ADVANCING EARTH AND SPACE SCIENCES

Water Resources Research®

RESEARCH ARTICLE
10.1029/2023WR035509

Integrating Tracers and Soft Data Into Multi-Criteria Calibration: Implications From Distributed Modeling in a Riparian Wetland

Songjun Wu[1,2], Doerthe Tetzlaff[1,2,3], Xiaoqiang Yang[1,4], Aaron Smith[1], and Chris Soulsby[1,3]

[1]Department of Ecohydrology, Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany, [2]Department of Geography, Humboldt University Berlin, Berlin, Germany, [3]Northern Rivers Institute, School of Geosciences, University of Aberdeen, Aberdeen, UK, [4]Department of Aquatic Ecosystem Analysis and Management, Helmholtz Centre for Environmental Research—UFZ, Magdeburg, Germany

Key Points:

- Multi-criteria calibration outperforms single-criterion calibration with improved process representation
- Isotopes are highlighted as appropriate auxiliary data to effectively constrain the model with relatively small weights (0.1)
- Checking against soft data allowed further model rejection and revealed incompatibilities between data and calibration

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

S. Wu,
songjun.wu@igb-berlin.de

Citation:

Wu, S., Tetzlaff, D., Yang, X., Smith, A., & Soulsby, C. (2023). Integrating tracers and soft data into multi-criteria calibration: Implications from distributed modeling in a riparian wetland. *Water Resources Research*, *59*, e2023WR035509. https://doi.org/10.1029/2023WR035509

Received 15 JUN 2023
Accepted 23 OCT 2023

Author Contributions:

Conceptualization: Songjun Wu, Doerthe Tetzlaff, Chris Soulsby
Methodology: Songjun Wu, Chris Soulsby
Supervision: Doerthe Tetzlaff, Chris Soulsby
Writing – original draft: Songjun Wu
Writing – review & editing: Songjun Wu, Doerthe Tetzlaff, Xiaoqiang Yang, Aaron Smith, Chris Soulsby

Abstract Calibrating distributed hydrological models often leads to equifinality due to complex model structures, which can be further exacerbated in wetlands due to spatio-temporal heterogeneity in ecohydrological processes. Here, step-wise calibrations of the physically-based distributed model EcH$_2$O-iso was conducted in a data-rich wetland by minimizing a weighted average of the errors on discharge, stream isotopes, groundwater (GW) isotopes, and soil moisture. Results showed multi-criteria calibration outperformed single-criterion calibration as it strongly increased the overall performance, yet only marginally degraded performance of each calibration target. Isotopes were highlighted as appropriate auxiliary data as they effectively constrained the model with relatively small weights (0.1). However, those parameter sets that minimize the errors could still lead to physically implausible simulations of uncalibrated internal states or fluxes. This was further demonstrated by an approach developed to check internal fluxes based on soft data (transpiration and lateral flow), suggesting 54% of optimized models gave “right answers for the wrong reasons.” By excluding those models against soft data, such an approach further constrained equifinality, and unraveled potential inconsistencies between observations and calibration. Modeling represented the wetland as a slow-draining system mainly fed by GW, but also influenced by near-surface flow during winter or summer convectional events. Further, heterogeneity in hydrological functioning was partly attributed to distinct evapotranspiration patterns between contrasting vegetation communities. Therefore, this study not only provided insights into wetland functioning, but also revealed potential equifinality even with abundant data for calibration, and potential solutions based on the integration of isotopes and soft data.

1. Introduction

Wetlands have been widely acknowledged as vital elements of the landscape due to the ecosystem services they provide through the storage and slow release of water resources that also sustains rich biodiversity and significant carbon sequestration (Aumen & Keddy, 2001). From a hydrological perspective, wetlands can buffer the water cycle by regulating water flow paths (infiltration, percolation, overland flow, groundwater (GW) flow, etc.), often attenuating flood peaks through internal storage dynamics, and subsequently sustaining dry-weather baseflows. Consequently, they have long been identified as a “hot-spot” and focus for hydrological research (Acreman et al., 2007). However, the underlying processes that govern the hydrological function of specific wetlands are generally characterized by marked spatio-temporal heterogeneity with numerous local factors interacting in complex ways (e.g., climate, topography, soils, etc.) (Musolff et al., 2015). This complexity is further exacerbated in riparian wetlands due to highly diverse vegetation communities along hydrological gradients and active exchange between different water sources (soil water, GW, streamflow, etc.) (Acreman et al., 2007; Zedler & Kercher, 2005). Therefore, substantial knowledge gaps still exist regarding how to quantify hydrological pathways in wetlands despite decades of efforts in geographically diverse systems (e.g., Bam & Ireson, 2019; Hayashi et al., 2016).

Distributed hydrological modeling is one way to further investigate these dynamic patterns and processes due to its capacity to leverage spatial information in the forcing data, via regional parameterization and thus better identify dynamic flow paths in space and time (Wellen et al., 2015). However, while spatial disaggregation of a model domain brings extra information, it poses additional challenges. Equifinality, one of the most common and intractable issues in lumped modeling, is exacerbated by distributed models, evidenced by multiple behavioral parameter sets being able to produce the same levels of performance for the observational data during model calibration (Beven, 2006). This can substantially increase uncertainty in distributed models, as it not only originates from the compensatory effects between different model components (i.e., various pathways and storages), but also from their contributions from different geographic areas (hill slopes, variable source areas, and sub-catchments) (Cao et al., 2006). In the other words, the simulated internal fluxes could be highly uncertain in both magnitude and spatial patterns, even though outlet streamflow was successfully reproduced (Beven, 2006), which has been summarized as models giving "right answer for the wrong reasons" (Kirchner, 2006). This remains a central challenge in distributed modeling, considering that many applications are based only on the streamflow at outlet (reviewed in Fatichi et al., 2016; Wellen et al., 2015 for hydrological and water quality modeling).

One way to mitigate equifinality and increased confidence in model results is by multi-criteria calibration; that is, constraining a model's degrees of freedom in simulating internal fluxes by incorporating more observational data during calibration. This method has been increasingly applied in recent years due to the increasing data availability, with many different types of auxiliary data used for model calibration, for example, snowpack volume (Berezowski et al., 2015), soil moisture (Smith et al., 2021), GW head (Jing et al., 2018), evaporation (Winsemius et al., 2008), transpiration (Douinot et al., 2019), etc. Crucially, water stable isotopes (hydrogen ^{2}H and oxygen ^{18}O) have also been increasingly used in distributed modeling to trace water sources, flow paths and transit times (Birkel et al., 2014; Holmes et al., 2020; Smith et al., 2021; Soulsby et al., 2015), because they are only mediated by water mixing and fractionation but remain independent from biogeochemical reactions, and naturally integrate field-scale heterogeneity (Tetzlaff et al., 2015).

When reviewing various applications of multi-criteria calibrations using hydrological models, many of them showed benefits in terms of an increased number of diagnostics brought by auxiliary data (Birkel et al., 2014; Clark et al., 2011; Kuppel et al., 2018a; Piovano et al., 2018; Seibert & McDonnell, 2002), which help reject infeasible models, and thus reduce the dispersion in optimized parameter sets and simulated fluxes within the remaining behavioral models. This is generally accompanied by a better overall model performance at the expense of relatively marginal degradation in the simulation of different observations, which substantially enhances the overall consistency of the modeling results, as the optimized model is closer to the global solution within the parameter space rather than the local optima achieved via single-criterion calibration (Piovano et al., 2018).

However, some studies have shown multi-criteria calibration is not always a panacea, as the trade-off between the performances of different observations can be pronounced. For example, Fenicia et al. (2008) saw a significant degradation in streamflow performance when isotopes were added to model calibration. A similar degraded performance after including isotopes was also found in Scudeler et al. (2016). Besides, poor predictive ability was found for the internal processes in catchment function though various observations were used for model optimization (Cao et al., 2006). Other challenges include reduced parameter identifiability, which is a common corollary of introducing tracer-based metrics into calibration (Birkel & Soulsby, 2015; Holmes et al., 2020). Many reasons have been suggested in previous studies, which can be roughly grouped into three aspects: (a) the information contained in observational data, (i.e., data sets have overlapping information that inform the calibration process with conflicting or inconsistent information, and thus pulling the model in different directions; Clark & Vrugt, 2006; Kuppel et al., 2018a), (b) errors in model structure or inappropriate process conceptualization (Beven, 2006; McDonnell et al., 2007), and (c) incommensurability between data and model (e.g., the scale difference in model conceptualization and point-scale measurements, Piovano et al., 2018; Weiler & Naef, 2003). In this context, "soft" data (qualitative information or measured data that are not directly comparable to model variables) are sometimes used to further constrain model uncertainty by specifying additional criteria to judge model simulations or the selection of model parameters (Seibert & McDonnell, 2002; Winsemius et al., 2008). However, soft data may themselves reflect some considerable uncertainties (Sherlock et al., 2000), and their inclusion often introduces subjectivity (such as the specification of evaluation rules and the weighing of the different objective functions; Seibert & McDonnell, 2002). Consequently, clear and consistent guidance for multi-criteria calibration that integrates both hard and soft data is still not available; and there remains a need to further study the pros and cons of the approach via more applications in data-rich areas to better constrain the modeling uncertainty in the future.

A data-rich riparian wetland located in the experimental catchment of Demnitzer Mill creek in north-east Germany provided a unique opportunity for such analyses. Identified as a "hot-spot" for hydrological and biogeochemical processes in the catchment (Smith et al., 2021; Wu et al., 2022a), different types of data have been

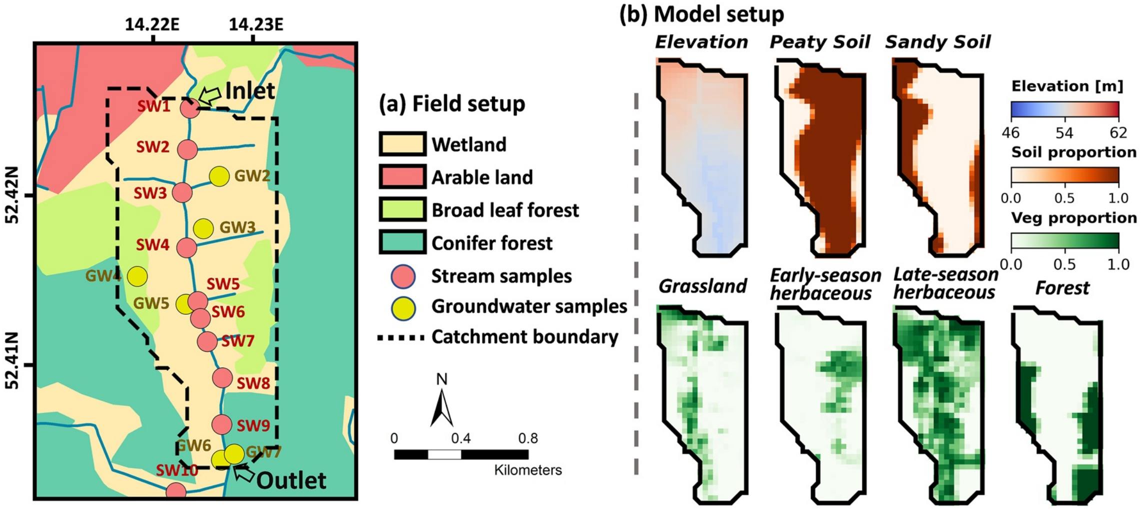

Figure 1. The field setup (a) and model conceptualization (b) of the riparian wetland.

extensively monitored in the wetland for at least 2 years, including (a) water level/discharge at the inlet and outlet, (b) stream isotopes at nine sites, (c) GW isotopes at six sites, and (d) surface soil moisture content at ~30 sites. Moreover, multispectral images acquired from monthly unmanned aerial vehicle (UAV) flights and leaf area index (LAI) times series acquired from remote sensing products provided a detailed representation of the topography of the model domain plus vegetation dynamics in space and time. Based on this thorough consideration of the spatio-temporal heterogeneity in forcing data sets, the grid-and-physics-based model EcH$_2$O-iso was setup and calibrated in this wetland. To thoroughly investigate the roles of weights in multi-criteria calibration, a total of 286 sets of weights were assigned to discharge, stream isotopes, GW isotopes, and soil moisture. The overarching goal of this application was to unravel the heterogenous spatio-temporal patterns of hydrological processes in the riparian wetland over a period of 2 years. Meanwhile, by summarizing the results from different calibration weights, we also sought to investigate the mechanisms behind the multi-criteria calibration (i.e., *how the different constraining data sets interact to influence the model calibration*, and *how soft data can further help constrain the equifinality*) with the following three research questions explored:

- How different constraining data sets interact to influence the model calibration?
- What are the challenges and uncertainties in quantifying wetland ecohydrological processes via multi-criteria calibration?
- How can we use soft data to reduce such uncertainty?

2. Methods and Materials

2.1. Study Site

The wetland studied in this investigation is located in the long-term experimental Demnitzer Millcreek catchment (DMC), 55 km SE of Berlin, Germany. The study area has a relatively flat topography (slope <2%) with an area of ~1.5 km^2 and is traversed by a stream of 2.1 km long, which also received drainage from several ditches to the east (Figure 1). Both land use and soil properties correlate with proximity to the stream: the riparian area is characterized by a wetland with peaty soils; while forests with more sandy soils dominate the areas further away from the stream (Figure 1a). Broad-leaved and conifer forests dominate most sandy areas, while vegetation communities in riparian wetland are highly complex and heterogenous. They were classified into grassland, early-season herbaceous, and late season herbaceous communities (Figure 1b) using a random forest model from monthly UAV imagery (for details please refer to Section 2.4 and Texts S1 and S2 in Supporting Information S1).

The study area experiences a typical continental climate with low annual precipitation (~570 mm/year since 1992) and high potential evapotranspiration (PET, 650–700 mm/year; Smith et al., 2021). The rainfall shows a distinct seasonality, with winter and summer rainfall characterized by low-intensity frontal events and intense convective events, respectively. Following an ongoing drought that started in 2018 (Kleine et al., 2020), the streamflow is intermittent within the study period, with annual average discharge of only 0.02 and 0.06 m^3/s at the inlet and outlet respectively (Wu et al., 2021).

In mid-2020, a beaver dam was built downstream of the outlet, increasing the inundated area and slowing streamflow upstream. A subsequent propagation of beaver habitat upstream was also observed in 2021 with several small dams built along the stream. However, these dams were generally temporary and small in size.

2.2. EcH$_2$O-Iso

EcH$_2$O-iso was recently developed by integrating an isotopic tracking module (Kuppel et al., 2018b) into the process-based ecohydrological model EcH$_2$O (Maneta & Silverman, 2013). As a fully distributed model, grid-based fluxes are simulated at each timestep based on the energy balance, water balance, and tracer balance (Figure S1 in Supporting Information S1). Here, we present a brief introduction on the model structure/parameterization and the adaption for this application; for detailed model descriptions please refer to Kuppel et al. (2018b) and Maneta and Silverman (2013).

2.2.1. Model Conceptualization

In each grid, the model resolves the mass balance of energy, water and tracer concentration sequentially for the vegetation canopy, soil surface, and three soil layers (Figure S1 in Supporting Information S1). The energy balance, driven by incoming shortwave and longwave radiation, as well as air temperature (minimum, maximum, and average), relative humidity, and wind speed, is simulated for each vegetation type at the canopy and soil surface level, where sensible heat, latent heat, net radiation, and ground heat fluxes are solved. Then the simulated heat fluxes are further used to estimate the heat-driven hydrological fluxes: evaporation from the canopy and first soil layer (estimated from the latent heat and available water storages), transpiration based on the transpiration-associated latent heat (regulated by the canopy temperature and the canopy conductance) and a Jarvis-type stomatal conductance model driven by maximum stomatal conductance and vapor pressure deficit, light availability, soil water availability, and root distribution in the three soil layers.

The water balance in EcH$_2$O-iso follows a typical multi-layer, top-down approach, with conceptualized bucket storages for canopy, surface, and three soil layers (Figure S1a in Supporting Information S1). Following a vertical sequence, the precipitation is first intercepted by the vegetation canopy, whose amount is determined by the LAI and maximum canopy storage (mm/LAI). Then throughfall or direct incident rainfall (on bare soils) reaches the surface and ponds. Water is infiltrated into soil layer 1 using the Green-Ampt model. Then it is further vertically redistributed from the upper to lower layers using a gravitational drainage model based on exceedance of field capacity, after which the seepage leaving the soil layer 3 is estimated via a leakage parameter.

After moving vertically, the lateral flow starts to route excess water across the catchment domain. Overland flow routes any ponded surface storage at the end of each timestep downslope following a steepest descent approach, until it re-infiltrates in a downstream cell or reaches the channel. GW flow, conceptualized as the water above field capacity in soil layer 3, is translocated downslope using a linear kinematic model driven by the local slope. Streamflow in the channel is routed using a non-linear kinematic wave model regulated by a scaled Manning's n to attenuate the hydrograph. Note that return flow from soil layers can occur as excess water in storage moving vertically to the surface when the entire soil profile is saturated. In this case, the overland flow represented the mixed water from surface ponding and upward water from the shallow soil layers, thus is called near-surface flow in this study.

The isotopic simulation follows a complete mixing assumption—inflow is fully mixed with storage whilst outflow shares the same composition with the storage (Kuppel et al., 2018b). Fractionation is allowed during all evaporation processes (from canopy, soil layer 1, and channel) using Craig-Gordon model (Craig & Gordon, 1964), while transpiration is assumed to be non-fractionating.

2.2.2. Model Adaptions

In the original EcH$_2$O-iso model, the lower boundary is set as soil layer 3, which excluded interactions with deeper, slower moving GW, and has been shown to be inappropriate for catchments with a strong interaction with such deeper GW stores (Yang et al., 2021). This leads to an adaption of additional groundwater layer (DeepGW; Yang et al., 2021), which was further revised and used for the studied wetland, as a strong surface-groundwater exchange in the area has been shown to be important in previous studies (Kleine et al., 2020). The initial DeepGW storage was set to 3 m in this study based on previous modeling results in DMC catchment (Smith et al., 2021) and recent geophysical surveys; as well as a preliminary analysis of stream and GW isotopes in this wetland (Wu, Tetzlaff, Goldhammer, et al., 2022). This storage is also similar to the values adopted by applications in nearby lowland catchments in Germany (e.g., Yang et al., 2021). The storage receives vertical leakage from soil layers and any lateral influx from upstream grid (also routed by a linear kinematic model), while its exfiltration to the channel is determined by the available storage and a weighting parameter. The weighting parameter considers both channel geometries and current GW level, regulating DeepGW recharge by its connecting area (A_t) with the channel bed:

$$A_t = \left[(W - D) + \sqrt{5} \times (D - \mathrm{GW}_t)\right] \times L \tag{1}$$

where D, L, and W denote the channel depth, channel length, and the width of river banks, while GW_t is the GW level below the surface at t time step in each grid. Note that here, we assume transects along the stream channel are isosceles trapezoid in shape with a bank slope of 200%, which matches our field observations since channels in DMC were artificially deepened for drainage in 1990s.

Similarly, the channel evaporation is linearly correlated to the channel surface within each grid ($W \times L$).

2.3. Data Used in the Modeling

The climatic forcing data, including precipitation, temperature, relative humidity, wind speed, air pressure, and net radiation (calculated from short wave and long wave radiation measured in both downward and upward directions), was monitored at an automatic weather station (Environmental Measurement Limited, UK) within the catchment at 15-min intervals. The precipitation was also collected for isotopic measurements on a daily basis, using paraffin in autosampler (ISCO) bottles to avoid evaporation effects.

As for hydrological data, water level was monitored at 15-min intervals at the inlet and outlet of the wetland, where discharge was measured using a sonTek-IQ-plus sonde (YSI, USA) to construct rating curves based on levels and manual gaugings. Six GW wells were also established across the study area, with water levels monitored by pressure transducers at GW3, 4, 5 and 7 (Figure 1a).

For model calibration and validation, four different data sets were collected within the studied wetland since 2020, including (a) the discharge at the outlet at a 15-min frequency, (b) water stable isotopes (δ^{18}O and δ^{2}H) measured from grab samples taken at nine stream water sampling sites (SW1-9) and (c) six groundwater wells (GW2-7, Figure 1a) on a biweekly basis, and (d) soil moisture content in the upper soils (first 10 cm) measured every month at ~30 sites that are evenly distributed over the northern wetland (Figure S2 in Supporting Information S1). The corresponding components in EcH$_2$O-iso are in-stream discharge, in-stream isotopes, isotopes in soil layer 3 (taken as GW samples), and soil moisture in soil layer 1, respectively.

Additionally, an important auxiliary data set was the multispectral imagery from monthly UAV flights, whose high spatial resolution (10 cm) provided detailed representation on the topography of the model domain, channel geometries, and vegetation dynamics in space and time.

2.4. Model Setup

The riparian wetland was delineated into 50 × 50 m grids. Most channel information (coordinates, length, and width) was determined from the UAV imagery and embedded into the grid-based model domain. The channel depth was interpolated from manual measurements at ~50 sites along the stream. Due to the relatively small area, all the climatic forcing data were regarded spatially homogenous within the model domain. Similarly, the isotopic composition in precipitation was assumed to be spatially uniform according to our previous distributed isotopic sampling across the catchment (Kleine et al., 2020).

To reflect the spatial heterogeneity, soils were classified as sandy and peaty soils that were uniformly distributed over the vertical soil columns, while vegetation was differentiated into four communities (grassland, early-season herbaceous, late-season herbaceous and forest) using a random forest model trained by multispectral imagery from monthly UAV flights (see details in Text S1 in Supporting Information S1). These soil and vegetation maps were further resampled into 50 × 50 m grids, from which soil or vegetation parameters were weight-averaged in each grid by their proportions (Table S1 in Supporting Information S1). Additionally, LAI was required as an indicator of vegetation dynamics, which was derived from Moderate Resolution Imaging Spectroradiometer (MODIS) product (MOD15A2) as a reference time series, and then differentiated into community-level based on the summarized normalised difference vegetation index (NDVI) for each community (see details in Text S2 in Supporting Information S1).

The model was set up to run on daily timesteps with an entire modeling period of 2 years (1 January 2020 to 31 December 2021). The calibration was conducted from the first summer (1 July 2020 to 31 August 2021) for ~1 year, while the remainder of the time series was used for validation (1 September 2021 to 31 December 2021). Note that the first half year was excluded from calibration because of the increased residence time resulting from the beaver dam construction and its effect on isotope dynamics.

The initial boundary conditions of key states and fluxes were slightly modified from the optimized results in previous EcH_2O-iso application in DMC (Smith et al., 2021). For each model run, a 1-year spin-up period was adopted via a repetitive use of the forcing data in 2020. As it was the first-time that EcH_2O-iso was applied specifically to this wetland, most parameters that cannot be directly measured were selected for optimization, including 10 soil-type-dependent parameters, 8 vegetation-dependent parameters, and 7 global (uniformly distributed) parameters (Table S1 in Supporting Information S1). This resulted in a total of 59 parameters to calibrate. The ranges of parameters were mostly specified from previous catchment-scale modeling in DMC (Smith et al., 2021), with the range of max canopy storages modified based on our monitored LAI (*MaxCanStorage*). The use of narrower parameter ranges accelerated the search of parameter space exponentially given the relatively high parameter number.

2.5. Model Calibration

Latin-Hypercube sampling was selected to generate samples of parameters from the initial space. According to previous modeling experience (Gillefalk et al., 2021; Smith et al., 2021) 200,000 samples could lead to a satisfactory calibration with similar number of calibrated parameters. Here, to increase the robustness of this study, the number of samples was further increased to 500,000 based on the availability of computation resources. We also monitored the mean parameter values of the posterior models selected from different numbers of model runs (Figure S3 in Supporting Information S1). The values of key parameter are generally stable after 10,000 model runs (though smaller changes could happen with respect to the prior parameter distribution). This, to some extent, supported the robustness of Latin-Hypercube sampling in this application.

After completing 500,000 model runs, the performance of each parameter set was evaluated. Here, an error function consisting of normal and logarithmic forms of Nash-Sutcliffe efficiency (NSE) was selected to calculate the deviation of simulations from each calibration target (i.e., discharge, stream isotopes, GW isotopes, and soil moisture). Such error function (Equation 2a) has been well tested and demonstrated to effectively mitigate the over-sensitivity towards high values or outliers when using the conventional NSE approach (Wu et al., 2022a):

$$\mathrm{err}_{i,j} = \min\left\{ \sqrt[6]{(1-\mathrm{NSE}_{i,j})^6 + (1-\mathrm{lnNSE}_{i,j})^6} \right\} \tag{2a}$$

$$\mathrm{lnNSE}_{i,j} = \mathrm{NSE}(\ln(\mathrm{sim}_{i,j}), \ln(\mathrm{obs}_i)) \tag{2b}$$

$$\mathrm{NSE}_{i,j} = \mathrm{NSE}(\mathrm{sim}_{i,j}, \mathrm{obs}_i) = 1 - \frac{\sum_{t=1}^{\mathrm{Nt}} (\mathrm{sim}_{i,j,t} - \mathrm{obs}_{i,t})^2}{\sum_{t=1}^{\mathrm{Nt}} \left(\mathrm{obs}_{i,t} - \overline{\mathrm{obs}_{i,t}}\right)^2} \tag{2c}$$

$\mathrm{sim}_{i,j}$ denotes the simulation of ith type of data ($i = 1, 2, 3, 4$ means discharge, stream isotopes, GW isotopes, and soil moisture) for jth model run ($j = 1, \ldots, 500{,}000$), while obs_i means the observation of ith type of data. Nt is the number of timesteps. Note that for each type of observations, all data points from multiple locations were used simultaneously (i.e., concatenating all available time series into a single time series) when calculating NSE and ln NSE. In other words, the simulation/observations at different locations were aggregated into $\mathrm{sim}_{i,j}/\mathrm{obs}_i$ in Equation 2b.

The calculated errors were then turned from their absolute values into rank values for each calibration target ($\text{err}_{\text{norm},i,j}$) to avoid the impact from different magnitudes or distributions of the errors between calibration targets. Then the overall performance was evaluated as a weighted average of normalized errors from four calibration targets for each (jth) model run:

$$\text{err}_{\text{norm},j} = \sum_{i=1}^{4} \text{err}_{\text{norm},i,j} * w_i;\ \text{where} \sum_{i=1}^{4} w_i = 1 \tag{3}$$

Here to thoroughly investigate how different observations affect the model calibration, 10 individual weights (w_i, ranging from 0 to 1 with an interval of 0.1) were assigned for each type of observations, resulting in a total of 286 combinations of weights (i.e., weight settings). Finally for each weight setting, the most 30 behavioral parameter sets were selected as those with highest ranking, and were used for post-analyses.

2.6. Model Evaluation

To evaluate the simulation performance of each calibration setting, two additional metrics (mean relative error MRE and predictive uncertainty PU*, both modified from Kuppel et al., 2018a) were used to estimate the relative deviation from observations (Equation 4a) and the simulation uncertainty (Equation 4b) for each observation type in both calibration and validation period (1 July 2020 to 31 December 2021):

$$\text{MRE}_{i,j} = \frac{\sum_{t=1}^{\text{Nt}} |\text{MRE}_{i,j,t}|}{\text{Nt}};\quad \text{where}\ \text{MRE}_{i,j,t} = \frac{\sum_{k=1}^{N_{\text{eval}}} \text{sim}_{i,j,t,k}/N_{\text{eval}} - \text{obs}_{i,t}}{\text{obs}_{i,t}} \tag{4a}$$

$$\text{PU}^{*}_{i,j} = \frac{\sum_{t=1}^{\text{Nt}} \text{PU}_{i,j,t}}{\text{Nt}};\quad \text{where}\ \text{PU}_{i,j,t} = \frac{M_{95,k}(\text{sim}_{i,j,t}) - M_{5,k}(\text{sim}_{i,j,t})}{\sum_{k=1}^{N_{\text{eval}}} \text{sim}_{i,j,t,k}/N_{\text{eval}}} \tag{4b}$$

where i and j denote one of the 4 observation types and 286 calibration settings; Nt is the number of timesteps; N_{eval} is the number of optimized models for post-analysis (i.e., the 30 most behavior models); $M_{5,k}$ and $M_{95,k}$ denote the 5th and 95th percentile of simulations of the 30 behavior models. Note that for each observation type, simulations/observations at different locations were also aggregated before evaluation (the same as Equation 2b). The calculated MRE (Equation 4a) further allows to evaluate the overall performance for all observation types: the observation-specific $\text{MRE}_{i,j}$ was normalized among the 286 calibration settings and then assigned with identical weights, which finally leads to the overall MRE for all observation types ($\text{MRE}_{\text{overall}}$):

$$\text{MRE}_{\text{norm},i,j} = \frac{\text{MRE}_{i,j} - \min_j(\text{MRE}_{i,j})}{\max_j(\text{MRE}_{i,j}) - \min_j(\text{MRE}_{i,j})} \tag{5a}$$

$$\text{MRE}_{\text{overall},j} = \frac{\sum_{i=1}^{\text{No}} \text{MRE}_{\text{norm},i,j}}{\text{No}} \tag{5b}$$

where $\min_j(\text{MRE}_{i,j})$ and $\max_j(\text{MRE}_{i,j})$ respectively represent the minimum and maximum values within the ensemble of 286 MRE for ith observations; No denotes the number of observation types, which is 4 in this study. This eventually leads to the overall MRE ($\text{MRE}_{\text{overall},j}$) for each of the 286 calibration settings.

In addition, 11 different sets of weights for four observations (10 sets that are most commonly used in previous hydrological modeling—*CS*1–10, and the one with best overall performance in Equation 3—*CS*11, see details in Table 1) were selected, so as to better visualize the spatio-temporal patterns of simulated fluxes between calibration settings.

2.7. Soft Data Filtering

Due to the field sampling and monitoring over the past 30 years (Kleine et al., 2020), as well as recent modeling applications in DMC (Smith et al., 2021), additional empirical knowledge is available for this wetland (besides the direct measurements). For example, forest transpiration is higher than the grass communities (inferred from adjacent grassland and forest plots in DMC, see Kleine et al., 2020); overland flow is rarely generated over the wetland, while GW is the major source of stream water (Smith et al., 2021; Wu, Tetzlaff, Goldhammer,

Table 1
The Weight Settings Used for Post-Analysis (Numbers Denote the Weights for Corresponding Observations)

	CS1	CS2	CS3	CS4	CS5	CS6	CS7	CS8	CS9	CS10	CS11
Discharge	1				0.5	0.5	0.5	0.4	0.4	0.3	0.4
Stream isotopes		1			0.5			0.3	0.3	0.3	0.1
GW isotopes			1			0.5		0.3		0.2	0.1
Soil moisture				1			0.5		0.3	0.2	0.4

Note. *CS*11 is the weight setting resulting in the best overall simulation performance (selected based on Equation 4a).

et al., 2022). However, such knowledge is so-called "soft" data (in contrast to "hard" data that can be directly used for calibration) because these former measurements were close to, but not located in, the modeling domain; while the latter information cannot be quantitively represented. Therefore, unlike some previous studies which quantitively incorporated soft data into calibration (e.g., Seibert & McDonnell, 2002), here we conducted a posterior check based on consistency with this empirical knowledge. More specifically, after selecting the 30 best-performing models (for each weight setting), the internal fluxes of each model were checked against the field knowledge (forest transpiration > non-forest transpiration; near-surface flow < GW flow + DeepGW flow); only models that fulfill both criteria could pass the soft data check and be retained as "plausible models."

3. Results

3.1. Simulation Performance With Different Constraining Data sets

3.1.1. Single-Criterion Calibration

The discharge at outlet could be simulated well when calibrating against its observation (*CS*1, NSE = 0.76, Figure 2a), while calibration based on other observations only captured the general seasonal patterns yet performed worse in peaks given the relatively low NSE (0.38–0.5). This is clearer for the validation period, as NSE remained relatively high (0.62) for discharge-based calibration but dropped to 0.1–0.4 for other observations. The gradual enrichment of isotopes along the stream was correctly simulated under all calibration settings for both calibration and validation periods; however, their exact values could only be adequately captured when calibrating against the corresponding isotopic observation (*CS*2), while strong overestimation was found when using the other data sets for optimization (Figure 2b). The differences in GW isotopic simulation were not very marked between calibration settings, but incorporating GW isotopes did slightly improve the performance for both calibration and validation periods (*CS*3, Figure 2c). Finally, soil moisture simulation was clearly improved when it was calibrated against observed soil moisture, while strong deviations (mostly overestimation) were observed under other calibration settings (*CS*4, Figure 2d). However, this best simulation of soil moisture that was directly calibrated against its observation still showed certain deviation with MRE of 0.24.

3.1.2. Multi-Criteria Calibration

More insights on the performance of multi-criteria calibrations were gained when the deviation and uncertainty of simulations are compared between calibration settings with not only a single constraining data set but also combination of multiple data sets. In Figure 3, we explored the simulation deviations (Figure 3b) and PU (Figure 3c) of different types of observations (*y*-axis) when calibrating against different sets of weight combinations (*x*-axis). In general, the simulation deviation and uncertainty of a specific variable was significantly reduced when calibrating against its observation, especially the soil moisture whose performance shows highest dependency on the weights (Figure 3d). This is further demonstrated when checking the full distribution of weights and normalized deviations of observations (Figures 3a and 3b): a negative relationship with weights was found for soil moisture and discharge. Interestingly, however, such a linear relationship was not observed when calibrating against isotopes, as their deviations remained relatively low until an abrupt increase when weights dropped below 0.1 (Figure 3b). In terms of the predictively uncertainty, negative relationships (to weights) were also found (Figure 3c). The only difference is that discharge reacted to weights less linearly.

Further checking the correlation between different observations, we found that the performance of most observations degraded when more weights were assigned for other observations. However, such degradation was relatively mild between discharge, stream isotopes and GW isotopes. For example, the deviation of GW isotope

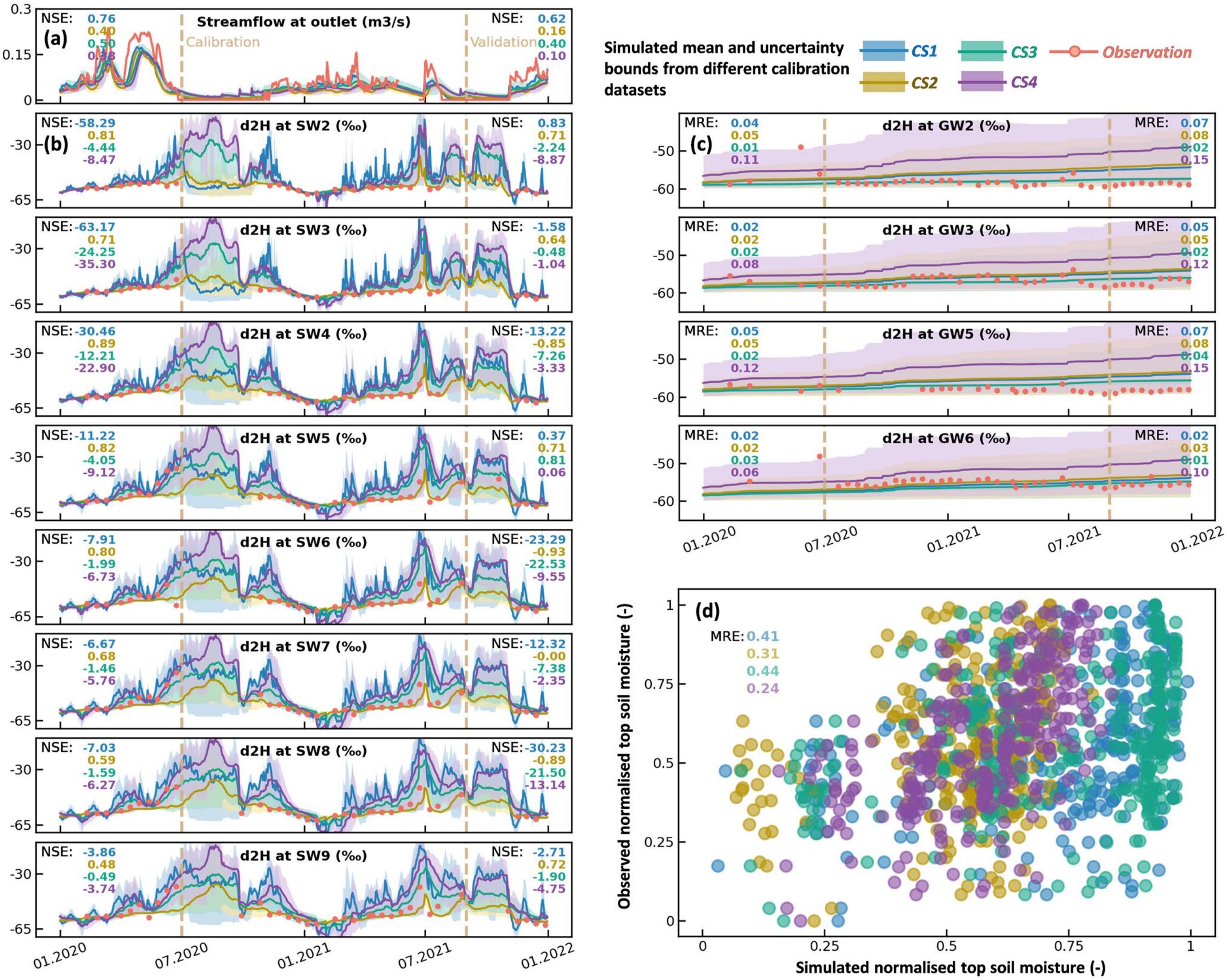

Figure 2. The mean and 90% bounds of simulated streamflow (a), streamwater ^{2}H (b), groundwater ^{2}H (c), and normalized soil moisture (d) from the best 30 parameter sets. The performances of parameter sets selected from different observation data sets are shown in different colors. The performance metrics were respectively evaluated for calibration (1 July 2020 to 31 August 2021, on the left side) and validation period (1 September 2021 to 31 December 2021, on the right side). Note that the period prior to dam construction was excluded from validation.

simulation either did not or only slightly increased when discharge was weighted more (Figures 3a and 3b). Similarly, discharge simulations also remained relatively consistent with increased weights of stream and GW isotopes. However, when incorporating soil moisture into calibration, the degradation of simulation performance on the remaining observations was much stronger, especially for GW isotopes and discharge (Figure 3b).

The compensation between information brought by different types of observations is also reflected by the parameters from the selected behavioral models. As is shown in Figure S8 in Supporting Information S1, the posterior distribution of most parameters was relatively discrete give their wide ranges. Moreover, the parameter distribution showed considerable differences between different weight settings. Apparently, information embedded in observations are different under the current calibration scheme and model setup.

3.2. Simulation of Internal Fluxes With Different Constraining Data Sets

Generally, the simulated internal fluxes showed strong discrepancies in both magnitude and spatial patterns between calibration settings (Figure 4a), and reacted very differently toward changes of weights (Figure 4b). Among the four observations, the internal fluxes were most diverse when discharge weighted increased; or in the other words, there were many pathways to achieve an ideal simulation of outlet discharge. This is because the

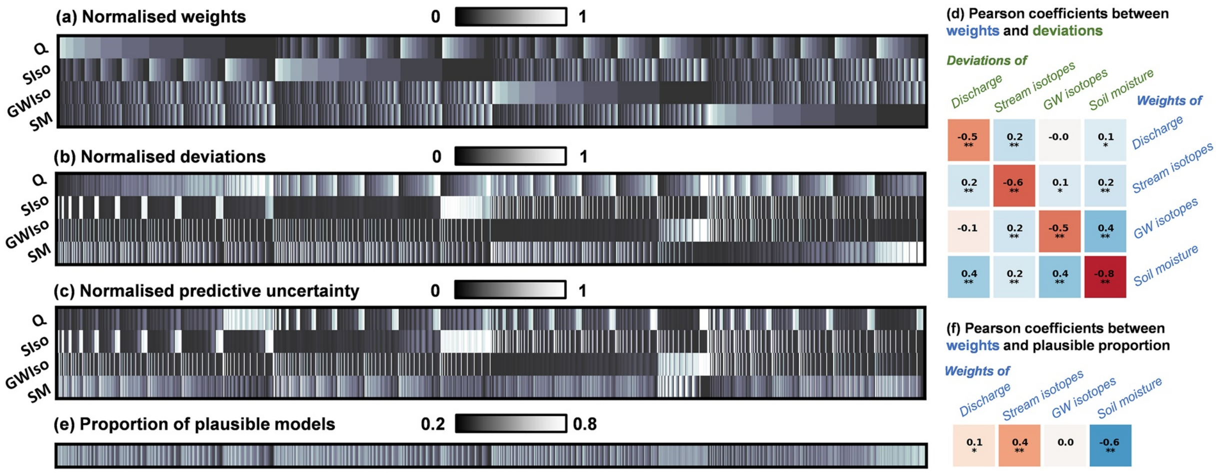

Figure 3. The simulation performances between different sets of weights. For all heatmaps, each column on *X* axis represents one of the 286 weight combinations, while *Y* axis corresponds to four types of measured data, that is, discharge (Q), stream isotopes (SIso), groundwater isotopes (GWIso), and soil moisture (SM) from top to bottom. Subplot (a), (b), and (c) respectively show the normalized weights, and the resulting simulation deviations and predictive uncertainties for each type of data, while subplot (d) shows the correlation between weights and simulation deviations (*/**: significant/highly statistically significant). Further, subplot (e) shows how many optimized models were plausible after posterior-check against soft data, while subplot (f) shows how that plausible proportion related to calibration weights.

discharge-oriented models (e.g., *CS*1, 5, and 6 in Figure 5) inherently produced relatively high total lateral flow (thus easily compensating between different components) across the model domain in order to generate enough streamflow to fill the gaps between discharge at the inlet and outlet of the wetland. In contrast, stream isotopes offered more consistent solutions—reducing the soil and channel evaporation, or decreasing the near-surface flow while increasing GW and deep GW flow—to reach a better isotopic performance (Figure 4b). This is because high lateral flow from discharge-oriented calibration often led to overestimation of near-surface flow, which was generally isotopically-enriched due to soil evaporative fractionation, and thus resulted in the over-enrichment of simulated stream isotopes. These solutions are clearly shown in selected models in Figure 5, for example, the depressed near-surface flow in *CS*2, 8, and 9 and lower soil evaporation in non-forest area in *CS*2, 9, and 10. This further led to reduced contribution of near-surface flow to stream water balance (*CS*2, 9 in Figure S5 in Supporting Information S1).

Similarly, calibrations based on GW isotopes also offered a simple and directional solution to mitigate the isotopic overestimation, that is, reducing the infiltration through soil layers and percolation to GW (Figure 4b as well as *CS*3 in Figure 5), so that enriched isotopic inputs from upper layers were limited. In terms of the soil moisture

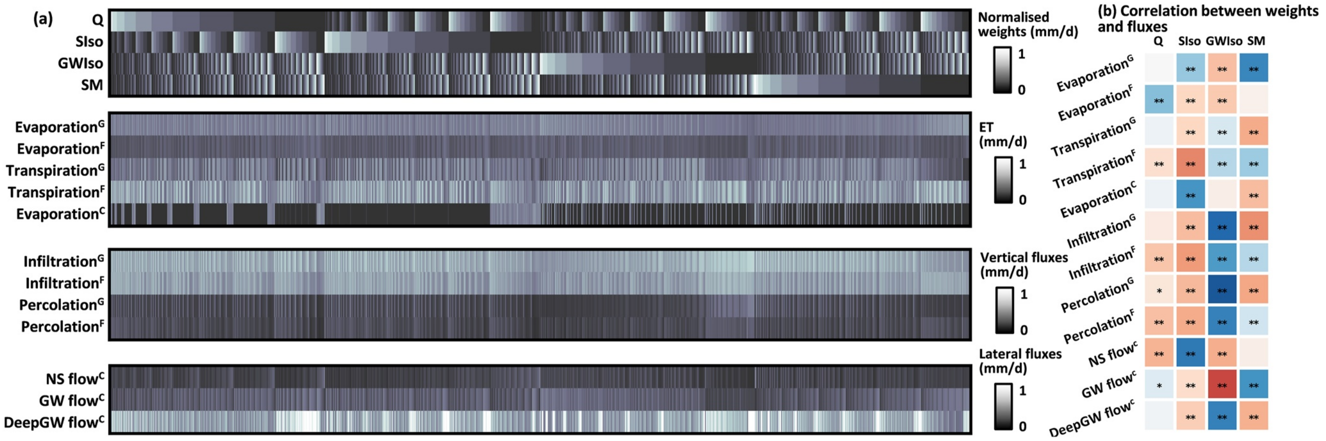

Figure 4. The simulated fluxes (ET, vertical fluxes, and lateral flows in daily mean, mm/day) of 30 most behavioral models selected by different calibration weights (in *X* axis) were shown in subplot (a). Subplot (b) shows the relationship between weights and fluxes (*/**: significant/highly statistically significant). Suffix G, F, C denote grassland, forest, and channel, respectively. The color blue and red in subplot (b) respectively mean the positive and negative slope between variables.

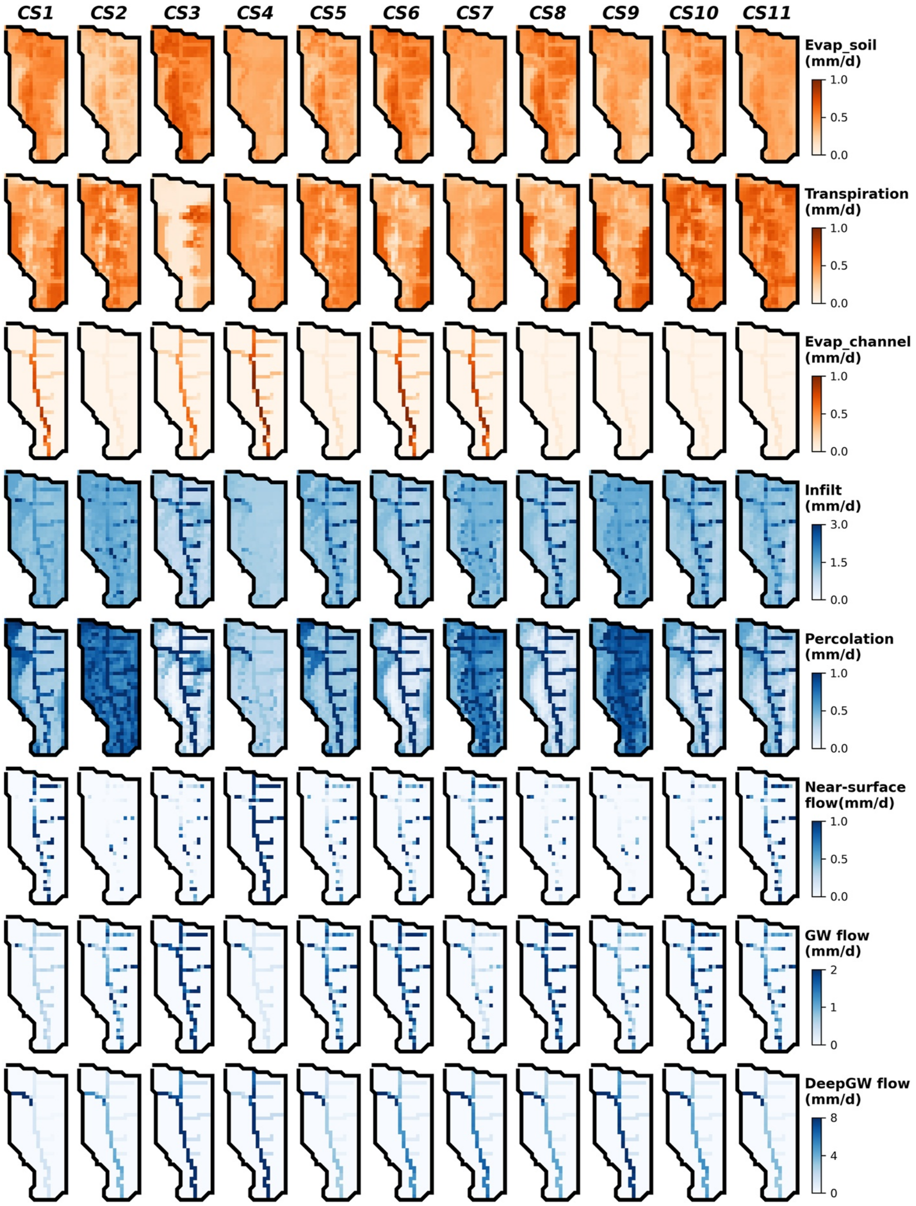

Figure 5. The spatial patterns of hydrological fluxes (averaged from the 30 most behavioral models into daily mean, mm/day) under different weight settings. For difference maps between weight settings please refer to Figure S4 in Supporting Information S1.

simulation, the main difficulty is to capture the extreme spatial heterogeneity in the wetland (see measurements in Figure S2 in Supporting Information S1), which is inherently related to transpiration discrepancies between four different vegetation communities. Therefore, the model driven by soil moisture calibration relied more on transpiration rather than soil evaporation to reproduce such soil moisture heterogeneity that were driven by the vegetation (Figure 4b).

The spatial contrasts between alternative calibration settings were also pronounced after further differentiating the internal fluxes at soil type/vegetation community level (see Figure 6 where fluxes were averaged among the grid cells where the proportion of a specific soil or vegetation type exceeds 90%). For example, while most constraining data sets showed ET fluxes across the vegetation communities with transpiration gradually decreasing in grassland < early-season herbaceous < late-season herbaceous < forest (and the opposite for soil evaporation), the GW isotope-based calibration (*CS*3 in Figure 6) suggested the strongest transpiration rate for early-season

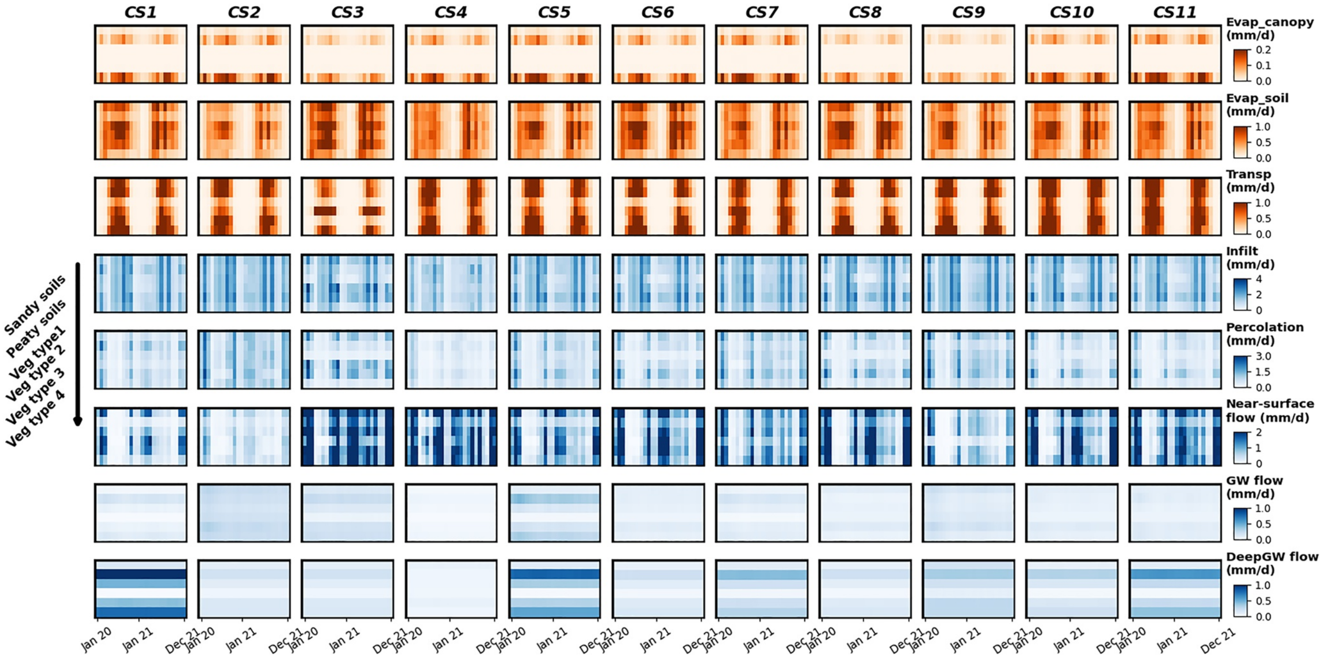

Figure 6. The monthly hydrological fluxes under different calibration settings (summarized by different soil and vegetation types). Vegetation type 1–4 denote grassland, early-season herbaceous, late-season herbaceous and forest, respectively. The fluxes were averaged into daily mean (mm/day) among the grid cells where the proportion of a specific soil or vegetation type exceeds 90%.

herbaceous species. Interestingly, like the spatial maps in Figure 4, internal fluxes were again more homogenous across the soil types or vegetation species when using all the data sets for calibration (*CS*10 and *CS*11), especially for the soil evaporation and transpiration.

In contrast to the substantial discrepancies in spatial patterns, the temporal dynamics of simulated internal fluxes followed similar seasonal patterns in hydroclimatic drivers for all 10 calibration settings (Figure 6). For example, all the ET fluxes peaked around July, with transpiration mainly in the growing season (April to September) while soil evaporation extended over a longer period (March to October). Similarly, when screening the fluxes for hydrological flow paths under different calibration settings, they also shared relatively similar temporal trends (though behaved very differently regarding the magnitude, Figure 6). In general, the activation timing of infiltration, percolation, and near-surface flow was similar and physically realistic despite being constrained by different data sets, both related to the soil saturation (mostly in winter) or convective rainfall events (in summer); while the lateral flows within deeper layers (GW and DeepGW flow) were much more consistent.

3.3. Posterior-Check on Internal Fluxes Using Soft Data

By checking against soft data, 10%–70% of selected behavioral models fulfilled both criteria (i.e., *forest transpiration* > *non-forest transpiration* and near-*surface flow* < *GW flow* + *DeepGW flow*) and regarded as "plausible models." The average pass rate of the soft data check for the 286 calibration settings was 46% (Figure 3e). By examining the correlation between calibration weights and pass rates (Figure 3f), we found that assigning more weights to stream isotopes could significantly increase the proportion of plausible models, while the increase of soil moisture weights would drive model into less physical-realistic directions, leading to lower pass rates. Discharge and GW isotopes did not show significant correlation to the pass rate (Figure 3f).

Further investigation of the internal fluxes simulated by the selected plausible models (Figure 7; also see full distributions of fluxes in Figure S6a in Supporting Information S1) showed a clear difference compared to the ones without any soft data validation (Figure 5; see full distributions in Figure 4a). The main differences resulted from the soft-data check were related to ET and lateral flow components: the transpiration was constantly higher in forest than in riparian wetland, while sources from deeper layer was now the main component of lateral flow (Figure 7). Interestingly, the other internal fluxes produced by plausible models (e.g., infiltration, percolation,

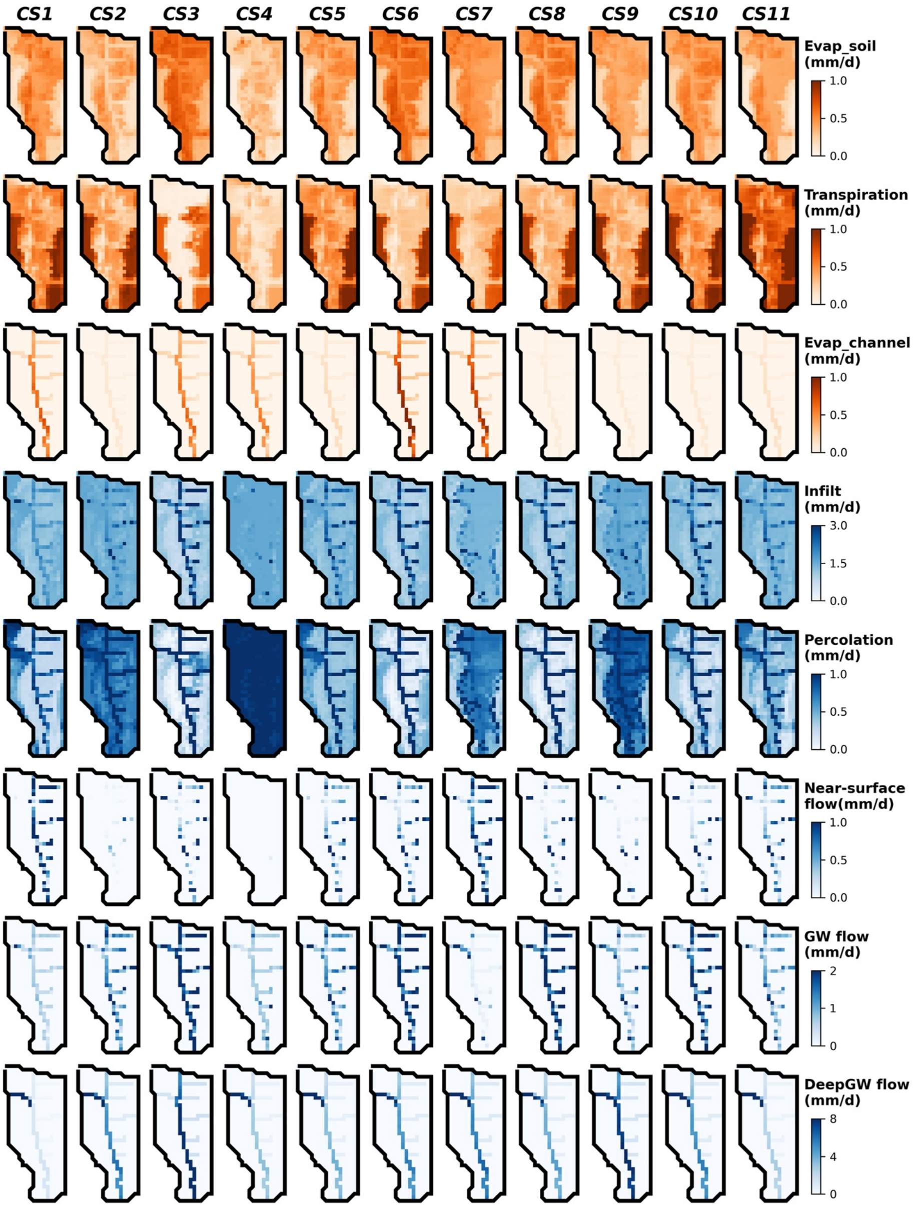

Figure 7. After filtered by soft data, the spatial patterns of hydrological fluxes under different weight settings (averaged from the 30 most behavioral models into daily mean, mm/day) were shown. For difference maps between weight settings please refer to Figure S7 in Supporting Information S1.

and channel evaporation) remained relatively unchanged, at least regarding the magnitude and spatial patterns (Figure 7). This is further demonstrated when checking the correlations between weights and internal fluxes, we obtained similar results for almost all the fluxes except evapotranspiration before (Figure 4b) and after soft data check (Figure S6b in Supporting Information S1).

4. Discussion

4.1. Benefits and Implications of Multi-Criteria Calibration

Equifinality has been long recognized as a major challenge for calibration in hydrological modeling (Beven & Freer, 2001). Especially when nowadays physics-based distributed models are increasingly developed and applied, their complex model structure, detailed process representation, and spatial disaggregation of the parameterization make them prone to equifinality (McDonnell et al., 2007; Wellen et al., 2015). Here, we further investigated this central issue by alternatively calibrating a physics-and-grid based model EcH$_2$O-iso in a riparian wetland using different calibration settings, either with single constraining data sets or the combination of multiple data sets.

Starting with discharge only as the calibration target, the calibrated model produced relatively high fluxes of all lateral flows (near-surface flow, GW flow, and DeepGW flow) in order to provide the flow increment between the wetlands' inlet and outlet (Figure 5), which resulted in a relatively good performance for discharge simulations. However, this also posed large uncertainties as all the variables/fluxes, except the outlet discharge, had large degrees of freedom regarding both their magnitude and spatial distributions, and thus could easily compensate one and other. This was evidenced by a strong overestimation of stream and GW isotope ratios, and the poor performance of spatial distributed soil moisture (Figure 2), indicating that while the optimized parameter sets had successfully calibrated discharge, they had failed to capture the other variables/fluxes in a plausible manner. This is consistent with the general perception that models are highly uncertain when calibrated against only outlet streamflow (Kirchner, 2006). As the celerity of the rainfall-runoff response, the outlet discharge integrates the effects of all the processes from upstream networks, whose divergent nature makes it difficult to uniquely backtrack the velocity of water, that is, *where* (at which locations within the model domain) and *how* (by which processes) the flow is generated (Birkel & Soulsby, 2015; Guse et al., 2016; Piovano et al., 2018).

From this perspective, multi-criteria calibration effectively reduced the uncertainty, as information contained in these auxiliary data sets helps to diagnose the simulated internal fluxes. For example, over-enriched stream isotopes indicated overestimated isotopic inputs via near-surface inflow; while the overestimation of GW isotopes suggested that inputs via infiltration and percolation from surface were too high (Figure 5). Based on this information, the model then either reduced the corresponding fluxes (near-surface flow in lateral transport or infiltration/percolation in vertical transport) or mitigated soil evaporation to constrain the isotopic enrichment through fractionation, thus pulling the process representation in a more physically realistic direction. Therefore, multi-criteria calibration did increase the credibility of calibration by rejecting the models with unfeasible simulations of internal fluxes, as has been reported previously (Clark et al., 2011; Kuppel et al., 2018a; Piovano et al., 2018; Smith et al., 2021).

Here we would like to emphasize the benefits of incorporating water isotopes into model calibration, because while most auxiliary hydrological variables used for calibration (soil moisture, evaporation, transpiration, etc.) are linked to the upper soil profiles (Efstratiadis & Koutsoyiannis, 2010; Wellen et al., 2015), the flow dynamics in deeper layers are often missing due to their poor measurability (Beven, 2001). In the other words, commonly used auxiliary variables mainly help constrain modeling of near-surface processes (e.g., evapotranspiration, infiltration, etc.), while simulation of deeper processes (percolation into deeper layers and exchange with GW) still has a certain degree of freedom and thus higher uncertainties. From this aspect, water isotopes are powerful because their concentrations reflect the cumulative effects of water mixing between all available storages, which inherently integrates the field heterogeneity and thus helps identify the velocity of different pathways, as was also demonstrated previously (Birkel et al., 2014; Holmes et al., 2020; Tetzlaff et al., 2015). More importantly, including isotopes could significantly increase the credibility of overall simulation with only small weights (given the weights of models producing the best overall performance were 0.4, 0.1, 0.1, 0.4 for discharge, stream and GW isotopes, and soil moisture in this study). Note that equal weights were used when evaluating the overall performance of different observations. In other words, each type of observations was regarded as an independent aspect of the physical realism of the model. Of course, a priority could be set for an observation of specific interest (e.g., discharge in flood prediction), but the calibration in this study favors the models that can reproduce as many types of observation as possible to avoid giving "the right answer for the wrong reasons" (Kirchner, 2006), which also fits the general concepts of many other multi-criteria calibrations (e.g., the use of Pareto front; Efstratiadis & Koutsoyiannis, 2010). In this context, isotopes are ideal auxiliary data for calibration, because they could effectively nudge the modeling into a more physically realistic direction with a relatively small expense of predictive accuracy for the other observations. Such implication could potentially be generic, because the model EcH$_2$O-iso has been tested not only in DMC (e.g., Smith et al., 2021), but also in many catchments spanning different climatic, soil, and vegetation backgrounds (e.g., Neill et al., 2021; Yang et al., 2021). Moreover, its structure (multiple soil layers plus conceptual reservoirs in deeper layers), though with a relatively simple, computationally efficient module for saturated subsurface flow compared to traditional GW-oriented models (e.g., ParFlow; Maxwell, 2013), has been widely adopted and tested for many hydrological models, for example, SWAT and mHM-Nitrate (Wellen et al., 2015). Therefore, sharing similar model structures and process conceptualization, the implications from EcH$_2$O-iso could be potentially informative to those models and contribute to the modeling community that focus more on near-surface hydrological processes.

Notably however, assigning weights toward different observations could still be difficult in multi-criteria calibration, even with the knowledge that isotopes probably need very limited focus according to our results. This is due to the fact that the performance of other hydrological observations (i.e., discharge and soil moisture) was linearly correlated to assigned weights in this study (Figures 3a and 3b), which is highly likely for future applications given our similar experience on calibration against these types of catchment-scale hydrological models (Smith et al., 2021; Wu et al., 2022a; Yang et al., 2023). In the other words, regular hydrological observations would require a certain proportion of weights for an acceptable representation, and that proportion would be unpredictable as it is inherently related to detailed catchment characteristics. Technically speaking, such uncertainty in weights can only be constrained by testing and comparing the performance of different strategies for weight assignment. In this context, calibration based on random sampling has benefits, because differing from the optimization-based algorithms where parameters are evolved iteratively (e.g., Dynamically Dimensioned Search; Tolson & Shoemaker, 2007), here the parameter set of each run was randomly sampled and thus independent, which allows the modeler to split the model runs and post analysis (e.g., checking the calibration performance under different weights). Therefore, though one could argue about the inefficiency of random sampling-based calibration in searching high-dimensional parameter space, such calibration would still likely be more suitable or efficient (see the stabilization of parameter values in Figure S3 in Supporting Information S1) than optimization-based calibration which needs to be repeatedly iterated (e.g., 286 times for all sets of weights in this case), when a thorough check on weight assignment is required. However, it is still important to recognize the potential drawbacks of residue-based calibrations where the uncertainties from inputs, observations, and model structure were not considered. A potential solution would be using the limits of acceptability approach which considers the observations uncertainty with a random sampling scheme (Beven, 2006), or other iterative approaches (e.g., the Bayesian total error analysis to estimate input uncertainty (Marshall Price et al., 2007) or multi-model averaging techniques to assess the structural uncertainty (Raftery et al., 2005)). For those calibrations bypassing such analysis, we suggest incorporating isotopes if possible with relatively small weights (<0.2), and assigning the remaining weights to regular hydrological observations depending on specific research questions.

4.2. Challenges From Equifinality and Soft Data Check as a Solution

In an optimal expectation for multi-criteria calibration, the auxiliary data sets would correct and refine the model step by step toward a unique simulation representing the dominant wetland hydrological processes. However, the reality is inevitably more ambiguous.

Encouragingly, if we focus on most monitored variables (i.e., outlet discharge, stream isotopes, and GW isotopes), including each time series in the calibration significantly improved its simulation while only marginally degrading performance measures of the others (Figure 3). Such increased overall performance has also been observed previously (Holmes et al., 2020; Kuppel et al., 2018a; Piovano et al., 2018; Yang et al., 2023), which seems to indicate that multi-criteria calibration has driven the model toward more consistent simulations with adequate process representation in the studied wetland. Accordingly, the "best" model might be expected to be the one calibrated with all the observations as it performed overall well with acceptable trade-off (e.g., *CS*11 in Figure 5).

However, such an assessment changes when examining the uncalibrated internal fluxes, as strong discrepancies in their magnitudes and spatial (Figure 5) and temporal patterns (Figure 6) were found between weight settings. The most likely reason is that these different types of data had overlapping footprints that inform the calibration process with conflicting or inconsistent information (Clark et al., 2011; Kuppel et al., 2018a), which was evidenced by the discrepancies in the vertical, lateral and ET fluxes under *CS*1–4 summarized in Figures 8a–8d. Therefore, when adding auxiliary data sets into the calibration, they no longer improved the model on the basis of the parent/previous calibration, but led to a restructuring of the simulated hydrological function due to the conflicting information embedded in different constraining data sets. In other words, the convergence of monitored variables/fluxes was realized at the expense of the compensation between the other (unmonitored) variables/fluxes that are highly diverse and uncertain. This is further evidenced by the most calibrated parameters (Figure S8 in Supporting Information S1), whose range still remained quite wide after adding more constraining data sets, indicating that the models still suffered from strong equifinality rather than converging toward a unique solution. Some exceptions of parameters achieving constrained posterior distributions could be explained by lower conflicts

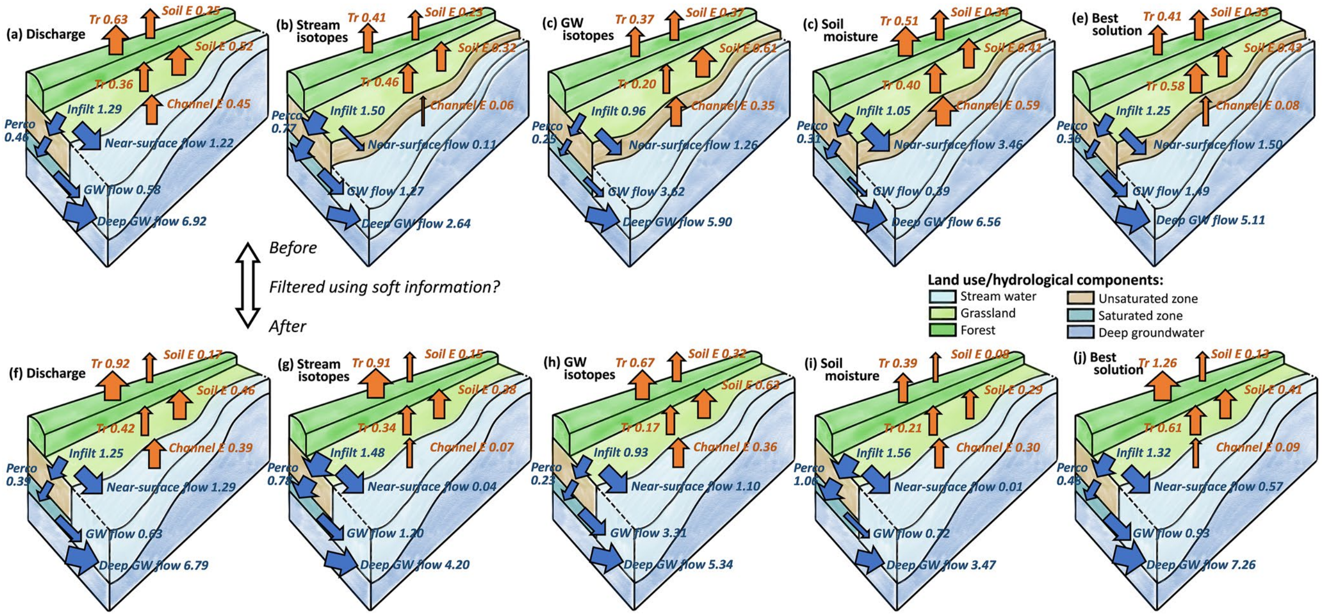

Figure 8. The key internal fluxes calibrated separately using (a) discharge, (b) in-stream isotopes, (c) groundwater (GW) isotopes, and (d) top soil moisture, or jointly using (e) discharge, in-stream isotopes, GW isotopes, and soil moisture (with weights of 0.6, 0.1, 0.1, and 0.2). Subplots (f)–(j) shows the internal fluxes of plausible models filtered by soft data check.

between observation-specific information. For example, the parameter *channelE_weight* had a good relatively good convergence because the channel evaporation exerted a dominant impact on in-stream isotopic composition (streamwater ^{2}H), while had little (or no) influence on the water balance of stream (discharge), GW (GW ^{2}H), or soil water (soil moisture). Such a finding indeed contrasts to some previous findings that the performance range for certain parameters were significantly reduced in multi-criteria calibration compared to those calibrated with single criterion (Kuppel et al., 2018a; Piovano et al., 2018). The potential reason is that while most of those studies are based on relatively parsimonious models (with <20 parameters), the number of parameters to calibrate reaches 59 after parameterization in this EcH$_2$O-iso application, leading to high dimensionality in the parameter space. Strictly speaking, the difficulty in fully covering such high-dimensional parameter space hindered the convergence of all parameters, and a sensitivity analysis might help to further reduce the uncertain parameters as well as the dimension of the search space. However, most parameters were retained for calibration given it is the first-time of EcH$_2$O to be applied to a heterogenous wetland. According to previous EcH$_2$O modeling experience (Gillefalk et al., 2021; Smith et al., 2021), a total of 200,000 samples could already yield a satisfactory calibration on the same parameters against multiple targets. Here, to further constrain the calibration, the number of samples was further increased to 500,000, which took over a week of computation time with parallel use of 200 cores from the cluster in Humboldt University. The relatively high sample number and stabilization of key parameter values (Figure S3 in Supporting Information S1) support the credibility of this random-sampling-based calibration. Such reduced identifiability of calibrated parameters when introducing more criteria into calibration process were also observed in Birkel and Soulsby (2015) and Holmes et al. (2020).

Therefore, significant equifinality seems to be unavoidable even with abundant observations (four different types of data in up to 30 locations in this study), which leads to marked uncertainty in internal fluxes, the so called getting "right answer for the wrong reasons" (Kirchner, 2006). For instance, in the model producing the best overall performance for all observations (*CS*11 in Figures 5 and 8e), transpiration exhibited higher values in grassland than forest, which was contrary to our field perception based on recent monitoring and modeling in DMC (Kleine et al., 2020; Smith et al., 2021) or nearby Berlin (Gillefalk et al., 2021). Such implausible quantification of unmonitored fluxes is notable as it shows potential dangers in previous applications which favored more balanced models calibrated with all available observations (e.g., Birkel et al., 2014; Piovano et al., 2018). More importantly, the lack of physical realism would significantly reduce the credibility of such optimized models, because the uncertainty could be marked when models were further used for prediction, especially for data scarce areas where models are often calibrated against a relatively short period (e.g., 2 years in this study).

Therefore, checking the plausibility of internal fluxes can be invaluable, using either detailed field knowledge or other "soft" data (Arnold et al., 2015). Here we posted a simple approach to filter the models based on transpiration and lateral flow, and results showed that the average pass rate of optimized models was only 46% (Figure 3e). In the other words, more than half of the optimized models gave the "right answer for the wrong reasons," which underlined the importance of such a soft data check. By excluding the ones with implausible simulation of internal fluxes based on field knowledge, now the discrepancies between different calibration weights were significantly reduced as they all gave feasible results (Figure 7). The remaining models are thus more reliable, especially the ones that still captured the all types of observations (i.e., *CS*11 in Figures 7 and 8j). They represent the wetland as a slow-draining system mainly fed by lateral inflow from deeper layers but was also influenced more by near-surface inflow when soils were saturated in winter or after strong convective events in summer. Infiltration and percolation were higher in forests with sandy soils while near-surface flow was more frequently generated in wetlands due to the high soil moisture content and lower infiltration capacity of peaty soils. The vegetation communities also contributed to the heterogeneity of hydrological fluxes, given the different transpiration amounts between communities (grassland < early-season herbaceous < late-season herbaceous < forest), and contrary spatial patterns of soil evaporation (forest > non-forest communities). Note that here we only did a binary test on the internal fluxes rather than some other studies that quantitatively incorporated soft data into optimization metrics (e.g., Seibert & McDonnell, 2002). In this context, our positive result is encouraging for distributed modeling community that is constantly plagued by equifinality, because it shows that even with simple field knowledge that can be only qualitatively described, soft data can still be informative for model calibration and significantly constrain the equifinality.

In addition, the soft data check could potentially identify the weakness in model calibration by examining how the proportion of plausible models reacts to different weight settings. For example, we found that calibration with more weights to soil moisture favored the models producing physically unrealistic results (Figure 3f). This potentially points to the incompatibility between soil moisture observations and the current calibration schemes, which is likely due to the scale differences between the model setup (50 m grids) and point-scale measurements (Figure S2 in Supporting Information S1), as the model cannot reproduce observations with such high spatial resolutions (Beven, 2001; McDonnell et al., 2007). Moreover, one needs to consider the uncertainties during the measurement through the high heterogeneity in top soil moisture, which could be significantly different even within a few meters of distance. Accordingly, soil moisture data probably generated additional uncertainty during the calibration and thus needs to be further aggregated spatially rather than just normalized as in the current calibration scheme. Therefore, our example further demonstrates the soft data check as a useful tool to refine the calibration process, which is likely to be beneficial in future applications. Such refinement was also reported previously using hydrologically relevant "signature measures" instead of direct use of time series data (Yilmaz et al., 2008), which would be of further interest for future comparative studies.

4.3. Other Potential Steps to Constrain Equifinality

Although assisted by water isotopes and a further soft data test, a more plausible set of constrained models capturing dominant wetland processes was obtained in this study, a definitive model that fulfills both spatio-temporal pattern of all monitored variables/fluxes and physical realism of unmonitored internal fluxes remains elusive. Facing such challenges which isotopes and soft data are no longer able to overcome, there are still pathways to improve model results.

4.3.1. Model Development

When reviewing our multi-criteria calibration, the main challenges originated from the different information embedded in different observations, leading to the discrepancies between calibrations. However, such information does not inherently come from the data itself, but instead more depends on how the corresponding component is conceptualized and parameterized in the model. As stated in Holmes et al. (2020), "*the extent to which introducing auxiliary data into calibration degrades model performance likely depends on a model's ability to alter internal flow path contributions and still preserve total streamflow contribution*." One would aim toward a model that enables a global optimum producing the same levels of performance for all observations compared to their local optimum, while in contrast, failing to find such a model points to the incompatibility between the model and observations. Therefore, using these modeling insights to inform future model development (in terms of process conceptualization and parameterization) for better data-model compatibility would be an important next step.

Here, we take the subsurface processes in this study as an example, whose importance has been demonstrated here and previously in the wetland (Smith et al., 2021; Wu, Tetzlaff, Goldhammer, et al., 2022). In EcH$_2$O-iso, the lateral flow of GW is conceptualized as a linear kinematic model driven by slope (Maneta & Silverman, 2013), which is responsible for the strong discrepancies between the riparian area and remote areas due to the slope difference. However, it remains a matter of debate over such conceptualization because for our application in the wetland, the dominating role of slope on GW transport in deeper layers seems to be not physically realistic enough, which is mostly attributed to the consistent soil depths of the three layers. Therefore, further development of EcH$_2$O-iso toward dynamical soil depths across the model domain and better conceptualization of lower boundary conditions could lead to improved subsurface simulations, though extra computational resources and information/parameters (e.g., depth and slope of impeding layers) would be required.

However, arguing that many of the newly introduced parameters and processes cannot be appropriately measured, some would fear that evolution of more complex models would simply layer up uncertainty (Beven & Freer, 2001; McDonnell et al., 2007). In this case, changing the model parameterization is a safer option. Taking top soil moisture as an example, we sought to answer why the model failed to capture the top soil moisture in high spatial resolution in our application? This can be very likely attributed to the inappropriate parameterization from a mechanistic perspective (Kumar et al., 2013), apart from scale differences and measurement uncertainty stated above. Like many distributed hydrological models, our parameterization is realized based on soil and vegetation types (Wellen et al., 2015); however, it is possible that such coarse parameterization cannot reproduce the measured sub-grid scale heterogeneity in soil moisture, leading to simulation error and hindering the model in extracting information from the observations. To mitigate this, a two-step solution could be potentially adopted: extrapolating the possible optimal parameterization scheme via diagnostic analysis (e.g., spatial sensitivity analysis described in Wu et al., 2022b), and collecting corresponding forcing data with finer resolution to reference the new parameterization.

4.3.2. Improved Data Acquisition

Apart from the model development, multi-criteria calibration can also be used as a learning framework to guide data collection and monitoring, whose pragmatic scheme should direct experimental efforts toward collection of data that is most informative for model development and evaluation (McGuire et al., 2007; Soulsby et al., 2008).

The discrepancies in simulations between calibration settings suggest the lack of information on the magnitude of the internal fluxes, whose high degree of freedom easily led to the compensation between each other. Therefore, expanding the monitoring for more hydrological variables/fluxes would be helpful to gain extra diagnostic information, which has been increasingly carried out in recent studies (e.g., transpiration estimated via sapflow in Gillefalk et al., 2021). Such expansion will be especially informative for physics-based models, as they inherently seek to explicitly represent the state variables and fluxes that are theoretically observable in reality (Fatichi et al., 2016; Kuppel et al., 2018a).

Apart from the inter-variable uncertainty, equifinality was marked within each variable given the strong dynamics in their spatial patterns. In the other words, it is difficult for models to appropriately differentiate the unmonitored fluxes across the model domain due to the lack of information. This highlights the importance of increasing the spatial resolution for monitoring, as observation at a single point could not help tackle such uncertainty in spatial distribution, though it is common in distributed modeling (Wellen et al., 2015).

5. Conclusion

In this study, a multi-criteria calibration was conducted with the physically-based, fully distributed model EcH$_2$O-iso to unravel the spatio-temporal patterns of hydrological functions in a riparian wetland. To investigate the role of weights in multi-criteria calibration, a total of 286 sets of weights were assigned to discharge, stream isotopes, GW isotopes, and soil moisture.

The results show the benefits of using multi-criteria calibration, as it strongly increased the overall performance of all observations while only marginally degraded the performance of each. Notably, isotopes were highlighted as appropriate auxiliary data as they effectively constrained the model with relatively small weights (0.1). However, strong discrepancies in magnitudes and spatial patterns of the uncalibrated internal states/fluxes were still found between different calibration weights, and many simulated internal fluxes were not physically plausible. This indicates the ongoing equifinality issues and points out the fact that more observations do not necessarily lead to more convincing results. This is potentially attributed to the conflicting information embedded in observations (discharge: amount of lateral flow; stream isotopes: proportions from different lateral pathways; GW isotopes: infiltration and percolation), thus pulling the calibration into different directions.

Therefore, an approach was developed as a posterior check on the internal fluxes based on soft data (transpiration and sources of lateral flow) in order to identify the models producing physically-implausible simulation of internal fluxes. Results show that more than half (54%) of optimized models gave "right answers for the wrong reasons." By excluding those models, the approach effectively constrained equifinality, while meanwhile unraveling the potential incompatibility between observations and calibration process (e.g., here physical realism decreased when soil moisture was weighted more). The remaining models reflected the wetland as a slow-draining system mainly fed by lateral inflow from deeper layers but can be influenced more by near-surface inflow when soils are saturated in winter or during strong convective events in summer. Vegetation also plays an important role: forest shows a higher transpiration rate than riparian grass/herbaceous species, while higher soil evaporation is observed in non-forest area due to limited canopy cover. All ET fluxes peaked in ~July, with transpiration mainly in the growing season while soil evaporation lasting for longer. The activation timing of infiltration, percolation, and near-surface flow was closely related to the soil saturation (mostly in winter) or convective rainfall events (in summer), while the GW flow were more stable temporally. Overall, this study not only provided direct insights into wetland functioning, but also revealed the risk of equifinality even with abundant data for calibration; however, such equifinality could be effectively constrained by directly including isotopes into calibration and a posterior check on soft data.

Data Availability Statement

The source codes of EcH_2O-iso are available in Zenodo repository (Wu et al., 2023). The data used as model forcing and calibration (catchment geography, climate, discharge, and other observations) were also available in the same repository.

Acknowledgments

Songjun Wu is funded by the Chinese Scholarship Council (CSC). This research has been supported by the BMBF (funding code 033W034A), which supported the IGB stable isotope Laboratory. Contributions from Soulsby are supported by the Leverhulme Trust through the ISO-LAND project (Grant RPG 2018 375). Tetzlaff's contribution was partly funded through the Einstein Research Unit "Climate and Water under Change" from the Einstein Foundation Berlin and Berlin University Alliance. We deeply thank Jonas Freymüller and David Dubbert for the help on field work and laboratory analysis of the water isotopes. Hauke Dämpfling is acknowledged for conducting monthly UAV flights. We also thank Ke Chen for discussion on the paper. The valuable comments from associate editor and three anonymous reviewers are highly appreciated.

References

Acreman, M. C., Fisher, J., Stratford, C. J., Mould, D. J., & Mountford, J. O. (2007). Hydrological science and wetland restoration: Some case studies from Europe. *Hydrology and Earth System Sciences*, *11*(1), 158–169. https://doi.org/10.5194/HESS-11-158-2007

Arnold, J. G., Youssef, M. A., Yen, H., White, M. J., Sheshukov, A. Y., Sadeghi, A. M., et al. (2015). Hydrological processes and model representation: Impact of soft data on calibration. *Transactions of the American Society of Agricultural and Biological Engineers*, *58*, 1637–1660. https://doi.org/10.13031/TRANS.58.10726

Aumen, N. G., & Keddy, P. A. (2001). Wetland ecology: Principles and conservation. *Journal of the North American Benthological Society*, *20*(4), 683–685. https://doi.org/10.2307/1468096

Bam, E. K. P., & Ireson, A. M. (2019). Quantifying the wetland water balance: A new isotope-based approach that includes precipitation and infiltration. *Journal of Hydrology (Amsterdam)*, *570*, 185–200. https://doi.org/10.1016/J.JHYDROL.2018.12.032

Berezowski, T., Nossent, J., Chormaåski, J., & Batelaan, O. (2015). Spatial sensitivity analysis of snow cover data in a distributed rainfall-runoff model. *Hydrology and Earth System Sciences*, *19*(4), 1887–1904. https://doi.org/10.5194/HESS-19-1887-2015–1904

Beven, K. (2001). How far can we go in distributed hydrological modelling? *Hydrology and Earth System Sciences*, *5*, 1–12. https://doi.org/10.5194/HESS-5-1-2001

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology (Amsterdam)*, *320*(1–2), 18–36. https://doi.org/10.1016/J.JHYDROL.2005.07.007

Beven, K., & Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology (Amsterdam)*, *249*(1–4), 11–29. https://doi.org/10.1016/S0022-1694(01)00421-8

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

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

Cao, W., Bowden, W. B., Davie, T., & Fenemor, A. (2006). Multi-variable and multi-site calibration and validation of SWAT in a large mountainous catchment with high spatial variability. *Hydrological Processes*, *20*(5), 1057–1073. https://doi.org/10.1002/HYP.5933

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*(9), W09301. https://doi.org/10.1029/2010WR009827

Clark, M. P., & Vrugt, J. A. (2006). Unraveling uncertainties in hydrologic model calibration: Addressing the problem of compensatory parameters. *Geophysical Research Letters*, *33*(6), L06406. https://doi.org/10.1029/2005GL025604

Craig, H., & Gordon, L. (1964). Deuterium and oxygen 18 variations in the ocean and marine atmosphere, stable isotopes in oceanographic studies and paleotemperatures.

Douinot, A., Tetzlaff, D., Maneta, M., Kuppel, S., Schulte-Bisping, H., & Soulsby, C. (2019). Ecohydrological modelling with EcH_2O-iso to quantify forest and grassland effects on water partitioning and flux ages. *Hydrological Processes*, *33*(16), 2174–2191. https://doi.org/10.1002/HYP.13480

Efstratiadis, A., & Koutsoyiannis, D. (2010). One decade of multi-objective calibration approaches in hydrological modelling: A review. *Hydrological Sciences Journal*, *55*(1), 58–78. https://doi.org/10.1080/02626660903526292

Fatichi, S., Vivoni, E. R., Ogden, F. L., Ivanov, V. Y., Mirus, B., Gochis, D., et al. (2016). An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. *Journal of Hydrology (Amsterdam)*, *537*, 45–60. https://doi.org/10.1016/J.JHYDROL.2016.03.026

Fenicia, F., McDonnell, J. J., & Savenije, H. H. G. (2008). Learning from model improvement: On the contribution of complementary data to process understanding. *Water Resources Research*, *44*(6), W06419. https://doi.org/10.1029/2007WR006386

Gillefalk, M., Tetzlaff, D., Hinkelmann, R., Kuhlemann, L. M., Smith, A., Meier, F., et al. (2021). Quantifying the effects of urban green space on water partitioning and ages using an isotope-based ecohydrological model. *Hydrology and Earth System Sciences*, *25*(6), 3635–3652. https://doi.org/10.5194/HESS-25-3635-2021

Guse, B., Pfannerstill, M., Gafurov, A., Fohrer, N., & Gupta, H. (2016). Demasking the integrated information of discharge: Advancing sensitivity analysis to consider different hydrological components and their rates of change. *Water Resources Research*, *52*(11), 8724–8743. https://doi.org/10.1002/2016WR018894

Hayashi, M., van der Kamp, G., & Rosenberry, D. O. (2016). Hydrology of Prairie Wetlands: Understanding the integrated surface-water and groundwater processes. *Wetlands*, *36*(S2), 237–254. https://doi.org/10.1007/S13157-016-0797-9

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 Resources Research*, *56*(9), e2020WR027447. https://doi.org/10.1029/2020WR027447

Jing, M., Heße, F., Kumar, R., Wang, W., Fischer, T., Walther, M., et al. (2018). Improved regional-scale groundwater representation by the coupling of the mesoscale Hydrologic Model (mHM v5.7) to the groundwater model OpenGeoSys (OGS). *Geoscientific Model Development*, *11*(5), 1989–2007. https://doi.org/10.5194/GMD-11-1989-2018

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

Kleine, L., Tetzlaff, D., Smith, A., Wang, H., & Soulsby, C. (2020). Using water stable isotopes to understand evaporation, moisture stress, and re-wetting in catchment forest and grassland soils of the summer drought of 2018. *Hydrology and Earth System Sciences*, *24*(7), 3737–3752. https://doi.org/10.5194/HESS-24-3737-2020

Kumar, R., Samaniego, L., & Attinger, S. (2013). Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resources Research*, *49*(1), 360–379. https://doi.org/10.1029/2012WR012195

Kuppel, S., Tetzlaff, D., Maneta, M. P., & Soulsby, C. (2018a). What can we learn from multi-data calibration of a process-based ecohydrological model? *Environmental Modelling & Software*, *101*, 301–316. https://doi.org/10.1016/J.ENVSOFT.2018.01.001

Kuppel, S., Tetzlaff, D., Maneta, M. P., & Soulsby, C. (2018b). EcH$_2$O-iso 1.0: Water isotopes and age tracking in a process-based, distributed ecohydrological model. *Geoscientific Model Development*, *11*(7), 3045–3069. https://doi.org/10.5194/GMD-11-3045-2018

Maneta, M. P., & Silverman, N. L. (2013). A spatially distributed model to simulate water, energy, and vegetation dynamics using information from regional climate models. *Earth Interactions*, *17*(11), 1–44. https://doi.org/10.1175/2012EI000472.1

Marshall Price, L., Nott, D., & Sharma, A. (2007). Towards dynamic catchment modelling: A Bayesian hierarchical mixtures of experts framework. *Hydrological Processes*, *21*(7), 847–861. https://doi.org/10.1002/hyp.6294

Maxwell, R. M. (2013). A terrain-following grid transform and preconditioner for parallel, large-scale, integrated hydrologic modeling. *Advances in Water Resources*, *53*, 109–117. https://doi.org/10.1016/J.ADVWATRES.2012.10.001

McDonnell, J. J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., et al. (2007). Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. *Water Resources Research*, *43*(7), W07301. https://doi.org/10.1029/2006WR005467

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

Musolff, A., Schmidt, C., Selle, B., & Fleckenstein, J. H. (2015). Catchment controls on solute export. *Advances in Water Resources*, *86*, 133–146. https://doi.org/10.1016/J.ADVWATRES.2015.09.026

Neill, A. J., Birkel, C., Maneta, M. P., Tetzlaff, D., & Soulsby, C. (2021). Structural changes to forests during regeneration affect water flux partitioning, water ages and hydrological connectivity: Insights from tracer-aided ecohydrological modelling. *Hydrology and Earth System Sciences*, *25*(9), 4861–4886. https://doi.org/10.5194/HESS-25-4861-2021

Piovano, T. I., Tetzlaff, D., Ala-aho, P., Buttle, J., Mitchell, C. P. J., & Soulsby, C. (2018). Testing a spatially distributed tracer-aided runoff model in a snow-influenced catchment: Effects of multicriteria calibration on streamwater ages. *Hydrological Processes*, *32*(20), 3089–3107. https://doi.org/10.1002/HYP.13238

Raftery, A., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*(5), 1155–1174. https://doi.org/10.1175/MWR2906.1

Scudeler, C., Pangle, L., Pasetto, D., Niu, G.-Y., Volkmann, T., Paniconi, C., et al. (2016). Multiresponse modeling of variably saturated flow and isotope tracer transport for a hillslope experiment at the Landscape Evolution Observatory. *Hydrology and Earth System Sciences*, *20*(10), 4061–4078. https://doi.org/10.5194/hess-20-4061-2016

Seibert, J., & McDonnell, J. J. (2002). On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. *Water Resources Research*, *38*(11), 23-1–23-14. https://doi.org/10.1029/2001WR000978

Sherlock, M. D., Chappell, N. A., & Mcdonnell, J. J. (2000). Effects of experimental uncertainty on the calculation of hillslope flow paths. *Hydrological Processes*, *14*(14), 2457–2471. https://doi.org/10.1002/1099-1085(20001015)14:14<2457::AID-HYP106>3.0.CO;2-I

Smith, A., Tetzlaff, D., Kleine, L., Maneta, M., & Soulsby, C. (2021). Quantifying the effects of land use and model scale on water partitioning and water ages using tracer-aided ecohydrological models. *Hydrology and Earth System Sciences*, *25*(4), 2239–2259. https://doi.org/10.5194/HESS-25-2239-2021

Soulsby, C., Birkel, C., Geris, J., Dick, J., Tunaley, C., & Tetzlaff, D. (2015). Stream water age distributions controlled by storage dynamics and nonlinear hydrologic connectivity: Modeling with high-resolution isotope data. *Water Resources Research*, *51*(9), 7759–7776. https://doi.org/10.1002/2015WR017888

Soulsby, C., Neal, C., Laudon, H., Burns, D. A., Merot, P., Bonell, M., et al. (2008). Catchment data for process conceptualization: Simply not enough? *Hydrological Processes*, *22*(12), 2057–2061. https://doi.org/10.1002/HYP.7068

Tetzlaff, D., Buttle, J., Carey, S. K., Mcguire, K., Laudon, H., & Soulsby, C. (2015). Tracer-based assessment of flow paths, storage and runoff generation in northern catchments: A review. *Hydrological Processes*, *29*(16), 3475–3490. https://doi.org/10.1002/HYP.10412

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*(1), W01413. https://doi.org/10.1029/2005WR004723

Weiler, M., & Naef, F. (2003). An experimental tracer study of the role of macropores in infiltration in grassland soils. *Hydrological Processes*, *17*(2), 477–493. https://doi.org/10.1002/HYP.1136

Wellen, C., Kamran-Disfani, A. R., & Arhonditsis, G. B. (2015). Evaluation of the current state of distributed watershed nutrient water quality modeling. *Environmental Science & Technology*, *49*(6), 3278–3290. https://doi.org/10.1021/ES5049557

Winsemius, H. C., Savenije, H. H. G., & Bastiaanssen, W. G. M. (2008). Constraining model parameters on remotely sensed evaporation: Justification for distribution in ungauged basins? *Hydrology and Earth System Sciences*, *12*(6), 1403–1413. https://doi.org/10.5194/HESS-12-1403-2008

Wu, S., Tetzlaff, D., Goldhammer, T., Freymueller, J., & Soulsby, C. (2022). Tracer-aided identification of hydrological and biogeochemical controls on in-stream water quality in a riparian wetland. *Water Research*, *222*, 118860. https://doi.org/10.1016/J.WATRES.2022.118860

Wu, S., Tetzlaff, D., Goldhammer, T., & Soulsby, C. (2021). Hydroclimatic variability and riparian wetland restoration control the hydrology and nutrient fluxes in a lowland agricultural catchment. *Journal of Hydrology*, *603*, 126904. https://doi.org/10.1016/J.JHYDROL.2021.126904

Wu, S., Tetzlaff, D., Yang, X., Smith, A., & Soulsby, C. (2023). The model EcH$_2$O-iso and catchment data used as model forcing and calibration. [Model codes and data]. https://zenodo.org/record/7025071

Wu, S., Tetzlaff, D., Yang, X., & Soulsby, C. (2022a). Disentangling the influence of landscape characteristics, hydroclimatic variability and land management on surface water NO$_3$-N dynamics: Spatially distributed modelling over 30 years in a lowland mixed land use catchment. *Water Resources Research*, *58*(2), e2021WR030566. https://doi.org/10.1029/2021WR030566

Wu, S., Tetzlaff, D., Yang, X., & Soulsby, C. (2022b). Identifying dominant processes in time and space: Time-varying spatial sensitivity analysis for a grid-based nitrate model. *Water Resources Research*, *58*(8), e2021WR031149. https://doi.org/10.1029/2021WR031149

Yang, X., Tetzlaff, D., Müller, C., Knöller, K., Borchardt, D., & Soulsby, C. (2023). Upscaling tracer-aided ecohydrological modeling to larger catchments: Implications for process representation and heterogeneity in landscape organization. *Water Resources Research*, *59*(3), e2022WR033033. https://doi.org/10.1029/2022WR033033

Yang, X., Tetzlaff, D., Soulsby, C., Smith, A., & Borchardt, D. (2021). Catchment functioning under prolonged drought stress: Tracer-aided ecohydrological modeling in an intensively managed agricultural catchment. *Water Resources Research*, *57*(3), e2022WR029094. https://doi.org/10.1029/2020WR029094

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 Resources Research*, *44*(9), W09417. https://doi.org/10.1029/2007WR006716

Zedler, J. B., & Kercher, S. (2005). Wetland resources: Status, trends, ecosystem services, and restorability. *Annual Review of Environment and Resources*, *30*(1), 39–74. https://doi.org/10.1146/ANNUREV.ENERGY.30.050504.144248