The Utility of Information Flow in Formulating Discharge Forecast Models: A Case Study From an Arid Snow-Dominated Catchment

Christopher Tennant1, Laurel Larsen1, Dino Bellugi1, Edom Moges1, Liang Zhang1, and Hongxu Ma1

1Department of Geography, University of California at Berkeley, Berkeley, CA, USA

Abstract Streamflow forecasts often perform poorly because of improper representation of hydrologic response timescales in underlying models. Here, we use transfer entropy (TE), which measures information flow between variables, to identify dominant drivers of discharge and their timescales using sensor data from the Dry Creek Experimental Watershed, ID, USA. Consistent with previous mechanistic studies, TE revealed that snowpack accumulation and partitioning into melt, recharge, and evaporative loss dominated discharge patterns and that snow-sourced baseflow reduced the greatest amount of uncertainty in discharge. We hypothesized that machine learning models (MLMs) specified in accordance with the dominant lag timescales, identified via TE, would outperform timescale-agnostic models. However, while lagged-variable random forest regressions captured the dominant process—seasonal snowmelt—they ultimately did not perform as well as the unlagged models, provided those models were specified with input data aggregated over a range of timescales. Unlagged models, not constrained by timescales of the dominant processes, more effectively represented variable interactions (e.g., rain-on-snow events) playing a critical role in translating precipitation into streamflow over long, intermediate, and short timescales. Meanwhile, long short-term memory (LSTM) models were effective in internally identifying the key lag and aggregation scales for predicting discharge. Parsimonious specification of LSTM models, using only daily unlagged precipitation and temperature data, produced the highest performing predictions. Our findings suggest that TE can identify dominant streamflow controls and the relative importance of different mechanisms of streamflow generation, useful for establishing process baselines and fingerprinting watersheds. However, restricting MLMs based on dominant timescales undercuts their skill at learning these timescales internally.

Plain Language Summary Hydrologic forecast models play an important role in predicting flooding or water scarcity. However, uncertainty in model inputs such as the spatial distribution of precipitation or the amount of underground water storage all reduce the accuracy of discharge forecasts. One way to simplify the complex processes that affect riverflows is to quantify “hydrologic response times” and ensure these timescales are represented in hydrologic models. Here, we used techniques from information theory to identify important drivers of riverflows and quantify response times between water inputs and riverflows. Based on the dominant timescales identified, we formulated “informed” models to forecast riverflows but found that the informed models outperformed baseline models only for a limited set of conditions. Rather, specialized machine learning models that “learn” a spectrum of response times between predictors and riverflow, representing timescales of both dominant and subordinate processes affecting riverflow, achieved the best predictive performance. Moreover, the models performed well using only precipitation and temperature data—both widely available—as inputs. This finding suggested that these models could perform well for forecasting streamflow in both gauged and ungauged basins.

1. Introduction

1.1. “Critical Timescales” for Hydrologic Modeling

Forecasting riverflows is a cornerstone of operational hydrology and an important research topic due to the substantial impacts of flooding on life and the economy. Forecasting is also a critical aspect of water resources planning such as predicting summer low flows (Godsey et al., 2014), anticipating drought conditions (Kapnick et al., 2018), or managing reservoir systems (Rhoades et al., 2018). A salient challenge that
exists for all discharge forecasting approaches is the identification and representation of the dominant processes and timescales controlling discharge response. In this work, we examined the use of information metrics to quantify critical timescales between hydrologic inputs and discharge response and evaluated if hydrologic forecasts based on information flow analyses (i.e., transfer entropy [TE]) could outperform “uninformed” forecasts.

Hydrologic forecasts can be generated from models that are physically based or from data-driven/regression approaches. Physically based models are grounded in first principles and attempt to describe the physics of water transport throughout the atmosphere, biosphere, bedrock, soil, and river channels (e.g., Beven, 1989; Kollet & Maxwell, 2006, 2008). Both the parameterization of catchment properties and representation of hydrologic processes span a range of complexity in physically based models. Lumped models and tipping bucket representations exemplify the simpler end of the spectrum (see models reviewed in Seiller et al., 2012), and spatially distributed parameters with detailed flow routing represent the complex end (e.g., Khakbaz et al., 2012; Moreda et al., 2006; Vieux, 2001). This representational complexity participates in a trade-off with model computational and developmental tractability, performance for an individual catchment, and generalizability to a range of catchments (Larsen et al., 2016). Presently, agency-sponsored efforts to develop continental-extent models for producing hydrologic forecasts at the catchment scale are grounded in physical modeling approaches with high representational complexity (Hipsey et al., 2015; Newman et al., 2015; Xia et al., 2012).

Another approach to hydrologic forecasting with a long history is regression-based models (Kratzert et al., 2018; Mulvaney, 1851). Data-driven, regression, or more recently, machine learning strategies (also referred to as artificial intelligence or deep-learning), have been shown to perform as well, if not better, than process-based models (de Vos & Rientjes, 2005; Nearing et al., 2016; Shrestha & Nestmann, 2009; Todini, 2007). While data-driven models can achieve high performance, they can be quite sensitive to the data set used for training, and performance often suffers if there is a shift in system state or if the model encounters events not contained within the training set. Furthermore, data-driven models can lose sight of first principles and lack physical interoperability, though Young and Beven (1994) derived data-based models that replicate the highly nonlinear patterns of rainfall-runoff processes and demonstrated that these models can be directly interpreted in the context of dynamic conservation equations (i.e., mass balance).

Because of their ability to replicate events in complex, highly nonlinear systems, neural networks have also been used to generate hydrologic forecasts. Early applications were focused on the use of feed-forward artificial neural networks (ANNs) to inform/predict runoff processes (e.g., Daniell, 1991; Halff et al., 1993 and discussed in Kratzert et al., 2018) and were added to with the use of recurrent neural networks (RNNs) which can be quite effective at capturing temporal dynamics in time series (e.g., Kumar et al., 2004). More recently, Kratzert et al. (2018) demonstrated how long short-term memory (LSTM) models, a specialized type of RNN, can be used to generalize hydrologic forecasts and make streamflow predictions across a range of catchments. Because LSTMs include input, output, and memory gates which dynamically control the flow and retention of time series information, they are particularly adept at capturing longer term hydrologic dependencies and variations in flow path connectivity which influence discharge response across a range of timescales (Benettin et al., 2015; Godsey et al., 2014; Nippgen et al., 2016).

Regardless of whether a physically based or data-driven approach is taken, forecasting models need to capture the temporal dynamics of hydrologic inputs, stores, and runoff. Hydrologists generally recognize several types of timescales as critically important for discharge forecasting. Transit time lags represent the time parcels of water spend in storage (e.g., snowpack, in soils, or groundwater) before reaching the catchment outlet, whereas hydraulic response timescales, governed by the celerity of the pressure wave, represent the time between a precipitation signal and a discharge response (McDonnell & Beven, 2014). Transit time lags are orders of magnitude larger than hydraulic response timescales (Beven, 1982). For example, McGuire and McDonnell (2006) reported transit time lags ranging from less than 1–5 years, though these lags may be highly variable within single catchments (Benettin et al., 2015; Godsey et al., 2014; Nippgen et al., 2016).

An inability to accurately represent transit times may be why estimates of catchment storage derived from continental-extent models exhibit substantial mismatches to satellite-derived estimates of storage (Scanlon et al., 2018). Further, Dawson and Wilby (2001) found that determination of interaction time lags between predictors and target variables (e.g., hydraulic response timescales) is one of the critical factors governing
performance of data-driven models, while Yu et al. (2006) showed improved performance in a support vector machine model by considering timescales that are associated with hydrological response times. These studies highlight the need for a systematic guide in identifying interaction timescales in model development.

Analysis techniques, in particular TE (Schreiber, 2000), recently introduced to the earth sciences (e.g., Goodwell & Kumar, 2017; Goodwell et al., 2018; Ruddell & Kumar, 2009a, 2009b) with roots in information theory could help identify the dominant processes and critical timescales over which hydrologic inputs and stores are translated into runoff (Benettin et al., 2015; Gibson et al., 2002; McGuire & McDonnell, 2006; McNamara et al., 2011; Skøien et al., 2003). TE is an information statistic that provides a directional, non-symmetric measure of information transfer from a source variable $X$, lagged in time by an amount $\tau$, to a target variable $Y$ at time $t$, where $Y$'s conditional dependence on its own history is removed. Mathematically, information transfer is defined as a reduction in uncertainty (i.e., entropy) of the target variable given knowledge of the source, formulated through joint and conditional probability distributions (Schreiber, 2000, as modified by Larsen & Harvey, 2017):

$$T_{x_t \rightarrow y_t} = \sum_{l, k, r} p(y_t, y_{t-r}, x_{t-r}) \log \left( \frac{p(y_t, y_{t-r}, x_{t-r})}{p(y_t, y_{t-r})} \right),$$

where $y_t$ and $x_{t-r}$ are the realizations of $Y$ and $X$ at times $t$ and $t - r$, respectively, and $l$ and $k$ are the block lengths used in the delineation of the probability distributions, typically set to one due to prohibitive data requirements for longer block lengths (Ruddell & Kumar, 2009a, 2009b). TE helps identify which variables provide independent information regarding the future values of $Y$. In practice, Equation 1 is evaluated over a range of lags ($\tau$) to evaluate which timescales are associated with the greatest amounts of information transfer. Here, we postulate that these timescales are consistent with hydraulic response timescales that characterize lags between input hydrologic signals (e.g., precipitation) and discharge response.

1.2. Identification of Dominant Hydrological Processes With TE

As an example of how a TE analysis can provide insights into dominant hydrological processes, consider the relationship between a daily time series of snow-water equivalent (SWE) and discharge in a generic snow-dominated catchment (Figure 1a). In this catchment, knowledge of SWE in the middle of winter would reduce uncertainty in the value(s) of discharge during the spring snowmelt; high SWE at the height of winter would likely yield high discharge in spring, whereas low SWE might indicate a snow drought and low spring runoff. Meanwhile, because this hypothetical catchment is snow dominated, discharge is low...
during the rest of the year, such that low or zero SWE in summer and early fall implies that discharge will also be low in fall and winter respectively, again reducing uncertainty in discharge with a one-season lag. Thus, the TE from daily SWE, measured at a spatially representative location, to daily discharge would peak at a one-season time lag (see SWE annotation in Figure 1b). However, the peak would be broad, given spatial heterogeneity in SWE and in the initiation and progression of melt.

Other monitored variables may reflect snowmelt as a dominant process slightly differently in the TE analysis. Whereas SWE is integrative of the balance between past snowfall and evaporation or sublimation within the season, other variables such as snowmelt reflect instantaneous contributions that are highly variable within one season (Figure 1a). Hence, a single day's melt rate during winter may not greatly reduce uncertainty in the peak discharge in spring; small or large snowpacks might melt rapidly or slowly depending on energy inputs. However, the cumulative snowmelt aggregated over winter (Figure 1a) would substantially reduce uncertainty in peak spring discharge, resulting in high TE from multiweek aggregated melt time series to daily discharge at a shorter lag than that between daily SWE and daily discharge (see “snowmelt” annotation in Figure 1b). Deep moisture on north-facing slopes often exhibits a strong connection to melt (Hinckley et al., 2014) but is less variable than melt rate and would likely reduce uncertainty in discharge over lag timescales that are similar to snowmelt but possibly shorter aggregation scales during times of high catchment connectivity. Other variables that reflect total evaporation or sublimation from the snowpack, such as wind speed or air temperature aggregated over multiple weeks, would likewise reduce uncertainty in peak spring discharge and have high TE in a similar region to snowmelt (Figure 1b).

Additionally, variables such as air temperature that are important controls on evapotranspiration (ET) might in fact have high TE to discharge over multiple data aggregation timescales, as ET serves as an important control on streamflow generation mechanisms over multiple timescales. While multiweek aggregated (i.e., averaged) air temperature may be indicative of total evaporation from snow storage, multimonth aggregated air temperature—together with multimonth precipitation inputs—may indicate water inputs that result in a change in groundwater storage and baseflow inputs (yellow region in Figure 1b). In the baseflow season, for example, today’s or tomorrow’s streamflow is likely low if last year was exceedingly hot, resulting in high ET. Mathematically, this conditional statement can be represented as a high TE from air temperature to discharge at a relatively short lag. Other variables such as soil moisture (Figure 1a) that may be directly or indirectly indicative of multiple processes may also have high TE in multiple combinations of lag and aggregation timescales.

In this manner, Figure 1b presents a conceptual model, or fingerprint, of how dominant hydrological processes, as represented in different sensor time series, may be reflected in peak TE values over multiple lag and source-variable aggregation timescales. It is to be interpreted as a composite across multiple monitored variables, with individual variables exhibiting peak TE only in the region(s) representative of hydrologic processes that are sensitive to that variable. Across all variables, consistently weak TE in one particular region (e.g., quickflow/rapid response) of an empirical fingerprint would suggest that the process to which that region of the fingerprint is sensitive is a subordinate mechanism of streamflow generation in that catchment.

### 1.3. Study Objectives and Hypothesis

In this work, we examined the potential for critical timescale detection through TE analyses to improve hydrologic forecasts in a catchment that has been a focus of detailed hydrological process studies. Publicly available time series from the Dry Creek Experimental Watershed (DCEW), Idaho were used to evaluate information flows (TE) from meteorological variables to discharge and to infer the dominant watershed controls on discharge. Results from this analysis were used to determine model specifications for a number of machine learning approaches and to evaluate if forecasts using timescales selected through information analyses outperformed models that lacked these constraints. We hypothesized that models specified using TE would outperform models lacking information-based specification as they would better represent the dominant drivers and timescales controlling catchment discharge.

### 2. Site Description and Data

The DCEW is in southwestern Idaho just north of the city of Boise and within the northern portion of the Snake River Basin (Figure 2). The DCEW drains 27 km² of semiarid mountainous terrain that ranges...
from ~1,000 m up to 2,100 m in elevation. Most of the precipitation occurs during winter months, with snow being dominant at higher elevations and rain at lower elevations; approximately 54% of the basin’s annual average precipitation of 57 cm is in the form of snow (Stratton et al., 2009; Williams et al., 2009). Snow can persist at high elevations from November through March, though interannual variability is high, and in warm years, precipitation can fall dominantly as rain (Tyler et al., 2008). Rain-on-snow events are common in late autumn and early spring, while summers are hot and dry, with infrequent thunderstorms (Williams et al., 2009). According to the van Koppen climate classification system, upper elevations are characterized as moist continental climate and dry summers (Dsa), and lower elevations as steppe summer dry climate (Bsa) (Henderson-Sellers & Robinson, 1986). Vegetation within the catchment varies with aspect, elevation, and soil type, with grass and shrubland being the most common at lower elevations and Douglas-fir (Pseudotsuga menziesii), lodgepole pine (Pinus contorta), and aspen (Populus tremuloides) dominating the upper elevations (>1,500 m). Riparian areas feature dense brush along perennial portions of Dry Creek and its tributaries and around low-elevation seeps, consisting of cottonwoods (Populus fremontii), water birch (Betula occidentalis), yellow willow (Salix lutea), mountain alder (Alnus viridis), and mountain maple (Acer spicatum). The geology of DCEW is predominantly biotite granodiorite of the Atlanta Lobe of the Idaho Batholith which is Cretaceous in age (Johnson et al., 1988). Soils tend to be relatively thin (max depth of 1.2 m) (Williams et al., 2009) and coarse-grained and are classified as loamy sands and sandy loams (Gribb et al., 2009); a thin veneer of wind-blown loess covers portions of the basin (McNamara et al., 2018). Soil characteristics also vary with aspect, with steeper north-facing slopes having thicker soil depths, more organic matter and silt, higher porosity, and greater water storage than gentler south-facing slopes (Geroy et al., 2011). Because of limited soil moisture storage capacity (Smith et al., 2011) and no groundwater flow contribution (Yenko, 2003), high-elevation portions of Dry Creek are intermittent and lose flow (Aishlin, 2006) to deep
(>100 m) groundwater (Miller et al., 2008), though one tributary (Shingle Creek, with a drainage area of 8.6 km²) and the lower elevation mainstem are perennial, sustained by groundwater (Stratton et al., 2009).

The DCEW has been operational since 1998, with hydrometeorological and soil moisture sensors supplying data at an hourly time increment. Hydrometeorological sensors are positioned at a treeline (TL) and lower elevation gauge (LG) station. Soil moisture sensors are distributed at four locations across the catchment, on predominantly north-facing slopes adjacent to each hydrometeorological station. Discharge is monitored at the catchment outlet near the LG station (Figure 2).

2.1. Data

2.1.1. General Characteristics

Raw hourly data for the TL and LG meteorological stations (Figure 2) from 1 January 2001 through 19 July 2017 were obtained from Boise State University (http://boisestate.edu/drycreek; also see McNamara et al., 2017) for the following set of variables: air temperature, precipitation, relative humidity, solar radiation, snow water equivalent (SWE), wind speed, and wind direction (see Table S1 in the supporting information for further details). Hourly soil moisture and soil temperature were also obtained from north-facing soil pits at both the LG and TL station at depths that range from 5 to 100 cm (specific depths provided in Table S1). Hourly discharge for the DCEW catchment was obtained for the same period as the meteorological and soil moisture time series from a station just downstream of the LG meteorological station (Figure 2).

All data were carefully quality controlled, and any outliers or spurious patterns in the data were identified and removed by hand. Gaps in the data were filled for training the MLMs using interpolation, multiple linear regression, and autoregressive models. When possible, multiple linear regression was used to fill gaps as this provided a synthetic record based on observations from within the catchment. When gaps occurred over small timescales at both meteorological stations (LG and TL) autoregressive or linear interpolation was used to in-fill the records. Overall, only a small portion of the time series (<6% on average) required gap infilling.

Snowmelt was computed from decreases in SWE, and a temperature threshold of 0°C was used to parse pre-precipitation into rainfall or snowfall. We also estimated evaporation at the LG and TL gauges using the Priestley-Taylor method (Priestley & Taylor, 1972), which is based on radiation and is a simplification of the Penman-Monteith combination equation (Monteith, 1981; Penman, 1948). An α of 1.72 was used instead of the commonly used 1.26 to reflect the higher moisture stress of the arid conditions within the catchment.

We compared mean monthly estimates of evaporation between the Priestley-Taylor and Penman-Monteith methods with long-term (1916–2005) pan observations from the nearby Arrowrock Dam, Boise River, Idaho (https://wrcc.dri.edu/Climate/comp_table_show.php?stype=pan evap_avg) and found better qualitative agreement with the Priestly-Taylor estimates.

The quality-controlled data and derived parameters were aggregated to 1-, 7-, 14-, 30-, 60-, 90-, and 183-day timescales. The 1-day timescale was the minimum time stamp evaluated in this study and was computed by taking the daily mean of all meteorological variables except for ET, precipitation, rainfall, snowfall, and snowmelt which were computed as a daily total. The 7-, 14-, 30-, 60-, 90-, and 183-day aggregation scales were computed with a back-looking moving mean. To remove periodic/seasonal trends from the time series, we computed an anomaly for each variable at each aggregation length by taking the day of water year (DOWY) mean (based on the full period of record) and then differencing the DOWY mean from the aggregated values. These anomaly time series were used for our TE analysis and allowed us to detect interactions between hydrologic variables that were not driven by synoptic changes in seasonal conditions. Thus, the full suite of candidate variables includes the 61 primary meteorological variables between the two weather stations and four soil pits, with each of these variables having seven aggregated time series (1-day, 1-week, 2-week, 1-month, etc.) for a total of 427 time series. Example daily and longer timescale aggregated time series are provided for the LG and TL stations in Figures S1–S4.

3. Methods

3.1. Information Statistics

To evaluate information transfers (i.e., TE) from meteorological and hydrological predictor variables to discharge, we computed Equation 1 for τ ranging from lags of 1 day up to 183 days, using block lengths (k and l) of one. The reduction in uncertainty from a meteorological variable to discharge was deemed statistically
significant when the TE value exceeded the 95th percentile of a distribution of TE values computed from 500 randomly shuffled versions of the input data matrices (Ruddell & Kumar, 2009a, 2009b). The full length of the time series used for calculating TE was 6,044 days (~16.5 years). Thus, at the maximum lag of 183 days, there were still >5,800 overlapping data points with which to compute the joint and marginal entropy values for estimating TE. We quantified the relative significance in TE at each lag \( \tau \) as

\[
T_{rel, \tau} = \frac{(T_\tau - T_{0, \tau})}{H_Q}
\]

where \( T_{0, \tau} \) is the significance threshold at that time lag and \( H_Q \) is the total uncertainty (Shannon entropy; Shannon, 1948) in the sink variable, discharge. \( T_{rel, \tau} \) is a normalized version of the TE that quantifies the significant reduction in the uncertainty of discharge relative to the total uncertainty in discharge. The \( \tau \) associated with the highest \( T_{rel, \tau} \) within the first 183 lags was selected as the critical timescale (i.e., most significant lag) to be used in the forecasting models. The use of a single time lag in these analyses only captures the timing of when the greatest amount of information is transferred from the meteorological variables to discharge, which potentially ignores multiyear and/or less dominant processes; however, the decision to apply one time lag to each variable allowed us to cleanly test the hypothesis that incorporating information about critical timescales improves a model’s forecast skill, without applying arbitrary thresholds for peak selection. Further, by limiting the \( \tau \) selected for the machine learning analysis to 183 days, we minimize the number of training days that must be discarded (i.e., since the first portion of the time series up to the maximum number of lag days considered cannot be modeled).

### 3.2. Forecasting Models

In the following sections, we discuss the discharge forecasting models used in this study and outline our experimental design. We evaluated a number of forecasting models including feed-forward ANNs, LSTM networks, random forest regression (RFR), and support vector regression (SVR), and as the simplest end member, we used the DOWY mean discharge (\( Q_{DOWY} \)) as a predictor; collectively, we refer to these models as the “discharge forecasting methods” or machine learning models (MLMs). We focus our presentation on results from the RFR, LSTM, and \( Q_{DOWY} \) (results for the ANNs and SVRs are provided in the Supporting Information S1). We chose to focus on the LSTM and RFR as both approaches are capable of making accurate predictions in complex nonlinear systems, but their architecture and training are markedly different. Because LSTMs use sequence-based training where time series patterns are learned through explicit cell memory architecture, we expected they would perform well regardless of using lagged inputs. In contrast, RFRs do not learn patterns through sequence- or temporal-based training; instead, patterns are learned by recursively splitting the data into smaller subsets and minimizing covariance between input and response variables. Thus, we expected that the RFRs would experience enhanced performance by incorporating lagged inputs in model specifications.

#### 3.2.1. LSTM Networks

LSTM networks are a type of artificial RNN used in deep learning that use feedback connections to determine temporal dependencies between predictor and response variables (Gers et al., 1999; Hochreiter & Schmidhuber, 1997). A common LSTM unit consists of a cell, an input gate, an output gate, and a forget gate. LSTM gates regulate the flow of information (including the state of the network) into and out of each cell (Kratzert et al., 2018; Sak et al., 2014), and the forget gate resets the cell’s internal state over longer time intervals to prevent cell values from growing indefinitely (Gers et al., 1999). We adopt the common hyperbolic tangent and the logistic sigmoidal functions as state and gate activation functions, respectively (Hochreiter & Schmidhuber, 1997; Kratzert et al., 2018).

As with most deep networks, LSTMs have several parameters that must be determined during design and training. These include the number of layers, the number of hidden units in each layer, the presence/absence of dropout layers, parameters that control the number of training passes, the learning rate, and the back-propagation sensitivity to residuals between predicted and observed outputs. LSTMs also can operate in two different modes: sequence-to-sequence and sequence-to-one. In the former, an input sequence of arbitrary length produces an output sequence of the same length, while in the latter, an input sequence of fixed length is used to make an individual prediction.
The architecture utilized in this application consists of an input layer with as many neurons as input variables (model specifications are discussed in section 3.2.3), one or two LSTM layers, a fully connected layer, a dropout layer, and a regression layer with a single unit for the response variable (discharge). Following Kratzert et al. (2018), we chose to operate in sequence-to-one mode, using a 1-year back-looking window applied to the input vector \( x = [x_1, \ldots, x_n] \) (all or subsets of meteorological variables, section 3.2.3) which is passed through the recurrent cells of each network layer to predict an individual response variable \( y \) (daily discharge). While this choice implicitly assumes independence between each 1-year sequence, it results in orders of magnitude more input sequences in the training phase which results in an increase of computational cost and the lack of predictions for the first 364 days of input data. We used a limited number of Bayesian optimization iterations (Mockus, 2012) to determine working ranges for each parameter, which were subsequently adjusted with further “trial and error” tests. These tests showed that, while optimal performance is strongly dependent on the number of input features and the various tuning parameters, good performance is relatively insensitive to LSTM architecture.

### 3.2.2. Random Forest Regression

RFR is an ensemble learning technique that combines several decision trees into an additive model to generate a predicted value (here, streamflow). Each base or tree model is trained on a random sample of the features included in the training data set. The prediction is the average of all the outputs from the trees, which controls for overfitting. RFR is typically valued for the accuracy and stability of the output models (Breiman, 2001).

Here, we used the Statistics and Machine Learning Toolbox Matlab (Matlab, 2018) package to implement the RFR (Pedregosa et al., 2011), splitting the branches based on the Gini impurity coefficient (Liaw & Wiener, 2002). Five algorithmic parameters were optimized through Bayesian optimization (Mockus, 2012) with fivefold cross-validation, using root mean square error (RMSE) as the loss function: (1) the number of individually trained regression tree models in the ensemble, (2) the number of features to be considered when searching for the best split strategy, (3) the minimum number of samples needed to generate leaf nodes, (4) the minimum number of samples needed for a leaf node, which will make sure the model will have enough historical data to learn for a certain scenario, and (5) the threshold to stop the growth of the tree. The node will stop splitting if this split induces a decrease of the impurity smaller than this threshold.

### 3.2.3. Model Specifications

We used several different specifications to evaluate how feature selection using TE affected model performance. The greatest complexity in variable inputs is represented using all available meteorological variables (Table S1). We evaluated models that used daily values for all meteorological variables as well as specifications that included all aggregation scales (1, 7, 14, 30, 60, 90, and 183 days). Next in complexity was the “expert’s choice” model which was based on meteorological variables known to be important in controlling the water balance and catchment runoff. We used pairwise correlation analysis and principal components analysis (Figures S5–S8 and Text S1) to ensure that the “expert’s choice” variables did not exhibit multicollinearity. The “expert’s choice” variables included air temperature, deep and shallow soil moisture, ET, precipitation, snowmelt, and SWE for both the LG and TL stations. At the simpler end of variable complexity, we used air temperature and precipitation (P only) from the LG and TL stations because these variables are measured widely and they remain the primary inputs to many physically based and regression models. Finally, the mean DOWY discharge \( Q_{DOWY} \) was used as the simplest “benchmark” model with respect to variable complexity.

Notably, unlike many other machine learning-based forecasts of discharge discussed in the literature (Besaw et al., 2010; Erdal & Karakurt, 2013; Hsu et al., 1995; Jeong & Kim, 2005; Kisi & Cigizoglu, 2007), we did not consider past discharge as a predictor. Our rationale was that our analysis was intended to evaluate the type of forecasts that might be generated from projections of hydroclimatic information for ungauged basins from land-surface models or available through remote sensing or sensors at neighboring basins (e.g., Kratzert et al., 2018), for longer forecast timescales than a single day in advance.

### 3.2.4. Assessment of Model Performance and Predictive Uncertainty

The use of multiple model performance measures is important to reveal model goodness of fit for the different segments of a hydrograph. Here, we considered four statistical performance measures: (1) RMSE, (2) the mutual information (MI) between observed and simulated discharge, (3) Nash-Sutcliffe efficiency (NSE),...
which emphasizes model performance in predicting high flow and is equivalent to an $R^2$, and (4) the log-transformed Nash-Sutcliffe (logNSE) metric, which emphasizes model performance in low flow regimes. The performance metrics are formulated as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (Q_i - \hat{Q}_i)^2}{N}}.$$  

(3)

$$MI(Q_i, \hat{Q}_i) = \sum_{Q_i, \hat{Q}_i} p(Q_i, \hat{Q}_i) \log \frac{p(Q_i, \hat{Q}_i)}{p(Q_i)p(\hat{Q}_i)}.$$  

(4)

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{N} (Q_i - \hat{Q}_i)^2}{\sum_{i}^{N} (Q_i - \bar{Q})^2}.$$  

(5)

$$\text{logNSE} = 1 - \frac{\sum_{i=1}^{N} (\log(Q_i) - \log(\hat{Q}_i))^2}{\sum_{i}^{N} (\log(Q_i) - \log(\bar{Q}))^2}.$$  

(6)

In the above equations, $Q_i$ and $\hat{Q}_i$ represent observed and simulated streamflow over a total record length of $N$ data points (or total number of bins for MI, Equation 4), while $\bar{Q}$ represents the mean of the discharge time series.

To assess uncertainty in the LSTM and RFR predictions, we leveraged the stochastic processes present in the two methods (e.g., dropout layers in LSTM and random feature subselection in RFR) and trained ensembles of identically parameterized models but with randomized input data stream sequence. The correspondence between instances of the response variable (discharge) and its predictors (forcing variables or sequence) is preserved, but the order in which the predictor/response pairs is used in training is different for each model. Figure S10 shows the mean ensemble prediction along with the 10–90% quantiles of all the predictions.

4. Results

4.1. Dominant Timescales and Mechanisms of Interaction Between Discharge and its Hydrometeorological Predictors

Many of the predictor variables (SWE, snowmelt, and P3 and P4 deep moisture) exhibited the J-shaped TE signatures characteristic of snowmelt-dominated streamflow generation processes (i.e., the pink region in Figure 1b) (Figure 3). Among these variables, SWE (Figure 3c) reduced the uncertainty in discharge the most, by about 15%, at data aggregation timescales of 1 week to 1 month and lags of 50–60 days. Both snowmelt (Figure 3d) and P4 deep moisture (Figure 3i) reduced uncertainty in discharge by about 10% over weekly to monthly aggregation scales (snowmelt) or daily to 3-week aggregation scales (P4 deep moisture). At the shorter aggregation timescales, these variables exhibited peak TE to discharge at lags of approximately 75 days. P3 deep moisture exhibited a less-defined snowmelt characteristic signature, with a wider spread in peak lags (60–100 days for weekly data) and lower $T_{rel}$ (<0.07). The shallow moisture stations (Figures 3f and 3h) also exhibit a snowmelt signature, but only faintly, with low $T_{rel}$ (~0.06 and 0.04 for P3 and P4, respectively).

Several of the variables with a pattern of snowmelt-dominated streamflow generation processes (snowmelt, P3 and P4 shallow moisture) exhibited a stronger signature in the baseflow region (i.e., the yellow region in Figure 1b), characterized by TE peaks at 3–6-month aggregation scales. Among these three variables, snowmelt transferred the most information to discharge, reducing its uncertainty by about 16% (the highest uncertainty reduction among all variables at the TL station) at the 6-month aggregation scale and a lag of approximately 90 days. For all three variables, information transfer within the baseflow region (i.e., via baseflow mechanisms) reduced uncertainty in discharge to a greater extent than information transfer within the snowmelt-dominated region. Meanwhile, air temperature (Figure 3b) exhibited a TE peak only within the baseflow-dominated region.
Of all the predictor variables at the TL station, precipitation (Figure 3a) and P4 shallow moisture (Figure 3h) reduced uncertainty in discharge the least, peaking at a relative TE of ~0.06, both in the baseflow-dominated region. Though it transferred low information to discharge overall, precipitation exhibited TE peaks over a wide range of lags (10–80 days) and aggregation scales (2 weeks to 6 months), occupying both the snowmelt-dominated (pink) and hillslope-dominated (green) streamflow generation regions in Figure 1b.

Last, ET (Figure 3e) exhibited a distinctive TE signature that did not clearly reside within the core regions defined in Figure 1b. It reduced the most uncertainty in discharge (10%) at a 2-month aggregation scale at lags of 60–100 days.

Patterns in the maximum relative significant TE across all variables generated a mechanistic, empirical fingerprint of the dominant processes in the study area (Figures 4a and 4c). By comparing regions with high $T'_{rel,c}$ to the conceptual figure of hydrologic process signatures (Figure 1b) and the TE fingerprint maps (Figures 4a and 4c), the relative role of different observed hydrologic processes in reducing uncertainty in discharge can be evaluated. Here, observations of baseflow-dominated mechanisms of discharge generation at the TL catchment have the strongest potential to reduce uncertainty in discharge at the catchment outlet, followed by observations of snowmelt-dominated mechanisms. Snowmelt-dominated mechanisms are also important at the LG station, but less so than at the TL station, and baseflow-dominated mechanisms at the LG station are shifted to shorter lags and timescales of aggregation compared to the TL station.
SWE reduces the most uncertainty in discharge in the snowmelt region at the TL station (Figure 4d), deep soil moisture does so at the LG station (Figure 4b), indicating that at lower elevations, deep soil moisture best reflects the catchment-scale effects of melt. At both subcatchments, hillslope-dominated processes also reduce uncertainty in discharge, though to a lesser extent than other hydrologic processes. At high elevations, SWE reduces the most uncertainty in the hillslope-processes region (Figure 4d), whereas at low elevations, deep soil moisture reduces the most uncertainty in this region (Figure 4b).

Figure 4. Composite signatures of information flow from predictor variables to discharge. Colors in (a) and (c) indicate the maximum relative significant transfer entropy ($T_{rel}$; Equation 3) observed across all predictor variables to discharge over all aggregation and lag timescales evaluated. Patterns can be compared to Figure 1b to identify the relative importance of observed mechanisms of streamflow generation for the reduction of uncertainty in discharge. Meanwhile, colors in (b) and (d) indicate which variable is responsible for peak information flow at each combination of lag and aggregation timescale.
4.2. Comparative Performance of MLMs

4.2.1. General Model Performance

While forecast performance varied substantially across the different MLMs, the LSTMs exhibited the best performance, capturing baseflow and peakflow, and achieved the best performance across all model evaluation metrics (Figure 5, Table 1, Figure S9, and Table S2). The other MLMs (ANN, RFR, and SVR) also performed well with respect to RMSE but never achieved the same performance as the LSTMs (Table S2). The best performing LSTM and RFR models both captured the timing of hydrologic events well, but the LSTM was more effective at capturing the hydrograph peaks (Figure 5). While the LSTMs captured annual peakflows quite closely, they underpredicted flashy early season runoff. The best RFR run substantially overpredicted water year 2013 runoff (a below average year) but underpredicted water years 2014–2016 (Figure 5b).

4.2.2. Forecast Uncertainty

The 10th and 90th percentiles from the prediction ensembles (section 3.2.4) for the LSTM runs provided lower and upper bounds that enveloped observed discharge (Figure S10) indicating that LSTMs can be used to generate prediction sets that would likely encapsulate the actual event response. The RFR uncertainty ensembles did not exhibit as strong of performance where the 90th percentile typically underpredicted peakflow. The lower bound (10th percentile) did capture baseflow however (Figure S10).

4.2.3. Increase in Performance Relative to DOWY Mean Discharge ($Q_{DOWY}$)

On average, the MLMs outperformed predictions based on $Q_{DOWY}$; however, the number of model runs that exhibited improvement and the amount of performance increase (or decrease) varied substantially between the LSTMs and RFRs (Table 1). The LSTMs exhibited the greatest improvement with 9 of 12 model runs showing lower RMSE than $Q_{DOWY}$; the best LSTM (model run 1) exhibited just over 100% improvement.

Figure 5. Measured (thick black lines) and modeled hydrographs (blue lines for training period (2002–2012), cyan lines for test period (2013–2017)) for the best-performing (a) long short-term memory (LSTM) network (model run 1) and (b) random forest regression (RFR; model run 7). Precipitation is displayed in the background of the hydrographs (top x-axis), and predictions for the null ($Q_{DOWY}$; orange dashed line) model are shown for the test period.
and the average increase for model runs 1–8, 10, and 11 was 45% (Figure 6a). The RFRs exhibited improvement compared to the Q_{DOWY} for 8 out of 12 model runs with the best run (7) exhibiting almost 60% improvement; the average improvement for model runs 3, 4, 6–8, and 10–12 was ~24% (Figure 6a). The increase/decrease in model performance also varied as a function of variable specifications (Figure 6a). The LSTMs exhibited substantial performance increases for unlagged inputs for the “P-T only” and the “expert’s choice” variables. The best RFR (run 7) used the “expert’s choice” variables and the RFR runs for the “all variables” category showed consistent improvement in relation to Q_{DOWY}.

The LSTMs and RFRs generally exhibited similar performance with respect to the evaluation metrics (3.1.4) but with some variability (Table 1). The LSTM runs showed the best performance across all metrics and in particular with respect to MI where all model runs showed improved performance compared to Q_{DOWY} predictions. The fact that the LSTM runs captured baseflow and peakflow well (particularly for the P-T only and expert’s choice runs) is reflected in the high values for NS and logNSE (Table 1). In contrast, the RFR runs did much poorer in capturing base and peakflows (Figure 5) which are reflected in the lower NS and logNSE values; in fact, the RFR runs that used the unlagged P-T only set performed worse than Q_{DOWY}.

### 4.2.4. Effects of Lagged Inputs and Aggregation Scales on Model Performance

Including lagged input variables and variables aggregated across timescales produced a variety of effects on model performance. While the LSTM models experienced almost no improvement with lagged inputs with respect to RMSE (only model run 10 exhibited limited increase performance compared to the unlagged version, model run 9), many of the RFR runs exhibited increased performance with lagged inputs (Figure 6b). Increases in performance between RFR unlagged and lagged runs varied from ~8% to 30%. Interestingly, most of the performance boosts occurred when the RFR model specification used lagged daily inputs but not lagged inputs aggregated across all timescales. The greatest increase in performance for the RFR runs were for the “expert’s choice” and “all variables”; both models exhibited ~30% increase in performance compared to their unlagged counterparts. In contrast, the RFR runs that used lagged aggregated variables instead of lagged daily inputs did not show increased performance (Figure 6b).

Including variables aggregated across 1–183-day timescales generally resulted in high performing models. This was especially notable for the “P-T only” and “expert’s choice” variable sets (Figure 6a). In fact, the best RFR and second-best LSTM runs was the “expert’s choice,” all timescales, unlagged (run 7). However, performance was much poorer for specifications that used aggregated variables that were lagged (Figure 6a and Table 1).
5. Discussion

5.1. Discharge Timing and Controls Inferred Through Information Flow (TE)

The TE analysis is consistent with the understanding of DCEW as a snowmelt-dominated watershed, with meltwaters contributing to streamflow via multiple pathways. Both the high peak magnitudes of $T_{\text{rel}}$ for variables that are indicators of melt-driven streamflow contributions, particularly SWE (Figure 3), and the approximately 2-month timescale over which that information transfer occurs (note that mean annual lags between peak SWE and peak discharge were 64 and 45 days for the LG and TL gages, respectively, Figure S11) point to snowmelt as a dominant mechanism for streamflow generation. While specific lag times comprising the TE fingerprint are undoubtedly influenced by sensor placement within the catchment, this approximate 2-month timescale agrees with the January-to-March difference in the average centroid of precipitation for all elevations and the centroid for streamflow (McNamara et al., 2018). This lag timescale can best be described as a “response timescale,” governed primarily by the celerity of the snowmelt impulse, which pushes older water into the stream (McNamara et al., 2018), rather than a “transit timescale,” governed by the velocity of molecules in the meltwater (Kirchner, 2016; McDonnell & Beven, 2014). While isotopic studies suggest some fast lateral movement of meltwater within the snowpack at high elevations (Evans et al., 2016), isotope-based residence times in a tributary of Dry Creek are much longer than the TE-based lag time (1–10 years, with mean age >4 years during nonmelt periods, decreasing to approximately 1 year during melt periods) (Ala-aho et al., 2017).

Previous studies in the DCEW illuminate the mechanisms through which melt generates streamflow. As melt begins, meltwater percolates downward to the shallow bedrock. Antecedent dry conditions may initially inhibit gravity drainage to the deepest soil layers, so streamflow often remains disconnected from the dynamics of shallow soil moisture (McNamara et al., 2005, 2011). However, when hillslope to stream connectivity of a near-bedrock, deep, saturated layer of soil is established, melt contributes to streamflow (McNamara et al., 2005). This mechanistic understanding of the primary driver of streamflow explains why integrated variables indicative of the water content of snow (SWE at the high-elevation catchment; Figure 4d) and deep soil moisture of north-facing slopes (which contain the thickest, most sheltered soils), such as at P2 at the lower elevation catchment, substantially reduce uncertainty in discharge in the snowmelt-dominated region of the TE fingerprint plots.

The shape of the snowmelt-dominated region of the TE fingerprint plots for soil moisture variables (Figure 3) provides further insight into spatially explicit mechanisms responsible for streamflow generation. P4 deep soil moisture (Figure 3i) exhibits a well-defined snowmelt signature, featuring high information transfer even at daily aggregation timescales, with a magnitude comparable to that of SWE and snowmelt. This pattern suggests that deep soil moisture at P4 contributes to streamflow only when the snowpack contributes to streamflow, under conditions of catchment connectivity. On the other hand, P3 deep moisture contributes less information to discharge, with almost no information transferred at the daily timescale, and its TE signature (Figure 3g) is less well defined. These observations suggest that deep moisture of P3 is not as good an indicator of the spatially connected conditions that generate streamflow as at P4. Instead, P3 deep soils may become saturated before the catchment is fully connected and generating streamflow, or percolation to deep soils at this location may be more subject to ET than at P4 at times when connectivity is established.
Meanwhile, the primary contribution of shallow moisture stations (Figures 3f and 3h) to reducing uncertainty in discharge occurs in the baseflow-dominated region of the plot, suggesting that shallow moisture on these north-facing slopes serves as a proxy for long-term ET within the catchment and its effects on groundwater recharge. This finding is consistent with Hinckley et al. (2014), who found that storage within the thick soils of north-facing slopes exposes that reservoir of water to ET. Notably, shallow soil moisture does not substantially reduce uncertainty in discharge through snowmelt mechanisms, consistent with the findings of McNamara et al. (2005), who noted a disconnectivity between shallow and deep soil moisture, attributed to ET.

Interestingly, the snowmelt variable itself reduces more uncertainty in discharge through baseflow mechanisms than through spring runoff processes (Figure 3d), which results in the potential for knowledge of baseflow-dominated mechanisms to reduce the most uncertainty in discharge overall (Figure 4). This finding is consistent with modeling studies calibrated to isotope measurements, which suggest that in Dry Creek, slow groundwater discharge is a more important contributor to streamflow than direct contributions of meltwater (Tetzlaff et al., 2015). Here, the reduction in uncertainty in streamflow from melt occurs at a 90-day (3-month) lag for the 6-month aggregated time series at the TL sensor. The 6-month timescale of aggregation is sufficient to capture the entire cumulative melt, and the 3-month lag is consistent with the concept that knowledge of the total melt generated from the snowpack at high elevations primarily reduces uncertainty late in the dry season, after spring runoff peaks have subsided. Intuitively, this makes sense, as high snowpack years would be expected to generate high late-season baseflow, and vice-versa. In contrast, at the lower elevation LG station, the 3-month timescale of aggregation is sufficient to capture the cumulative melt and its effects on baseflow-dominated streamflow generation processes (Figure 4).

Patterns of TE from ET (Figure 3e) further reinforce the idea that snowpack is the primary source of dry-season baseflow. While the lag in peak TE is consistent with that of snowmelt that contributes to baseflow, the fact that the peak occurs in the 2-month aggregation timescale rather than the 6-month timescale suggests that only the ET that occurs during the period when substantial melt is being generated has a dominant influence on deep percolation to groundwater. The 6-month aggregation scale, in contrast, may be too long for ET to reduce uncertainty in discharge, because it combines periods of active melt generation with periods with overwhelmingly high ET (i.e., summer) but limited supply of water. Shallow soil moisture serves (Figures 3f, h, and 4) as an additional proxy for ET’s influence on baseflow, given its TE peak in the baseflow region but not in regions representative of streamflow generation through other mechanisms.

The insight into baseflow-dominated mechanisms of streamflow generation provided by the TE analysis is consistent with results from previous hydrological process studies. For example, the greater uncertainty reduction in streamflow resulting from knowledge of variables relevant to baseflow processes (compared to knowledge of variables relevant to snowmelt processes) is consistent with a water balance model of DCEW that suggested that up to 36% of precipitation may percolate deeply to groundwater, with only 11–16% transformed over shorter timescales into streamflow (Kelleners et al., 2010). Meanwhile, previous chloride mass balance studies conducted at DCEW revealed that at the catchment scale, up to 11% of annual precipitation recharges deep groundwater (Aishlin, 2006). These studies also underscored the importance of ET, which consumes up to 53% of annual precipitation (Kelleners et al., 2010), in determining the balance of water available for recharge.

However, the literature also illuminates how other, less dominant processes also contribute to streamflow generation in the DCEW. Late summer rains may generate streamflow if deep moisture connectivity is established prior to freezing (McNamara et al., 2005). Similarly, depending on the timing of spring rain events following the melt and whether they coincide with connected conditions in deep soil moisture, these rain events can also contribute to streamflow (Williams et al., 2009), often via rapid lateral flow through the snowpack (Eirikkson et al., 2013; Evans et al., 2016). These processes are spatially heterogeneous and highly variable interannually (McNamara et al., 2005; Williams et al., 2009) but may be partially represented in the low but significant transfer of information from water balance-relevant variables (e.g., precipitation, SWE, P3 deep moisture in Figures 3a, 3c, and 3g) to discharge at short timescales such as those in the hillslope and quickflow/rapid response regions in Figure 1. In particular, the greater importance of SWE compared to other variables in the hillslope and quickflow regions of the TE fingerprint for the TL station (Figure 4d) underscores the importance of lateral flow through snowpack to discharge generation,
performed the unlagged models in MI, NSE, and logNSE, which is the case for the variables (such as soil moisture) that are integrative of past hydrologic behavior, the lagged models outperformed the unlagged daily RFR models with respect to these metrics suggests that the previous day's data may be more important for predicting nonpeak events. Interestingly, while the lagged daily RFR models generated more accurate peak magnitudes than historical data. However, if the previous day's data do not contain variables (such as soil moisture) that are integrative of past hydrologic behavior, the lagged models outperform the unlagged models in MI, NSE, and logNSE, which is the case for the “PT-only” model.

A primary reason why the LSTM models outperformed the RFR models is that the LSTM architecture enables the models to “learn” the memory timescales of the input variables from the training data set. Importantly, the LSTMs achieved strong performance relative to the RFRs because they were sufficiently flexible to incorporate multiple timescales of hydrologic memory. Unlagged LSTM models consistently

5.2. Importance of Appropriate Timescale Representation and Sensor Network Representativeness for Model Performance

As a state-of-the-art deep learning model, the performance of an LSTM is interpretable as a measure of the appropriateness or representativeness of a sensor/data network in capturing the controls on a process of interest (Nearing et al., 2018). Here, the best performing LSTM benchmarks captured only approximately 50% of the total information present in observed discharge time series at best (Table 1). Lack of performance of this benchmark is attributable at least in part to an inability to sample spatial and temporal variability in melting and precipitation inputs with just two sampling stations distributed over a large catchment with a large range in elevation.

On top of the shortcomings of the input data for making predictions about discharge, a performance discrepancy between the lagged and unlagged MLMs suggested that models that are restricted to using only data from the timescales of dominant processes for streamflow generation miss shorter timescale but less dominant phenomena that are nonetheless critical for prediction. In other words, results of the MLMs underscored the importance of representing the critical timescales over which all discharge generation mechanisms occur, though in ways contrary to our initial hypothesis that lagging input data series by the timescale of dominant mechanisms would produce the best forecasts.

Indeed, for the RFR models that used daily data only, lagging the input series by the TE-selected lags improved model performance by RMSE metrics (Table 1), but for all the RFR models that used data aggregated over multiple scales except for the PT-only models, the unlagged versions produced superior performance. These models produced better performance because the maximum data aggregation scales (3 months and 6 months) were long enough to incorporate signals generated during the snow buildup and melt period into the spring runoff discharge prediction, while the shorter timescales of aggregation captured recent events that have a strong influence on individual peak magnitudes (Eirikkson et al., 2013; Evans et al., 2016; Williams et al., 2009). Whether these processes generate streamflow and how much streamflow they generate may be a function of preexisting catchment connectivity (Smith et al., 2011), whether temperatures hit the freezing point before or after connectivity is established (McNamara et al., 2005), whether and for how long temperatures drop below freezing during the early snowmelt season (McNamara et al., 2011; Williams et al., 2009), and whether a snowpack exists during a rain event (Eirikkson et al., 2013; Williams et al., 2009). In contrast, while the lagged RFR models would capture the general spring runoff, they would vastly miss the magnitude of individual peaks influenced by processes occurring at timescales shorter than the TE-identified lag.

Interestingly, while the lagged daily RFR models generally outperformed the unlagged daily models in RMSE, the opposite was true for the other performance metrics (Table 1). While RMSE is strongly impacted by the model's ability to predict peaks and may be strongly influenced by a single point that is substantially overpredicted or underpredicted, the other performance metrics are more strongly influenced by prediction over the whole time series. Poor performance of the lagged daily RFR models with respect to these metrics suggests that the previous day's data may be more important for predicting nonpeakflow than historical data. However, if the previous day's data do not contain variables (such as soil moisture) that are integrative of past hydrologic behavior, the lagged models outperform the unlagged models in MI, NSE, and logNSE, which is the case for the “PT-only” model.
outperformed lagged models because, again, lagged input data sets eliminate information from recent events in favor of capturing only the snowmelt response timescale. In addition to learning appropriate lag timescales, LSTM models were able to learn appropriate data aggregation scales, as evidenced by the strong performance of the daily unlagged LSTM models with respect to all metrics.

5.3. LSTM Models Appropriately Capture Critical Timescales of Hydrologic Processes in Data-Limited Catchments

Surprisingly, the best performing model overall, across all four metrics, was the unlagged “PT-only” LSTM model. Intuitively, this result is pleasing, as precipitation and temperature are the ultimate drivers of all catchment water balance processes. Furthermore, it is convenient, as precipitation and temperature are available across far more catchments than snow or soil moisture data. However, it is surprising in that it outperformed models that assimilate far more comprehensive hydrologic data.

The conceptual (Figure 1) and empirical (Figures 3 and 4) TE fingerprints yield insight into why this result might arise. First, these fingerprints reinforce the notion that incorporating multiple lag and data aggregation timescales is essential for comprehensive representation of the hydrologic processes that generate streamflow. However, they also show that there is redundant information across timescales of aggregation (as evidenced by the colored regions in Figure 1b that are vertically high) and across variables (as many variables represent the same set of hydrologic processes). This redundancy may lead to problems with model fitting. For example, only when few variables (i.e., precipitation and temperature) were considered was the LSTM able to “learn” appropriate timescales of aggregation, causing the daily model to outperform the all-scale model. Further, as more redundant variables were added to the LSTM in going from the PT models to the expert and all-variable models, performance declined. This finding raises a compelling question that merits future investigation: when LSTMs are trained with variables beyond the ultimate drivers of a hydrologic system, is it generally true that performance declines?

5.4. Information Theory-Informed Modeling of Catchment Behavior

Overall, the TE-informed identification of critical timescales is consistent with the conceptualization of the catchment as a nonlinear filter for signals from precipitation (Kirchner et al., 2000, 2001). Predictor variables tend to transfer significant amounts of information to discharge over a majority of timescales examined, reflecting both a range of mechanisms for streamflow generation and heterogeneous flow paths that encompass both active and passive storage (McNamara et al., 2005, 2011; Tetzlaff et al., 2011, 2014). This study demonstrated that when only the timescales of peak information transfer are used to guide the selection of inputs for MLMs, only the dominant streamflow-generating process or processes will be represented in the model. The result is predictions that capture the broad, low-frequency characteristics and timing of the annual hydrograph but miss higher frequency components that are affected by interactions among several of the predictor variables. While a TE analysis may thus serve as a quantitative guide for ensuring representation of dominant processes in the simplest of models (e.g., here, those with very short histories of the variables considered and a memory-agnostic architecture), it may undercut the full predictive power of an MLM when only peaks of information transfer are considered in the curation of input data sets and timescales.

Fortunately, new deep learning models such as LSTMs are capable of learning appropriate timescales over which hydrologic processes influence discharge. Optimally designing sensor networks to maximize their representativeness of spatially and temporally heterogeneous processes will improve the predictive power of these models.

While TE-informed model specification does not appear to benefit the performance of MLM-based forecasts, TE analyses are highly complementary to the development of predictive models. For one, TE analyses produce process insight that exceeds what can be learned from a linear correlation analysis. For example, although the TE signatures of many monitored variables (Figure 3) are indicative of snowmelt-dominated streamflow generation processes, these variables themselves are poorly correlated to snowmelt. Our correlation analysis (Figure S5) shows that the correlation between soil moisture and snowmelt or SWE is just under 0.50 for the shallow soil moisture pits, declining to about 0.30 for deep soil moisture pits. However, the TE signatures suggest that it is deep soil moisture that reduces the most uncertainty in discharge through snowmelt mechanisms, consistent with the findings of McNamara et al. (2005).
Secondly, the process insight obtained from TE analyses can be used to evaluate whether a model performs well for the “right” reasons (sensu Kirchner, 2006). Namely, predictive models should exhibit similar patterns of information transfer among modeled variables as those resolved among sensor variables. In comparing timescales and relative magnitudes of information flow between models and data, a hydrologist takes advantage of a powerful tool for model selection and diagnostics (e.g., Nearing et al., 2018; Ruddell et al., 2019).

6. Summary and Conclusions

In this work, we evaluated the ability of TE to reveal the dominant catchment processes that drive streamflow and analyzed how the inclusion of these variables and the timescales over which they reduce uncertainty in discharge affected model performance. The empirical TE fingerprints suggested that the seasonal accumulation of snowpack and the interaction of snowmelt with soil moisture stores, groundwater, and energy fluxes driving evaporative losses were the dominant controls on seasonal discharge patterns, which agrees well with previous mechanistic-based studies that have evaluated controls on catchment discharge.

For MLMs that do not account for variable “memory,” lagging the input time series in accordance with the dominant time lags identified through the TE analysis improved model performance but only for models based on daily time series. The lagged formulations improved the ability of the model to capture the dominant processes (i.e., accumulation, snowmelt, evaporative losses, and percolation to groundwater) but limited the ability of the models to capture a wider range of processes known to be important for streamflow generation (e.g., catchment connectivity and higher frequency variations driven by daily snowmelt and rainfall). Models that included a wider range of timescales—even without lags—were better able to capture the range of processes relevant to local peak flows. The best-performing MLMs were the LSTM models, which were able to effectively “learn” the wide range of critical lag and aggregation timescales revealed by the TE fingerprints to be controls on discharge. LSTM performance was highest for parsimoniously specified models (e.g., with just precipitation and temperature input time series and daily aggregation scales), potentially because the risk of overfitting was minimized. In other words, more data are not necessarily better for machine learning discharge forecasts. This finding bodes well for the widespread development of hydrologic forecasts from widely available daily precipitation and temperature data.

Our analyses showed that recurrent MLMs are highly effective at identifying both the lag and data aggregation timescales that are most useful for forecasting and are not improved by information theory-informed model specification. Nonetheless, TE fingerprints yield insight into mechanisms through which hydrologic processes reduce uncertainty in streamflow and may provide a tool for catchment classification and process benchmarking (e.g., Ruddell et al., 2019). Here, the TE fingerprints showed that observations of hydrologic processes (primarily the balance between snowpack accumulation and ET) at the TL subcatchment reduced uncertainty in discharge more than at the lower elevation LG station and that melt processes reduced more uncertainty in discharge rather than spring runoff mechanisms. These findings imply that the observational data network may be inherently better suited for predicting low flows than peakflows and that additional sensors placed high in the catchment may improve discharge forecasts to a greater degree than low in the catchment.

Data Availability Statement

All data and code used in this analysis are available as a publicly available repository at the following links, https://doi.org/10.6078/D1CH64 and https://gitlab.com/esdl/dry-creek.

References

Aishlin, P. S. (2006). Groundwater recharge estimation using chloride mass balance, Dry Creek Experimental Watershed (PhD Thesis). Boise State University.

Ala-aho, P., Tetzlaff, D., McNamara, J. P., Laudon, H., Kormos, P., & Soulsby, C. (2017). Modeling the isotopic evolution of snowpack and snowmelt: Testing a spatially distributed parsimonious approach. Water Resources Research, 53(7), 5813–5830. https://doi.org/10.1002/2017wr020650

Benettin, P., Kirchner, J. W., Rinaldo, A., & Botter, G. (2015). Modeling chloride transport using travel time distributions at Plynlimon, Wales. Water Resources Research, 51, 3259–3276. https://doi.org/10.1002/2014WR016600

Besaw, L. E., Rizzo, D. M., Bierman, P. R., & Hackett, W. R. (2010). Advances in ungauged streamflow prediction using artificial neural networks. Journal of Hydrology, 386(1-4), 27–37. https://doi.org/10.1016/j.jhydrol.2010.02.037
Beven, K. (1982). On subsurface stormflow: Predictions with simple kinematic theory for saturated and unsaturated flows. Water Resources Research, 18(6), 1627–1633. https://doi.org/10.1029/WR018i006p01627

Beven, K. (1989). Changing ideas in hydrology—The case of physically-based models. Journal of Hydrology, 105(1–2), 157–172. https://doi.org/10.1016/0022-1694(89)90011-7

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1007555808488

Daniell, T. M. (1991). Neural networks. Applications in hydrology and water resources engineering. In Conference Publication-Institute of Engineers. Australia.

Dawson, C. W., & Wilby, R. L. (2001). Hydrological modelling using artificial neural networks. Progress in Physical Geography, 25(1), 80–108. https://doi.org/10.1191/0309133301tp00104

de Vos, N. J., & Rietjens, T. H. M. (2005). Constraints of artificial neural networks for rainfall-runoff modelling: Trade-offs in hydrological state representation and model evaluation. Hydrology and Earth System Sciences, 9(1/2), 111–126. https://doi.org/10.5194/hess-9-111-2005

Erikkson, D., Whitson, M., Lace, C. H., Marshall, H. P., Bradford, J., Benner, S. G., et al. (2013). An evaluation of the hydrologic relevance of lateral flow in snow at hillslope and catchment scales. Hydrological Processes, 27(5), 640–654. https://doi.org/10.1002/hyp.9666

Erdal, H. I., & Karakurt, O. (2013). Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. Journal of Hydrology, 477, 119–128. https://doi.org/10.1016/j.jhydrol.2012.11.015

Evans, S. L., Flores, A. N., Heilig, A., Kohn, M. J., Marshall, H. P., Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(4), 1049–1073. https://doi.org/10.1162/neco.1997.9.4.1049

Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Continual prediction using LSTM with forget gates. Neural nets, IJCNN Vietri-99 (pp. 133–138). Vietri sul Mare, Salerno, Italy: Proceedings of the 12th Italian Workshop On Neural Nets.

Godsey, S. E., Kirchner, J. W., & Tague, C. L. (2014). Effects of changes in winter snowpacks on summer low flows: Case studies in the Sierra Nevada, California, USA. Hydrological Processes, 28(19), 5048–5064. https://doi.org/10.1002/hyp.9943

Goodwell, A. E., & Kumar, P. (2017). Temporal information partitioning: Characterizing synergy, uniqueness, and redundancy in interacting environmental variables. Water Resources Research, 53, 5920–5942. https://doi.org/10.1002/2016WR020216

Goodwell, A. E., Kumar, P., Fellows, A. W., & Flerchinger, G. N. (2018). Dynamic process connectivity explains ecohydrologic responses to rainfall pulses and drought. Proceedings of the National Academy of Sciences, 115(37), E8604–E8613. https://doi.org/10.1073/pnas.1803261115

Henderson-Sellers, A., & Robinson, P. J. (1986). Contemporary climatology. London, New York: Longman Scientific & Technical. Wiley.

Hinckley, E. L. S., Ibel, B. A., Barnes, R. T., Anderson, R. S., Williams, M. W., & Anderson, S. P. (2014). Aspect control of water movement on hillslopes near the rain–snow transition of the Colorado Front Range. Hydrological Processes, 28(1), 74–85. https://doi.org/10.1002/hyp.9549

Hipsey, M. R., Hamilton, D. P., Hanson, P. C., Carey, C. C., Coletti, J. Z., Read, J. S., et al. (2015). Predicting the resilience and recovery of aquatic systems: A framework for model within environmental observatories. Water Resources Research, 51, 7023–7043. https://doi.org/10.1002/2015WR017175

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

I., & Kim, Y. (2019). Evolving water model. Civil Engineering and Environmental Survey.

Kapnick, S. B., Yang, X., Vecchi, G. A., Delworth, T. L., Gudgel, R., Malyshev, S., et al. (2018). Potential for western US seasonal snowpack prediction. Proceedings of the National Academy of Sciences, 101, 716, 760.

Kelleners, T. J., Chandler, D. G., McNamara, J. P., Gribb, M. M., & Seyfried, M. S. (2010). Modeling runoff generation in a small snow-dominated mountainous catchment. Vadose Zone Journal, 9(3), 517–527. https://doi.org/10.2136/vzj2009.0033

Khakbaz, B., Imam, B., Hsu, K., & Sorosohnian, S. (2012). From lumped to distributed via semi-distributed: Calibration strategies for semi-distributed hydrologic models. Journal of Hydrology, 418, 61–77.

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. Water Resources Research, 42, W05S04. https://doi.org/10.1029/2005WR004362

Kirchner, J. W. (2016). Aggregation in environmental systems- Part 2: Catchment mean transit times and young water fractions under hydrologic nonstationarity. Hydrology and Earth System Sciences, 20(1), 299–328. https://doi.org/10.5194/hess-20-299-2016

Kirchner, J. W., Feng, X., & Neal, C. (2000). Fractal stream chemistry and its implications for contaminant transport in catchments. Nature, 403(6769), 524–527. https://doi.org/10.1038/35000357

Kirchner, J. W., Feng, X., & Neal, C. (2001). Catchment-scale advection and dispersion as a mechanism for fractal scaling in stream tracer concentrations. Journal of Hydrology, 254(1–4), 82–101. https://doi.org/10.1016/S0022-1694(01)00487-5

Kissi, O., & Cigizoglu, H. K. (2007). Comparison of different ANN techniques in river flow prediction. Civil Engineering and Environmental Systems, 24(3), 211–231. https://doi.org/10.1080/1028660060088565

Kollet, S. J., & Maxwell, R. M. (2006). Integrated surface–groundwater flow modeling: A free-surface overland flow boundary condition in a parallel groundwater flow model. Advances in Water Resources, 29(7), 945–958. https://doi.org/10.1016/j.advwatres.2005.08.006

Kollet, S. J., & Maxwell, R. M. (2008). Capturing the influence of groundwater dynamics on land surface processes using an integrated, distributed watershed model. Water Resources Research, 44, W02002. https://doi.org/10.1029/2007WR006904
Skøien, J. O., Blöschl, G., & Western, A. W. (2003). Characteristic space scales and timescales in hydrology. *Water Resources Research, 39*(10), 1304. https://doi.org/10.1029/2002WR001736

Smith, T. J., McNamara, J. P., Flores, A. N., Gribb, M. M., Aishlin, P. S., & Benner, S. G. (2011). Small soil storage capacity limits benefit of winter snowpack to upland vegetation. *Hydrological Processes, 25*(25), 3858–3865. https://doi.org/10.1002/hyp.8340

Stratton, B. T., Sridhar, V., Gribb, M. M., McNamara, J. P., & Narasimhan, B. (2009). Modeling the spatially varying water balance processes in a semiarid mountainous watershed of Idaho. *JAWRA Journal of the American Water Resources Association, 45*(6), 1390–1408. https://doi.org/10.1111/j.1752-1688.2009.00371.x

Tetzlaff, D., Birkel, C., Dick, J., Geris, J., & Soulsby, C. (2014). Storage dynamics in hydropedological units control hillslope connectivity, runoff generation, and the evolution of catchment transit time distributions. *Water Resources Research, 50*, 969–985. https://doi.org/10.1002/2013WR014147

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

Tetzlaff, D., McNamara, J. P., & Carey, S. K. (2011). Measurements and modelling of storage dynamics across scales. *Hydrological Processes, 25*(25), 3831–3835. https://doi.org/10.1002/hyp.8396

Tyler, S. W., Burak, S. A., McNamara, J. P., Lamontagne, A., Selker, J. S., & Dozier, J. (2008). Spatially distributed temperatures at the base of two mountain snowpacks measured with fiber-optic sensors. *Journal of Glaciology, 54*(187), 673–679. https://doi.org/10.3189/002214308786570827

Vieux, B. E. (2001). *Distributed hydrologic modeling using GIS*. Boston, MA USA: Springer. https://doi.org/10.1007/978-94-015-9710-4

Williams, C. I., McNamara, J. P., & Chandler, D. G. (2009). Controls on the temporal and spatial variability of soil moisture in a mountainous landscape: The signature of snow and complex terrain. *Hydrology and Earth System Sciences, 13*(7), 1325–1336. https://doi.org/10.5194/hess-13-1325-2009

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., et al. (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research, 117*, D03109. https://doi.org/10.1029/2011JD016048

Yenko, M. K. (2003). Hydrometric and geochemical evidence of streamflow sources in the upper Dry Creek Experimental Watershed, Southwestern Idaho (PhD Thesis). Boise State University, Boise, ID.

Young, P. C., & Beven, K. J. (1994). Data-based mechanistic modelling and the rainfall-runoff non-linearity. *Environmetrics, 5*(3), 335–363. https://doi.org/10.1002/env.3170050311

Yu, P.-S., Chen, S.-T., & Chang, I.-F. (2006). Support vector regression for real-time flood stage forecasting. *Journal of Hydrology, 328*(3–4), 704–716. https://doi.org/10.1016/j.jhydrol.2006.01.021