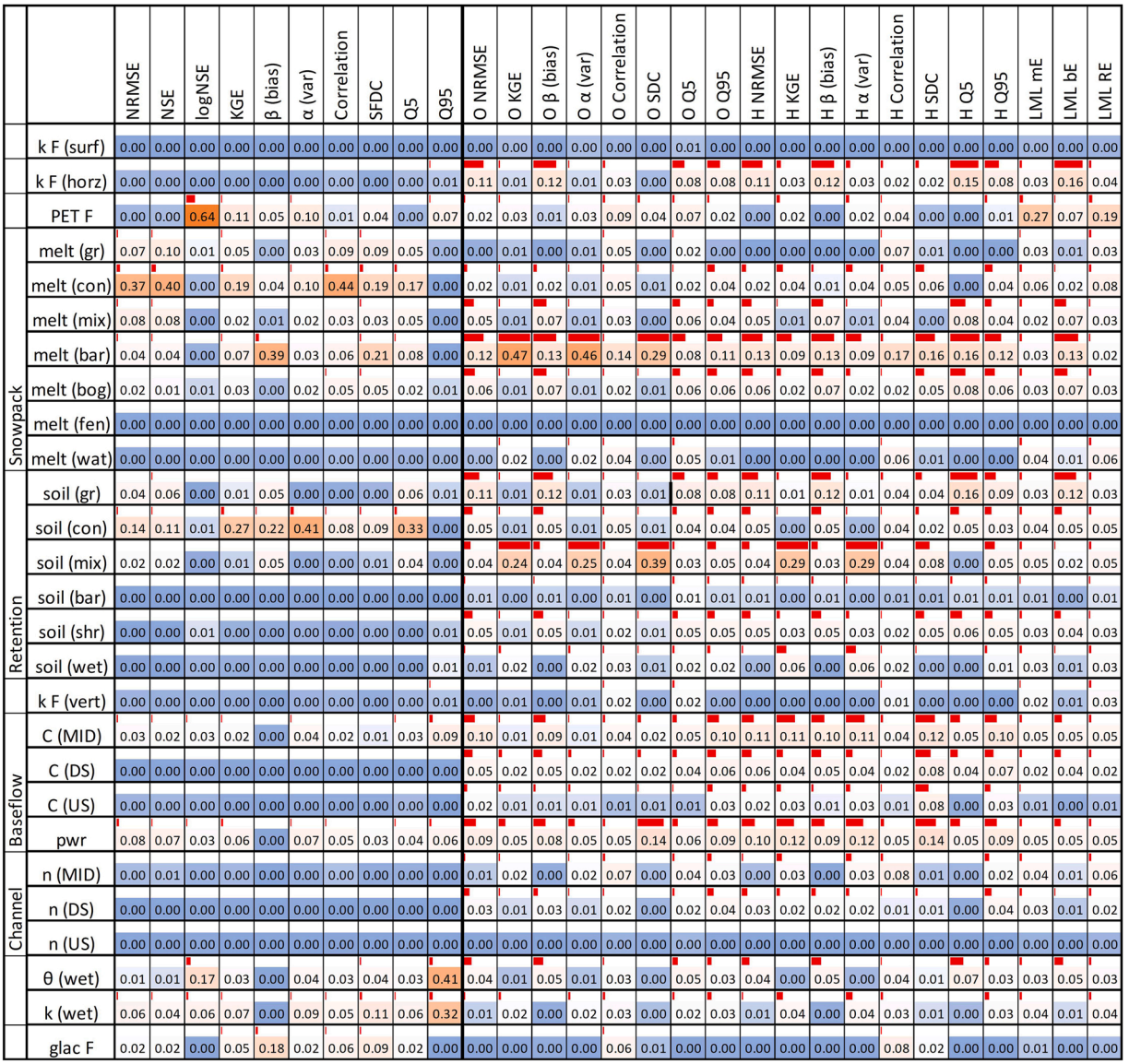

Fig. 3. Relative parameter sensitivity for assessed metrics; insensitive parameters are highlighted in blue and highly sensitive parameters shaded in orange (darker shading is more sensitive/insensitive). Red bars summarize 90% uncertainty range in sensitivity values (displaying 0–0.5 relative sensitivity, where sensitivity values with higher uncertainty have longer bars). Parameter names and descriptions are provided in Table 1, and performance metric information in Table 4.

2016a). The 90% confidence intervals on the sensitivity results were estimated via the internal VARS bootstrap procedure, using 1000 sampling iterations. All sensitivity results in our study were normalized, with values relative to the total sensitivity of a given metric (i.e. the individual IVARS50 values are normalized using the sum of the IVARS50 for all parameters). Parameter sensitivities were calculated for all gauges and isotope sampling sites, and the corresponding sensitivity indices were averaged, either for the entire watershed, or for all observation points within a defined region. Parameter sensitivities quantify the responsiveness of metrics to changes in parameter values and are not intended to identify goodness of fit or optimal parameter values. Sensitivity analyses were similarly performed on the data that was split seasonally (December-January-February (DJF), March-April-May (MAM), June-July-August (JJA), and September-October-November (SON)) to improve the temporal resolution of the analysis and assess seasonal changes in parameter and process sensitivity.

The individual sensitivities for parameters in a process-based hydrologic model can inform our understanding of the model and basin function, but they can also be grouped to illuminate the sensitivity of the model to specific simulated hydrologic processes. Changing a parameter value changes the simulated process that depends on that parameter value within the model, and hence the contribution of that process to total streamflow. If changing a parameter value, and therefore the magnitude or timing of a process, has no effect on the value of a performance metric, then that metric is considered insensitive to both the parameter and the process. Multiple parameters can influence a process, thus only if the metric is insensitive to all the parameters controlling a process is it truly insensitive to the process. The number of analyzed parameters varied between processes, ranging from one to seven individual parameters per process (the parameters associated with each process are provided in Table 2). The overall process sensitivity is summarized as the average sensitivity of all individual parameters controlling the process simulation; assuming that if all parameters were equally sensitive, all processes would have equal shares of sensitivity in the visualizations.

3. Results

3.1. Parameter sensitivity

The relative parameter sensitivities for all analyzed metrics are presented in Fig. 3; insensitive parameters are highlighted in blue, and parameters which dominate the metric response are highlighted in orange shades (darker shading is more sensitive/insensitive). The 90% confidence intervals for the relative sensitivity values are indicated by the red bars above the sensitivity values; precise values for the confidence intervals are presented in Table C.1 in the supplementary material.

The majority of flow simulation performance metrics (left) are sensitive to a consistent subset of parameters and are likewise insensitive to another subset of parameters. In general, flow metrics are highly sensitive to the snowmelt and soil retention parameters in the dominant land class (coniferous forest). Flow metrics are also sensitive to snowmelt parameters in other prevalent land classes, wetland, evaporation and glacier parameters, and baseflow parameters in the dominate river class. All of these parameters have significant control over peak streamflow volume and timing. There are a few outliers to these general tendencies: logNSE, bias and Q95. The logNSE and Q95 metrics are largely insensitive to snowmelt and soil retention parameters; however, logNSE is extremely sensitive to the evaporation parameter (PET F), while the Q95 metric is highly sensitive to wetland parameters (k_{wet} and θ). The flow bias metric has a higher relative sensitivity to glacier and barren ground snowpack melt and is insensitive to baseflow and wetland parameters.

Flow performance metrics are uniformly insensitive to all three soil conductivity parameters (k_{surf} F, k_{horz} F, k_{vert} F), channel roughness parameters, snowmelt parameters for connected wetlands and open water, and soil retention and baseflow parameters in non-dominant classes. Combined, these trends in relative sensitivity results mean that most flow metrics have correlated sensitivities (correlation values for the sensitivities of all metrics are provided in Table C.2). Only the logNSE and Q95 metrics are not highly correlated with any other flow metrics. Confidence intervals on parameter sensitivities are narrow, meaning that parameter sensitivity rankings for flow metrics are not highly dependent on the sampled parameter space (i.e. are reliable, such that a small subset of simulations with extreme results are not determining the overall sensitivity estimate). These narrow confidence intervals are indicative of a smoother response surface, where gradual changes in parameter values do not result in discontinuous changes in performance metric values.

In contrast to flow performance metrics, isotope simulation performance metrics are at least somewhat sensitive to most parameters (fewer blue cells in the right of Fig. 3). There was no parameter to which all isotope metrics were highly sensitive, but all isotope metrics were at least moderately sensitive to the barren ground snowmelt parameter and the power parameter determining baseflow. Unlike the flow metrics, the majority of isotope metric sensitivities have wide confidence intervals, meaning that the relative sensitivities are often more uncertain for isotope metrics. The isotope simulations' correlations, and the LML slope and R^2 errors have the most reliable sensitivity estimates, with confidence bounds comparable to those of the flow metrics. As isotope performance metrics are all somewhat sensitive to nearly all parameters, the relative sensitivities for isotope metrics are generally correlated with each other. The LML slope and R^2 errors are least correlated with other isotope metrics due to their high sensitivity to the PET adjustment factor.

Flow and isotope performance metrics have distinct parameter sensitivities, with only a few parameters having similar relative sensitivities for both metric types. The surface conductivity, snowmelt in connected wetlands, and the alpine channel roughness parameters were uninfluential for all metrics; no parameters were sensitive for all metrics. Isotope metric sensitivities were not generally correlated with flow metric sensitivities; of the flow simulation metrics, the bias sensitivities were most similar to isotope metric sensitivities, due to the relative importance of the barren snowmelt parameter for both the flow bias and isotope metrics.

3.2. Process sensitivity

The parameter sensitivities from Fig. 3 were grouped based on the process they affect, in order to evaluate what metrics in this study are sensitive to specific simulated processes. Overall process sensitivities for all sampling sites or hydrometric gauges (i.e. what would be used in a conventional model calibration) are shown in Fig. 4.

Only six of the ten modeled processes dominate flow performance metric sensitivities, with some variation in the relative shares for each metric. The channel and upper zone soil fluxes have small or negligible shares of the overall flow metric sensitivity. Many of the evaluated metrics' sensitivities are dominated by a single process, particularly evaporation for logNSE, and wetlands for Q95, and (to a lesser degree) glacier for bias, and snowmelt for NRMSE, NSE and correlation. The KGE metric has the most balanced shares of process

sensitivity, for the six processes dominating the streamflow response.

There is a substantial variation in the process sensitivities for isotope simulation performance metrics, possibly due in part to the uncertainty in the parameter sensitivity values. Isotope metrics are consistently sensitive to soil retention, snowmelt and baseflow, and generally sensitive to interflow (i.e., the horizontal upper soil zone flux). No single isotope metric is sensitive to all processes, but there at least one isotope metric is slightly sensitive for all ten processes. The second isotope simulation, ^{2}H, was not needed to cover all ten processes, since the ^{18}O simulation alone is sensitive to all processes. Given the uncertainty in many of the sensitivity estimates for isotope tracer metrics, the differing sensitivities of the two tracers is likely insignificant.

In comparison to the flow metrics, isotope metrics cover sub-surface soil processes better and emphasize the channel roughness more. However, only the isotope correlations are highly sensitive to glacier melt, meaning isotope simulations do not generally respond to changes in simulated glacier melt. The isotope metrics have smaller shares of the overall sensitivity dedicated to wetlands, evaporation, and snowmelt than flow metrics, although they are significant processes for both types of metric.

Performance metrics can be calculated on a seasonal basis, and observation locations can likewise be geographically separated for a more detailed analysis of process sensitivity across the Athabasca region (Figs. 5 and 6, for flow and isotope metrics respectively).

As would be expected from the overall annual results in Fig. 4, the KGE, NSE and SFDC metrics are sensitive to six of the ten simulated hydrologic processes for either the average of all sites in the basin (ALL), or the basin-wide annual (A) sensitivities, with some variation in the ranking of these processes. NSE is generally more sensitive to snowmelt, KGE is more consistently sensitive to evaporation, and glacier melt and wetlands are most influential over the annual SFDC. It is important to note that cumulative error metrics have 'overall' sensitivities that can be estimated from subsets, with weighting (e.g. overall snowmelt sensitivity for the NSE will be between the maximum and minimum seasonal sensitivity). This is not the case for population-based error metrics: the SFDC calculated from seasonally separated data may have different sensitive parameters than the SFDC calculated from unseparated data (e.g. SFDC more sensitive to glaciers when considering whole years, than seasonally separately data).

Glacier melt is frequently the most influential process in the mountainous upstream, headwater region of the Athabasca basin in the summer and fall but is largely irrelevant to flow metrics in the mid- or downstream basins. Among flow metrics, snowmelt is an influential process across the entire basin, but it is most sensitive in the period covering the freshet (MAM and to some extent, JJA). Similarly, evaporation is most influential in summer and fall, when the bulk of evaporation loss occurs. In contrast to these seasonally varying processes, flow metrics are sensitive to wetland fluxes year-round, more so in the lower slope mid- and downstream regions of the basin where wetlands are more prevalent. Sub-dividing data (from Fig. 4) geographically and temporally does render some soil flux sensitivities noticeable. All flow metrics in Fig. 5 are sensitive to interflow and lower zone recharge in the lowest flow data sub-division: winter flows in the upstream basin. Other low flow times and locations are likely to be sensitive to interflow and

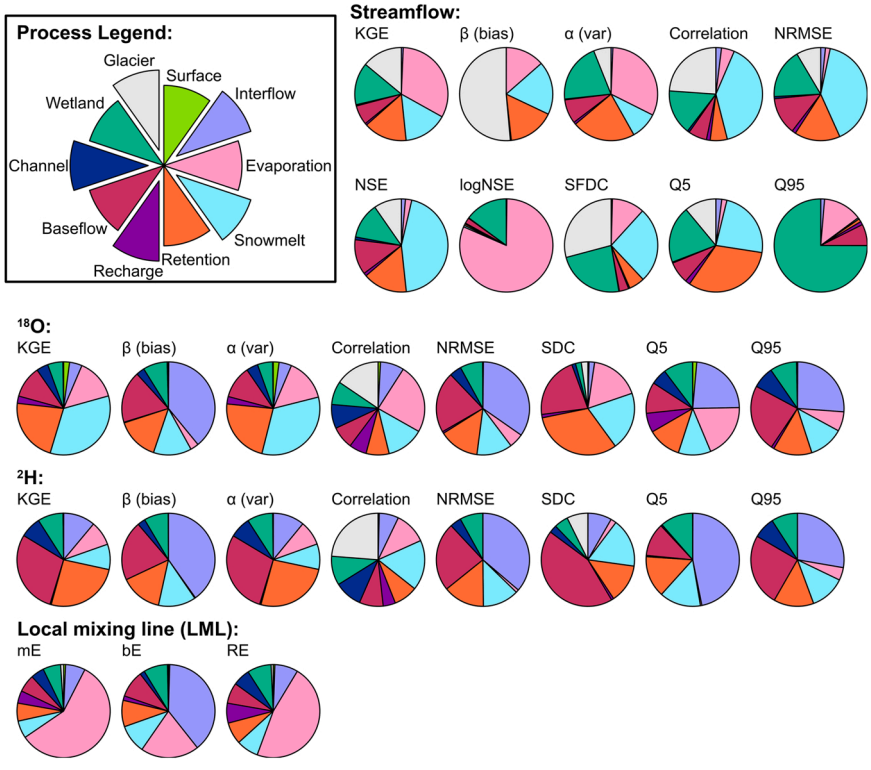

Fig. 4. Overall relative process sensitivities for all evaluated performance metrics (Table 4), averaged for all observation locations over the entire simulation period.

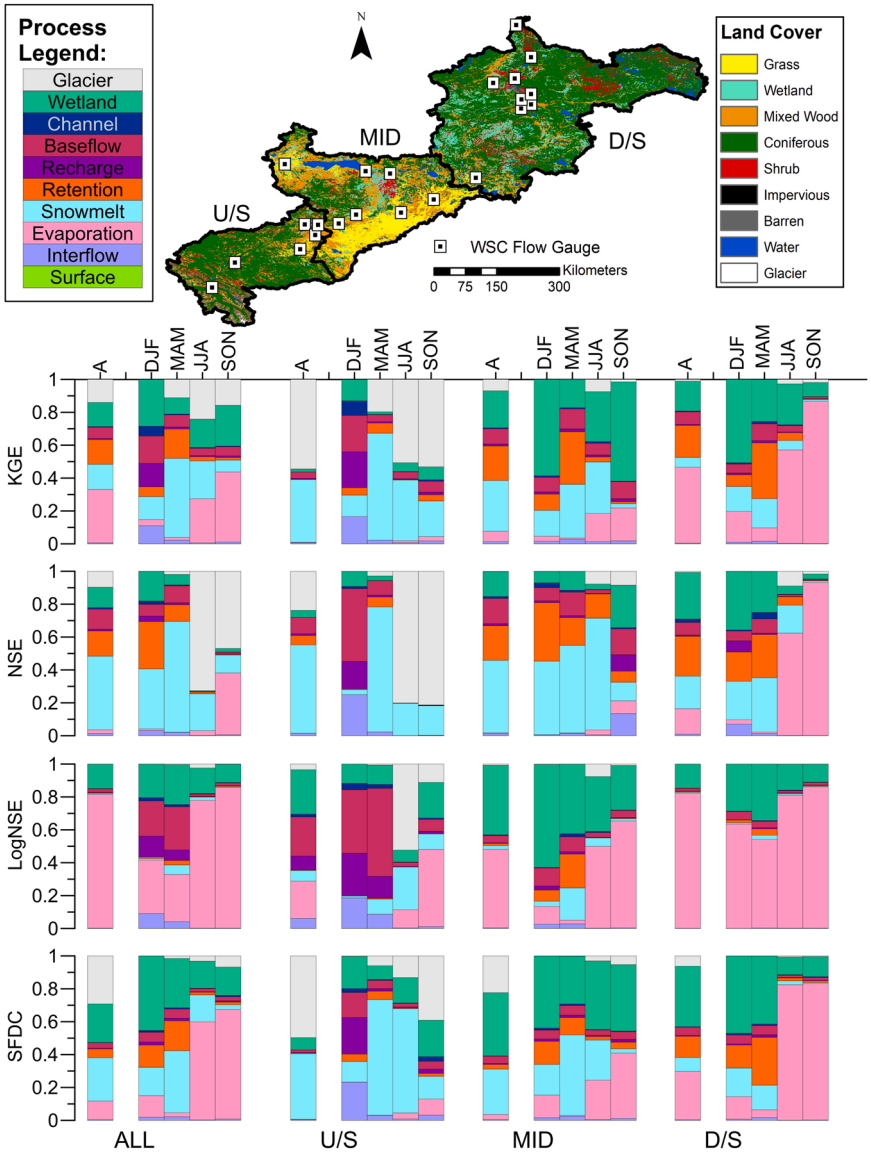

Fig. 5. Regional flow metric process sensitivity for the upstream (U/S), mid-basin (MID) and downstream (D/S) reaches, and basin-wide (ALL), temporally aggregated by season (DJF-December, January, February; MAM-March, April, May, JJA-June, July, August, SON-September, October, November) and full year (A).

recharge; these times and places are also likely to have baseflow as a relatively sensitive process. Soil retention is an influential process but has the highest relative sensitivity in spring (MAM), coinciding with most snowmelt, indicating that the primary cause of flow metrics' sensitivity to retention is the soil's ability to absorb snowmelt. No flow metric is sensitive to surface infiltration during any season or region; therefore, these metrics are completely insensitive to whether water on the soil surface infiltrates or runs off directly to wetlands. Flow metrics can be somewhat sensitive to channel parameters, in very low-flow, or high (peak) flow periods. Parameters controlling flow velocity in the channel can influence flow simulation timing, but wetlands have a much larger influence on the magnitude and timing of streamflow in this watershed.

Isotope sampling locations are biased toward the downstream portion of the basin, in the oil sands region of the Athabasca River basin (reference map, Fig. 6). While the data sampling resolution in the upstream and mid-basin regions are of the same quality as the best locations in the downstream region, the smaller number of sites limits confidence in the generalizability of the sensitivities for the upstream and mid-basin regions. The large confidence intervals on isotope metric sensitivity (Fig. 3) likewise adds to the uncertainty in regional and seasonal process sensitivity results.

Some general observations from the isotope metric sensitivities may still, however, be drawn. Soil fluxes are the main theme of isotope sensitivity: all isotope performance metrics are sensitive to some combination of infiltration, interflow, recharge, soil retention

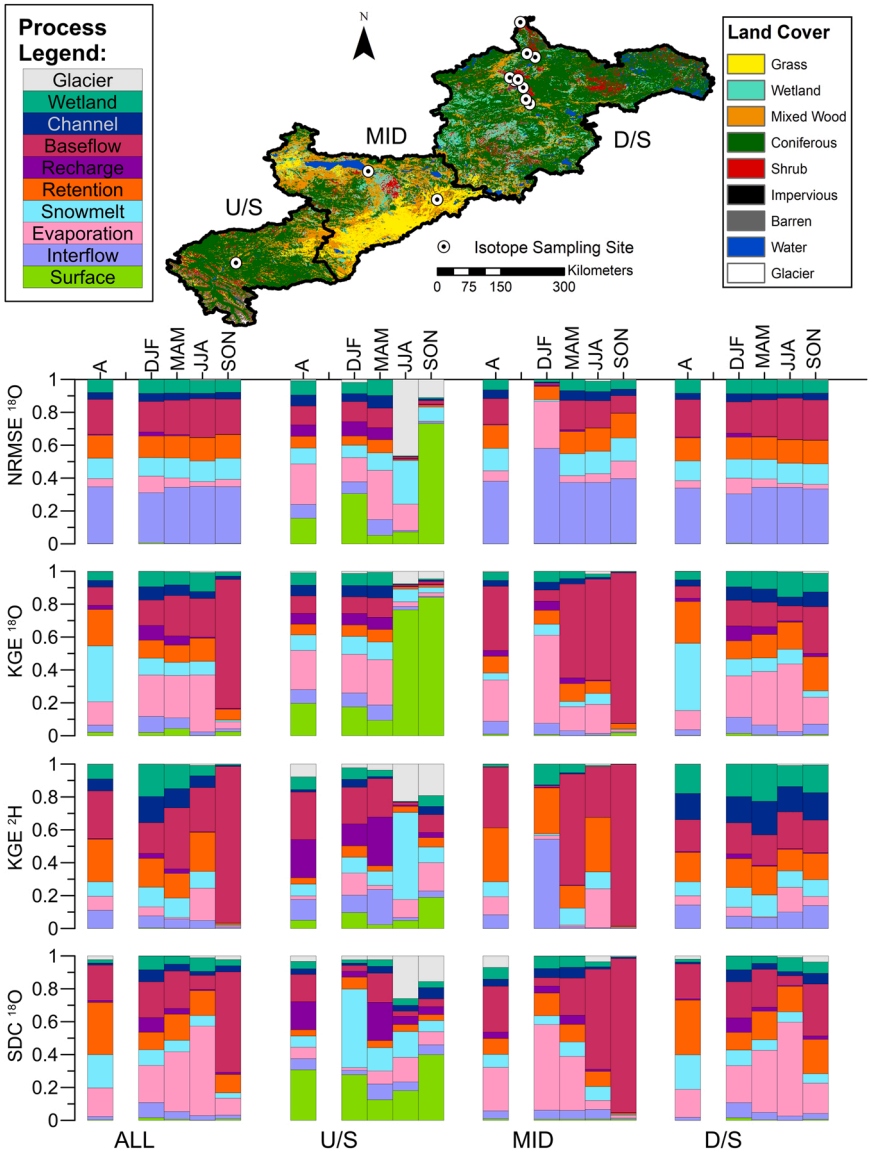

Fig. 6. Regional isotope tracer metric process sensitivities for the upstream (U/S), mid-basin (MID) and downstream (D/S), and basin-wide (ALL), temporally aggregated by season (DJF-December, January, February; MAM-March, April, May, JJA-June, July, August, SON-September, October, November) and full year (A).

and baseflow, with these soil processes dominating the sensitivity in most seasons and regions. Baseflow is approximately the only sensitive process in the mid-basin during fall (SON). The isotope simulation at the upstream sampling site is clearly sensitive to surface infiltration, and other downstream sites are not completely insensitive. Isotope performance metrics are more sensitive to recharge in winter (DJF) and spring (MAM), but there is no clear seasonal pattern to soil water retention sensitivity for isotope metrics, unlike flow metrics. Overall, there is considerable variation in the most sensitive processes for the various metrics, locations, and time periods, and every modeled process is significant in at least one relative sensitivity sub-division.

Interestingly, snowmelt and evaporation are not generally the most influential processes for isotope metrics, even though both have distinctive signals in the isotope data; these processes are also sensitive processes outside of their main seasons of occurrence. Isotope metrics are much less sensitive to glacier melt than flow metrics, though it remains an influential process in the headwater basin during the summer and fall. On the other hand, compared to flow metrics, isotope metrics are more sensitive to channel velocity.

The downstream transfer of process sensitivity for flow and isotope tracer metrics is illustrated in Fig. 7, with a pair-wise comparison of sensitivities for nested watersheds along the Athabasca River mainstem, from upstream to downstream.

The flow metric sensitivity shows significant downstream transfer of process sensitivity along the mainstem of the Athabasca River; glacier melt is the dominant process in the Athabasca headwaters and remains a significant process in the simulation at Fort McMurray,

over 1000 km downstream. The flow gauges on the Athabasca River itself are outliers from the regional sensitivities (Fig. 5) due to influence of upstream areas: glacier and snowmelt have much larger shares of the relative sensitivity. Isotope tracer sensitivities, in contrast, have limited downstream transfer of process sensitivity. At the downstream end of the Athabasca River, isotope tracer sensitivities closely resemble those of local tributaries rather than upstream gauges. Isotope tracer sensitivities for the largest watershed areas also have limited seasonal variation, unlike flow sensitivities for the same observation location.

4. Discussion

4.1. Evaluating with blinders: what flow-based metrics 'see'

Flow simulation performance metrics respond primarily to processes with substantial influence on either water volume (i.e. evaporation and glacier melt) or peak flow timing (i.e. snowmelt, wetland and soil retention). The processes with the largest influence on flow metric values vary significantly within the Athabasca River basin. Glacier melt is hugely significant to most flow simulation performance metrics in the mountain headwaters of the river, but averaged flow metrics are less sensitive to the magnitude of glacial melt outside the alpine region. Conversely, wetland retention and evaporation make only a modest contribution to flow metric responses in the headwaters, where there is a limited area of wetlands and open water, but these two processes are among the most important to flow metrics in the downstream oil sands region. This finding is supported by Gibson et al. (2019a, 2019b) who found similar relationships between headwater and lowland regions using isotope-derived estimates of water yield. Snowmelt and soil water retention are sensitive processes across the entire Athabasca River basin, but the scope of their influence on flow metric response is somewhat limited temporally: the spring freshet (MAM) is most sensitive to both melt rates and the upper soil zone water retention capacity. From the timing of maximum sensitivity to soil water retention, it is clear that flow metrics respond primarily to the capacity of the upper zone to absorb runoff and damp peak flows, rather than its ability to retain water in the longer term and affect evapotranspiration. Based on the flow simulation alone, it would appear that spatial variation in the most influential processes within the Athabasca basin depends only on the prevalence of glaciers or wetlands.

Although the model was not calibrated in this study, sensitivity results indicate that tuning the simulation of just six of the ten modeled processes in the model would be sufficient to generate a 'good' flow simulation, assuming a 'good' simulation is considered to be one where the simulated hydrographs closely resemble the observed hydrographs. If this simple measure of accuracy is sufficient for the intended application of the hydrologic model, a simulation that is well-calibrated to optimize KGE or some other combination of flow simulation performance metrics can be considered fit for its purpose (e.g., short-term peak flow forecasting). In such situations, however, it would be unclear why a process-based hydrologic model is being used in the first place. The blind spots of flow metrics are of concern for potential model applications where soil fluxes are important, or where the fidelity of process representation matters. For example, when total basin storage, water age or flow paths are relevant outputs (i.e. for water supply assessments or long-term climate change studies), calibrating the model to optimize only flow performance will likely prove to be inadequate (Kirchner, 2006). In fact, this study demonstrates that increasing structural complexity (i.e. more parameters and processes) is likely to result in a larger decision space of 'acceptable' solutions derived from different hydrologic partitions, or proportional flow path contributions (Figs. 3 and 4). It should be noted that many applications of hydrologic models and scenarios in the Canadian Oil Sands region require the accurate simulation of both water storage and flow paths for contaminant tracing or climate and land use impact assessment.

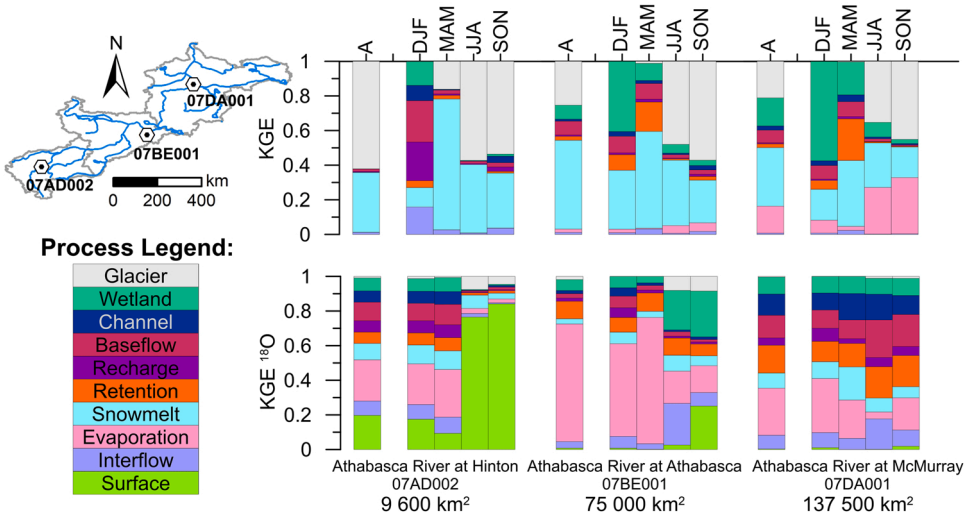

Fig. 7. Flow and isotope tracer KGE process sensitivities for sites along the Athabasca River mainstem demonstrating the downstream transfer of process sensitivity aggregated seasonally (DJF, MAM, JJA, SON) and full year (A).

4.2. The added value of isotope-aided metrics

Isotope simulation performance metrics are sensitive to a wider variety of model processes than flow metrics (Fig. 3). It is well-known from previous research that some processes are only locally or periodically hydrologically significant, and that parameter sensitivities therefore change depending on the time period or location within the watershed (Herman et al., 2013; Höllering et al., 2018). We show here that there are, in fact, *no* processes that flow performance metrics are sensitive to that isotope performance metrics are not (Fig. 4). In addition to those processes which flow metrics are sensitive to, isotope metrics are additionally sensitive to soil water fluxes (*k F, C,* and *pwr*) and channel roughness (*n*), and are more sensitive to soil water storages (*soil*). In our modeling, isotope metrics were sensitive to infiltration rates in association with exposed or barren ground, lower and upper soil zone flows in association with grassland and mixed deciduous and coniferous forest, and a mix of soil and wetland properties with channel velocities in association with wetlands and coniferous forest (Fig. 6). Snowmelt and evaporation were less influential than soil processes, in spite of their importance to the water balance, although this may be an artifact of the sampling resolution. Model sensitivity to sub-surface processes is a reflection of the significance of mixing volumes in isotope tracer simulations; the variability of the isotope tracer simulation in other models is also largely determined by the volume of simulated water in storage (the water age) (Birkel et al., 2011; Klaus et al., 2015; Rodriguez and Klaus, 2019). Flow simulations are dependent on the amount and timing of flow, while an isotope tracer simulation, with a concentration output, is dependent on the age of flow (flow path length) and fractionation processes (surface versus subsurface flow paths). Flow metrics are therefore better at detecting flow volume or timing errors, such as glacier melt rate errors (Fig. 5), and isotope tracer metrics respond most to errors in flow paths, such as soil water flux rates (Fig. 6).

There is little seasonal variation in process sensitivity for isotope tracers in the downstream region of the Athabasca River basin; the inherent mixing within large upstream areas, sub-surface storages or extensive regional wetlands limits the temporal variation in isotope data (isotope data provided in Appendix A). Isotope tracers are therefore better at providing information on processes in smaller basins than larger ones. The damped isotope signals in larger watersheds provides useful information on long-term process contributions, but a high-resolution isotope dataset from a smaller watershed can clarify individual process contributions for specific events or types of events, as illustrated by Fig. 7. Regions of low water yield have been reported within the middle and downstream portions of the Athabasca River basin, where it is believed there may be buried channels and a shift toward more vertical flow exchange as opposed to lateral surface runoff (Gibson et al., 2019a, 2019b). In fact, isotope tracer process sensitivities reflect this finding from surface water dominance in the headwater basin toward more soil storage dominated processes in the mid and downstream reaches; this is, however, not reflected in flow-based process sensitivities (Fig. 7). To specifically diagnose lateral and vertical flow exchange processes, other tracer types or enhanced sampling resolution may be necessary to fully delineate the geographical and temporal significance of these processes. Isotope tracer datasets can be better leveraged when sub-basin scale or type are explicitly considered: a single headwater sampling site, or a site with different land cover or topography, can add far more valuable information than adding more gauges along the mainstem of a river.

Isotope tracer metrics in conjunction with flow metrics provide a more complete picture of the influential processes in the Athabasca River basin than flow metrics alone. When both data types are considered, the alpine headwaters are affected not only glacial and snow melt, but also infiltration and surface runoff (Fig. 6). In the central portion of the Athabasca watershed, isotope tracer data can highlight the importance of soil storage and fluxes year-round, which flow performance metrics ignore. Both isotope tracer and streamflow simulations agree on the critical importance of wetlands and evaporation rates in the downstream regions of the Athabasca River, but isotope tracer metrics are also responsive to the path through the soil that water takes to reach those wetlands (Fig. 6), which is intrinsically linked to residence time or water age. There is also an interesting possibility that for processes both flow and isotope performance metrics are sensitive to, the different simulation types may have contradictory optimization outcomes. Evaporation, for example, reduces simulated flow but increases both the magnitude and annual variability of isotope tracer concentrations (i.e. seasonally enriched streamflow); what may appear to be equifinality when streamflow alone is considered, may not be equivalent when both tracers and streamflow are evaluated (Beven, 2006; Kirchner, 2006). For example, a model optimized with a flow performance metric could therefore have a much higher evaporative loss than the same model optimized with an isotope performance metric (Holmes et al., 2020).

This study utilized one hydrologic model with multiple simulated outputs to produce a suite of flow *and* isotope tracer-based simulations. The exact proportions we report for process sensitivity in relation to various metrics and proportion of simulated flow are specific to the hydrologic model used (here, isoWATFLOOD) as they are a reflection of the model's internal structure and the algorithms that numerically define each process. Our findings, however, are model agnostic in terms of the cautionary tale they tell of over-reliance on flow data for model evaluation, or rather the missing information content when calibration is based on flow data alone. The value of adding isotope tracer data is that water age and flow paths are directly incorporated into model evaluation, which correlates to internal process function and model structure. This outcome would occur generally for physically based models, as it is a reflection of adding metrics and data capable of diagnosing such storage and flux interactions. Some findings relating to snowmelt sensitivity are only transferrable to watersheds under similar climates (i.e., seasonal basins in mid- to high-latitude regions), and outcomes would differ for lower latitude regions experiencing exclusively rainfall and much higher proportions of evaporative loss. The actual value added by the isotope tracers in any particular application ultimately depends on the degree to which the isotopes fractionate throughout the regional hydrologic cycle, and the isotope concentration distinctness of processes (or end members) in the watershed.

4.3. On the selection of performance metrics for model calibration

The various isotope and flow performance metrics have different relative advantages in the context of hydrologic model parameter calibration. Flow metrics consistently have reliable parameter sensitivities (Fig. 3), due to a regular response surface for flow performance. These reliable sensitivities are advantageous in model calibration as they identify consistently insensitive parameters and remove them from the calibration; a smoother performance response surface also facilitates searching in optimization. KGE is the flow performance metric with the broadest range of process sensitivities, and it is therefore the best choice for a stand-alone flow metric in a process-based optimization (out of those evaluated in this study). The NSE is a possible alternative, although unlike the KGE, it is skewed toward snowmelt (i.e. the primary peak flow generating mechanism, with high magnitude residual error, in the Athabasca basin). Using more specialized metrics, such as logNSE or flow signatures, can highlight particular processes, but these metrics did not respond strongly to processes also not covered by the KGE metric (Fig. 4). Juggling different metrics or data subsets (e.g. Q_{95} or alpine flow gauges) can highlight particular processes (e.g. wetland fluxes or glacial melt) far better than averaged general response metrics such as KGE, but does little to expose the internal soil processes. The components of KGE may be just as useful as specialized metrics for rebalancing process sensitivities. When only streamflow is evaluated, process-based hydrologic models can behave as something like a black box for simulated flow pathways, since streamflow simply tracks how much water comes out of the landscape (Blöschl et al., 2019). Streamflow performance is therefore unresponsive to changes in water flow paths alone; flow metrics are indifferent to how much water is stored internally within a simulation, or how long precipitation takes to reach the channel network, as long as the correct volume of water reaches the river at the right time.

In contrast to flow metrics, many isotope performance metrics have low reliability for parameter sensitivity (Fig. 3), in that a small region of the parameter space can have an outsized influence over the relative sensitivity of a parameter. As an example, the isotope concentration simulation can perform extremely poorly if one combination of soil conductivity and soil water retention parameters results in the desiccation of a fractionating storage unit, yet desiccation (and poor simulation performance) can be avoided by slight changes in any one of three parameters. Every isotope performance metric included in the analysis was sensitive to soil fluxes and storage, indicating that responsiveness to internal flow paths is an inherent property of the isotope tracer simulation, which the flow simulation alone does not have. Isotopic sensitivity to subsurface flow paths has been identified previously in the literature (Delavau et al., 2017; Stadnyk and Holmes, 2020) which is broadly why isotopes are considered excellent hydrologic tracers (Klaus and McDonnell, 2013). Therefore, unlike flow signatures, isotope tracer performance metrics can cover processes missed by streamflow KGE. An isotope tracer simulation can produce a better-informed hydrologic model, but the utility of the added information is dependent on the application.

Isotope performance metrics have a larger number of sensitive parameters than flow performance metrics (Fig. 3); including isotope metrics in model optimization therefore increases the scope of the optimization: more parameters need to be included in the optimization, but more parameters will actually be optimized. Both the isotope tracer and seasonal flow metric sensitivity results oppose the common practice of removing parameters from calibration based on simple sensitivity analyses: the considerable variation in sensitive processes for the various metrics, locations, and time periods meant every modeled process is significant to the model at some place or time. The most reliable isotope parameter sensitivity metrics were the LML errors and the correlations between simulated and observed isotope data; they are relatively unaffected by desiccation events in the simulation which can lead to substantially different model responses in highly localized parts of the decision space (Sahraei et al., 2020).

The more reliable isotope sensitivity estimates (i.e., correlation and NRMSE) have similar process sensitivities for both tracers, which is anticipated under similar atmospheric forcing. However, the advantage of simulating both isotopes is that it allows the calculation of a simulated LML, and therefore LML errors; of the isotope metrics, LML error metrics were the most sensitive to evaporation. No evidence was found to support using KGE for isotope simulation evaluation in place of the traditional residual error metrics (e.g., NRMSE). KGE sensitivities were no more reliable, and the same processes were influential for both NRMSE and KGE metrics. Furthermore, the KGE sensitivity was dominated by its variability component, however, trying to evaluate the variability error of a simulation based on sporadic observations is highly dubious. Just as using the variability in observations to normalize the squared error (i.e. using the NSE) is not recommended for discontinuous data because the data sample may not be representative of the true population variability, the variability error of the KGE metric is not recommended for calibrating tracer simulations when only sparse observations are available. The correlation or bias components of the KGE are better supplements to a residual error metric in isotope simulation evaluation (e.g. in a multi-objective calibration problem formulation) with discontinuous or sparse observed datasets.

Simulating both isotope tracers does not increase the number of sensitive processes, as all processes are sensitive to some degree to either of the two tracers. The differing sensitivities of the two tracers for some metrics cannot be attributed to the properties of the tracers due to the uncertainties in the tracer sensitivity results. It must also be noted that this analysis has not been extended to include either uncertainty in observed data values, or from sampling (analytic uncertainties for isotope data are relatively low, however observations are sparse both spatially and temporally). Multi-objective optimization methods are highly suitable for calibrating hydrologic models with both tracer and flow data, as they allow a transparent choice in the trade-off between simulation qualities; the importance and uncertainty of an accurate tracer simulation can be balanced by the modeler.

5. Conclusions

This study highlights the important regional hydrologic differences between the upper, middle, and lower basins of the Athabasca River. The Oil Sands Monitoring program is concerned with cumulative effects assessment, which requires knowledge of the impacts to more than just streamflow (or total volume), and accurate projections of future water supply depend on the accurate partitioning of

processes controlling the overall water balance. A 'black box' model calibrated without specific consideration of these process can – and likely will – result in inaccurate partitioning of water in soils, which directly influences projection of evapotranspiration (air-land), and infiltration or baseflow (land-subsurface) flow paths, skewing future projections of streamflow.

The scope of this study was limited to sensitivity analyses in the Athabasca River basin, but there are some conclusions applicable to model calibration or evaluation more generally:

- Flow simulation performance metrics alone provide an incomplete picture of hydrologic process regional variation and significance.
- KGE is the best stand-alone flow performance metric for process-based optimization as it exhibited the broadest range of process sensitivity.
- Flow signature metrics can highlight specific processes already covered by the generalized KGE but do not add new ones.
- Including an isotope tracer simulation expands the number of processes which can be evaluated.
- Residual error metrics, or bias and correlation are all reasonable measures of simulation performance for isotope tracers.

In conclusion, we suggest that a process-based hydrologic model cannot be considered fully calibrated if the performance of the model is only evaluated with streamflow metrics, because it is not possible for an optimization to tune parameters or processes to which the calibration objective is insensitive. Either the streamflow-insensitive internal flux simulation should be ignored as unreliable, or the model calibration should be expanded to include relevant datasets. Isotope tracers have demonstrable value for informing process-based hydrologic model calibration, although further research is needed on isotope-enabled calibration methodologies and the effects of metric choice on simulated streamflow-generating processes.

CRediT authorship contribution statement

Tegan Holmes: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - Original Draft, Visualization. **Tricia Stadnyk**: Conceptualization, Methodology, Resources, Writing - Review & Editing, Funding acquisition. **Masoud Asadzadeh**: Conceptualization, Methodology, Writing - Review & Editing. **John J. Gibson**: Resources, Writing - Review & Editing, Funding acquisition.

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.

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 origin of the meaning behind 'Athabasca River' is derived from the Woodland Cree word *aeapaskāw* meaning" where there are plants one after another". 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. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ejrh.2022.101088.

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. 578 https://doi.org/10.1016/j.jhydrol.2019.124032.

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

Bajracharya, A., Awoye, H., Stadnyk, T., Asadzadeh, M., 2020. Time variant sensitivity analysis of hydrological model parameters in a cold region using flow signatures. Water 12. https://doi.org/10.3390/W12040961.

Belachew, D.L., Leavesley, G., David, O., Patterson, D., Aggarwal, P., Araguas, L., Terzer, S., Carlson, J., 2016. IAEA Isotope-enabled coupled catchment–lake water balance model, IWBMIso: description and validation†. Isot. Environ. Health Stud. 52, 427–442. https://doi.org/10.1080/10256016.2015.1113959.

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. Environ. Model. Softw. 40, 1–20. https://doi.org/10.1016/j.envsoft.2012.09.011.

Beven, K., 2006. A manifesto for the equifinality thesis, in: Journal of Hydrology. pp. 18–36. ⟨https://doi.org/10.1016/j.jhydrol.2005.07.007⟩.

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.

Birkel, C., Soulsby, C., Tetzlaff, D., 2011. Modelling catchment-scale water storage dynamics: Reconciling dynamic storage with tracer-inferred passive storage. Hydrol. Process. 25, 3924–3936. https://doi.org/10.1002/hyp.8201.

Blöschl, G., Bierkens, M.F.P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J.W., McDonnell, J.J., Savenije, H.H.G., Sivapalan, M., Stumpp, C., Toth, E., Volpi, E., Carr, G., Lupton, C., Salinas, J., Széles, B., Viglione, A., Aksoy, H., Allen, S.T., Amin, A., Andréassian, V., Arheimer, B., Aryal, S.K., Baker, V., Bardsley, E., Barendrecht, M.H., Bartosova, A., Batelaan, O., Berghuijs, W.R., Beven, K., Blume, T., Bogaard, T., Borges de Amorim, P., Böttcher, M.E., Boulet, G., Breinl, K., Brilly, M., Brocca, L., Buytaert, W., Castellarin, A., Castelletti, A., Chen, X., Chen, Yangbo, Chen, Yuanfang, Chifflard, P., Claps, P., Clark, M.P., Collins, A.L., Croke, B., Dathe, A., David, P.C., de Barros, F.P.J., de Rooij, G., di Baldassarre, G., Driscoll, J.M., Duethmann, D., Dwivedi, R., Eris, E., Farmer, W.H., Feiccabrino, J., Ferguson, G., Ferrari, E., Ferraris, S., Fersch, B., Finger, D., Foglia, L., Fowler, K., Gartsman, B., Gascoin, S., Gaume, E., Gelfan, A., Geris, J., Gharari, S., Gleeson, T., Glendell, M., Gonzalez Bevacqua, A., González-Dugo, M.P., Grimaldi, S., Gupta, A.B., Guse, B., Han, D., Hannah, D., Harpold, A., Haun, S., Heal, K., Helfricht, K., Herrnegger, M., Hipsey, M., Hlaváčiková, H., Hohmann, C., Holko, L., Hopkinson, C., Hrachowitz, M., Illangasekare, T.H., Inam, A., Innocente, C., Istanbulluoglu, E., Jarihani, B., Kalantari, Z., Kalvans, A., Khanal, S., Khatami, S., Kiesel, J., Kirkby, M., Knoben, W., Kochanek, K., Kohnová, S., Kolechkina, A., Krause, S., Kreamer, D., Kreibich, H., Kunstmann, H., Lange, H., Liberato, M.L.R., Lindquist, E., Link, T., Liu, J., Loucks, D.P., Luce, C., Mahé, G., Makarieva, O., Malard, J., Mashtayeva, S., Maskey, S., Mas-Pla, J., Mavrova-Guirguinova, M., Mazzoleni, M., Mernild, S., Misstear, B.D., Montanari, A., Müller-Thomy, H., Nabizadeh, A., Nardi, F., Neale, C., Nesterova, N., Nurtaev, B., Odongo, V.O., Panda, S., Pande, S., Pang, Z., Papacharalampous, G., Perrin, C., Pfister, L., Pimentel, R., Polo, M.J., Post, D., Prieto Sierra, C., Ramos, M.-H., Renner, M., Reynolds, J.E., Ridolfi, E., Rigon, R., Riva, M., Robertson, D.E., Rosso, R., Roy, T., Sá, J.H.M., Salvadori, G., Sandells, M., Schaefli, B., Schumann, A., Scolobig, A., Seibert, J., Servat, E., Shafiei, M., Sharma, A., Sidibe, M., Sidle, R.C., Skaugen, T., Smith, H., Spiessl, S.M., Stein, L., Steinsland, I., Strasser, U., Su, B., Szolgay, J., Tarboton, D., Tauro, F., Thirel, G., Tian, F., Tong, R., Tussupova, K., Tyralis, H., Uijlenhoet, R., van Beek, R., van der Ent, R.J., van der Ploeg, M., van Loon, A.F., van Meerveld, I., van Nooijen, R., van Oel, P.R., Vidal, J.-P., von Freyberg, J., Vorogushyn, S., Wachniew, P., Wade, A.J., Ward, P., Westerberg, I.K., White, C., Wood, E.F., Woods, R., Xu, Z., Yilmaz, K.K., Zhang, Y., 2019. Twenty-three unsolved problems in hydrology (UPH) – a community perspective. Hydrol. Sci. J. 64, 1141–1158. https://doi.org/10.1080/02626667.2019.1620507.

Bowen, G.J., Cai, Z., Fiorella, R.P., Putman, A.L., 2019. Isotopes in the water cycle: regional-to global-scale patterns and applications. Annu. Rev. Earth Planet. Sci. https://doi.org/10.1146/annurev-earth-053018.

Buttle, J.M., Allen, D.M., Caissie, D., Davison, B., Hayashi, M., Peters, D.L., Pomeroy, J.W., Simonovic, S., St-Hilaire, A., Whitfield, P.H., 2016. Flood processes in Canada: regional and special aspects. Can. Water Resour. J. 41, 7–30. https://doi.org/10.1080/07011784.2015.1131629.

Carlisle, D.M., Wolock, D.M., Meador, M.R., 2011. Alteration of streamflow magnitudes and potential ecological consequences: a multiregional assessment. Front. Ecol. Environ. 9, 264–270. https://doi.org/10.1890/100053.

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.

Clark, M.P., Kavetski, D., Fenicia, F., 2011. Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resour. Res. 47. https://doi.org/10.1029/2010WR009827.

Coulibaly, P., Samuel, J., Pietroniro, A., Harvey, D., 2013. Evaluation of Canadian national hydrometric network density based on WMO 2008 standards. Can. Water Resour. J. 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., Birks, J., 2011. Model based spatial distribution of oxygen-18 isotopes in precipitation across Canada. Can. Water Resour. J. 36 https://doi.org/10.4296/cwrj3604875.

Delavau, C.J., Stadnyk, T., Holmes, T., 2017. Examining the impacts of precipitation isotope input δ18Oppt) 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.

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

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

Gibson, J.J., Birks, S.J., Moncur, M., 2019a. Mapping water yield distribution across the South Athabasca Oil Sands (SAOS) area: baseline surveys applying isotope mass balance of lakes. J. Hydrol. Reg. Stud. 21 https://doi.org/10.1016/j.ejrh.2018.11.001.

Gibson, J.J., Yi, Y., Birks, S.J., 2019b. 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 https://doi.org/10.1016/j.ejrh.2019.100643.

Gibson, J.J., Yi, Y., Birks, S.J., 2016. 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.

Government of Canada, 2021. Canada-Alberta oil sands environmental monitoring [WWW Document]. URL ⟨https://www.canada.ca/en/environment-climate-change/services/oil-sands-monitoring.html⟩ (accessed 5.30.21).

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

Haghnegahdar, A., Razavi, S., Yassin, F., Wheater, H., 2017. Multicriteria sensitivity analysis as a diagnostic tool for understanding model behaviour and characterizing model uncertainty. Hydrol. Process. 31, 4462–4476. https://doi.org/10.1002/hyp.11358.

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. 571, 332–348. https://doi.org/10.1016/j.jhydrol.2019.01.048.

Herman, J.D., Kollat, J.B., Reed, P.M., Wagener, T., 2013. From maps to movies: high-resolution time-varying sensitivity analysis for spatially distributed watershed models. Hydrol. Earth Syst. Sci. 17, 5109–5125. https://doi.org/10.5194/hess-17-5109-2013.

Höllering, S., Wienhöfer, J., Ihringer, J., Samaniego, L., Zehe, E., 2018. Regional analysis of parameter sensitivity for simulation of streamflow and hydrological fingerprints. Hydrol. Earth Syst. Sci. 22, 203–220. https://doi.org/10.5194/hess-22-203-2018.

Holmes, T., 2016. isoWATFLOOD Stable water isotope simulation in the WATFLOOD hydrologic model.

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.

Intsiful, A., Ambinakudige, S., 2021. Glacier cover change assessment of the Columbia Icefield in the Canadian rocky mountains, Canada (1985–2018). Geosciences 11, 1–9. https://doi.org/10.3390/geosciences11010019.

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.
Klaus, J., Chun, K.P., McGuire, K.J., McDonnell, J.J., 2015. Temporal dynamics of catchment transit times from stable isotope data. Water Resour. Res. 51, 4208–4223. https://doi.org/10.1002/2014WR016247.
Klaus, J., McDonnell, J.J., 2013. Hydrograph separation using stable isotopes: review and evaluation. J. Hydrol. 505, 47–64. https://doi.org/10.1016/J.JHYDROL.2013.09.006.
Knoben, W.J.M., Freer, J.E., Woods, R.A., 2019. Technical note: inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. Hydrol. Earth Syst. Sci. 23, 4323–4331. https://doi.org/10.5194/hess-23-4323-2019.
Kouwen, N., 2018. WATFLOOD/WATROUTE Hydrological Model Routing & Flood Foresting System [WWW Document]. URL ⟨www.watflood.ca⟩.
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.
Nash, J.E., Sutcliffe, J. v, 1970. River flow forecasting through conceptual models part I—A discussion of principles. J. Hydrol. 10, 282–290.
Nenzén, H.K., Price, D.T., Boulanger, Y., Taylor, A.R., Cyr, D., Campbell, E., 2020. Projected climate change effects on Alberta's boreal forests imply future challenges for oil sands reclamation. Restor. Ecol. 28, 39–50.
Newman, A.J., Mizukami, N., Clark, M.P., Wood, A.W., Nijssen, B., Nearing, G., 2017. Benchmarking of a physically based hydrologic model. J. Hydrometeorol. 18, 2215–2225. https://doi.org/10.1175/JHM-D-16-0284.1.
Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation, and confirmation of numerical models in the earth sciences Naomi Oreskes; Kristin Shrader-Frechette; Kenneth Belitz. Science 263, 641–646.
Peters-Lidard, C.D., Clark, M., Samaniego, L., Verhoest, N.E.C., van Emmerik, T., Uijlenhoet, R., Achieng, K., Franz, T.E., Woods, R., 2017. Scaling, similarity, and the fourth paradigm for hydrology. Hydrol. Earth Syst. Sci. 21, 3701–3713. https://doi.org/10.5194/hess-21-3701-2017.
Pfannerstill, M., Guse, B., Reusser, D., Fohrer, N., 2015. Process verification of a hydrological model using a temporal parameter sensitivity analysis. Hydrol. Earth Syst. Sci. 19, 4365–4376. https://doi.org/10.5194/hess-19-4365-2015.
Razavi, S., Gupta, H. v, 2016a. 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., Gupta, H. v, 2016b. A new framework for comprehensive, robust, and efficient global sensitivity analysis: 2. Application. Water Resour. Res. 52, 440–455. https://doi.org/10.1002/2015WR017559.
Razavi, S., Gupta, H. v, 2015. What do we mean by sensitivity analysis? the need for comprehensive characterization of "global" sensitivity in Earth and Environmental systems models. Water Resour. Res. 51, 3070–3092. https://doi.org/10.1002/2014WR016527.
Razavi, S., Sheikholeslami, R., Gupta, H. v, Haghnegahdar, A., 2019. VARS-TOOL: A toolbox for comprehensive, efficient, and robust sensitivity and uncertainty analysis. Environ. Model. Softw. 112, 95–107. https://doi.org/10.1016/j.envsoft.2018.10.005.
Rodriguez, N.B., Klaus, J., 2019. Catchment travel times from composite storage selection functions representing the superposition of streamflow generation processes. Water Resour. Res. 55, 9292–9314. https://doi.org/10.1029/2019WR024973.
Rosa, L., Davis, K.F., Rulli, M.C., D'Odorico, P., 2017. Environmental consequences of oil production from oil sands. Earth's Fut. 5, 158–170. https://doi.org/10.1002/2016EF000484.
Sahraei, S., Asadzadeh, M., Unduche, F., 2020. Signature-based multi-modelling and multi-objective calibration of hydrologic models: application in flood forecasting for Canadian Prairies. J. Hydrol. 588, 125095 https://doi.org/10.1016/J.JHYDROL.2020.125095.
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 https://doi.org/10.1002/2013MS000293.
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. https://doi.org/10.1016/j.jhydrol.2015.02.013.
Spangenberg, J.E., 2012. Caution on the storage of waters and aqueous solutions in plastic containers for hydrogen and oxygen stable isotope analysis. Rapid Commun. Mass Spectrom. 26. https://doi.org/10.1002/rcm.6386.
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. Hydrol. Sci. J. 65, 1525–1538. https://doi.org/10.1080/02626667.2020.1751847.
Stahl, K., Moore, R.D., Shea, J.M., Hutchinson, D., Cannon, A.J., 2008. Coupled modelling of glacier and streamflow response to future climate scenarios. Water Resour. Res. 44 https://doi.org/10.1029/2007WR005956.
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.
van Huijgevoort, M.H.J., Tetzlaff, D., Sutanudjaja, E.H., Soulsby, C., 2016. Using high resolution tracer data to constrain water storage, flux and age estimates in a spatially distributed rainfall-runoff model. Hydrol. Process. 30, 4761–4778. https://doi.org/10.1002/hyp.10902.
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. Can. J. For. Res. 30, 283–287.
Wagener, T., McIntyre, N., Lees, M.J., Wheater, H.S., Gupta, H.V., 2003. Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis. Hydrol. Process. 17, 455–476. https://doi.org/10.1002/HYP.1135.
Wan, W., Zhao, J., Popat, E., Herbert, C., Döll, P., 2021. Analyzing the impact of streamflow drought on hydroelectricity production: a global-scale study. Water Resour. Res. 57, e2020WR028087 https://doi.org/10.1029/2020WR028087.
Westerberg, I.K., Sikorska-Senoner, A.E., Viviroli, D., Vis, M., Seibert, J., 2020. Hydrological model calibration with uncertain discharge data. Hydrol. Sci. J. 1–16. https://doi.org/10.1080/02626667.2020.1735638.
Yamanaka, T., Ma, W., 2017. Runoff prediction in a poorly gauged basin using isotope-calibrated models. J. Hydrol. 544, 567–574. https://doi.org/10.1016/j.jhydrol.2016.12.005.
Yilmaz, K.K., Gupta, H. v, Wagener, T., 2008. A process-based diagnostic approach to model evaluation: application to the NWS distributed hydrologic model. Water Resour. Res. 44 https://doi.org/10.1029/2007WR006716.