Water Resources Research

RESEARCH ARTICLE
10.1029/2023WR034420

Identifying Structural Priors in a Hybrid Differentiable Model for Stream Water Temperature Modeling

Farshid Rahmani[1], Alison Appling[2], Dapeng Feng[1], Kathryn Lawson[1], and Chaopeng Shen[1]

[1]Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, USA, [2]U.S. Geological Survey, Reston, VA, USA

Abstract Although deep learning models for stream temperature (T_s) have recently shown exceptional accuracy, they have limited interpretability and cannot output untrained variables. With hybrid differentiable models, neural networks (NNs) can be connected to physically based equations (called structural priors) to output intermediate variables such as water source fractions (specifying what portion of water is groundwater, subsurface, and surface flow). However, it is unclear if such outputs are physically meaningful when only limited physics is imposed, and if structural priors have enough impacts to be identifiable from data. Here, we tested four alternative structural priors describing basin-scale water temperature memory and instream heat processes in a differentiable stream temperature model where NNs freely estimate the water source fractions. We evaluated models' abilities to predict T_s and baseflow ratio. The four priors exhibited noticeably different behaviors in these two metrics and their tradeoffs, with some dominating others. Therefore, the better structural priors can be identified. Moreover, testing different priors yielded valuable insights: having a separate shallow subsurface flow component better matches observations, and a recency-weighted averaging of past air temperature for calculating source water temperature resulted in better T_s and baseflow prediction than traditionally employed simple averaging. However, we also highlight the limitations when insufficient physical constraints are implemented: the internal variables (water source fractions) may not be adequately constrained by a single target variable (stream temperature) alone. To ensure the physical significance of the internal fluxes, one can either employ multivariate data for model selection, or include more physical processes in the priors.

Plain Language Summary A new framework called differentiable modeling combines the benefits from neural networks (NNs) and process-based models. This framework can learn from big data while the process-based model components (called prior knowledge, or priors) are intended to output intermediate physical variables. However, do such priors matter, can we tell if one set of priors is better than another, and do the intermediate outputs represent the intended physical concepts? We explore these questions with a differentiable stream temperature model where the NN replaces the hydrologic component and estimates parameters pertaining to the stream temperature module. The strong optimizing capability of NNs allows us to avoid some complexities and attribute the differences in model outcomes to the assumed priors. Testing different priors thus yielded many important lessons, for example, the need for having a separate shallow groundwater "bucket," the benefit of placing more importance on recent air temperature when estimating groundwater temperature, and the importance of describing in-stream temperature. The results show lots of untapped potential with differentiable modeling and the data we have available.

Key Points:

- We can identify better structural priors (process equations) using differentiable models, circumventing intertwined parameter issues
- Considering a separate bucket for shallow subsurface water improves both stream temperature and baseflow simulation
- The models selected by the multivariate evaluation produce physically meaningful estimates of water source fractions

Supporting Information:

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

Correspondence to:

C. Shen,
cshen@engr.psu.edu

Citation:

Rahmani, F., Appling, A., Feng, D., Lawson, K., & Shen, C. (2023). Identifying structural priors in a hybrid differentiable model for stream water temperature modeling. *Water Resources Research*, *59*, e2023WR034420. https://doi.org/10.1029/2023WR034420

Received 6 JAN 2023
Accepted 17 NOV 2023

Author Contributions:

Conceptualization: Chaopeng Shen
Data curation: Alison Appling
Formal analysis: Farshid Rahmani
Funding acquisition: Alison Appling, Chaopeng Shen
Investigation: Farshid Rahmani, Dapeng Feng
Methodology: Farshid Rahmani, Chaopeng Shen
Project Administration: Chaopeng Shen
Resources: Chaopeng Shen
Software: Farshid Rahmani
Supervision: Chaopeng Shen
Validation: Dapeng Feng
Visualization: Farshid Rahmani

1. Introduction

Stream temperature (T_s, temperature of water in a river) is an important variable: it not only exerts a strong influence on ecosystem health (Chapin et al., 2014; Marcogliese, 2001; Martins et al., 2012), water quality (Ducharne, 2008; Morrill et al., 2005; Zhi et al., 2021), and human water uses (Förster & Lilliestam, 2010; Madden et al., 2013; van Vliet et al., 2013), but also carries information about the source of water contributions in a hydrologic system (Du et al., 2020; Michel et al., 2020). In the face of climate change and widespread alteration of streamflow (Virkki et al., 2022), understanding how stream temperature will change can be important for management of fish habitat and planning for power generation and other human uses (Du et al., 2022; Fellman et al., 2014; Vliet et al., 2016). Moreover, in headwater basins, T_s is a manifestation of hydrologic processes: groundwater contribution to streams has a low annual thermal variability and is often the critical source of cool water during base flow periods (Hare et al., 2021); surface runoff temperature is closer to air temperature (or snowmelt) and is more temporally dynamic; shallow subsurface flow is between these two extremes. These water

Writing – original draft: Farshid Rahmani
Writing – review & editing: Alison Appling, Kathryn Lawson, Chaopeng Shen

sources mix in the channel and further undergo instream thermal processes such as radiative heating and cooling, shading, evaporation, and streambed friction and conduction. A significant departure of stream temperature from air temperature typically suggests influences from groundwater (Bundschuh, 1993; Hare et al., 2021) or reservoir operations (Amirkhani et al., 2016; Webb & Walling, 1995). Stream temperature data thus contain the signatures of different streamflow sources and may be useful in inferring difficult-to-measure variables such as thermal memory (the time frame over which the air temperature influences the groundwater) and the volume of groundwater influx to a stream reach.

Water temperature has often been predicted using process-based models (Dugdale et al., 2017; Markstrom, 2012; Meisner et al., 1988; Wanders et al., 2019) or statistical models (Benyahya et al., 2007; Detenbeck et al., 2016; Graf et al., 2019; Mohseni & Stefan, 1999; Siegel et al., 2022) with simplifying assumptions that are difficult to prove true or false. For example, the process-based stream temperature model component of the Precipitation-Runoff Modeling System (PRMS) (hereafter "SNTEMP") (Markstrom, 2012; Regan & Markstrom, 2021), can assume that the temperatures of the surface runoff, shallow subsurface water, and groundwater are the averages of the air temperature in the most recent 1, 30, and 365 days, respectively (Sanders et al., 2017). The temperature of lateral flow entering the stream is then assumed to be a weighted average of these different sources based on their volumetric contributions to streamflow, and then instream thermal processes are considered to estimate the downstream water temperature. The assumption in Meisner et al. (1988), is similar to the assumption in SNTEMP but assumes no shallow subsurface component and sets the averaging length of air temperature impact on groundwater temperature to multiple years (e.g., one or 2 years). The model from Wanders et al. (2019) introduced a different assumption for the water temperature lateral flow, adding bias terms and threshold functions to the surface flow and groundwater temperature calculations to compensate for the cooling of rain as it falls.

The uncertainty with these structural assumptions is substantial, and it is challenging to assess their quality because the structural uncertainty is intertwined with parametric uncertainty. If models perform poorly, the issue could be due to inadequate structures, suboptimal parameters, a combination of both (Draper, 1995; Loucks & Beek, 2017), or data uncertainty. Characterizing structural uncertainty is traditionally harder because for every set of model structures, we need to either find the optimal parameter set (highly ambiguous) or identify a range of functional parameters and assess their uncertainty (highly computationally expensive). Moreover, stream temperature models are seldom benchmarked on the same data set (please refer to Table S1 in Supporting Information S1 in Rahmani, Lawson, et al., 2021; Rahmani, Shen, et al., 2021), making the conclusions sometimes data set-dependent and difficult to generalize.

Since 2017, it has been demonstrated that deep learning (DL) networks (Shen, 2018; Shen et al., 2018), like long short-term memory (LSTM; Hochreiter & Schmidhuber, 1997), can learn to predict environmental variables including stream temperature with exceptionally high accuracy (Rahmani, Lawson, et al., 2021; Rahmani, Shen, et al., 2021; Rehana & Rajesh, 2023; Sadler et al., 2022; Weierbach et al., 2022; S. Zhu & Piotrowski, 2020; Zwart et al., 2023). LSTM has shown its strength and versatility in simulating hydrologic variables such as soil moisture (Fang et al., 2017, 2019; J. Liu et al., 2022; O & Orth, 2021), streamflow (Feng et al., 2020, 2021; Khoshkalam et al., 2023; Xiang et al., 2020), dissolved oxygen (Heddam et al., 2022; Zhi et al., 2023), snow water equivalent (Broxton et al., 2019; Meyal et al., 2020), stream nitrate concentration (Saha et al., 2023; Samarinas et al., 2020), and radiation (Y. Liu et al., 2020; F. Zhu et al., 2021). However, mostly used as a forward simulator in hydrology, LSTM is also limited by its black-box nature: it does not provide an interpretable explanation of internal processes, and its intermediate variables lack physical meaning and thus cannot be compared against observations to diagnose the model's internal logic (Appling et al., 2022). Apart from LSTM, other machine learning models have also been leveraged to simulate stream temperature (Feigl et al., 2021; Sohrabi et al., 2017), but in most cases they similarly serve as black boxes.

To overcome DL's limitations while benefiting from its ability to learn from big data, a new class of physics-informed machine learning models—"differentiable models"—has emerged. They harness the core technology behind DL, differentiable programming, while including process-based equations as model priors or constraints of the system (Shen et al., 2023). These models enable efficient and accurate calculations of gradients of the outputs with respect to the variables used in the model, and these gradients are used to update weights in the connected neural networks (NNs) or parameters in the model. Differentiable models can take basic model structures and assumptions from existing process-based models to serve as the backbone (i.e., structural priors) and then insert NNs to either provide parameter estimation or replace existing process descriptions. Varying degrees

of constraining physics-based structural priors can be imposed, ranging from a full physical structure or graph connectivity (Bindas et al., 2023) with NN-based parameterization, to limited physical constraints. The training is done in an "end-to-end" fashion using gradient descent, and the loss is only computed for the model's final outputs on a large collection of data simultaneously. Because the supervising signal can be backpropagated from the loss function related to the output of the physical components, we do not need target data to directly supervise the outputs of the NNs, nor do we need to pretrain NNs as surrogate models— the NN training and physically based simulation can be done in one single stage.

Differentiable models can, to an extent, circumvent the intertwined parametric issues mentioned above because NNs can produce highly performant, robust, and well-generalized parameters by efficiently learning from big data, reducing the need for complex parameter interrogation. Within this paradigm, we first demonstrated the advantages of differentiable modeling in a method we call differentiable parameter learning (dPL) (Aboelyazeed et al., 2023; Tsai et al., 2021). Another advantage of the dPL framework over current parameter optimization methods is that it has orders of magnitude savings in computational power (because it leverages commonalities between sites and data points, as well as parallel computing). When NNs learn to predict physical parameters directly based on some raw input information and these parameters simultaneously impact the modeling results, there is no need for training data (or "ground truth" measurements) for these physical parameters or other intermediate variables (which is highly beneficial, since such data almost never exist). It has been further shown that we can approach the state-of-the-art prediction performance of LSTM for streamflow using an interpretable, mass-conservative model (Feng et al., 2022) and even potentially obtain better performance with spatial extrapolation and projection of future trends (Feng, Beck, Lawson, & Shen, 2023). The untrained variables (those that were not provided as training targets but are calculated and output due to the physical priors) like evapotranspiration also compared well with alternative estimates (Feng et al., 2022). Differentiable models leverage the advantages of speed and scale (parameters can be rapidly estimated for many sites at once) offered by modern artificial intelligence development (Shen et al., 2023).

Although differentiable models have demonstrated strong potential, their configuration can be highly flexible so their ability to derive process insights remains to be explored. As mentioned above, differentiable models can be set up to be "NN-dominant" in that NN components estimate the bulk of the dynamical variables. For example, Kraft et al. (2022) sought to estimate physical fluxes including evaporation, runoff, and recharge using LSTM, and connected them using mass balance equations. However, for such systems, the physical significance of these fluxes are not guaranteed and the variance of such estimates may be high. In addition, because the deep network components are powerful at fitting to observation data, it is not yet clear if the structural priors matter to the prediction accuracy or if better priors can be identified within a hybrid framework. Unlike with a purely process-based model, there is a risk that the NN components of the model will more-than-competently adapt to and compensate for the unique weaknesses of each alternative structural prior such that accuracy of the combined system is always high and thus the best structural prior cannot be identified. This problem may occur when the implementations of physical constraints in a hybrid model lack balance, and those specific components of the model exhibit an excessive utilization of available freedom. For example, most water quality models depend on hydrologic simulations. A hybrid model which only relies on the physical constraints of water quality equations and excludes the hydrological constraints may be at risk of having large uncertainty.

In this work, we evaluate an NN-dominant hybrid model (an NN model constrained by structural priors of a stream temperature process-based model) with the purpose of not only predicting daily stream temperature, but also estimating the volumes and temperatures of streamflow sources (surface flow, shallow subsurface flow, and groundwater flow) and identifying the most effective model structure for predicting all of these variables. The viability of this approach hinges on the fact that stream temperature data contain information about hydrologic pathways, and with some structural assumptions (priors), the estimation of fluxes can be solved at large scales. Another key idea is that, based on our past experiences with the strong performance of differentiable models joined with LSTM, we hypothesize that the hybrid model will nearly optimally estimate parameters for a given structural prior and configuration, allowing us to reveal the limitations of the structural priors. The present system is "NN-dominant" but can be expanded to include more priors later. We apply multidimensional assessment to four different sets of structural priors, drawn from three existing process-based stream temperature models, and compare their performance and impacts on the learned relationships. We seek to address the following research questions.

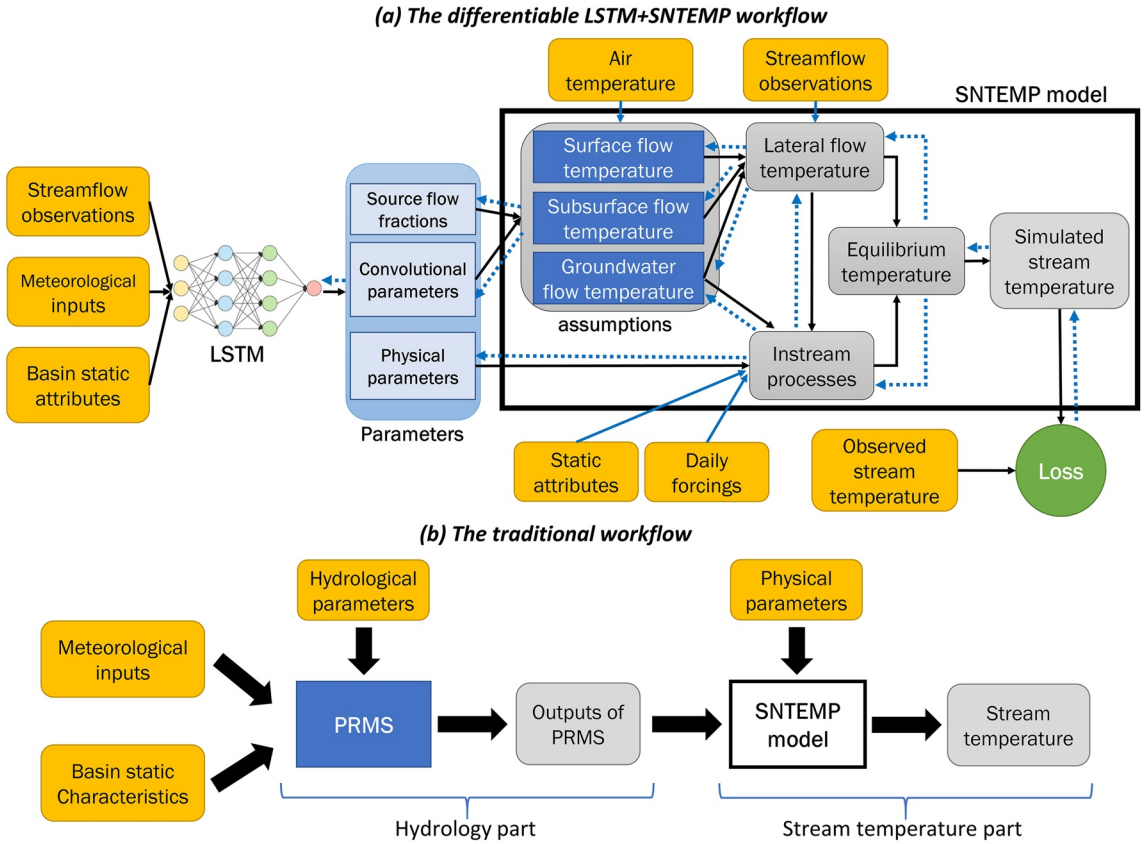

Figure 1. (a) The differentiable parameter learning workflow. Here, parameters include water source fractions, convolutional parameters for temperature calculation in water sources, and physical parameters for the stream temperature (SNTEMP) instream processes module. The blue dashed lines are the backpropagation paths for updating long short-term memory (LSTM) weights. "Assumptions" (blue boxes inside SNTEMP model) shows where different structural priors for thermal memory were implemented. (b) The traditional workflow in PRMS-SNTEMP, of which we only use the SNTEMP part in this work.

1. How well can NN-dominant hybrid models trained only on T_s data predict T_s compared to purely data-driven models, and how much variance do they have in predicting baseflow fractions (which are internal untrained variables) given that they have only limited physics?
2. Are instream processes necessary to achieve high simulation quality for daily stream temperature, and is the previously employed air temperature averaging approach the most suitable for approximating inflow temperature?
3. Given the strong adaptive capability of LSTM, do different structural priors matter to the results of the hybrid model? That is, can we identify better priors?

2. Methods

2.1. Overall Framework

As a summary, our hybrid model (LSTM + mixing assumptions + physical instream thermal processes) followed the dPL approach (Figure 1a) first described in Tsai et al. (2021) and then in Feng et al. (2022). First, given forcing and attribute data for a basin, an LSTM unit predicts (a) daily fractions of water sources into the stream (lateral inflow); (b) convolutional parameters (θ_c) for calculating water temperature from air temperature for different water sources; and (c) some physical parameters (θ). Since this gives LSTM and the overall model lots of flexibility, we call it an NN-dominant system. Next, given various transit time assumptions, we estimate water source temperatures by convolving daily air temperature with kernels (defined by θ_c)—at the risk of oversimplifying the description, this convolution can be regarded as a recency-weighted average of air temperature over certain lengths of time. Lastly, after mixing water sources with a flow-weighted mean of source temperatures to obtain lateral flow water temperature, we used physical parameters (θ) and instream heat processes from

SNTEMP to predict stream temperature at a downstream gauge (Figure 1a). This framework is most applicable to smaller, headwater basins where the river network can be represented by a single conceptual reach, otherwise a routing section should be added to the framework to convey runoff from upstream reaches to downstream. In practice, this conceptual representation also seems to be effective for medium-sized basins of around a few tens of thousands of square kilometers (Figure S1 in Supporting Information S1). This NN-dominant system does not require specification of the structural priors for the hydrologic component. This setup saves developmental effort, but subjects the system to larger uncertainty, which we are trying to understand in this work.

Regarding the structural priors for thermal memory, we considered three sets of basic assumptions existing in the literature: the original assumption in SNTEMP (Sanders et al., 2017) (hereafter SMRA17), the assumption in Meisner et al. (1988) (hereafter MRR88), and that of Wanders et al. (2019) (hereafter WVWBB19). They differ in the number of water storage components included, the length of the air temperature record that is convolved or averaged to estimate the temperature of each storage component, and the thresholds and biases that are applied. For example, both SMRA17 and WVWBB19 have three sources: surface runoff (sr), shallow subsurface water (ss), and groundwater (gw), whereas MRR88 only has surface runoff and groundwater. WVWBB19 applies offsets (biases) to air temperature to calculate sr temperature and assumes 5°C as the minimum gw temperature threshold, whereas the other two studies do not. The necessity of the instream processes is not entirely clear. Thus we also evaluated a model option (called "LSTM + mixing") that did not involve the SNTEMP instream module and thus omitted radiative heating and cooling, shading, evaporation, and streambed friction and conduction (therefore only including the mixing assumptions). Besides these differences, we also considered variants with different convolutional lengths (n) and treatment of some SNTEMP instream parameters as either static or dynamic (θ). For each prior, we tested many configurations (Section 2.4) to exhaustively explore its limits.

All the hybrid models were trained on daily water temperature observations for 415 stations across the conterminous United States (CONUS) (Section 2.3). We evaluated the model with two metrics: the daily temperature simulation accuracy against observed data, and the correlation between the baseflow estimation of our model with a published alternative estimate based on baseflow recession analysis (Section 2.6.2). The remainder of this section describes each part of the model in detail.

2.2. Differentiable Water Temperature Model

The main difference between the differentiable model and many other previous physics-guided frameworks is that the process-based model is written in a differentiable platform like those used for NNs. This approach makes it possible for the framework to seamlessly integrate both machine learning and process-based model parts and be trained as a unified model (Shen et al., 2023). The training process is "end-to-end," with any parameters within the model being learnable by gradient descent, just like when a pure LSTM model is trained (indicated by the blue dashed-line arrows in Figure 1a for the gradient descent path from the loss function to the process-based part and to the NN part). Thus, there is no need to calibrate each component independently. In other words, the process-based model becomes a part of the NN with the added advantage of process transparency. Equations 1–4 (Table 1) serve as the key assumptions of the estimation framework by providing hypothesized constraints (structural priors) that force the LSTM to estimate physically meaningful water source fractions. We will test the effects of these constraints.

All these process-based and NN components were implemented in a machine-learning platform that supports automatic differentiation (thus, the model is programmatically differentiable) to enable end-to-end training of the hybrid model. We used the PyTorch platform (Paszke et al., 2017); however, similar platforms, such as Tensorflow (Abadi et al., 2016), JAX (Bradbury et al., 2021), and Julia (Bezanson et al., 2012), are also options for differentiable modeling. In theory, it is even possible to write the different parts in different differentiable languages; however, the gradient connectivity should be maintained.

2.2.1. Long Short-Term Memory

The LSTM estimates water source fractions and optionally other parameter values for the source water temperature estimation, mixing module, and instream module (Figure 1a). This recurrent NN is a powerful tool for retaining information for long timescales. LSTM equations, gates, and cells have been described in previous works in detail (Fang et al., 2017; Hochreiter & Schmidhuber, 1997) and are thus omitted here.

To standardize the LSTM inputs, streamflow values were divided by long-term mean annual precipitation and watershed drainage area to become a dimensionless variable to decrease the differences between large and small

Table 1
Major Equations for the Models With Different Structural Priors

Model	Equations	
LSTM-SMRA17	$\left\{f'^{1:t}_{ss}, f'^{1:t}_{gw}, \theta^{1:t}, \theta^{1:t}_{c}\right\} = \mathrm{LSTM}\left(x^{1:t}, A_1\right)$	(1)
	$\left\{f^{1:t}_{sr}, f^{1:t}_{ss}, f^{1:t}_{gw}\right\} = \mathrm{MA\_adj}\left(f'^{1:t}_{ss}, f'^{1:t}_{gw}, Q^{1:t}\right)$	
	$T^{t}_{gw} = \max\left(0, \mathrm{conv}\left(T^{t-n_{gw}:t}_{air}, \theta^{t}_{c,gw}\right)\right)$	
	$T^{t}_{ss} = \max\left(0, \mathrm{conv}\left(T^{t-n_{ss}:t}_{air}, \theta^{t}_{c,ss}\right)\right)$	
	$T^{t}_{sr} = \max\left(0, T^{t}_{air}\right)$	
	$T^{t}_{lat} = f^{t}_{sr}T^{t}_{sr} + f^{t}_{ss}T^{t}_{ss} + f^{t}_{gw}(T^{t}_{gw} + b_{opt})$	
	$T^{t}_{s} = \mathrm{SNTEMP}(T^{t}_{lat}, T^{t-1}_{s}, \theta^{t}, A_2, x^{t}_{2})$	
LSTM-MRR88	$\left\{f'^{1:t}_{gw}, \theta^{1:t}, \theta^{1:t}_{c}\right\} = \mathrm{LSTM}\left(x^{1:t}, A_1\right)$	(2)
	$\left\{f^{1:t}_{sr}, f^{1:t}_{gw}\right\} = \mathrm{MA\_adj}\left(f'^{1:t}_{gw}, Q^{1:t}\right)$	
	$T^{t}_{gw} = \max\left(0, \mathrm{conv}\left(T^{t-n_{gw}:t}_{air}, \theta^{t}_{c,gw}\right)\right)$	
	$T^{t}_{sr} = \max\left(0, T^{t}_{air}\right)$	
	$T^{t}_{lat} = f^{t}_{sr}T^{t}_{sr} + f^{t}_{gw}(T^{t}_{gw} + b_{opt})$	
	$T^{t}_{s} = \mathrm{SNTEMP}\left(T^{t}_{lat}, T^{t-1}_{s}, \theta^{t}, A_2, x^{t}_{2}\right)$	
LSTM-WVWBB19	$\left\{f'^{1:t}_{ss}, f'^{1:t}_{gw}, \theta^{1:t}, \theta^{1:t}_{c}\right\} = \mathrm{LSTM}\left(x^{1:t}, A_1\right)$	(3)
	$\left\{f^{1:t}_{sr}, f^{1:t}_{ss}, f^{1:t}_{gw}\right\} = \mathrm{MA\_adj}\left(f'^{1:t}_{ss}, f'^{1:t}_{gw}, Q^{1:t}\right)$	
	$T^{t}_{gw} = \max\left(5.0, \mathrm{conv}\left(T^{t-n_{gw}:t}_{air}, \theta^{t}_{c,gw}\right)\right)$	
	$T^{t}_{ss} = \max\left(0, T^{t}_{air}\right)$	
	$T^{t}_{sr} = \max\left(0, T^{t}_{air} - 1.5\right)$	
	$T^{t}_{lat} = f^{t}_{sr}T^{t}_{sr} + f^{t}_{ss}T^{t}_{ss} + f^{t}_{gw}(T^{t}_{gw} + b_{opt})$	
	$T^{t}_{s} = \mathrm{SNTEMP}\left(T^{t}_{lat}, T^{t-1}_{s}, \theta^{t}, A_2, x^{t}_{2}\right)$	
LSTM + mixing	$\left\{f'^{1:t}_{ss}, f'^{1:t}_{gw}, \theta^{1:t}, \theta^{1:t}_{c}\right\} = \mathrm{LSTM}\left(x^{1:t}, A_1\right)$	(4)
	$\left\{f^{1:t}_{sr}, f^{1:t}_{ss}, f^{1:t}_{gw}\right\} = \mathrm{MA\_adj}\left(f'^{1:t}_{ss}, f'^{1:t}_{gw}, Q^{1:t}\right)$	
	$T^{t}_{gw} = \max\left(0, \mathrm{conv}\left(T^{t-n_{gw}:t}_{air}, \theta^{t}_{c,gw}\right)\right)$	
	$T^{t}_{ss} = \max\left(0, \mathrm{conv}\left(T^{t-n_{ss}:t}_{air}, \theta^{t}_{c,ss}\right)\right)$	
	$T^{t}_{sr} = \max\left(0, T^{t}_{air}\right)$	
	$T^{t}_{s} = f^{t}_{sr}T^{t}_{sr} + f^{t}_{ss}T^{t}_{ss} + f^{t}_{gw}(T^{t}_{gw} + b_{opt})$	

Note. Explanations: $x^{1:t}$ are the basin-averaged atmospheric forcings time series, from day 1 to day t that consist of daily precipitation, minimum and maximum air temperature, day length, shortwave solar radiation, vapor pressure, and observed streamflow. A_1 are the static attributes representing basin characteristics. $f'^{1:t}_{ss}$, $f'^{1:t}_{gw}$ are the raw outputs of LSTM. MA is the moving average (MA) filter. $f^{1:t}_{sr}$, $f^{1:t}_{ss}$, and $f^{1:t}_{gw}$ are the daily fractions of streamflow contributions coming from surface flow, subsurface flow, and groundwater flow. T^{t}_{sr}, T^{t}_{ss}, T^{t}_{gw} are the surface, subsurface, and groundwater temperatures. $\theta^{1:t}$ are the SNTEMP parameters (can be time-dependent or static), which are: light reduction factors due to riparian shading and topographic shading and no shading (these sum to 1 each day), two parameters for estimating stream width w from discharge Q ($w = a \times Q + b$, in which a and b are the parameters), the Hamon coefficient for evaporation heat flux, and an optional bias term (b_{opt}) for groundwater temperature. θ_c are the parameters for the convolutional filter (conv(), a cumulative gamma distribution) used to compute source water temperatures from recent air temperatures. n_{gw} and n_{ss} are the number of days used in the convolutional filter for groundwater and subsurface flows. T_{lat} is the lateral flow temperature, which is calculated as a flow-weighted average of the temperatures of all water sources. A_2 are the basin attributes that SNTEMP needs in order to model stream temperature at each time step: slope, elevation, and stream length. x_2 are the dynamic daily forcings that SNTEMP needs, such as mean air temperature, solar radiation, flow rate, and vapor pressure. T^{t}_{s} is the final stream temperature at time t. MA_adj is a MA adjustment operator to improve the smoothness of the subsurface and groundwater fractions prediction. It involves applying a MA smoothing operator to f_{gw} and f_{ss} with different lengths (MA_length), and then adjusting the other fractions so they sum up to one. For priors that contain all three components, f_{gw}, f_{ss}, and f_{sr} can written as the following: $f^{1:t}_{gw} = \min\left(\frac{\mathrm{MA}\left(f'^{1:t}_{gw} * Q^{1:t}, \mathrm{MA}_{\mathrm{length}}\right)}{Q^{1:t}}, 1\right)$, $f^{1:t}_{ss} = \min\left(\frac{\mathrm{MA}\left(f'^{1:t}_{ss} * \left(1 - f^{1:t}_{gw}\right) * Q^{1:t}, \mathrm{MA}_{\mathrm{length}}\right)}{Q^{1:t}}, 1 - f^{1:t}_{gw}\right)$, $f^{1:t}_{sr} = \left(1 - f^{1:t}_{gw} - f^{1:t}_{ss}\right)$, while for the prior without $f^{1:t}_{ss}$, the subsurface fraction is equal to zero.

basins. This variable, along with precipitation, was redistributed by a logarithmic formula to look more like a Gaussian distribution (Equation 5) (please see Feng et al., 2020). These two variables, along with the remaining inputs, were then standardized using Equation 6. The standardization process is similar to our previous work (Rahmani, Shen, et al., 2021).

$$v^* = \log_{10}(\sqrt{v} + 0.1) \tag{5}$$

$$x_{i,\text{stand}} = \frac{\left(x_i - \overline{x}\right)}{\sigma} \tag{6}$$

where v is the dimensionless streamflow or precipitation value, v^* is the value transformed to a Gaussian distribution, x_i is the initial values of streamflow or precipitation (after processing by Equation 6), $\overline{x}$ and σ are the mean and standard deviation of the initial values (calculated based on the training data set only), and $x_{i,\text{stand}}$ is the standardized value that has been used as an input for the LSTM NNs.

The LSTM hyperparameters were determined based on our previous experiences with running the LSTM model for predicting daily stream temperature across the contiguous United States (Rahmani, Lawson, et al., 2021; Rahmani, Shen, et al., 2021). In this study, we ran all experiments with 600 epochs. One forward run for all training basins is considered one epoch. The number of features in LSTM, called hidden layer size, was 256. The length of the time series for training and testing, known as ρ, was selected as 365 days. Each training sample had 208 basins. We used Adadelta (Zeiler, 2012) with adaptive learning rates as the optimizer to update the parameters based on gradients.

2.2.2. Structural Priors for Source Water Temperature and Mixing

The main structural priors for source water temperature and mixing are described in Table 1. In the original assumptions by SMRA17, WVWBB19, and MRR88, the source water temperature is an average of the air temperatures from a number of recent days. MRR88 and WVWBB19 assume the groundwater temperature equals the annual mean air temperature. SMRA17 gives the freedom to select different time spans for averaging air temperature for the groundwater temperature calculation. The default value is 365 days (Regan & Markstrom, 2021); however, this value is adjustable from 1 to 365. SMRA17 makes the same assumptions for shallow subsurface flow temperature but with a default value of 30 days. The WVWBB19 model assigns daily temperature to subsurface flow.

In our work, to accommodate gradient-based training and to make the model more realistic, we chose to convolve daily air temperature with a cumulative gamma function distribution (defined by θ_c parameters, and conv() in Table 1), resulting in an S-type curve similar to Mohseni and Stefan (1999). Earlier days were assigned lower weights (Figure S2 in Supporting Information S1). This curve may be flat, in which case it reverts to simple averaging just as in the original SMRA17. It can also have a large gradient to give higher weights to the most recent days. Moreover, as to be discussed in Section 2.2.4, we also explored adding an optional bias term b_{opt} to the groundwater temperature part to compensate for additional heat fluxes that are not described, such as the heat conduction from geothermal heat flux (Burns et al., 2016; Yasukawa et al., 2009). The initial temperature of lateral flow entering the stream (before instream thermal exchanges occur) was calculated as the flow-weighted average of surface, subsurface, and groundwater source waters. We applied a moving average (MA) convolutional filter on the raw LSTM outputs for groundwater flow ($f'_{\text{gw}} \times$ total_Q) and subsurface flow ($f'_{\text{ss}} \times$ total_Q), and subsequently obtained f_{gw} (groundwater flow fraction) and f_{ss} (subsurface flow fraction), respectively. Surface flow fraction f_{sr} is determined by the subtraction of the combined proportion of groundwater and subsurface flow fractions from the total value of 1 (Table 1).

2.2.3. The SNTEMP Instream Module

SNTEMP contains a heat transport model that can simulate daily mean and daily maximum temperatures based on the concept of equilibrium temperature (Sanders et al., 2017) and net heat flux calculation. SNTEMP is a one-dimensional in-stream temperature model that calculates the transfer of energy to or from a stream segment by heat flux equations or advection. As mentioned earlier, in each basin we only simulate one conceptual reach to represent all the channel flow in the basin. This lumped representation may not capture the local heterogeneities in large basins, but greatly simplifies the modeling. The energy balance is computed for each time step for each stream segment based on net heat flux from shortwave solar radiation, evaporation heat flux from the latent

heat of vaporization, streambed heat flux, heat flux from friction, and longwave radiation emitted by the water, riparian vegetation, and the atmosphere. Daily cloud coverage is calculated based on the average watershed slope, aspect, and the ratio of actual shortwave solar radiation and maximum shortwave solar radiation on each Julian day of the year (in our case, we assumed aspect and slope to be equal to zero for simplicity). SNTEMP assumes that the initial water temperature tends to reach the equilibrium temperature in a given meteorological condition (Edinger et al., 1968; Theurer et al., 1984). The final stream temperature (T_s) in the conceptual reach segment is a second-order Taylor expansion (Hazewinkel, 2001) between the equilibrium temperature and the incoming source water temperature, in our case T_{lat}. This Taylor expansion mainly relies on the above-mentioned heat flux equations. Here we are testing different structural priors (SMRA17, MRR88, and WVWBB19) in the SNTEMP model (Table 1).

2.2.4. Structural Configurations

With each prior there are still many configurations that control the behaviors and best achievable outcomes of the model. Given different structural configurations, the training may produce models with different performance or internal diagnostic outputs. To understand the variance of performance, we ran experiments with a multitude of structural configurations, including:

1. The length of the convolutional filter in groundwater temperature (n_{gw}) was set to 365 days for all models.
2. The length of the convolutional filter for subsurface flow temperature (n_{ss}) in LSTM-SMRA17 was 30 days.
3. To increase the smoothness of LSTM-predicted daily groundwater flow, we added a MA convolution filter to groundwater and subsurface flow (the MA() operations in Table 1). After this filter, we rescaled the total water source fractions to 1. We ran the experiments with and without the MA filter. MA was tried with different lengths (MA_length) of 7, 15, 30, 40, 60, 75, and 90 days for LSTM-SMRA17, LSTM-MRR88, and LSTM-WVWBB19.
4. One parameter provides an optional bias adjustment (b_{opt}) for groundwater temperature. We ran experiments with this parameter set to 0 or nonzero values for all models except LSTM + mixing.
5. Most parameters can be either dynamic or static in our settings (see Table S1 in Supporting Information S1 for the list of combinations). Dynamic parameters are different from day to day whereas static parameters do not change over time. We ran experiments with both static and dynamic options for all parameters except temperature convolutional filters that were fixed as static.

In total, we ran 144 experiments with different configurations of structural hyperparameters to understand whether they induced substantial variance in the performance and how they impacted simulated stream temperature and water source fractions.

2.3. Data Sets

We selected 415 headwater basins from the 9,017 basins in the Geological Attributes of Gages for Evaluating Streamflow data set, version-II (GAGES-II) (Falcone, 2011) across the CONUS. Basin attributes from GAGES-II were used as inputs or evaluation targets as described in Sections 2.4 and 2.6.2, respectively. We selected those 415 basins for which observed streamflow data were fully available from October 2010 to September 2016, and stream temperature data were available for at least 10% of the days in that window (Rahmani, Shen, et al., 2021). Mean daily observed streamflow and stream temperature data were downloaded from the U.S. Geological Survey (USGS) National Water Information System (U.S. Geological Survey, 2016). We collected daily meteorological forcings for these basins in the above-mentioned time range by extracting the overlaps of the basins' shapefiles and the 1-km grid cells of the Daymet V4 data set (Thornton et al., 2022).

2.4. Inputs

In this study, we used 33 attributes and seven forcings as inputs to the LSTM (Table S2 in Supporting Information S1). The attributes were mostly related to topography (e.g., area, slope), long-term climate variables (e.g., mean and maximum annual air temperatures), dam and reservoir information (e.g., number of dams), land coverage (e.g., forest coverage percentage in watershed), geological aspects (e.g., dominant geology type), and soil types (e.g., silt and clay percentages). Meteorological forcings for the LSTM were maximum and minimum daily air temperature, vapor pressure, observed daily streamflow, day length, shortwave solar radiation,

and precipitation. SNTEMP needs a different set of attributes and forcings (Table S2 in Supporting Information S1). SNTEMP attributes are slope, watershed drainage area, and stream (segment) length. We calculated stream length based on basin area provided by GAGES-II.

Additionally, SNTEMP needs daily vapor pressure for the atmospheric longwave radiation heat flux equation, shortwave solar radiation to calculate the absorbed portion by the stream, and flow rate for each water source. In the original SNTEMP settings, daily flow information is modeled and provided by the hydrology part called PRMS (Figure 1b); however, we provided this information to SNTEMP by defining the water source fractions as outputs by the NN (Figure 1a).

2.5. Outputs

Our framework predicts daily stream temperature in an end-to-end fashion. LSTM provides the parameters that the process-based model (SNTEMP) component needs, and the process-based model component simulates daily stream temperature. Intermediate values, including the water source fractions and evapotranspiration rate, can also be tracked.

2.6. Experiments and Metrics

2.6.1. Training and Testing Settings

The training time range in all experiments was from 1 October 2010 to 30 September 2014. The testing period was the next 2 years, between 1 October 2014 and 30 September 2016. We chose these time ranges to be comparable to another study that used LSTM alone to model daily stream temperature prediction (Rahmani, Shen, et al., 2021). We ran 144 different model configurations, as described in Section 2.2.4.

2.6.2. Model Evaluation Against Literature

Apart from comparing daily T_s to observations, we estimated one baseflow index (BFI) per basin by summing the daily baseflow (groundwater and subsurface flow fractions each multiplied by streamflow for each day) and dividing it by the total streamflow in the same period. For SMRA17, "baseflow" includes flow from both the deep and shallow subsurface. For MRR88 and WVWBB19, there is only one deep subsurface compartment which is treated as baseflow. We calculated the spatial correlation between this estimate and the BFI from hydrograph separation (R_{BFI}), which combines a local minimums approach with a recession slope test (Wolock, 2003). Based on Wolock's work, GAGES-II (Falcone, 2011) calculated a long-term estimation of the baseflow percentage of total streamflow by inverse-distance-weighted interpolation of the baseflow ratios that were calculated at USGS stream gauges and made 1-km resolution grid points. Although none of the above methods can be considered the ground truth, the comparison can be a soft criterion to provide a sense of our model's capability to capture the hydrological signals.

2.6.3. Performance Metrics

We computed metrics for each basin, treated for simplicity as a single stream reach. The metric used in the loss function for error minimization between simulated and observed stream temperature was the root-mean-square error (RMSE) (Equation 7). However, we also calculated bias (the average error between simulation and measurement) (Equation 8), Nash-Sutcliffe efficiency (NSE) (Nash & Sutcliffe, 1970) (Equation 9), and Kling-Gupta efficiency (Gupta et al., 2009) (Equation 10) for some of the selected experiments to compare with each other.

Additionally, one of the aims of this study was to investigate whether it is possible to get a low error in stream temperature simulation (the target of the model) and a high correlation between the simulated source baseflow ratio (intermediate output of the model) and an alternative baseflow estimate (BFI, in GAGES-II). We calculated the Pearson correlation (Equation 11) between simulated baseflow and GAGES-II values. These metrics measure whether the model accurately predicts stream temperature and utilizes physically meaningful intermediate parameters. The metrics can be expressed as:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (T_{i,\mathrm{sim}} - T_{i,\mathrm{obs}})^2}{n}} \tag{7}$$

$$\text{Bias} = \frac{\sum_{i=1}^{n}(T_{i,\text{sim}} - T_{i,\text{obs}})}{n} \tag{8}$$

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n}(T_{i,\text{sim}} - T_{i,\text{obs}})^2}{\sum_{i=1}^{n}\left(T_{i,\text{obs}} - \overline{T}_{\text{obs}}\right)^2}, \qquad \overline{T}_{\text{obs}} = \frac{\sum_{i=1}^{n} T_{i,\text{obs}}}{n} \tag{9}$$

$$\text{KGE} = 1 - \sqrt{(\text{Corr} - 1) + \left(\frac{\sqrt{\frac{\sum_{i=1}^{n}\left(T_{i,\text{sim}} - \overline{T}_{\text{sim}}\right)^2}{n}}}{\sqrt{\frac{\sum_{i=1}^{n}\left(T_{i,\text{obs}} - \overline{T}_{\text{obs}}\right)^2}{n}}} - 1\right) + \left(\frac{\overline{T}_{\text{sim}}}{\overline{T}_{\text{obs}}} - 1\right)} \tag{10}$$

$$\text{Corr} = \frac{\sum_{i=1}^{n}\left[\left(T_{i,\text{obs}} - \overline{T}_{\text{obs}}\right)\left(T_{i,\text{sim}} - \overline{T}_{\text{sim}}\right)\right]}{\sqrt{\sum_{i=1}^{n}\left(T_{i,\text{obs}} - \overline{T}_{\text{obs}}\right)^2 \sum_{i=1}^{n}\left(T_{i,sim} - \overline{T}_{\text{sim}}\right)^2}}, \qquad \overline{T}_{\text{sim}} = \frac{\sum_{i=1}^{n} T_{i,\text{sim}}}{n} \tag{11}$$

In which $T_{i,\text{obs}}$ and $T_{i,\text{sim}}$ are observed and simulated stream temperatures and n is the number of observations during the time range the metrics are calculated. $\overline{T}_{\text{obs}}$ and $\overline{T}_{\text{sim}}$ are the average of observed and simulated stream temperatures. Index i indicates the ith day of the testing period. RMSE, Bias, NSE, KGE, and Corr are the RMSE, bias, Nash-Sutcliff efficiency, Kling-Gupta efficiency, and correlation metrics between simulations and observations, respectively. All results discussed in this manuscript are the performance results in the testing period, which were very close to those of the training period (please see Rahmani et al., 2023 for full data release).

3. Results and Discussion

We first show the mild tradeoff between temperature prediction accuracy and BFI correlation. Stream temperature serves as the target variable in the model, and baseflow is an intermediate untrained variable. Then we discuss the different tradeoff curves corresponding to different priors. We then selected several models to explore the predictive capability and behaviors of these models in more depth.

3.1. The Mild RMSE-R_{BFI} Pareto Front and the Significant Variance in R_{BFI}

With different structural configurations, the trained models are scattered in different places on the median RMSE-R_{BFI} plane (R_{BFI} represents the correlation coefficient of BFI), showing noticeable variance even for the same priors (Figure 2). For the LSTM-SMRA17 prior, many points (with different configurations) were spread out laterally near the bottom of the plot, which means these models have almost equally good temperature predictive performance but widely different R_{BFI} values (which also implies the simulated BFIs are varying widely). Apparently, near the bottom line, we have exhausted the room for optimization with respect to temperature RMSE for this prior. However, given that prior, various structural configurations with similar temperature prediction accuracies led to different R_{BFI} values. Similar patterns are found with LSTM-MRR88 and LSTM-WVWBB19. R_{BFI} is a useful second dimension; however, because the hydrograph separation approach used for comparison is also empirical and cannot clearly distinguish between shallow and deep subsurface outflow, R_{BFI} should only provide a soft constraint on model selection.

The tradeoffs between RMSE and R_{BFI} are mild, manifesting as sharp corners on the point clouds for each prior in Figure 2. Sharp Pareto fronts mean that the two dimensions do not strongly interact, and we do not have to lose performance by one metric to improve the other. The mild tradeoff may have resulted from the strong optimization capability of deep networks and the NN-dominant nature of this hybrid model. We hypothesize that the tradeoffs would be stronger with more rigid structural constraints. More constraints could be provided by integrating a process-based hydrology model into the current framework.

3.2. The Impacts of Different Priors and Process Insights

We can see that different structural priors demonstrate distinct behaviors in both the best achievable metrics and the tradeoff, and analyzing such behaviors gives us a wealth of insights about processes. Summarizing the many experiments we have run (Table 1), LSTM-SMRA17 gave the lowest error in terms of stream temperature

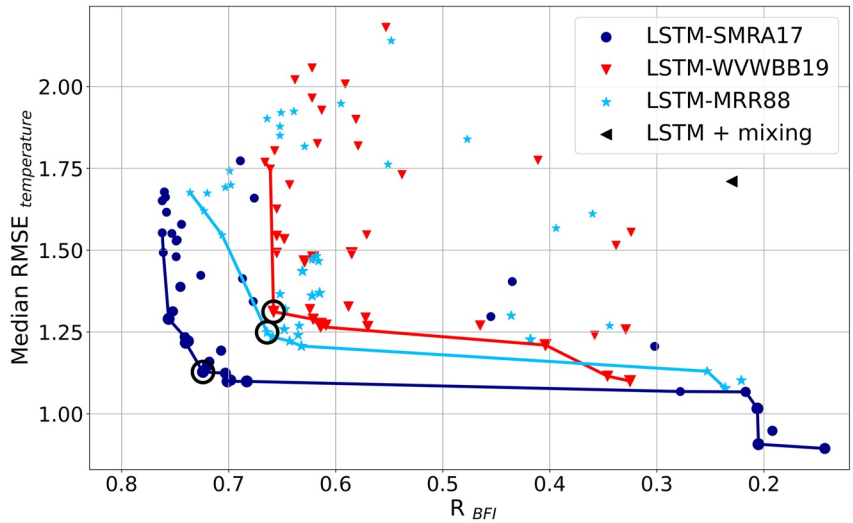

Figure 2. Scatterplot of the median root-mean-square error (RMSE) of stream temperature and the Pearson correlation of simulated baseflow with baseflow index ratio (R_{BFI}) in the Geological Attributes of Gages for Evaluating Streamflow data set, version-II for four model structures with different structural priors. Each point represents an individual model with different configurations. The points closer to the bottom left of XY coordinates are more desirable (higher R_{BFI} and lower median RMSE). The black circles at the Pareto fronts are the models that were selected for further analysis in Section 3.3. The size of the points increases with more dynamic parameters in the model.

as well as the best BFI correlation (Figure 2). LSTM-SMRA17 has the most complicated processes, enabling a low RMSE, but it may have too much flexibility, as evidenced by the wide range of R_{BFI} values that suggest a large variance in the internal representations. More than half of the LSTM-SMRA17 stream temperature RMSE values are between 0.89°C and 1.27°C, and between 0.74 and 0.80 for R_{BFI}. LSTM-MRR88 displays slightly worse stream temperature accuracy (more than half of RMSE values are between 1.08°C and 1.45°C), and R_{BFI} (more than half of R_{BFI} values are between 0.63 and 0.74). LSTM-WVWBB19, on the other hand, has a minimum RMSE of 1.10°C. In addition, as mentioned earlier, LSTM-SMRA17 has a sharp Pareto front, LSTM-MRR88 has a mild one, and LSTM-WVWBB19 has a moderate one. All results discussed in this manuscript are the performance results in the testing period, which are very close to the training period performance (please see Rahmani et al., 2023 for full data release).

These different behaviors show that the basic priors have noticeable control over the stream temperature predictions despite the presence of a highly capable ML component for parameterization—thus we can attribute these differences clearly to the priors and gain process insights. In this case, it seems SMRA17 and MRR88 are both sound priors to perform at the national scale, with the former being slightly preferred. However, the variance of the predicted BFI remains large even within a small range of the optimal RMSE, suggesting the internal dynamics are underconstrained by stream temperature alone. If LSTM-SMRA17 were to be developed further, its utility might be improved with additional data sets or process constraints. Notice that "being a good prior" is different from "having small variance"—a good prior defines the upper bound of performance and structural interpretation, but more data may be required to reduce the variance.

Comparing the results between LSTM-SMRA17 and LSTM-MRR88 implies that a separate bucket for shallow subsurface flow (the only difference between the two models) could improve both stream temperature accuracy and baseflow representation. Having this compartment is justified physically, as Briggs et al. (2018) associated the substantial phase lag between air temperature and stream temperature with the shallow subsurface water passing through the preferential water zones of groundwater. LSTM-SMRA17 and LSTM-WVWBB19 use different assumptions for temperature calculation of water source fractions that resulted in relatively better performance in both aspects in LSTM-SMRA17 along the Pareto front. For instance, LSTM-WVWBB19 assumes the temperature of the subsurface flow to be the same as air temperature because the flow is considered as an interflow/stormflow, not baseflow (Wanders et al., 2019). LSTM-SMRA17 assumes a longer averaging time for calculating subsurface flow temperature and considers 0°C as the minimum groundwater temperature, which seems more consistent with big data.

All the models with relatively high R_{BFI} and lower RMSE values (lower left corner of Figure 2) are regarded as functional, and LSTM-SMRA17 has many such models. Most of the models with lower accuracy in BFI correlations (the points on the left side of Figure 2, mostly) are the ones in which the flow MA operator was not activated. This MA operator serves to stabilize the water source fractions and can be interpreted as representing transient storages and pools along the flow paths that act as buffers. This aspect of simulation is not constrained by the temperature observations but is regulated by the BFI correlation. The benefit of MA operators in terms of BFI correlation suggests the water source fractions should not vary too rapidly.

Almost all experiments in the Pareto fronts consist of dynamic θ_c or θ, having the configuration of type 1 (all parameters except the convolutional filters are dynamic), type 2 (similar to 1 but with static groundwater temperature bias), or type 3 (similar to 2 but with static stream width coefficient factors) (Tables S1 and S3 in Supporting Information S1), which suggests different values exist for parameters in different seasons. The models with all static parameters (type 4 in Table S1 in Supporting Information S1) had relatively larger median RMSE values for stream temperature. This is consistent with our knowledge that these parameters can behave in a time-dependent manner. For instance, topographic and vegetation shading vary seasonally with the sun's height and leaf abundance. There is a tradeoff: overly dynamical parameterization could lead to overfitting and non-physical outputs, while over-static parameterization could lead to underfitting. Here, based on the multidimensional evaluation, it seems that dynamical parameterization is justified.

For the purpose of determining source water temperature (groundwater and shallow subsurface if applicable), a cumulative gamma function (recency-weighted) outperforms a simple averaging function in three models (WVWBB19, SMRA17, and MRR88). We observed that switching from a simple average function to convolutional parameters respectively improved the RMSE and R_{BFI} values by 11% and 13% on average (Table S4 in Supporting Information S1). Because both metrics were improved, it can be argued that the use of the recency-weighted scheme, rather than a simple average, is a better approximation of reality. Conceptually, a simple average implies a well-mixed groundwater reservoir, whereas the recency-weighted scheme could arise from flow systems where the water in the shallow layers plays a prominent role in the outflow (Freeze & Cherry, 1979; Wang et al., 2017; Zijl, 1999). Our results lend support to the latter, consistent with other groundwater modeling efforts (Maxwell et al., 2016; Zhang et al., 2021), and show that such signals of flow can be identified by the data and our model. Simple averaging has been employed in stream temperature models for decades (Anderson, 2005; Meisner et al., 1988; Pekárová et al., 2022), and our results clearly show that the alternative is better. This is not meant to disparage previous models, but rather to highlight the promise of new differentiable modeling methods which can estimate high-dimensional functions and learn from big data.

3.3. Additional Analysis of Selected Models

We selected three models out of all 144 experiments (the three black-circled points in Figure 2, Table S5 in Supporting Information S1). For additional comparison, we added a model that does not have instream processes and only describes the mixing processes (LSTM + mixing, black triangle in Figure 2), to understand the necessity of the instream component.

Overall, these selected differentiable models performed formidably well in stream temperature prediction, but there are also differences (Figure 3). The chosen LSTM-SMRA17 model demonstrates outstanding performance with median NSE and KGE values of 0.970 and 0.951. These metrics are very close to the LSTM model by Rahmani, Lawson, et al. (2021) and Rahmani, Shen, et al. (2021) with median NSE and KGE values of 0.98 and 0.956. The variability of metrics, indicated by the lengths of the whiskers on the boxes of Figure 3, was the smallest for LSTM-SMRA17 amongst the differentiable models, which indicates the model's performance was consistent across basins. The median bias was the closest to zero (bias = −0.23°C); however, the median RMSE was 0.23°C higher than the LSTM model's RMSE. LSTM-MRR88 and LSTM-WVWBB19 ranked third and fourth in terms of median RMSE, NSE, and KGE (Figure 3).

Instream processes played a crucial role in simulating stream temperature, judged by the considerable difference in RMSE values between LSTM + mixing and the other models (Figure 3, also notice the black triangle in Figure 2). The LSTM + mixing model incurred a much larger bias (mostly negative), suggesting instream processes such as solar insolation and stream-atmospheric heat exchange frequently served to impact stream temperature by warming up the water. Considering that, in the absence of instream processes, the bias cannot

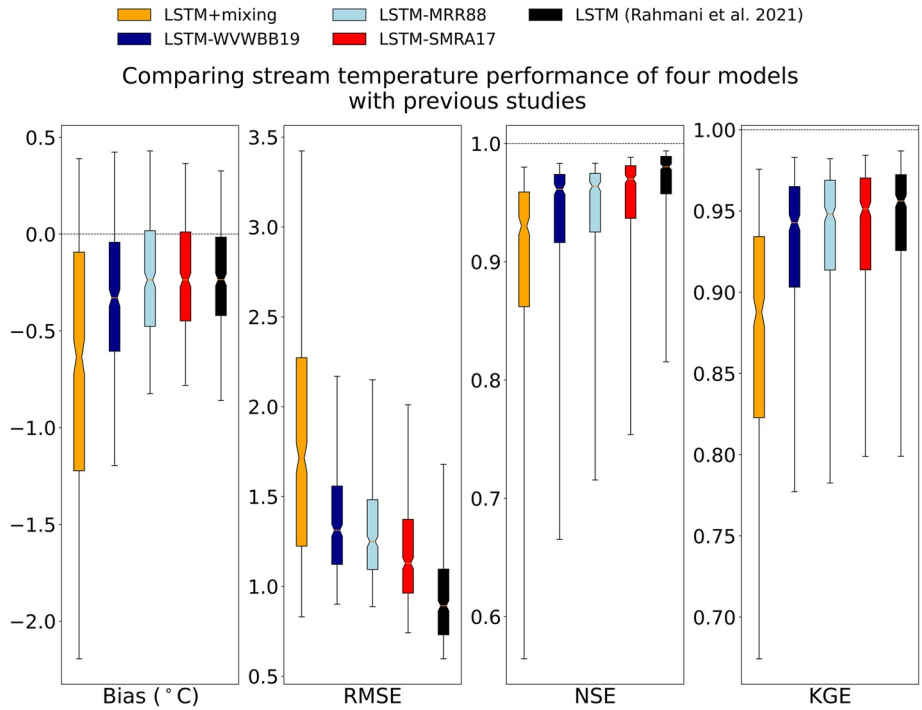

Figure 3. Conterminous United States-scale aggregated metrics (415 basins, tested between 1 October 2014 and 30 September 2016) of the selected stream temperature models using the differentiable parameter learning (dPL) framework, with several basin-scale water temperature memory functions as priors combined with the long short-term memory (LSTM) algorithm (Rahmani, Shen, et al., 2021). Except for LSTM-mixing, these priors are all connected to stream temperature as the in-stream temperature module, but they have different ways to calculate temperatures and fractions of water sources (Table 1). LSTM-mixing is a hybrid differentiable model that only contains mixing processes and excludes in-stream processes entirely. The selected models have only slightly lower performance than LSTM, while also providing process transparency and likely better reliability in cases of extrapolation. The lower whisker, lower box edge, center bar, upper box edge, and upper whisker represent 5%, 25%, 50%, 75%, and 95% of the sites, respectively.

even be mitigated by the strong adaptive capability of LSTM, processes that lead to shifts in channel temperature, in addition to those accounting for lags and mixing, are mandatory to fit the observations. If we examine equation set 6 in Table 1 (LSTM + mixing), there is no term that would introduce shifts in temperature into the system after source waters are mixed, which led to irreconcilable negative biases.

Visualizing stream temperature time series at a few stream gauges confirms that the issue with the LSTM + mixing model is due to omitting instream processes (Figure 4). For instance, in the Maine and Alabama sites, LSTM + mixing mostly showed a negative bias resulting from staying closer to the air temperature (e.g., the average absolute biases between air temperature and simulation in Maine for LSTM + mixing, LSTM-SMRA17, and LSTM-MRR88 were 4.32°C, 4.55°C, and 4.52°C, respectively)—presumably, the solar insolation processes not captured by LSTM + mixing caused this bias. Moreover, the shallow subsurface flow component let LSTM-SMRA17 better control the local peaks and valleys in temperature simulation, compared to LSTM-MRR88 in Alabama (Figures 4a and 4b). In summer and fall at the stream gauge in Oregon, the lack of a shallow subsurface flow temperature module in LSTM-MRR88 resulted in more bias compared to LSTM-SMRA17 (Figure 4c).

The selected models illustrated strong correlation with the baseflow estimates obtained from hydrograph separation compared to LSTM + mixing (Figure 5). Among the chosen models (circled in Figure 2), LSTM-SMRA17 showed the highest Pearson correlation (R_{BFI}) with GAGES-II BFI ($r = 0.724$). LSTM-WVWBB19 and LSTM-MRR88 were relatively less correlated, with R_{BFI} values of 0.658 and 0.664, respectively. The lowest R_{BFI} was for LSTM + mixing with a value of 0.234, suggesting that this model is compensating for the lack of instream processes using other processes like source flow temperatures in mixing processes. Other than the intercept of the LSTM + mixing model, the intercepts LSTM-MRR88 and LSTM-WVWBB19 were approximately 0.2, however

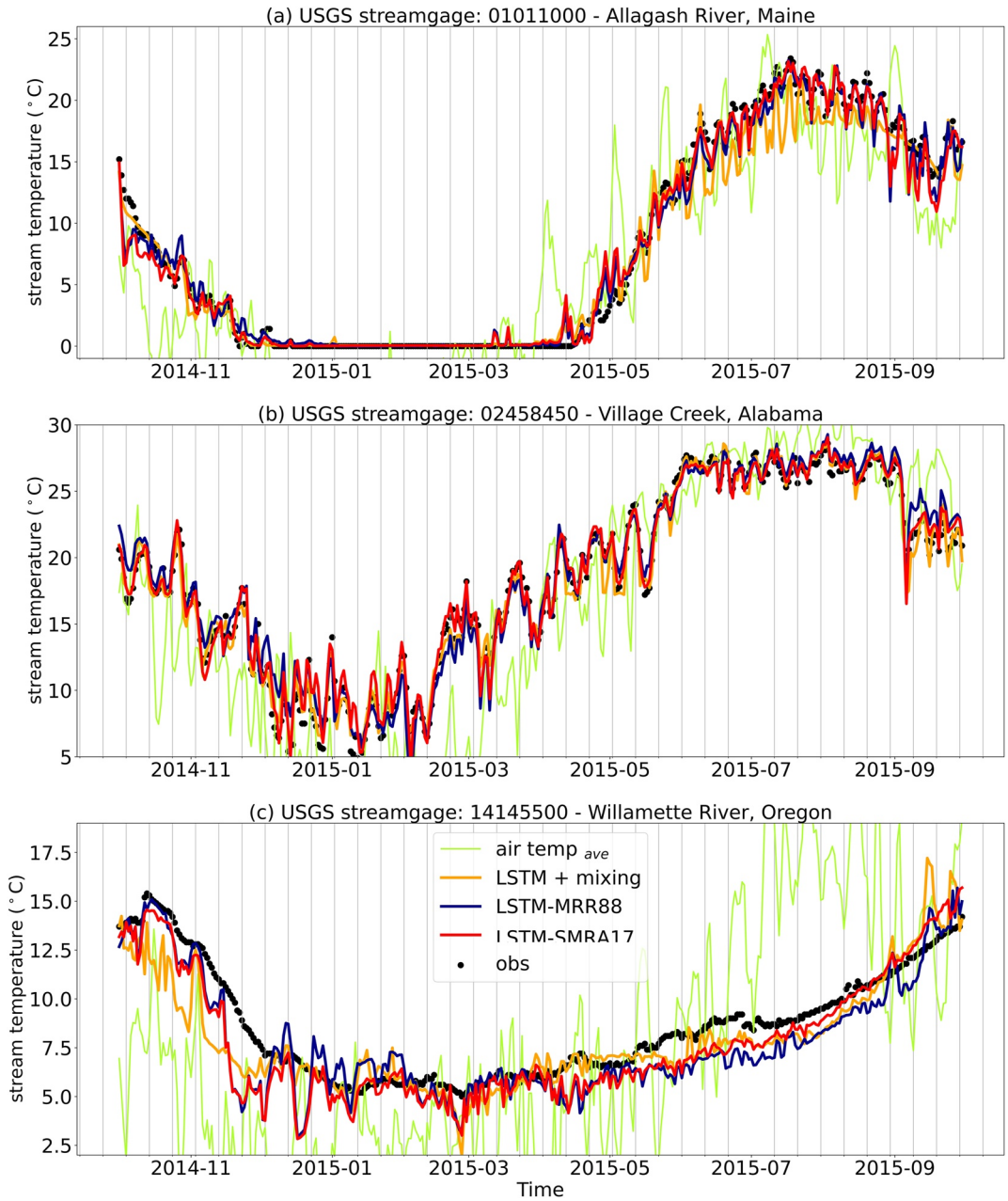

Figure 4. Stream temperature time series of observed values, LSTM-SMRA17, LSTM-MRR88, LSTM + mixing, and air temperature in testing time range for U.S. Geological Survey (USGS) stream gauges in (a) Allagash River in Maine, (b) Village Creek at Alabama, and (c) Willamette River in Oregon. Observed data from U.S. Geological Survey (2016).

the intercept in LSTM-SMRA was negligible which implies that the shallow subsurface module is necessary for achieving higher performance in both stream temperature and baseflow estimation (Figure 5). It also suggests there is enough information in stream temperature data to inform shallow and deep hydrologic paths, a conclusion that can be used for future large-scale modeling.

LSTM-SMRA17 showed consistent spatial BFI patterns with GAGES-II (Falcone, 2011; Wolock, 2003), in general (Figure 6). However, the lack of a subsurface module in LSTM-MRR88 resulted in an underestimation of BFI in many basins. BFI values were larger in the Rocky Mountains, Oregon, and Washington, and smaller in parts of the central lowlands (Texas, Oklahoma, Kansas, Missouri, Iowa, Indiana, Ohio) and Louisiana. LSTM-SMRA17 could capture these different behaviors. The main differences between LSTM-SMRA17

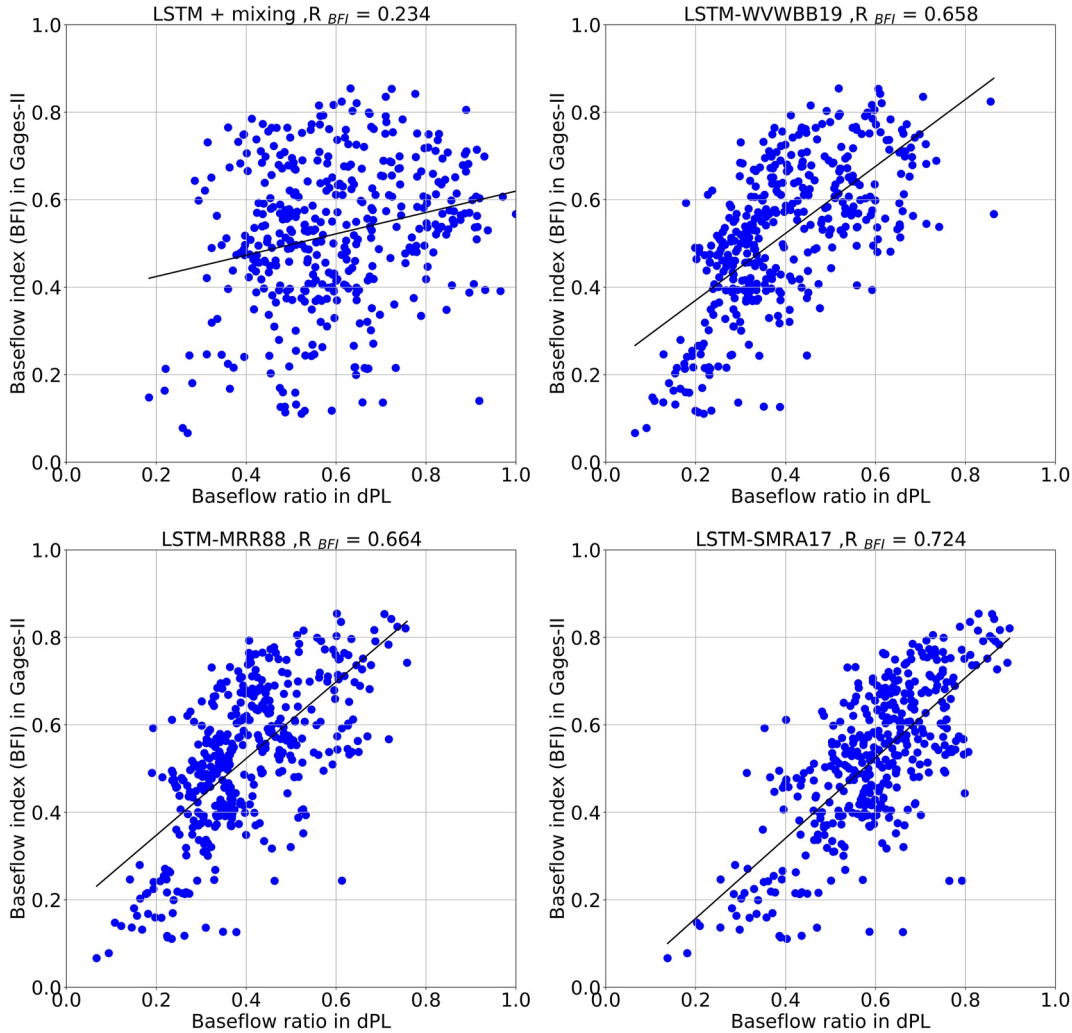

Figure 5. Scatterplots comparing baseflow ratio predicted by each of four models (our selected differentiable parameter learning [dPL] models in Figure 2) and Geological Attributes of Gages for Evaluating Streamflow dataset, version-II (GAGES-II) estimates of baseflow index (BFI) (Falcone, 2011). Each dot represents one of the 415 basins in our study data set during the testing time range, and the lines are the best linear regression fit to the scatterplots. R_{BFI} is the Pearson correlation between baseflow estimations which is stated in the title for each plot.

baseflow and GAGES-II BFI were in the northeastern United States and the Michigan Upper Peninsula (Figure 6), where our estimates tended to be higher than those from GAGES-II. In these regions, flows through the glacial deposit layers are dominant (Fang & Shen, 2017; Niu et al., 2014; Shen & Phanikumar, 2010; Shen et al., 2016) and may not follow the air temperature of the recent days, therefore, a significant shallow subsurface component with low temperature could also contribute toward BFI in the GAGES-II estimate. We hypothesize that although LSTM-SMRA17 separates flows into shallow subsurface and groundwater, it lacks a module for modeling snow-melt and its temperature.

3.4. Further Modeling Implications

In this work, we employed a multidimensional assessment to choose models with better physics, and our results indicate that caution is needed when interpreting results from NN-dominant systems where the NN has lots of flexibility. In the future, we can seek additional data and physics to co-constrain the system, for example, applying a differentiable model for the hydrologic component as in Feng et al. (2023) and Feng et al. (2022) and constraining it using daily discharge. That will potentially remove the need for a post-training multidimensional evaluation, but will require us to implement the hydrologic component as physics-based equations (rather than

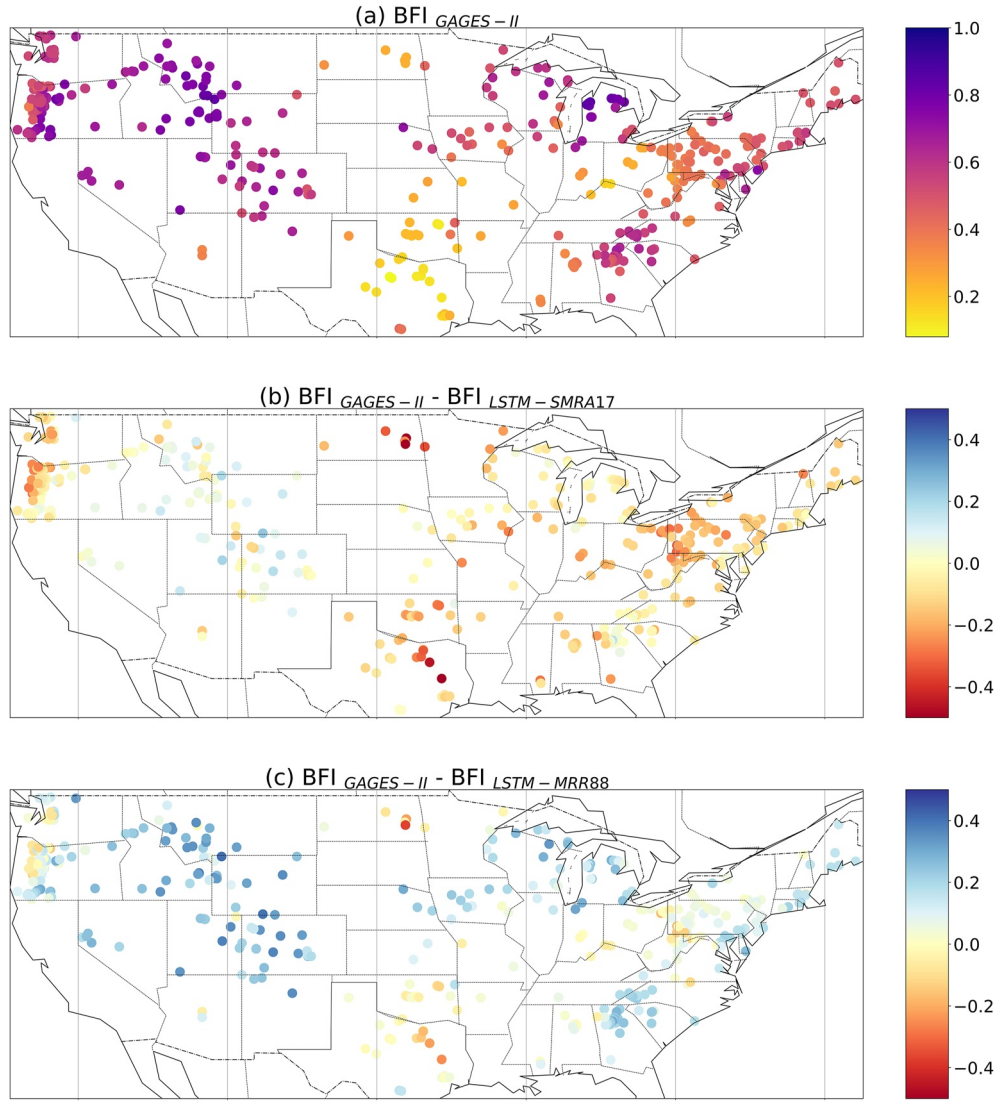

Figure 6. Spatial maps of average baseflow index (BFI) in the 415 basins of our data set for (a) Geological Attributes of Gages for Evaluating Streamflow dataset, version-II (GAGES-II) long-term estimation, as an alternative for comparison from literature, (b) deviation between our LSTM-SMRA17 model and GAGES-II BFI estimations from October 2014 to September 2016, and (c) deviation between our LSTM-MRR88 model and GAGES-II BFI estimations from October 2014 to September 2016. Base map layers are drawn using the database provided by Wessel et al. (2019) and Wessel and Smith (1996).

allowing LSTM to provide all the estimates). Temperature data can be used to provide meaningful constraints on parameterizations of the hydrologic cycle. As we incorporate more and better physics, we expect to maintain strong performance and sharp Pareto fronts as more realistic systems can satisfy all constraints at once. In addition, we will continue to seek data sets that can provide an objective estimate of the source water fractions to further validate the model.

This work demonstrates the advantage of differentiable modeling—the genre of model that allows gradient information to be propagated along the entire chain of calculations and thus support training of NNs anywhere in the model. We recommend that readers refer to Shen et al. (2023) and Tsai et al. (2021) for a more comprehensive discussion. Traditional calibration or response-surface approaches typically work on a site-by-site basis and cannot train a massive amount of weights in complex NNs on large data, and thus typically fall into the problem of non-uniqueness, also called equifinality. In contrast, the size of the NN allowed by differentiable modeling is practically unbounded and is only limited by the amount of training data available. This allowed us to constrain one NN using all the data from 415 basins (and more if data becomes available) and build a more complex but

generalizable relationship, mitigating the issue of equifinality. Unlike with calibration, the parameters obtained are constrained by all sites and in cases of unavailable data, can be reliably used for interpolation or extrapolation of predictions (Feng, Beck, de Bruijn, et al., 2023; Feng, Beck, Lawson, & Shen, 2023). While we need to further constrain these flexible NNs, learning from big data is essential to evolving the complexity and performance of our models—it provides an implicit spatial constraint among all the sites and we see beneficial scaling with data (Feng et al., 2022; Tsai et al., 2021). Differentiable modeling also allowed us to train a dynamical NN that can output daily parameter and flux estimates (based on a history of daily forcings), which were simply not possible before. The full differentiability further means the supervising signal can come indirectly from the loss function related to any observable variables simulated by the combined system, here water temperature measurements, thus removing the need for direct supervising data or pretraining for the NN. This is critical since in practice we do not know the ground truth for these parameters or fluxes. Compared to "physics-guided ML" which focuses on adding constraints to ML algorithms, differentiable modeling also enables the discovery of knowledge (unknown relationships) from data. Lastly but nontrivially, computing on the differentiable platforms using graphical process units is highly parallel and can be orders of magnitude faster than the original models. Once a model is differentiable, we have great latitude in learning and optimization, along with from performance improvements.

4. Conclusions

We employed an NN-dominant hybrid model to simulate not only temperature but also internal dynamics including baseflow, stream evaporation, and other instream processes. We obtained stream water temperature predictions, as well as estimates of a variety of internal states and fluxes, which exhibited substantial variance and must be mitigated by multidimensional assessments. It appears the current system is still under-determined by water temperature alone, but a further selection based on baseflow fraction can narrow down the admissible models. Therefore, we must be cautious with interpreting the internal dynamics of such NN-dominant models. This system may not be the ultimate model for large-scale deployment, but the physical parts of the hybrid model still provided physically meaningful and substantive constraints, and the role of deep networks was limited. For example, the LSTM could not compensate for the lack of instream processes, which also confirms the important role these processes play in stream water temperature. In addition, the different priors led to substantially different behaviors in the Pareto fronts.

The large differences in behaviors of models with different priors mean this hybrid modeling and parameter learning approach can be used to identify better physical assumptions and gain a wealth of process insights. Due to the strong optimization ability of deep networks, we can circumvent parametric issues and focus on the structural priors. Here, the LSTM-SMRA17 prior gave strong results with respect to predicting both water temperature and baseflow (temperature NSE = 0.970 and $R_{\mathrm{BFI}} = 0.724$ for the selected configuration, though a few configurations had R_{BFI} as low as 0.12, or as high as 0.77) compared to the other three priors tested (LSTM-MRR88, LSTM-WVWBB19, and LSTM + mixing). The available data allowed us to identify that (a) a separate bucket for the shallow subsurface improves both temperature and baseflow modeling, since stream temperature data contain useful information regarding hydrologic pathways; (b) groundwater and subsurface flow temperature are better simulated by recency-weighted averaging, rather than a well-mixed reservoir (implied by simple averaging); (c) dynamical parameters that reflect seasonal changes in instream processes such as evapotranspiration and shading are favorable; and (d) in-stream heat processes are necessary which means their impacts on stream temperature are substantial. In each evaluation, a more flexible model was preferred. Considering the reported metrics were from the test period rather than the training period (and thus the models were not substantially overfitted), our results suggest more complicated models that better approximate reality can be trained using big data, which was challenging before differentiable modeling existed. Despite these qualitative insights, as stated earlier, the quantitative estimation of internal fluxes still needs more constraints and more careful evaluation. Involving other hydrologic observations and differentiable modeling (adding a process-based model for the simulation of water source fractions) can jointly inform hydrologic and transport processes to investigate the contribution of stream temperature observations on large-scale hydrological predictions. Our differentiable modeling framework allows us to ask different scientific questions that were impossible using a pure NN, while achieving better accuracy than a purely process-based model.

Conflict of Interest

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. KL and CS have financial interests in HydroSapient, Inc., a company which could potentially benefit from the results of this research. This interest has been reviewed by the University in accordance with its

Individual Conflict of Interest policy, for the purpose of maintaining the objectivity and the integrity of research at The Pennsylvania State University.

Data Availability Statement

The core LSTM and hybrid modeling code, forcing and attribute data set, as well as the results of different experiments relevant to this work are available in the Sciencebase.gov data release by Rahmani et al. (2023), https://doi.org/10.5066/P9UDDHVD. All data used is publicly available from the sources cited. SNTEMP code was from Sanders et al. (2017). Basin information was from the Geological Attributes of Gages for Evaluating Streamflow data set, version-II (GAGES-II) (Falcone, 2011). Mean daily observed streamflow and stream temperature data were downloaded from the U.S. Geological Survey (USGS) National Water Information System (U.S. Geological Survey, 2016). We collected daily meteorological forcings for these basins in the relevant time ranges by extracting the overlaps of the basins' shapefiles and the 1-km grid cells of the Daymet V4 data set (Thornton et al., 2022).

Acknowledgments
FR and CS were supported by US Department of Interior Grant G21AC10563-00. APA was funded by the USGS Water Availability and Use Science Program. DF, CS, and KL were supported by subaward A22-0307-S003 from Cooperative Institute for Research to Operations in Hydrology (CIROH) through the National Oceanic and Atmospheric Administration (NOAA) Cooperative Agreement with the University of Alabama (Grant NA22NWS4320003). Computing was partially supported by U.S. National Science Foundation Award PHY No. 2018280.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). Tensorflow: A system for large-scale machine learning. In *Paper presented at 12th USENIX symposium on operating systems design and implementation (OSDI 16)* (pp. 265–283). USENIX Association. Retrieved from https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

Aboelyazeed, D., Xu, C., Hoffman, F. M., Liu, J., Jones, A. W., Rackauckas, C., et al. (2023). A differentiable, physics-informed ecosystem modeling and learning framework for large-scale inverse problems: Demonstration with photosynthesis simulations. *Biogeosciences*, *20*(13), 2671–2692. https://doi.org/10.5194/bg-20-2671-2023

Amirkhani, M., Bozorg-Haddad, O., Fallah-Mehdipour, E., & Loáiciga, H. A. (2016). Multiobjective reservoir operation for water quality optimization. *Journal of Irrigation and Drainage Engineering*, *142*(12), 04016065. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001105

Anderson, M. P. (2005). Heat as a ground water tracer. *Groundwater*, *43*(6), 951–968. https://doi.org/10.1111/j.1745-6584.2005.00052.x

Appling, A. P., Oliver, S. K., Read, J. S., Sadler, J. M., & Zwart, J. (2022). Machine learning for understanding inland water quantity, quality, and ecology. Retrieved from https://eartharxiv.org/repository/view/3565/

Benyahya, L., Caissie, D., St-Hilaire, A., Ouarda, T. B. M. J., & Bobée, B. (2007). A review of statistical water temperature models. *Canadian Water Resources Journal*, *32*(3), 179–192. https://doi.org/10.4296/cwrj3203179

Bezanson, J., Karpinski, S., Shah, V. B., & Edelman, A. (2012). Julia: A fast dynamic language for technical computing. ArXiv, abs/1209.5145.

Bindas, T., Tsai, W.-P., Liu, J., Rahmani, F., Feng, D., Bian, Y., et al. (2023). Improving large-basin river routing using a differentiable Muskingum-Cunge model and physics-informed machine learning. *ESS Open Archive*. https://doi.org/10.22541/essoar.168500246.67971832/v1

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., et al. (2021). JAX: Autograd and XLA. *Astrophysics Source Code Library*. ascl:2111.002.

Briggs, M. A., Johnson, Z. C., Snyder, C. D., Hitt, N. P., Kurylyk, B. L., Lautz, L., et al. (2018). Inferring watershed hydraulics and cold-water habitat persistence using multi-year air and stream temperature signals. *Science of the Total Environment*, *636*, 1117–1127. https://doi.org/10.1016/j.scitotenv.2018.04.344

Broxton, P. D., van Leeuwen, W. J. D., & Biederman, J. A. (2019). Improving snow water equivalent maps with machine learning of snow survey and lidar measurements. *Water Resources Research*, *55*(5), 3739–3757. https://doi.org/10.1029/2018WR024146

Bundschuh, J. (1993). Modeling annual variations of spring and groundwater temperatures associated with shallow aquifer systems. *Journal of Hydrology*, *142*(1), 427–444. https://doi.org/10.1016/0022-1694(93)90022-2

Burns, E. R., Ingebritsen, S. E., Manga, M., & Williams, C. F. (2016). Evaluating geothermal and hydrogeologic controls on regional groundwater temperature distribution. *Water Resources Research*, *52*(2), 1328–1344. https://doi.org/10.1002/2015WR018204

Chapin, T. P., Todd, A. S., & Zeigler, M. P. (2014). Robust, low-cost data loggers for stream temperature, flow intermittency, and relative conductivity monitoring. *Water Resources Research*, *50*(8), 6542–6548. https://doi.org/10.1002/2013WR015158

Detenbeck, N. E., Morrison, A. C., Abele, R. W., & Kopp, D. A. (2016). Spatial statistical network models for stream and river temperature in New England, USA. *Water Resources Research*, *52*(8), 6018–6040. https://doi.org/10.1002/2015wr018349

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B*, *57*(1), 45–70. https://doi.org/10.1111/j.2517-6161.1995.tb02015.x

Du, X., Goss, G., & Faramarzi, M. (2020). Impacts of hydrological processes on stream temperature in a cold region watershed based on the SWAT equilibrium temperature model. *Water*, *12*(4), 1112. https://doi.org/10.3390/w12041112

Du, X., Silwal, G., & Faramarzi, M. (2022). Investigating the impacts of glacier melt on stream temperature in a cold-region watershed: Coupling a glacier melt model with a hydrological model. *Journal of Hydrology*, *605*, 127303. https://doi.org/10.1016/j.jhydrol.2021.127303

Ducharne, A. (2008). Importance of stream temperature to climate change impact on water quality. *Hydrology and Earth System Sciences*, *12*(3), 797–810. https://doi.org/10.5194/hess-12-797-2008

Dugdale, S. J., Hannah, D. M., & Malcolm, I. A. (2017). River temperature modelling: A review of process-based approaches and future directions. *Earth-Science Reviews*, *175*, 97–113. https://doi.org/10.1016/j.earscirev.2017.10.009

Edinger, J. E., Duttweiler, D. W., & Geyer, J. C. (1968). The response of water temperatures to meteorological conditions. *Water Resources Research*, *4*(5), 1137–1143. https://doi.org/10.1029/WR004i005p01137

Falcone, J. A. (2011). GAGES-II: Geospatial attributes of gages for evaluating streamflow [Dataset]. USGS Unnumbered Series. https://doi.org/10.3133/70046617

Fang, K., Pan, M., & Shen, C. (2019). The value of SMAP for long-term soil moisture estimation with the help of deep learning. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(4), 2221–2233. https://doi.org/10.1109/TGRS.2018.2872131

Fang, K., & Shen, C. (2017). Full-flow-regime storage-streamflow correlation patterns provide insights into hydrologic functioning over the continental US. *Water Resources Research*, *53*(9), 8064–8083. https://doi.org/10.1002/2016WR020283

Fang, K., Shen, C., Kifer, D., & Yang, X. (2017). Prolongation of SMAP to spatiotemporally seamless coverage of continental U.S. using a deep learning neural network. *Geophysical Research Letters*, *44*(21), 11030–11039. https://doi.org/10.1002/2017gl075619

Feigl, M., Lebiedzinski, K., Herrnegger, M., & Schulz, K. (2021). Machine-learning methods for stream water temperature prediction. *Hydrology and Earth System Sciences*, *25*(5), 2951–2977. https://doi.org/10.5194/hess-25-2951-2021

Fellman, J. B., Nagorski, S., Pyare, S., Vermilyea, A. W., Scott, D., & Hood, E. (2014). Stream temperature response to variable glacier coverage in coastal watersheds of Southeast Alaska. *Hydrological Processes*, *28*(4), 2062–2073. https://doi.org/10.1002/hyp.9742

Feng, D., Beck, H., de Bruijn, J., Sahu, R. K., Satoh, Y., Wada, Y., et al. (2023). Deep dive into global hydrologic simulations: Harnessing the power of deep learning and physics-informed differentiable models (δHBV-globe1.0-hydroDL). *Geoscientific Model Development Discussions*. https://doi.org/10.5194/gmd-2023-190

Feng, D., Beck, H., Lawson, K., & Shen, C. (2023). The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment. *Hydrology and Earth System Sciences*, *27*(12), 2357–2373. https://doi.org/10.5194/hess-27-2357-2023

Feng, D., Fang, K., & Shen, C. (2020). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resources Research*, *56*(9), e2019WR026793. https://doi.org/10.1029/2019WR026793

Feng, D., Lawson, K., & Shen, C. (2021). Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data. *Geophysical Research Letters*, *48*(14), e2021GL092999. https://doi.org/10.1029/2021GL092999

Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. *Water Resources Research*, *58*(10), e2022WR032404. https://doi.org/10.1029/2022WR032404

Förster, H., & Lilliestam, J. (2010). Modeling thermoelectric power generation in view of climate change. *Regional Environmental Change*, *10*(4), 327–338. https://doi.org/10.1007/s10113-009-0104-x

Freeze, R. A., & Cherry, J. A. (1979). *Groundwater* (1st ed.). Prentice Hall.

Graf, R., Zhu, S., & Sivakumar, B. (2019). Forecasting river water temperature time series using a wavelet–neural network hybrid modelling approach. *Journal of Hydrology*, *578*, 124115. https://doi.org/10.1016/j.jhydrol.2019.124115

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hare, D. K., Helton, A. M., Johnson, Z. C., Lane, J. W., & Briggs, M. A. (2021). Continental-scale analysis of shallow and deep groundwater contributions to streams. *Nature Communications*, *12*(1), 1450. https://doi.org/10.1038/s41467-021-21651-0

Hazewinkel, M. (2001). Taylor series. *Encyclopedia of Mathematics*.

Heddam, S., Kim, S., Danandeh Mehr, A., Zounemat-Kermani, M., Malik, A., Elbeltagi, A., & Kisi, O. (2022). Chapter 1—Predicting dissolved oxygen concentration in river using new advanced machines learning: Long-short term memory (LSTM) deep learning. In H. R. Pourghasemi (Ed.), *Computers in Earth and Environmental Sciences* (pp. 1–20). Elsevier. https://doi.org/10.1016/B978-0-323-89861-4.00031-2

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Khoshkalam, Y., Rousseau, A. N., Rahmani, F., Shen, C., & Abbasnezhadi, K. (2023). Applying transfer learning techniques to enhance the accuracy of streamflow prediction produced by long Short-term memory networks with data integration. *Journal of Hydrology*, *622*, 129682. https://doi.org/10.1016/j.jhydrol.2023.129682

Kraft, B., Jung, M., Körner, M., Koirala, S., & Reichstein, M. (2022). Towards hybrid modeling of the global hydrological cycle. *Hydrology and Earth System Sciences*, *26*(6), 1579–1614. https://doi.org/10.5194/hess-26-1579-2022

Liu, J., Rahmani, F., Lawson, K., & Shen, C. (2022). A multiscale deep learning model for soil moisture integrating satellite and in situ data. *Geophysical Research Letters*, *49*(7), e2021GL096847. https://doi.org/10.1029/2021GL096847

Liu, Y., Caballero, R., & Monteiro, J. M. (2020). RadNet 1.0: Exploring deep learning architectures for longwave radiative transfer. *Geoscientific Model Development*, *13*(9), 4399–4412. https://doi.org/10.5194/gmd-13-4399-2020

Loucks, D. P., & van Beek, E. (2017). *Water resource systems planning and management: An introduction to methods, models, and applications*. Springer.

Madden, N., Lewis, A., & Davis, M. (2013). Thermal effluent from the power sector: An analysis of once-through cooling system impacts on surface water temperature. *Environmental Research Letters*, *8*(3), 035006. https://doi.org/10.1088/1748-9326/8/3/035006

Marcogliese, D. J. (2001). Implications of climate change for parasitism of animals in the aquatic environment. *Canadian Journal of Zoology*, *79*(8), 1331–1352. https://doi.org/10.1139/z01-067

Markstrom, S. L. (2012). P2S—Coupled simulation with the Precipitation-Runoff Modeling System (PRMS) and the Stream Temperature Network (SNTemp) Models (No. 2012–1116). *Open-File Report—U. S. Geological Survey*. https://doi.org/10.3133/ofr20121116

Martins, E. G., Hinch, S. G., Patterson, D. A., Hague, M. J., Cooke, S. J., Miller, K. M., et al. (2012). High river temperature reduces survival of sockeye salmon (*Oncorhynchus nerka*) approaching spawning grounds and exacerbates female mortality. *Canadian Journal of Fisheries and Aquatic Sciences*, *69*(2), 330–342. https://doi.org/10.1139/f2011-154

Maxwell, R. M., Condon, L. E., Kollet, S. J., Maher, K., Haggerty, R., & Forrester, M. M. (2016). The imprint of climate and geology on the residence times of groundwater. *Geophysical Research Letters*, *43*(2), 701–708. https://doi.org/10.1002/2015GL066916

Meisner, J. D., Rosenfeld, J. S., & Regier, H. A. (1988). The role of groundwater in the impact of climate warming on stream salmonines. *Fisheries*, *13*(3), 2–8. https://doi.org/10.1577/1548-8446(1988)013<0002:TROGIT>2.0.CO;2

Meyal, A. Y., Versteeg, R., Alper, E., Johnson, D., Rodzianko, A., Franklin, M., & Wainwright, H. (2020). Automated cloud based long short-term memory neural network based SWE prediction. *Frontiers in Water*, *2*. https://doi.org/10.3389/frwa.2020.574917

Michel, A., Brauchli, T., Lehning, M., Schaefli, B., & Huwald, H. (2020). Stream temperature and discharge evolution in Switzerland over the last 50 years: Annual and seasonal behaviour. *Hydrology and Earth System Sciences*, *24*(1), 115–142. https://doi.org/10.5194/hess-24-115-2020

Mohseni, O., & Stefan, H. G. (1999). Stream temperature/air temperature relationship: A physical interpretation. *Journal of Hydrology*, *218*(3), 128–141. https://doi.org/10.1016/S0022-1694(99)00034-7

Morrill, J. C., Bales, R. C., & Conklin, M. H. (2005). Estimating stream temperature from air temperature: Implications for future water quality. *Journal of Environmental Engineering*, *131*(1), 139–146. https://doi.org/10.1061/(ASCE)0733-9372(2005)131:1(139)

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Niu, J., Shen, C., Li, S.-G., & Phanikumar, M. S. (2014). Quantifying storage changes in regional Great Lakes watersheds using a coupled subsurface-land surface process model and GRACE, MODIS products. *Water Resources Research*, *50*(9), 7359–7377. https://doi.org/10.1002/2014WR015589

O, S., & Orth, R. (2021). Global soil moisture data derived through machine learning trained with in-situ measurements. *Scientific Data*, *8*(1), 170. https://doi.org/10.1038/s41597-021-00964-1

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). Automatic differentiation in PyTorch. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Retrieved from https://openreview.net/forum?id=BJJsrmfCZ

Pekárová, P., Tall, A., Pekár, J., Vitková, J., & Miklánek, P. (2022). Groundwater temperature modelling at the water table with a simple heat conduction model. *Hydrology*, *9*(10), 185. https://doi.org/10.3390/hydrology9100185

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., & Shen, C. (2021). Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. *Environmental Research Letters*, *16*(2), 024025. https://doi.org/10.1088/1748-9326/abd501

Rahmani, F., Shen, C., Lawson, K., Feng, D., & Appling, A. (2023). Identifying structural priors in a hybrid differentiable model for stream water temperature modeling at 415 U.S. basin outlets, 2010–2016 [Dataset]. U.S. Geological Survey Data Release. https://doi.org/10.5066/P9UDDHVD

Rahmani, F., Shen, C., Oliver, S., Lawson, K., & Appling, A. (2021). Deep learning approaches for improving prediction of daily stream temperature in data-scarce, unmonitored, and dammed basins. *Hydrological Processes*, *35*(11), e14400. https://doi.org/10.1002/hyp.14400

Regan, R. S., & Markstrom, S. L. (2021). *Precipitation-Runoff Modeling System (PRMS) (Version 5.2.0)*. U.S. Geological Survey Software Release.

Rehana, S., & Rajesh, M. (2023). Assessment of impacts of climate change on Indian riverine thermal regimes using hybrid deep learning methods. *Water Resources Research*, *59*(2), e2021WR031347. https://doi.org/10.1029/2021WR031347

Sadler, J. M., Appling, A. P., Read, J. S., Oliver, S. K., Jia, X., Zwart, J. A., & Kumar, V. (2022). Multi-task deep learning of daily streamflow and water temperature. *Water Resources Research*, *58*(4), e2021WR030138. https://doi.org/10.1029/2021WR030138

Saha, G. K., Rahmani, F., Shen, C., Li, L., & Cibin, R. (2023). A deep learning-based novel approach to generate continuous daily stream nitrate concentration for nitrate data-sparse watersheds. *Science of the Total Environment*, *878*, 162930. https://doi.org/10.1016/j.scitotenv.2023.162930

Samarinas, N., Tziolas, N., & Zalidis, G. (2020). Improved estimations of nitrate and sediment concentrations based on SWAT simulations and annual updated land cover products from a deep learning classification algorithm. *ISPRS International Journal of Geo-Information*, *9*(10), 576. https://doi.org/10.3390/ijgi9100576

Sanders, M. J., Markstrom, S. L., Regan, R. S., & Atkinson, R. D. (2017). *Documentation of a daily mean stream temperature module—An enhancement to the Precipitation-Runoff Modeling System* (Vol. 6-D4, p. 28). U.S. Geological Survey. https://doi.org/10.3133/tm6D4

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*(11), 8558–8593. https://doi.org/10.1029/2018wr022643

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., et al. (2023). Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth & Environment*, *4*(8), 552–567. https://doi.org/10.1038/s43017-023-00450-9

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., et al. (2018). HESS opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, *22*(11), 5639–5656. https://doi.org/10.5194/hess-22-5639-2018

Shen, C., & Phanikumar, M. S. (2010). A process-based, distributed hydrologic model based on a large-scale method for surface—Subsurface coupling. *Advances in Water Resources*, *33*(12), 1524–1541. https://doi.org/10.1016/j.advwatres.2010.09.002

Shen, C., Riley, W. J., Smithgall, K. M., Melack, J. M., & Fang, K. (2016). The fan of influence of streams and channel feedbacks to simulated land surface water and carbon dynamics. *Water Resources Research*, *52*(2), 880–902. https://doi.org/10.1002/2015WR018086

Siegel, J. E., Fullerton, A. H., & Jordan, C. E. (2022). Accounting for snowpack and time-varying lags in statistical models of stream temperature. *Journal of Hydrology X*, *17*, 100136. https://doi.org/10.1016/j.hydroa.2022.100136

Sohrabi, M. M., Benjankar, R., Tonina, D., Wenger, S. J., & Isaak, D. J. (2017). Estimation of daily stream water temperatures with a Bayesian regression approach. *Hydrological Processes*, *31*(9), 1719–1733. https://doi.org/10.1002/hyp.11139

Theurer, F. D., Voos, K. A., & Miller, W. J. (1984). *Instream water temperature model*. Western Energy and Land Use Team, Division of Biological Services, Research and Development, Fish and Wildlife Service, U.S. Department of the Interior.

Thornton, M. M., Shrestha, R., Wei, Y., Thornton, P. E., Kao, S.-C., & Wilson, B. E. (2022). Daymet: Daily surface weather data on a 1-km grid for North America, version 4 R1 (version 4.4) [NetCDF]. *ORNL Distributed Active Archive Center*. https://doi.org/10.3334/ORNLDAAC/2129

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., et al. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, *12*(1), 5988. https://doi.org/10.1038/s41467-021-26107-z

U.S. Geological Survey. (2016). USGS water data for the Nation [Dataset]. U.S. Geological Survey National Water Information System database. https://doi.org/10.5066/F7P55KJN

van Vliet, M. T. H., Sheffield, J., Wiberg, D., & Wood, E. F. (2016). Impacts of recent drought and warm years on water resources and electricity supply worldwide. *Environmental Research Letters*, *11*(12), 124021. https://doi.org/10.1088/1748-9326/11/12/124021

van Vliet, M. T. H., Vögele, S., & Rübbelke, D. (2013). Water constraints on European power supply under climate change: Impacts on electricity prices. *Environmental Research Letters*, *8*(3), 035010. https://doi.org/10.1088/1748-9326/8/3/035010

Virkki, V., Alanärä, E., Porkka, M., Ahopelto, L., Gleeson, T., Mohan, C., et al. (2022). Globally widespread and increasing violations of environmental flow envelopes. *Hydrology and Earth System Sciences*, *26*(12), 3315–3336. https://doi.org/10.5194/hess-26-3315-2022

Wanders, N., Vliet, M. T. H., van Wada, Y., Bierkens, M. F. P., & Rens van Beek, L. P. H. (2019). High-resolution global water temperature modeling. *Water Resources Research*, *55*(4), 2760–2778. https://doi.org/10.1029/2018WR023250

Wang, X.-S., Wan, L., Jiang, X.-W., Li, H., Zhou, Y., Wang, J., & Ji, X. (2017). Identifying three-dimensional nested groundwater flow systems in a Tóthian basin. *Advances in Water Resources*, *108*, 139–156. https://doi.org/10.1016/j.advwatres.2017.07.016

Webb, B., & Walling, D. (1995). The long-term thermal impact of reservoir operation and some ecological implications. *IAHS Publications*, 230.

Weierbach, H., Lima, A. R., Willard, J. D., Hendrix, V. C., Christianson, D. S., Lubich, M., & Varadharajan, C. (2022). Stream temperature predictions for river basin management in the Pacific Northwest and Mid-Atlantic regions using machine learning. *Water*, *14*(7), 1032. https://doi.org/10.3390/w14071032

Wessel, P., Luis, J. F., Uieda, L., Scharroo, R., Wobbe, F., Smith, W. H. F., & Tian, D. (2019). The generic mapping tools version 6. *Geochemistry, Geophysics, Geosystems*, *20*(11), 5556–5564. https://doi.org/10.1029/2019GC008515

Wessel, P., & Smith, W. H. F. (1996). A global, self-consistent, hierarchical, high-resolution shoreline database. *Journal of Geophysical Research: Solid Earth*, *101*(B4), 8741–8743. https://doi.org/10.1029/96JB00104

Wolock, D. M. (2003). *Base-flow index grid for the conterminous United States (USGS Numbered Series No. 2003–263)*. U.S. Geological Survey. https://doi.org/10.3133/ofr03263

Xiang, Z., Yan, J., & Demir, I. (2020). A rainfall-runoff model with LSTM-based sequence-to-sequence learning. *Water Resources Research*, *56*(1), e2019WR025326. https://doi.org/10.1029/2019WR025326

Yasukawa, K., Uchida, Y., Tenma, N., Taguchi, Y., Muraoka, H., Ishii, T., et al. (2009). Groundwater temperature survey for geothermal heat pump application in tropical Asia. *Bulletin of the Geological Survey of Japan*, *60*(9–10), 459–467. https://doi.org/10.9795/bullgsj.60.459

Zeiler, M. D. (2012). ADADELTA: An adaptive learning rate method. *ArXiv Preprint ArXiv:1212.5701* (pp. 1–6). Retrieved from http://arxiv.org/abs/1212.5701

Zhang, J., Wang, X.-S., Yin, L., Wang, W., Love, A., Lu, Z.-T., et al. (2021). Inflection points on groundwater age and geochemical profiles along wellbores light up hierarchically nested flow systems. *Geophysical Research Letters*, *48*(16), e2020GL092337. https://doi.org/10.1029/2020GL092337

Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., & Li, L. (2021). From hydrometeorology to river water quality: Can a deep learning model predict dissolved oxygen at the continental scale? *Environmental Science & Technology*, *55*(4), 2357–2368. https://doi.org/10.1021/acs.est.0c06783

Zhi, W., Ouyang, W., Shen, C., & Li, L. (2023). Temperature outweighs light and flow as the predominant driver of dissolved oxygen in US rivers. *Nature Water*, *1*(3), 249–260. https://doi.org/10.1038/s44221-023-00038-z

Zhu, F., Li, X., Qin, J., Yang, K., Cuo, L., Tang, W., & Shen, C. (2021). Integration of multisource data to estimate downward longwave radiation based on deep neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–15. https://doi.org/10.1109/TGRS.2021.3094321

Zhu, S., & Piotrowski, A. P. (2020). River/stream water temperature forecasting using artificial intelligence models: A systematic review. *Acta Geophysica*, *68*(5), 1433–1442. https://doi.org/10.1007/s11600-020-00480-7

Zijl, W. (1999). Scale aspects of groundwater flow and transport systems. *Hydrogeology Journal*, *7*(1), 139–150. https://doi.org/10.1007/s100400050185

Zwart, J. A., Oliver, S. K., Watkins, W. D., Sadler, J. M., Appling, A. P., Corson-Dosch, H. R., et al. (2023). Near-term forecasts of stream temperature using deep learning and data assimilation in support of management decisions. *JAWRA Journal of the American Water Resources Association*, *59*(2), 317–337. https://doi.org/10.1111/1752-1688.13093

References From the Supporting Information

Gorelick, N., Hancher, M., Dixon, M., Simon, I., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment, Big Remotely Sensed Data: Tools, Applications and Experiences*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

McCabe, G. J., & Wolock, D. M. (2009). Recent declines in western U.S. Snowpack in the context of twentieth-century climate variability. *Earth Interactions*, *13*(12), 1–15. https://doi.org/10.1175/2009EI283.1

Reed, J., & Bush, C. (2005). *Generalized Geologic Map of the Conterminous United State*. United States Geologic Survey.