

# Water Resources Research®



## RESEARCH ARTICLE

10.1029/2022WR034092

### Key Points:

- We utilize data-driven sparse sensing to predict daily streamflow and identify the optimal times for streamflow measurement across the contiguous United States
- Streamflow was more effectively predicted in watersheds dominated by snowmelt and baseflow than those dominated by rainfall and quickflow
- The optimal sampling times for streamflow prediction by data-driven sparse sensing are periods with large flow magnitudes and variances

### Supporting Information:

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

### Correspondence to:

K. Zhang,  
[kzhang16@connect.hku.hk](mailto:kzhang16@connect.hku.hk)

### Citation:

Zhang, K., Luhan, M., Brunner, M. I., & Parolari, A. J. (2023). Streamflow prediction in poorly gauged watersheds in the United States through data-driven sparse sensing. *Water Resources Research*, 59, e2022WR034092. <https://doi.org/10.1029/2022WR034092>

Received 10 NOV 2022

Accepted 25 MAR 2023

### Author Contributions:

**Conceptualization:** Kun Zhang, Anthony J. Parolari

**Data curation:** Kun Zhang

**Formal analysis:** Kun Zhang

**Funding acquisition:** Anthony J. Parolari

**Investigation:** Kun Zhang





**Methodology:** Kun Zhang

**Project Administration:** Anthony J. Parolari

**Resources:** Anthony J. Parolari

**Software:** Kun Zhang

## Streamflow Prediction in Poorly Gauged Watersheds in the United States Through Data-Driven Sparse Sensing

Kun Zhang<sup>1,2</sup> , Mitul Luhan<sup>3</sup> , Manuela I. Brunner<sup>4,5,6</sup> , and Anthony J. Parolari<sup>1</sup> 

<sup>1</sup>Department of Civil, Construction, and Environmental Engineering, Marquette University, Milwaukee, WI, USA,

<sup>2</sup>Department of Civil and Environmental Engineering, Seattle University, Seattle, WA, USA, <sup>3</sup>Department of Aerospace and Mechanical Engineering, University of Southern California, Los Angeles, CA, USA, <sup>4</sup>Institute for Snow and Avalanche Research SLF, Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Davos, Switzerland, <sup>5</sup>Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland, <sup>6</sup>Climate Change, Extremes and Natural Hazards in Alpine Regions Research Center CERC, Davos Dorf, Switzerland

**Abstract** Many rivers and streams are ungauged or poorly gauged and predicting streamflow in such watersheds is challenging. Although streamflow signals result from processes with different frequencies, they can be “sparse” or have a “lower-dimensional” representation in a transformed feature space. In such cases, if this appropriate feature space can be identified from streamflow data in gauged watersheds by dimensionality reduction, streamflow in poorly gauged watersheds can be predicted with a few measurements taken. This study utilized this framework, named data-driven sparse sensing (DSS), to predict daily-scale streamflow in 543 watersheds across the contiguous United States. A tailored library of features was extracted from streamflow training data in watersheds within the same climatic region, and this feature space was used to reconstruct streamflow in poorly gauged watersheds and identify the optimal timings for measurement. Among different regions, streamflow in snowmelt-dominated and baseflow-dominated watersheds (e.g., Rocky Mountains) was more effectively predicted with fewer streamflow measurements taken. The prediction efficiency in some rainfall-dominated regions, for example, New England and the Pacific coast, increased significantly with an increasing number of measurements. The spatial variability of prediction efficiency can be attributed to the process-driven mechanisms and the dimensionality of watershed dynamics. Storage-dominated systems are lower-dimensional and more predictable than rainfall-dominated systems. Measurements taken during periods with large streamflow magnitudes and/or variances are more informative and lead to better predictions. This study demonstrates that DSS can be an especially useful technique to integrate ground-based measurements with remotely sensed data for streamflow prediction, sensor placement, and watershed classification.

**Plain Language Summary** Many rivers and stream reaches are ungauged or poorly gauged because streamflow measurement is costly and resource intensive. Predicting the streamflow time-series in these ungauged or poorly gauged watersheds is still challenging. Here, we use a signal processing technique called data-driven sparse sensing on a national-scale streamflow data set across the contiguous United States. We predict streamflow time-series in each watershed based on existing streamflow data in watersheds nearby, and explore the best times during the year for measuring streamflow. Our analysis shows that data-driven sparse sensing is an effective tool to predict streamflow time-series in poorly gauged watersheds based on very few streamflow measurements. The streamflow in watersheds with high snowmelt and high baseflow can be more easily predicted than in other watersheds. Our analysis also shows that the streamflow measurements taken during periods with large streamflow peaks and variances contain more information and are beneficial for making predictions. We conclude that data-driven sparse sensing can be further used to classify watersheds and to identify the best locations for streamflow gauging.

## 1. Introduction

Many rivers and stream reaches are ungauged or poorly gauged because streamflow measurements are costly and resource intensive. The prediction of streamflow in these watersheds is necessary and beneficial to enhance our understanding of hydrological processes and to improve water resources management. Although the prediction in ungauged basins (PUB) program was initiated by the International Association of Hydrological Sciences (IAHS) in 2003 and many efforts have been made to improve predictions in ungauged watersheds (Hrachowitz

© 2023. The Authors.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs License](https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

**Supervision:** Mitul Luhar, Anthony J. Parolari

**Validation:** Kun Zhang

**Visualization:** Kun Zhang, Mitul Luhar

**Writing – original draft:** Kun Zhang

**Writing – review & editing:** Mitul Luhar, Manuela I. Brunner, Anthony J. Parolari

et al., 2013), PUB remains a challenging problem that requires complex physically-based or data-driven approaches (Besaw et al., 2010; Samaniego et al., 2010).

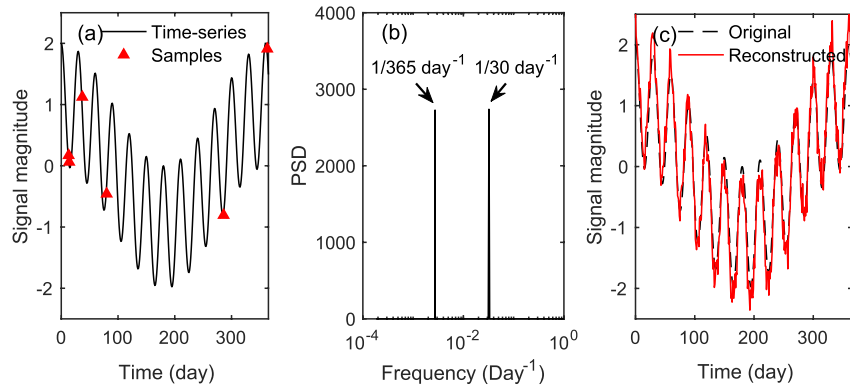
The most commonly used approaches to predict continuous streamflow include distributed physically-based hydrologic models, conceptual and semi-distributed models (e.g., HBV), and data-driven models (e.g., autoregressive moving average ARMA, artificial neural networks ANNs, and long short-term memory neural networks LSTM) (Razavi & Coulibaly, 2013). Data-driven models such as ANNs and LSTM have been increasingly used recently and can outperform some physically-based models in terms of predictive performance if appropriately utilized (Arsenault & Brissette, 2014). When using those physically-based or data-driven approaches, model parameters need to be estimated from independent data sources or calibrated against streamflow data. However, calibration becomes challenging or even impossible in ungauged watersheds in the absence of streamflow information.

Regionalization, which refers to the transfer of hydrologic features, model structures, or model parameters from gauged to ungauged or poorly gauged watersheds, is often needed to make streamflow predictions in ungauged or poorly gauged watersheds. However, regionalization is challenging. First, regionalization often requires watershed attributes (e.g., meteorological, and physiographic information), which may not be available in ungauged or poorly gauged watersheds. Second, due to spatial variability of watershed and streamflow characteristics, the predictive performance of each regionalization approach may vary substantially across watersheds (Yang et al., 2020). The regionalization is often performed on a case-by-case basis, and there is often no universally applicable approach (Arsenault & Brissette, 2014; Arsenault et al., 2019; Razavi & Coulibaly, 2013). Many studies have focused on the regionalization of flow metrics (e.g., peak flows, flow percentiles for flow duration curve) (Carlisle et al., 2010; Sanborn & Bledsoe, 2006) as it requires less parameters to be regionalized. In contrast, regionalizing continuous streamflow requires more parameters and is therefore even more challenging (Razavi & Coulibaly, 2013).

Existing regionalization techniques can be separated into hydrologic model -dependent and -independent approaches (Razavi & Coulibaly, 2013). Hydrologic model-dependent approaches transfer hydrologic model parameters from gauged watersheds to the target watershed for continuous streamflow simulations. Given larger amounts of parameters involved, complicated geospatial (e.g., spatial proximity, interpolation) and/or hydrologic (e.g., hydrological similarity) analyses are normally needed (He et al., 2011). Sometimes, optimization may be also needed (Li et al., 2010). In contrast, the hydrologic model-independent approaches transfer equation structures and relevant coefficients to the target watershed. Examples include the parameters of multiple linear regression models (MLR) (Chiang et al., 2002), or the architecture of ANNs, etc (Besaw et al., 2010).

Signal processing techniques, such as compressed sensing (CS), are potential tools to predict continuous natural signals such as streamflow in poorly gauged watersheds (K. Zhang, Bin Mamoon, et al., 2023). Many natural hydrological and environmental signals (e.g., soil moisture and temperature) are “sparse” (Katul et al., 2007; Parolari et al., 2021; Williams et al., 2018) when projected onto an appropriate space (Ebtehaj et al., 2015). In other words, these signals have a succinct (or “lower-dimensional”) representation in the frequency domain or show correlations in space. In such cases, with a few measurements taken, the signals can be accurately predicted with minimal loss of information (Candès et al., 2006; Donoho, 2006). For example, Figure 1a illustrates a synthetic signal with monthly and annual fluctuations. When transformed to a Fourier frequency space, the signal can be represented by only two active coefficients at frequencies of  $1/30 \text{ day}^{-1}$  and  $1/365 \text{ day}^{-1}$  (Figure 1b). With very few sparse measurements taken (Figure 1a), the signal can be well reconstructed (Figure 1c). This CS approach is used for signal compression and recovery in various contexts including medical imaging (Lustig & Donoho, 2008), radar imaging (L. Zhang et al., 2010), land cover data (Wei et al., 2017), and hydro-environmental signals such as soil moisture (Wu et al., 2014).

However, there are a few constraints in the application of CS for the prediction or regionalization of continuous streamflow in ungauged and poorly gauged watersheds. First, actual streamflow signals are much noisier than the sample signal given above, and it is unclear whether the streamflow signals are sufficiently sparse in Fourier or wavelet frequency spaces to allow for successful prediction. Second, CS requires taking measurements at random times, which is tricky, especially in ungauged watersheds. The technique of data-driven sparse sensing (DSS) proposed by Manohar et al. (2018) is a step toward addressing these deficiencies. First, instead of transforming the signals into Fourier or wavelet frequency spaces, DSS utilizes a Singular Value Decomposition (SVD), a dimensionality reduction tool which underpins techniques such as principal component analysis (PCA) or proper



**Figure 1.** (a) A sample signal with monthly and annual fluctuations. (b) Power spectral density (PSD) of the signal with two active frequencies. (c) The reconstructed time-series using compressed sensing based on sparse measurements.

orthogonal decomposition (POD), to identify the appropriate space where the signals show sparse dynamics. Second, instead of requiring taking measurements at random times, DSS can utilize a QR factorization algorithm with column pivoting to identify the optimal times for sampling. Therefore, DSS provides a potential tool to enable data-driven regionalization and streamflow prediction in poorly gauged watersheds. Specifically, to predict the streamflow in a poorly gauged watershed, streamflow time-series in gauged watersheds nearby can be collected as training data. Then, a SVD can be utilized to identify the lower-dimensional space—or basis—that best represents the streamflow time-series training data for this region. This lower-dimensional space varies for different regions. However, for a specific region, we expect it to serve as a reasonable basis since watersheds within the same geologic or climatic region can share similar hydrologic dynamics due to similar climate (e.g., precipitation) patterns and geospatial (e.g., land cover and soil) properties. Therefore, streamflow signals in poorly gauged watersheds may be predicted accurately by projecting available measurements (e.g., obtained using remotely sensed data from satellites or UAVs) onto the tailored basis functions appropriate for that region (Figure 1).

DSS has been successfully applied in image reconstruction, reconstruction of two-dimensional flow and temperature fields, as well as optimal sensor placement (Manohar et al., 2022; Ohmer et al., 2022). Here, we use DSS for the prediction of hydrologic signals and to identify optimal timings to measure streamflow in poorly gauged watersheds. Specifically, we ask: (a) How well can we predict stream flow time-series in poorly gauged watersheds using DSS? (b) In which types of watersheds does DSS-based regionalization work the best? (c) Is there any spatial pattern of streamflow predictability using DSS and how does it relate to watershed characteristics? And (d) How does the optimal sampling time relate to streamflow regimes? As detailed below, we utilized this approach on 543 daily-scale streamflow time series across the contiguous United States (CONUS) retrieved from the CAMELS data set (Newman et al., 2015). When predicting the streamflow time series in each watershed, we assumed the target watershed to be poorly gauged, and we used SVD to generate a tailored basis for that climate region from streamflow data in nearby gauged watersheds. In addition, a QR factorization algorithm with column pivoting is used to identify optimal timings to measure streamflow in that region. Streamflow time-series predictions from the optimally-timed measurements are compared against predictions made from measurements obtained at random and uniform intervals.

## 2. Data and Methods

### 2.1. Problem Formulation and Data-Driven Sparse Sensing

Hydrologic signals are normally collected as discrete time-series even though they are continuous in time. Here, we consider daily-scale time series of streamflow at a given station  $i$  over the course of a year, arranged into a column vector,  $\mathbf{x}_i \in \mathbb{R}^{365}$ . This vector can be represented by a linear combination of appropriate basis vectors arranged into a matrix  $\Psi \in \mathbb{R}^{365 \times 365}$ , such that

$$\mathbf{x}_i = \Psi \mathbf{s}_i \quad (1)$$

Each column  $\psi_i \in \mathbb{R}^{365}$  in the matrix  $\Psi = [\psi_1, \psi_2, \dots, \psi_{365}]$  is a temporal basis function. The vector  $\mathbf{s}_i \in \mathbb{R}^{365}$  contains amplitude coefficients corresponding to the individual basis functions. Like other natural signals, hydrologic signals can be sparse, meaning that when the discrete time series is represented in terms of an appropriate coordinate system or basis, only a few coefficients in  $\mathbf{s}_i$  have large amplitudes. Often, a generic or universal basis, such as Fourier modes or wavelets, can represent the signal sparsely even without prior knowledge about the properties of the signal (see Figure 1). However, with some physical understanding of the signal to be predicted, or with access to prior data, it is possible to obtain a basis that is tailored for the specific signal to be predicted, that is, a basis in which the signal can be more sparsely represented.

Since we are interested in predicting streamflow in poorly gauged watersheds, we assume that only a subset of measurements in  $\mathbf{x}_i$  can be obtained. In other words, the available data is in the form of a vector  $\mathbf{y}_i \in \mathbb{R}^p$  with  $p \ll 365$ , and

$$\mathbf{y}_i = \mathbf{C}\mathbf{x}_i = \mathbf{C}\Psi\mathbf{s}_i \quad (2)$$

where  $\mathbf{C} \in \mathbb{R}^{p \times 365}$  is a sampling matrix populated with ones and zeros, with non-zero entries in each of the  $p$  rows representing days on which streamflow measurements are available. If the coefficient vector  $\mathbf{s}_i$  can be estimated from the limited measurements  $\mathbf{y}_i$ , the original hydrologic time series  $\mathbf{x}_i$  can be reconstructed through inversion of Equation 1. However, for  $p < 365$ , Equation 2 represents an undetermined system of equations for  $\mathbf{s}_i$ , meaning that there are an infinite number of solutions. The standard least-squares solution to Equation 2 is not sparse and typically results in poor reconstruction due to overfitting. Since we expect natural hydrologic signals to be sparse in an appropriately chosen basis, we seek the sparsest coefficient vector  $\mathbf{s}_i$  that is consistent with the measurements  $\mathbf{y}_i$ , that is,

$$\mathbf{s}_i = \operatorname{argmin}_{\mathbf{s}_i'} \|\mathbf{s}_i'\|_0, \text{ such that } \mathbf{y}_i = \mathbf{C}\Psi\mathbf{s}_i', \quad (3)$$

where  $\|\mathbf{s}_i'\|_0$  represents the  $\ell_0$  pseudonorm of the vector  $\mathbf{s}_i'$ , or the number of non-zero entries in  $\mathbf{s}_i'$ . Unfortunately, the optimization problem in Equation 3 is not tractable since it involves a search over all possible sparse coefficient vectors. However, if the sampling matrix  $\mathbf{C}$  meets certain conditions (Candès et al., 2006; Donoho, 2006), the optimization problem in Equation 3 can be relaxed to a convex  $\ell_1$ -minimization problem of the form

$$\hat{\mathbf{s}}_i = \operatorname{argmin}_{\mathbf{s}_i'} \|\mathbf{s}_i'\|_1, \text{ such that } \mathbf{y}_i = \mathbf{C}\Psi\mathbf{s}_i'. \quad (4)$$

In the equation above,  $\|\mathbf{s}_i'\|_1$  represents the  $\ell_1$ -norm of the vector  $\mathbf{s}_i'$ , or the sum of absolute values of all entries in  $\mathbf{s}_i'$ . Equation 4 can be solved using standard algorithms. We also note that there are several alternative techniques that can be used to solve for the sparsest solution to Equation 2, many of which involve the use of so-called greedy algorithms. For further details, we point the reader to Manohar et al. (2018).

Two questions arise naturally from the preceding discussion. First, what is the most appropriate basis  $\Psi$  for representing streamflow time series? Second, what should be the form of the sampling matrix  $\mathbf{C}$  to ensure that the best possible estimate for  $\mathbf{s}_i$  (and therefore  $\mathbf{x}_i$ ) can be obtained from the measurements  $\mathbf{y}_i$ ? As noted previously, Fourier modes or wavelets can often be used to generate sparse representations of time series data. However, with access to prior time series data, more sophisticated data reduction techniques can be used to create a tailored basis,  $\Psi$ . In addition, to ensure that the sparsest  $\mathbf{s}_i$  with respect to  $\Psi$  can be found, the measurements taken should represent a wide range of the temporal basis functions in  $\Psi$ . In other words, the sampling matrix  $\mathbf{C}$ , which determines when the measurements are taken, should be “incoherent” to the representation basis  $\Psi$  (Candès & Wakin, 2008). For instance, if  $\Psi$  comprises periodic Fourier modes, random sampling is normally the best way to ensure maximum incoherence. However, this requirement limits its application for streamflow prediction because measurements are often taken periodically, and it may not be practical to take measurements at random timings.

To create the tailored basis, we assume the availability of a library of time series data  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$ , where different columns in the matrix  $\mathbf{X} \in \mathbb{R}^{365 \times n}$  can represent data available from different years at the same station and/or data available from measurement stations in close geographic proximity or with shared hydrologic characteristics. A singular value decomposition (SVD) of the data set

$$\mathbf{X} = \Psi\mathbf{\Sigma}\mathbf{V}^T \quad (5)$$

is then used to generate a set of orthonormal temporal basis functions (left singular vectors in  $\Psi = [\psi_1, \psi_2, \dots, \psi_{365}] \in \mathbb{R}^{365 \times 365}$ ) and spatiotemporal correlation functions across the  $n$  different time series (right singular vectors in  $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n] \in \mathbb{R}^{n \times n}$ ). The rectangular diagonal matrix  $\mathbf{\Sigma} \in \mathbb{R}^{365 \times n}$  contains the singular

values,  $[\sigma_1, \sigma_2, \dots]$ , which are sorted such that  $\sigma_1 > \sigma_2 > \dots$ . The SVD yields the *optimal* least-squares approximation to the data at a given rank. In other words,  $\mathbf{X} \approx \boldsymbol{\Psi}_1 \sigma_1 \mathbf{V}_1^T$  is the best rank-1 approximation to the data set  $\mathbf{X}$ . The best rank  $r$  approximation is

$$\mathbf{X} \approx \boldsymbol{\Psi}_r \boldsymbol{\Sigma}_r \mathbf{V}_r^T, \quad (6)$$

where the matrices  $\boldsymbol{\Psi}_r$  and  $\mathbf{V}_r$  contain the first  $r$  columns of  $\boldsymbol{\Psi}$  and  $\mathbf{V}$ , respectively, and  $\boldsymbol{\Sigma}_r$  is a diagonal matrix containing the first  $r \times r$  block of  $\boldsymbol{\Sigma}$ . Since the SVD yields an optimal orthonormal temporal basis (in a least squares sense) for the data in  $\mathbf{X}$ , the target streamflow time series can be represented as

$$\mathbf{x}_i = \boldsymbol{\Psi} \mathbf{a}_i \quad (7)$$

where  $\mathbf{a}_i$  is a vector representing coefficients corresponding to the different basis functions in  $\boldsymbol{\Psi}$ . Moreover, the coefficient vector  $\mathbf{a}_i$  can be estimated from the measurements  $\mathbf{y}_i$  as

$$\mathbf{y}_i = \mathbf{C} \mathbf{x}_i \approx \mathbf{C} \boldsymbol{\Psi}_r \hat{\mathbf{a}}_i \rightarrow \hat{\mathbf{a}}_i = (\mathbf{C} \boldsymbol{\Psi}_r)^+ \mathbf{y}_i \quad (8)$$

to yield the following estimate for the target time series

$$\hat{\mathbf{x}}_i \approx \boldsymbol{\Psi}_r \hat{\mathbf{a}}_i. \quad (9)$$

In Equation 8 above, the superscript  $+$  represents a Moore-Penrose pseudoinverse. Importantly, if the rank  $r$  is chosen such that  $r \leq p$  where  $p$  is the number of measurements in  $\mathbf{y}_i$ , Equation 8 is no longer undetermined (c.f., Equation 2). Thus, this approach involving a low-rank approximation to a tailored basis can be more efficient as it solves a standard least-squares problem instead of the convex optimization problem in Equation 4.

In addition, instead of taking measurements randomly, the  $r = p$  sampling points, referring to the times when streamflow measurements are recorded, can be optimized to best sample the  $r$  basis modes  $\boldsymbol{\Psi}_r$ . These optimal sampling points can be obtained using QR factorization with column pivoting (Manohar et al., 2018):

$$\boldsymbol{\Psi}_r^T \mathbf{C}^T = \mathbf{Q} \mathbf{R}, \quad (10)$$

where  $\mathbf{C}$  is a column permutation matrix,  $\mathbf{Q}$  is a unitary matrix, and  $\mathbf{R}$  is an upper-triangular matrix, all of which are obtained from the matrix  $\boldsymbol{\Psi}_r$ . Through an iterative process to maximize the  $\ell_2$  norm within all modes in the library, the QR factorization yields  $r$  column indices in  $\mathbf{C}$  that best sample the  $r$  basis modes  $\boldsymbol{\Psi}_r$ . These  $r$  columns represent the optimal timings to take measurements.

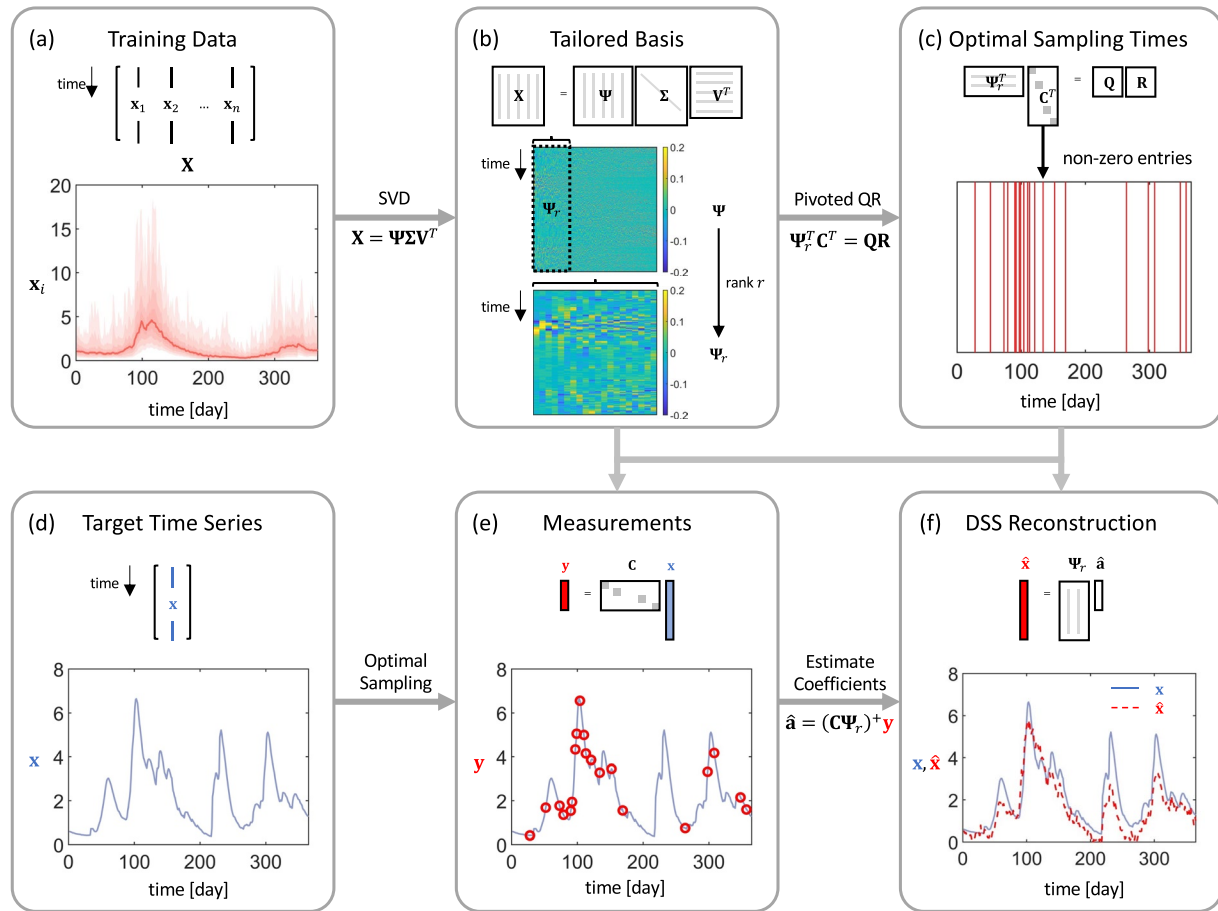
In summary, the DSS framework developed here obtains a tailored coordinate system or basis via an SVD from a training set or library of streamflow data. Furthermore, the optimal timings to take measurements are obtained using QR factorization of the tailored basis. If a small number of measurements can be taken at these optimal sampling times for ungauged watersheds, then the full streamflow time-series can be reconstructed or predicted (Figure 2). In this study, we assume the target watersheds to be poorly gauged, with sparse measurements extracted from the full data set. In real applications, sparse measurements in the watersheds could be obtained from remotely sensed data, from either satellites, for example, Surface Water and Ocean Topography, SWOT, or UAVs.

## 2.2. CAMELS Data Set

We collected daily-scale streamflow data at 671 gauges from 1981 to 2010 throughout CONUS from the CAMELS data set (Newman et al., 2015). The gauges span across the 18 hydrologic units (i.e., 2-digit HUCs) in the US, and they were further categorized into nine geologic regions (Figures 3a and 4a). The CAMELS data set includes streamflow time-series and hydrometeorological and geophysical watershed attributes for 671 watersheds across CONUS. For this analysis, we chose 543 watersheds out of the watershed pool by filtering out the gauges with continuous missing data that amounted to more than 1% of the time series (Figure 3a). There are nested and non-nested watersheds. These watersheds cover different geographic and climatic regions, and watershed sizes range from 4 to 14,269 km<sup>2</sup> (Figure 3b).

We normalized the streamflow by watershed area into specific streamflow (mm/d) for ease of prediction and cross-watershed comparison. In addition to continuous streamflow data, we also retrieved hydrometeorological and geophysical watershed attributes in these 543 gauges, including the fraction of precipitation falling as snow (referred to as snow fraction), baseflow index, and fractions of sand and clay. These watershed attributes were



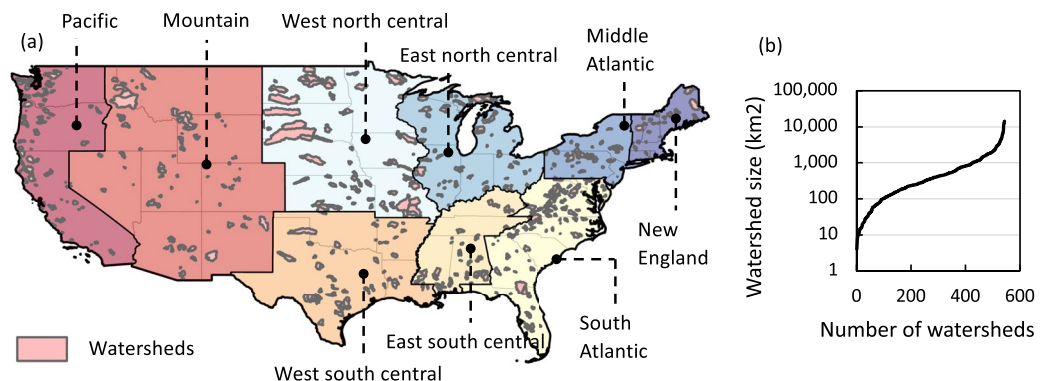


**Figure 2.** Methodologies and roadmaps of data-driven compressed sensing (DSS). The streamflow training data set  $X$  (a) (in nearby gauges) is used to identify the tailored basis through singular value decomposition (SVD) (b) and determine the optimal sampling times through QR factorization (c). With sparse measurements taken in the optimal times (d–e), the streamflow in the target gauge is reconstructed/predicted (f).

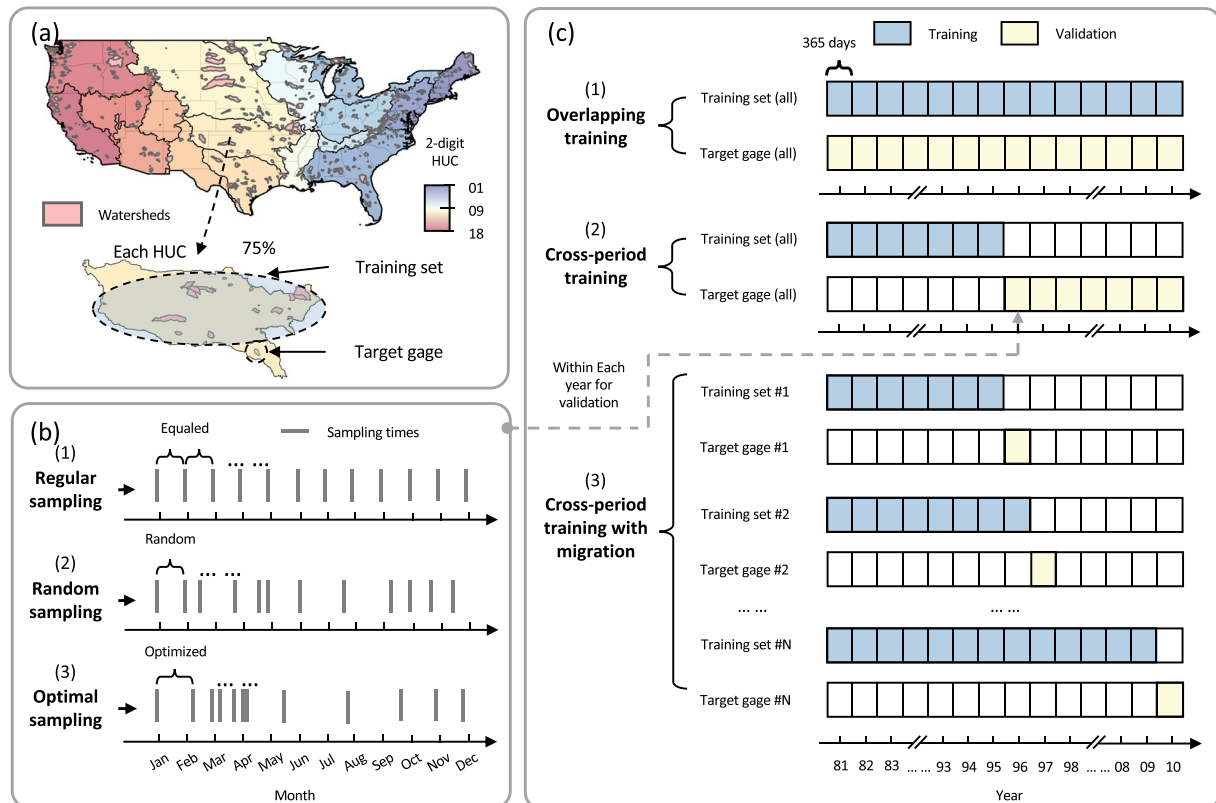
not used in the streamflow prediction but to provide some physical interpretation of the resulting prediction efficiency.

### 2.3. Scenarios of Analysis

We used the DSS to predict daily streamflow during 1981–2010 in each of the 543 gauges. MATLAB R2021a was used for the analysis. The MATLAB code for DSS was initially retrieved from the Github repository by



**Figure 3.** (a) Locations of 543 watersheds in the CAMELS data set analyzed in this study. (b) Distribution of watershed size.



**Figure 4.** (a) Selection of training set in each hydrologic unit (2-digit HUC). (b) Comparison of three sampling schemes in the watersheds with streamflow to be predicted. (c) Comparison of three training-validation strategies.

Krithika Manohar ([https://github.com/kmanohar/SSPOR\\_pub](https://github.com/kmanohar/SSPOR_pub)) and customized for this study. To fit the prediction model for each target gauge, we randomly selected 75% of the gauges in the same HUC (with the target gauge excluded) as the training set to construct the tailored bases (Figure 4a). To make sure that the prediction performance did not change by varying the training set, we repeated the selection of training gauges 10 times when predicting each target gauge. The final predicted time series is derived by averaging over the 10 predicted signals. A training-to-validation ratio of 75:25 is within the normal range adopted in most data-driven studies, especially in hydrology and environmental science (Nolan et al., 2015; Reza Md Towfiqul Islam et al., 2021). In addition, we checked that repeating the selection of training gauges 10 times is sufficient to cover most of the combinations in the training gauges.

Sparse measurements need to be taken in the target watersheds in order to predict streamflow. We tested three different sampling schemes and compared the prediction efficiency across these schemes (Figure 4b). First, a regular sampling scheme was adopted to mimic satellite overpasses, meaning that the measurements for signal reconstruction were taken periodically at a constant interval. The exact interval varies depending on the number of measurements ( $p = r$ ) to be acquired over the course of a year. The interval decreases proportionally when more measurements are taken. Second, a random sampling scheme was adopted to ensure incoherence between the measurement and the signal, meaning that the measurements were taken randomly within a year. Third, an optimal sampling scheme was adopted, meaning that the optimal times for measurement were determined using QR factorization (Equation 10) (Figure 4b).

In addition, we tested three different strategies for determining which period of streamflow data were used for training (i.e., for identifying the tailored basis and optimal sampling times from the training gauges) and validation (i.e., for testing time series predictions for the target gauge) (Figure 4c). We further compared the prediction efficiency across these strategies. First, an overlapping training-validation approach was adopted. The training and validation sets of streamflow data were both obtained from the period 1981–2010 (30 years) (training and validation from different gauges). Second, a cross-period training-validation approach was adopted. The training

data set was obtained from streamflow time series in the period 1981–1995 (15 years) in the training gauges, and the validation set was done on the period 1996–2010 (another 15 years) for the target gauge. The third method is similar to the second method because it does not involve an overlap between the training and validation datasets. However, this approach makes use of a migrating window of training data. For example, we obtained the training data set from 1980 to 1995 to predict streamflow data for the year 1996 at the target gauges. Then, data for the year 1996 at the training gauges were included into the training set when predicting streamflow data for the year 1997 at the target gauge, and so on. The first overlapping approach is expected to perform better than the other two approaches because the training data set includes information from the period for which predictions are made for the target gauges. In other words, this approach represents *offline* or a posteriori reconstruction of streamflow time series in poorly gauged watersheds. The second and third approaches could potentially be used for *online* reconstruction (and potentially, future prediction) because they only make use of historic (i.e., existing) streamflow measurements for training. Indeed, the third approach is perhaps the most realistic in practice and is expected to perform better than the second approach since it continuously updates the training data set as more streamflow data become available. With continuous updates of the training data set, the non-stationarity in the streamflow data, which could be caused by a change in climate and watershed characteristics and changes in flow measurement settings (e.g., sensor replacement, re-calibration of sensor), can be better handled, and the optimal times for streamflow measurements can be better identified. Note that, in all three cases, no data from the target gauge are used for identifying the tailored basis or optimal sampling times.

The Nash Sutcliffe Efficiency (NSE) (Nash & Sutcliffe, 1970) (Equation 11) was used to quantify the goodness of fit between the measured ( $\mathbf{x}_i$ ) and reconstructed/predicted time series of streamflow ( $\hat{\mathbf{x}}_i$ ):

$$\text{NSE} = 1 - \frac{\|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2}{\|\mathbf{x}_i - \bar{\mathbf{x}}_i\|_2^2} \quad (11)$$

NSE determines the relative magnitude of the residual variance between predictions and measurements, that is,  $\|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2$ , compared to the measured data variance, that is,  $\|\mathbf{x}_i - \bar{\mathbf{x}}_i\|_2^2$ . Since NSE is sensitive to extreme values, the modified NSE ( $\text{NSE}_m$ ) (Krause et al., 2005) was also calculated as a complementary indicator to represent the goodness of fit more comprehensively for both high and low flows:

$$\text{NSE}_m = 1 - \frac{\|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_1}{\|\mathbf{x}_i - \bar{\mathbf{x}}_i\|_1} \quad (12)$$

We calculated NSE and  $\text{NSE}_m$  values for different numbers of measurements taken to evaluate whether DSS can effectively predict streamflow time series in ungauged or poorly gauged watersheds. In addition, we evaluated spatial distributions in NSE and  $\text{NSE}_m$  across CONUS and analyzed the relationship between NSE and  $\text{NSE}_m$  values and watershed hydrometeorological and geophysical attributes to better understand what types of watersheds are more amenable for streamflow prediction using DSS. In addition, we compared the optimal timings for measurement (different months in each year) across watersheds and tried to identify relationships between these optimal sampling times and streamflow regime characteristics (e.g., peak streamflow and streamflow variance in each month) across watersheds. It should be noted that two watershed classification schemes were used, including (a) 18 HUCs and (b) nine geologic regions. The watersheds were classified into 18 HUCs when the training and prediction of streamflow were performed. In addition, the watersheds were classified into nine geologic regions when the results of streamflow prediction were discussed for ease of interpretation.

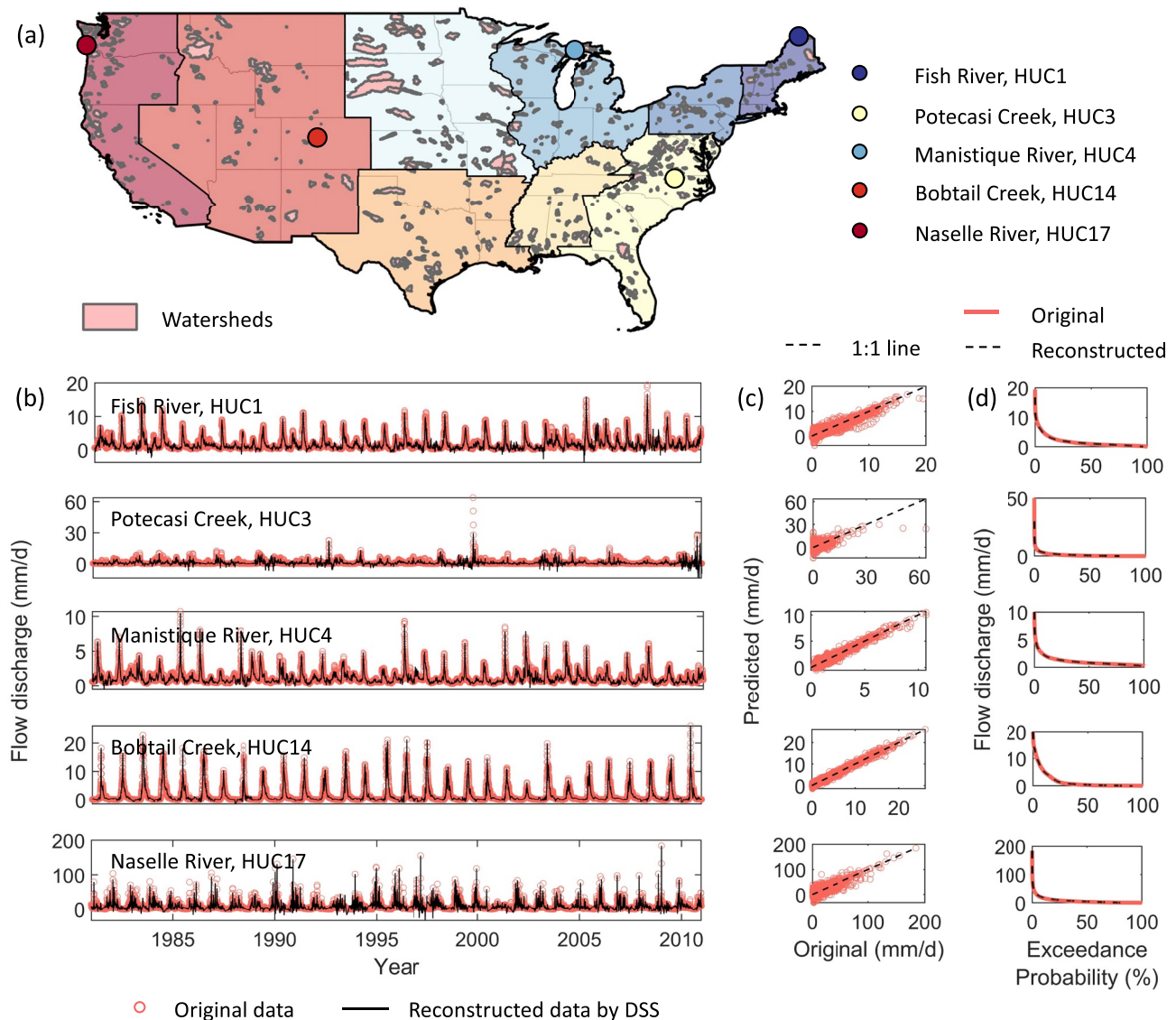
### 3. Results

#### 3.1. Prediction Efficiency of Different Sampling Schemes

DSS can effectively predict streamflow in poorly gauged watersheds if the optimal sampling scheme is adopted. Figure 5 shows the predicted streamflow in the period 1981 to 2010 in 5 representative gauges using DSS with an optimal sampling scheme (with 50 measurements taken, corresponding to 13.7% of daily data). Although the high flows at the Potecasi Creek were underestimated, the technique performed well in predicting the whole time-series (Figure 5b), including both low and high flows (Figure 5c), and in reproducing flow duration curves in all five gauges (Figure 5d).

DSS failed to predict streamflow with the regular and random sampling schemes. With an optimal sampling scheme and the number of streamflow measurements ( $p$ ) varying from 2 to 75 in a year, the median NSE ( $\text{NSE}_m$ )



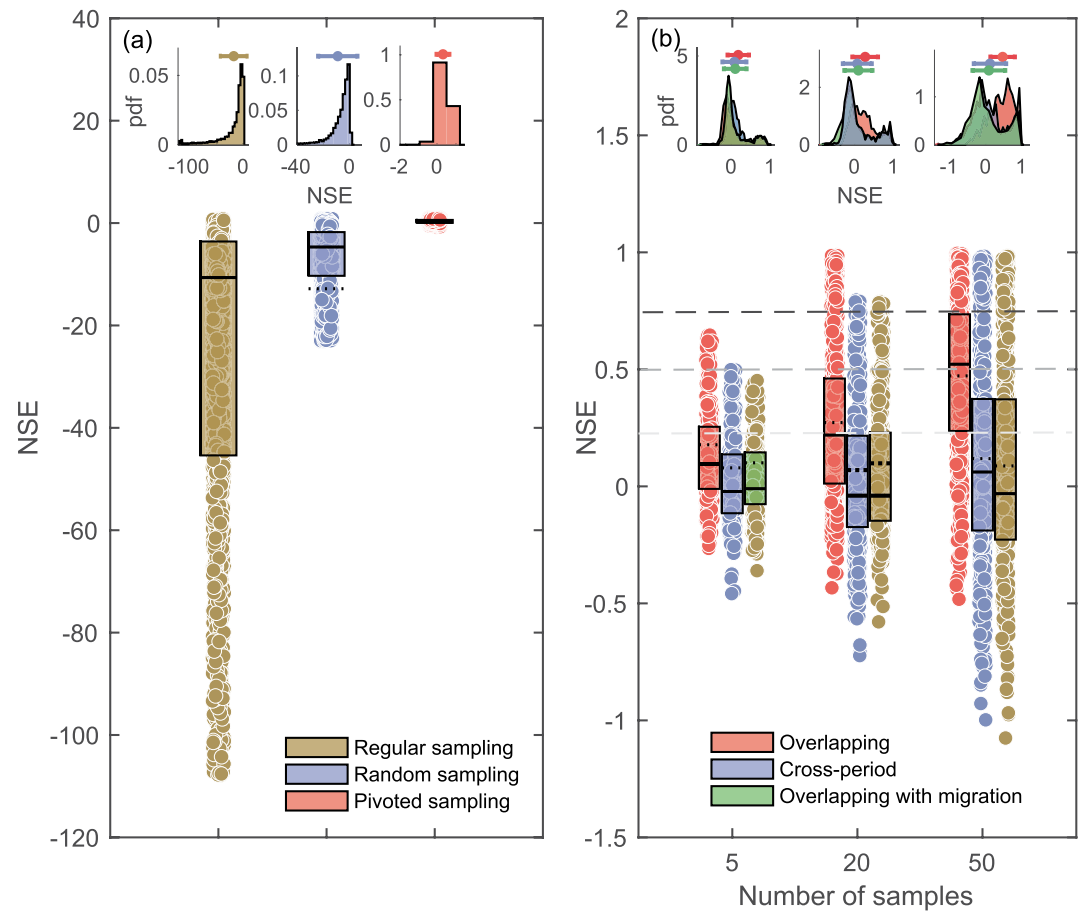


**Figure 5.** Predicted flow discharge time-series in selected watersheds using the scenarios with overlapping training and that considered 50 measurements in prediction as an example. DSS refers to data-driven sparse sensing. (a) Spatial distribution of the selected watersheds and the watersheds for comparison, that is, Fish River in HUC1, Potecasi Creek in HUC3, Manistique River in HUC4, Bobtail Creek in HUC14, and Naselle River in HUC17. (b) Observed versus predicted time series. (c) Scatter plots of observed versus predicted daily streamflow. (d) Observed versus predicted flow duration curves showing flow discharge versus exceedance probability.

was 0.25 (0.13) (Figure 6a and Figure S1 in Supporting Information S1). However, with a regular sampling scheme and the same number of daily streamflow measurements, the median NSE ( $NSE_m$ ) value was  $-10.6$  ( $-2.6$ ). With a random sampling scheme, the median NSE ( $NSE_m$ ) value was  $-4.7$  ( $-1.6$ ) (Figure 6a and Figure S1 in Supporting Information S1). The NSE and  $NSE_m$  represent the relative magnitude of residual variance compared to the measured data variance. Thus, the near-zero and negative NSE and  $NSE_m$  values obtained by the regular and random samplings mean that the DSS failed because the model predictions are not as accurate as the mean of the observed streamflow data. As a result, the following sections only consider reconstructions obtained using the optimal sampling scheme.

### 3.2. Prediction Efficiency of Different Training-Validation Approaches

The prediction effectiveness, represented by NSE and  $NSE_m$  values, increased with the number of measurements considered. With overlapping training and validation datasets  $p = 5$  measurements over a year (1.4% of data), the



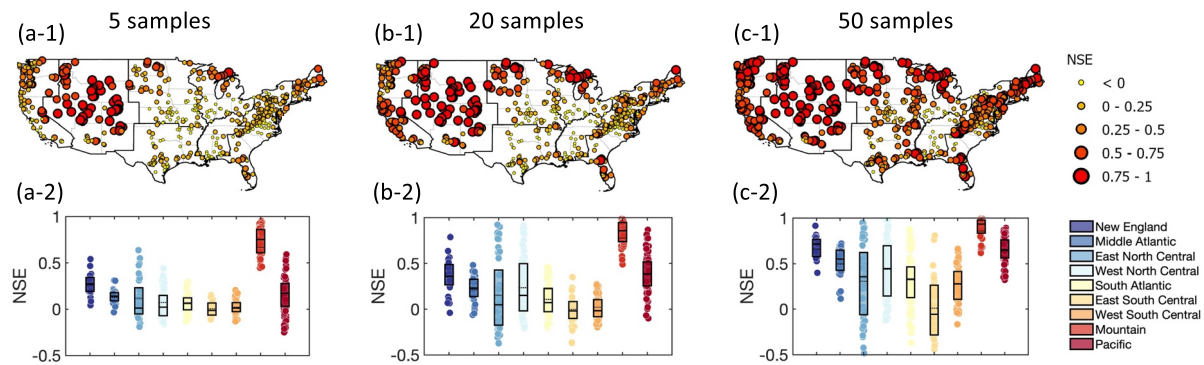
**Figure 6.** Nash-Sutcliffe efficiency (NSE) under (a) different sampling schemes and (b) different training-validation strategies and number of measurements considered for prediction. These plots only show results from overlapping training-validation datasets. In panel (a), the results obtained using different numbers of measurements (from 2 to 75 in a year) were combined. In panel (b), the measurements were taken based on the optimal sampling scheme, and the three gray solid lines from lighter to darker shades refer to NSE and/or  $NSE_m$  of 0.25, 0.5 and 0.75.

median NSE ( $NSE_m$ ) value was 0.09 (0.08) across all watersheds. With 50 measurements (13.7% of data), the median NSE ( $NSE_m$ ) increased to 0.52 (0.26) (Figure 6b and Figure S1 in Supporting Information S1).

Among the three training-validation strategies, the prediction efficiency was greater when the training and validation took place during the same period or overlapped (median NSE of 0.09–0.52; median  $NSE_m$  of 0.08–0.26) compared to the cross-period training and validation (median NSE of –0.02–0.06; median  $NSE_m$  of –0.01–0.02). DSS with the overlapping training was more sensitive to the number of measurements considered than DSS with cross-period training. For the overlapping training, when the number of measurements taken increased from 5 (1.4% of data) to 50 (13.7% of data), the median NSE ( $NSE_m$ ) increased from 0.09 to 0.52 (from 0.08 to 0.26). However, for the cross-period training, the median NSE (0.01–0.04) and  $NSE_m$  (–0.01–0.07) remained almost unchanged (Figure 6b and Figure S1 in Supporting Information S1).

### 3.3. Spatial Variability of Prediction Efficiency

The prediction efficiency of streamflow using DSS varied spatially. Comparatively, the prediction efficiency was the highest in the Rocky Mountain region in the West. With 5 measurements considered in prediction, corresponding to 1.37% of the data, the streamflow in most of the gauges in this region can be predicted with a median NSE > 0.75 ( $NSE_m$  > 0.5) (Figure 8 and Figures S1 and S3 in Supporting Information S1). In the Mountain region, the median NSE ( $NSE_m$ ) increased from 0.75 to 0.93 (from 0.56 to 0.76) when the number of measurements considered increased from 5 to 50 (Figure 7). The spatial distribution of  $NSE_m$  showed similar patterns



**Figure 7.** Spatial distribution of median Nash-Sutcliffe efficiency (NSE) over the validation period using data-driven sparse sensing with overlapping training and different numbers of measurements considered for prediction (a-1 to a-2) 5 measurements; (b-1 to b-2) 20 measurements; (c-1 to c-2) 50 measurements.

as the one for NSE (Figure S3 in Supporting Information S1), and those obtained from cross-period validations also showed similar patterns as the one obtained by overlapping validation (Figures S2 and S3 in Supporting Information S1).

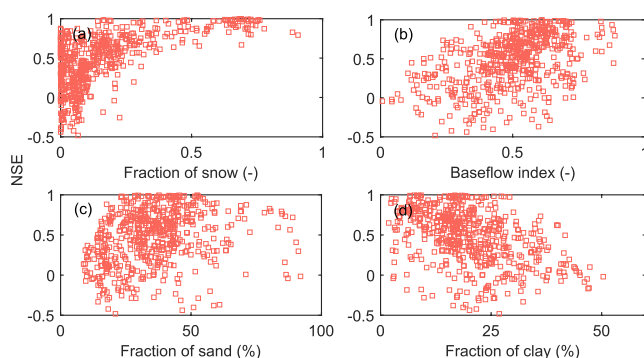
The streamflow in the New England, Middle Atlantic (i.e., northeast coast), and the Pacific (i.e., west coast) regions was not well predicted with only 5 measurements. But the streamflow in these regions showed a substantial increase in prediction efficiency with more measurements taken (Figure 7, Figures S2–S4 in Supporting Information S1). When the number of measurements increased from 5 to 50, in the New England region, the median NSE ( $NSE_m$ ) increased from 0.27 to 0.71 (from 0.17 to 0.43); in the Middle Atlantic region, the median NSE ( $NSE_m$ ) increased from 0.14 to 0.55 (from 0.11 to 0.32); and in the Pacific region, the median NSE ( $NSE_m$ ) increased from 0.17 to 0.65 (from 0.15 to 0.41) (Figure 7, Figures S2–S4 in Supporting Information S1).

The prediction efficiency was relatively low and did not significantly increase with more measurements taken in the central United States, especially the East-South and West-South Central regions. For most watersheds in these regions, streamflow was not well predicted even with 50 measurements considered for prediction (Figure 7, Figures S2–S4 in Supporting Information S1). More measurements may be needed to obtain a reasonable prediction in these regions. It should be noted that although the NSE values in these regions were lower, the method can still capture the general pattern of streamflow regimes.

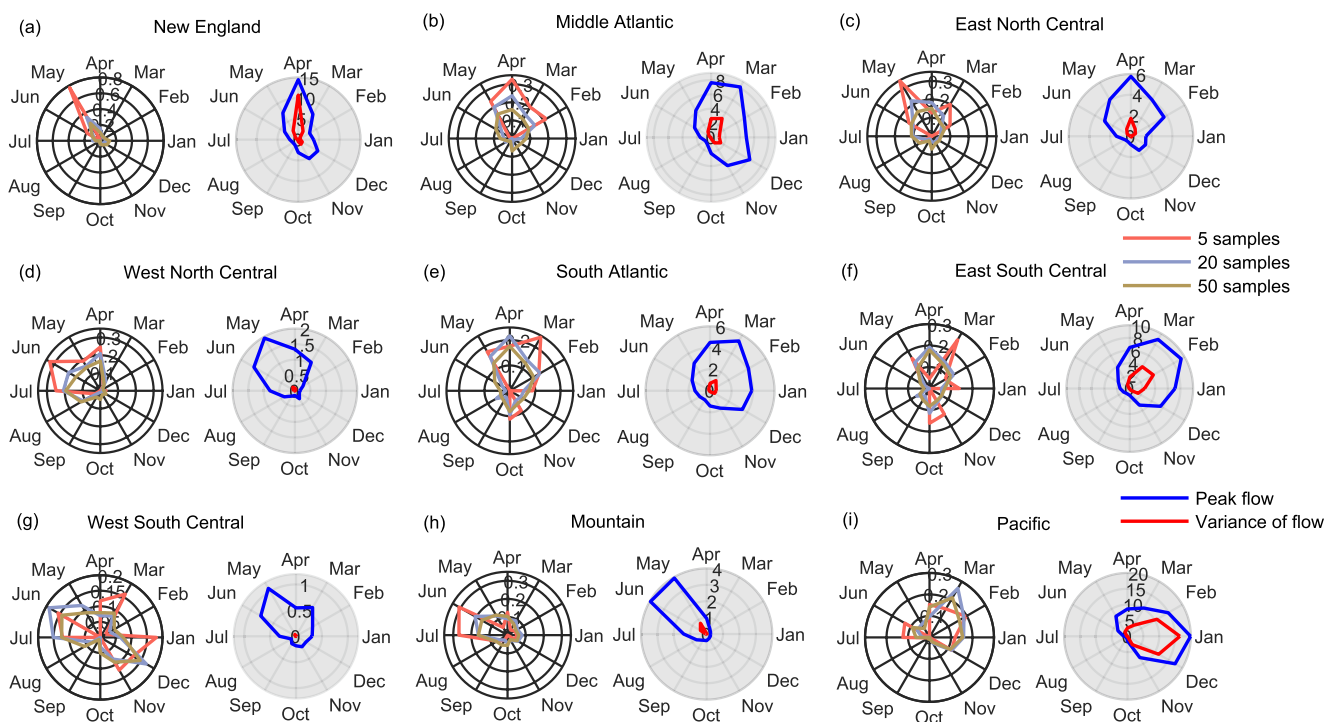
### 3.4. Relationship Between Prediction Efficiency and Watershed Geophysical Properties

The prediction efficiency using DSS was found to be greater in watersheds with a higher snow fraction than in rainfall-dominated watersheds. Although most of the watersheds have a low snow fraction (e.g., 0–0.2), we find a clear positive relationship between prediction efficiency and snow fraction ( $R^2 = 0.387$  for NSE;  $R^2 = 0.441$  for  $NSE_m$ ). For watersheds with snow fractions  $< 0.2$ , NSE values were scattered between 0 and 1 with most of the watersheds having efficiencies around 0–0.5. In contrast, NSE values were  $> 0.6$  for watersheds with a snow fraction  $> 0.5$  (Figure 8a and Figure S5 in Supporting Information S1). The prediction efficiency was high in some watersheds with low snow fractions possibly because the snow fraction is averaged spatially over the whole watershed area and temporally over each year while the streamflow regime is more dominated by flow coming from the upland and during months with snowmelt. Similar relationships were found between NSE values and the elevation of watersheds, the NSE value was higher for watersheds with higher elevations, corresponding to watersheds with more snowmelt (Figure S6 in Supporting Information S1).

Our results also show that the prediction efficiency using DSS is higher in watersheds with a higher baseflow index (Figure 8b and Figure S5 in Supporting Information S1). In addition, the prediction efficiency using DSS is slightly higher in watersheds with a higher sand fraction (Figure 8c and



**Figure 8.** Relationship between (a) fraction of precipitation falling as snow, (b) baseflow index, (c) sand fraction, and (d) clay fraction and median Nash-Sutcliffe efficiency (NSE) over the validation period using the scenarios that considered 50 measurements in prediction as an example.



**Figure 9.** (a–i) Distribution of optimal timing for taking measurements throughout the year relative to the characteristics of flow regimes characterized by monthly median peak (mm/d) and variance of streamflow (mm<sup>2</sup>/d<sup>2</sup>). The polar plots in white represent the optimal timing for taking measurements, while those in gray represent the characteristics of the flow regimes.

Figure S5 in Supporting Information S1) and a lower clay fraction (Figure 8d and Figure S5 in Supporting Information S1). No statistically strong relationship was observed between prediction efficiency and the other 47 watershed characteristics (Figures S6–S11 in Supporting Information S1).

### 3.5. Optimal Sampling Time

The optimal timings to take measurements obtained from DSS were distinct across regions and showed seasonal patterns. In most of the regions, the best sampling time was from late winter to early summer (i.e., January to July), represented by the large distances between the points and the origin during these months in the white polar diagrams in Figure 9. Measurements taken during summer and fall (e.g., August to December) add less predictive performance (Figure 9).

The variability of the ideal timing for measurement also differed among regions. The optimal timing for measurement was more concentrated in time in New England, Middle Atlantic, East North Central, West North Central, and Mountain regions, given that the polygons formed by the points protrude in a specific direction in the white polar diagrams in Figure 9. This trend is especially obvious when only 5 measurements are taken. Ideal sampling times are in May in the New England region, during February to May in the Middle Atlantic and East North Central regions, during April to July in the West North Central region, and during June to July in the Mountain region (Figures 9a–9d, and 9h).

The optimal timings to take measurements show some correlation with the flow regime, as characterized by the peak and variance of streamflow. Overall, our findings show that measurements are ideally taken during or approximately 1 month after the periods with greater peak flows and variances (Figure 9). For example, for the New England area, it is ideal to take measurements in May when the streamflow shows high peaks and pronounced variance (Figure 9a). In the Mountain area, sampling is most beneficial in June to July, the season with the highest peaks and flow variance (Figure 9h).



## 4. Discussion

### 4.1. Low-Dimensional Watershed Dynamics

The prediction of streamflow using data-driven sparse sensing (DSS) relies on the shared meteorological, hydrological, and/or geological properties between the watersheds in the training set and the target watershed for which predictions are made. Thus, the spatial variation of streamflow prediction efficiency and its relationship with watershed attributes, for example, snow fraction, baseflow index, fractions of sand and clay, observed in this study can be attributed to the presence or absence of shared low-dimensional dynamics between the watersheds. Streamflow in watersheds with high snow fractions such as the Rocky Mountains can be most efficiently predicted with a few measurements because snowmelt-dominated watersheds are less affected by temporal and spatial variability in precipitation than rainfall-driven watersheds. Better predictability in snowmelt-dominated watersheds is consistent with many existing studies relying on physically-based (Knoben et al., 2020; Pool et al., 2019; Pool & Seibert, 2021) and data-driven models (Pham et al., 2021). Flow regimes in these regions show more consistent seasonal patterns across watersheds (Brunner et al., 2020). Similarly, streamflow in watersheds with higher baseflow indices can be more efficiently predicted than streamflow in watersheds with lower baseflow, possibly because baseflow-dominated watersheds are more dependent on subsurface storage than precipitation which is more variable. The streamflow in these baseflow-dominated watersheds is also less affected by precipitation variability. In addition, streamflow in watersheds with higher sand fractions but lower clay fractions can be more efficiently predicted than streamflow in watersheds with low sand and high clay fractions. This might be related to their higher hydrologic connectivity (Tetzlaff et al., 2009). In summary, watersheds with temporally consistent (less intermittent) streamflow signals filter rainfall variability (i.e., randomness) and are more dominated by storage processes. Therefore, streamflow in storage-dominated systems is more predictable than streamflow in rainfall-dominated, higher-dimensional systems.

### 4.2. Optimal Sampling Times in Relation to Flow Regimes

DSS can predict streamflow effectively only if sampling times are optimized. Prediction does not work well with fixed intervals and random sampling. This corroborates findings by Manohar et al. (2018) who applied DSS to predict 2-dimensional flow fields. However, this observation is different from that obtained by Pool et al. (2017), who tested the impact of streamflow measurement times on the calibration of physically-based models and found that the model's performance in predicting flow duration curves of streamflow was not very sensitive to the sampling schedule.

The optimal sampling times for DSS are related to the flow regime. DSS performed best with sampling during, or approximately 1 month after, periods with greater peak flows and variances. This finding might be related to the wider range of streamflow magnitudes captured during these periods, that is, these periods contain more "information" than other periods with more consistent regimes. As observed by Etter et al. (2018), the streamflow prediction performance of a model can be higher if trained on observations distributed throughout the year. Although Pool et al. (2017) found the exact sampling time was not crucial for the performance of streamflow prediction (as mentioned above), they recognized that the streamflow can be better predicted if the measurements represent the full range of streamflow magnitudes.

### 4.3. Limitation and Potential Extension for Data-Driven Sparse Sensing

In this study, we show that DSS is a promising technique for the prediction of streamflow in poorly gauged watersheds. With the optimal sampling times identified, the prediction efficiency of DSS significantly improves compared to regular (i.e., fixed interval) and random sampling. Although DSS requires data for training, its requirement in training set is much smaller than other machine learning techniques and it can predict the full streamflow time-series with sparse measurements. Sparse measurements have been used to gain information for streamflow prediction in poorly gauged watersheds in previous studies, but mostly in a physically-based modeling context (Pool et al., 2017, 2019; Pool & Seibert, 2021). The present DSS approach replaces physically-based models with features extracted from existing data. In many watersheds (e.g., snowmelt-dominant watersheds; Rocky Mountain region), streamflow can be reasonably predicted with just 2–5 measurements (0.5%–1.4% of data) per year. DSS can not only predict the low flows (or flow medians, represented by the  $NSE_m$ ) but also the high flows (or high flow percentiles, represented by the  $NSE$ ). In addition, DSS can identify the optimal timings



for taking measurements. This could be beneficial for applied streamflow prediction when new gauges are to be planned or existing gauges are to be relocated.

However, DSS has some limitations. First, DSS showed limited effectiveness in predicting streamflow in rainfall-dominated watersheds. This is because the construction of the tailored basis (through SVD) depends on the shared features between the target gauge and the watersheds in the training set. These shared features are less significant across years and sites in rainfall-dominated watersheds. To predict streamflow in these watersheds, higher-frequency training data in watersheds closer to the target watershed may be needed. Future extensions of the technique developed here can also involve the inclusion of precipitation data for streamflow prediction in rainfall-dominated systems. In addition, other hydro-environmental data (e.g., soil moisture, water quality, peaks of streamflow) and forecasts (e.g., short-term precipitation forecast) can be integrated into the training set to increase the possibility of capturing the dominant streamflow features (Patil & Ramsankaran, 2017; Wyatt et al., 2020). Second, DSS needs some measurements in the target watersheds for prediction. In practice, it is not always feasible to obtain ground-based streamflow data in ungauged and poorly gauged watersheds. To apply DSS in totally ungauged watersheds, computational simulation data can be used for training purposes (Jayaraman & Al Mamun, 2020). With optimal times for measurement identified from training, remotely-sensed data, for example, derived from the Surface Water Ocean Topography (SWOT) product (Biancamaria et al., 2016) or from measurements obtained by flying unmanned aerial vehicles (UAVs) (Eltner et al., 2020), could be utilized to retrieve sparse streamflow information. Crowdsourced data from citizen science projects could be another way to get sparse streamflow measurements (Buytaert et al., 2014; Dickinson et al., 2010). Each of these data sources has some limitations. Remotely sensed data often have low spatial and temporal resolutions, and crowdsourced data can be irregular in availability and accuracy. Thus, information from multiple datasets might have to be assimilated. Such multiple-data or data-model fusion can be of great benefit for the characterization and prediction of nonlinear dynamics in hydrologic or other geophysical systems (Gettelman et al., 2022; Mazzoleni et al., 2017). Although many of these data sources require rating curves (e.g., water level-discharge relationship) to estimate discharge, obtaining discharge without field measurements has become possible with the development of many computational inverse methods to build rating curves from remotely sensed data (Gleason & Durand, 2020; Mahdadi et al., 2021; Pan et al., 2016).

There is great potential to extend the application of DSS beyond the temporal reconstruction of streamflow. The present paper focuses on temporal reconstruction of streamflow from a small number of measurements made over the course of a year. The framework developed here can also be used to identify optimal watersheds to gauge permanently in a given region, such that measurements from these gauged watersheds can be used to predict flows in ungauged watersheds. In addition, DSS can be used to optimize the spatial placement of gauges along streams. More specifically, process-based models, validated with ground-based measurements, can be used to calculate high-resolution spatial distributions of flow discharge along streams. Based on the simulated 2-dimensional flow discharge dynamics varying in space and time, DSS can be utilized to identify the optimal spatial points that can best represent the low-dimensional dynamics within the streams. These points would be the optimal locations for monitoring stations. Similar applications have been successfully used to identify optimal sensor locations for small-scale fluid dynamics, watershed-scale groundwater measurements (Ohmer et al., 2022), and global-scale water temperature measurements (Manohar et al., 2018). In addition, DSS can be also utilized for watershed classification. More specifically, instead of identifying the tailored basis that best preserves the variance in the training set as we did in this study, we can identify a basis in which the training set (e.g., streamflow) can be best classified into clusters. Then, we can project the sparse streamflow data in the target watershed to that basis and identify which cluster it most likely belongs to (Bai et al., 2017; Brunton et al., 2016).

## 5. Conclusions

Predicting streamflow in ungauged and poorly gauged watersheds is essential to develop water management strategies in regions with limited monitoring. This study used a novel data-driven signal processing approach, termed data-driven sparse sensing (DSS), to predict streamflow in poorly gauged watersheds across the contiguous United States (CONUS). Our results demonstrate that DSS is a promising approach to predict streamflow signals in poorly gauged watersheds by exploiting known patterns in streamflow data. DSS can effectively predict streamflow with sparse measurements when optimal sampling times are identified.

The prediction efficiency varied spatially. Streamflow in snowmelt-dominated watersheds and watersheds with a higher baseflow index was better predicted than streamflow in rainfall-dominated watersheds. The good

prediction performance in snowmelt-dominated regions may be related to the strong low-dimensional dynamics across watersheds and strong seasonality in their streamflow regime. The prediction efficiency was the highest in the Rocky Mountain region. With 5 measurements considered for prediction in a year (1.37% of the data), the streamflow in most of the gauges in this region can be predicted with a median NSE > 0.75 and  $NSE_m > 0.5$ . The prediction efficiency increased more strongly with more measurements taken in the New England, Middle Atlantic (i.e., northeast coast), and the Pacific (i.e., west coast) than in other regions. The prediction efficiency was relatively low and did not significantly increase with more measurements taken in the central United States, especially the East-south and West-south Central regions. The spatial variability of prediction efficiency can be attributed to the process-driven mechanisms and dimensionality of watersheds. Watersheds with more temporally consistent (less intermittent) streamflow signals filter rainfall variability (i.e., randomness) and are more “storage-dominated” and less “rainfall-dominated.” Streamflow in storage-dominated systems has less dimensions and is more predictable than streamflow in rainfall-dominated, higher-dimensional systems. In addition, this study also demonstrates that the optimal sampling time is related to the streamflow regimes. More “informative” measurements can be taken during periods that cover a wide range of streamflow magnitudes.

The DSS technique can be applied for streamflow prediction by taking a set of ground-based streamflow measurements, using simulated data or remotely sensed data set (e.g., Surface Water and Ocean Topography, SWOT, or flying UAVs). Going beyond temporal reconstruction of streamflow, DSS can be also applied to identify optimal watersheds to gauge permanently, optimize the spatial placement of gauges along streams within watersheds, or to classify watersheds.

## Data Availability Statement

All the data and code used in this study can be accessed from CUAHSI HydroShare (K. Zhang, Luhar, et al., 2023). The streamflow data used in this study was retrieved from the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) data set (<https://ral.ucar.edu/solutions/products/camels>). The MATLAB code used for data-driven sparse sensing was retrieved from the Github repository by Krithika Manohar ([https://github.com/kmanohar/SSPOR\\_pub](https://github.com/kmanohar/SSPOR_pub)) and customized for this study.

## Acknowledgments

This work relates to Department of Army award (“Novel Technologies to Mitigate Water Contamination for Resilient Infrastructure”; Federal Award Identification Number: W9132T2220001) issued by the U. S. Army Corps of Engineers (USACE) Engineer Research and Development Center (ERDC). The United States Government has a royalty-free license throughout the world in all copyrightable material contained herein. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of USACE-ERDC.

## References

- Arsenault, R., Breton-Dufour, M., Poulin, A., Dallaire, G., & Romero-Lopez, R. (2019). Streamflow prediction in ungauged basins: Analysis of regionalization methods in a hydrologically heterogeneous region of Mexico. *Hydrological Sciences Journal*, 64(11), 1297–1311. <https://doi.org/10.1080/02626667.2019.1639716>
- Arsenault, R., & Brissette, F. P. (2014). Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. *Water Resources Research*, 50(7), 6135–6153. <https://doi.org/10.1002/2013WR014898>
- Bai, Z., Brunton, S. L., Brunton, B. W., Kutz, J. N., Kaiser, E., Spohn, A., & Noack, B. R. (2017). Data-driven methods in fluid dynamics: Sparse classification from experimental data. In A. Pollard, L. Castillo, L. Danaila, & M. Glauser (Eds.), *Whither Turbulence and Big Data in the 21st Century?* (pp. 323–342). Springer. [https://doi.org/10.1007/978-3-319-41217-7\\_17](https://doi.org/10.1007/978-3-319-41217-7_17)
- 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>
- Biancamaria, S., Lettenmaier, D. P., & Pavelsky, T. M. (2016). The SWOT mission and its capabilities for land hydrology. 117–147. [https://doi.org/10.1007/978-3-319-32449-4\\_6](https://doi.org/10.1007/978-3-319-32449-4_6)
- Brunner, M. I., Melsen, L. A., Newman, A. J., Wood, A. W., & Clark, M. P. (2020). Future streamflow regime changes in the United States: Assessment using functional classification. *Hydrology and Earth System Sciences*, 24(8), 3951–3966. <https://doi.org/10.5194/HESS-24-3951-2020>
- Brunton, B. W., Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Sparse sensor placement optimization for classification. *SIAM Journal on Applied Mathematics*, 76(5), 2099–2122. <https://doi.org/10.1137/15M1036713>
- Buytaert, W., Zulkafli, Z., Grainger, S., Acosta, L., Alemie, T. C., Bastiaensen, J., et al. (2014). Citizen science in hydrology and water resources: Opportunities for knowledge generation, ecosystem service management, and sustainable development. *Frontiers of Earth Science*, 2, 26. <https://doi.org/10.3389/FEART.2014.00026/BIBTEX>
- Candès, E. J., Romberg, J., & Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2), 489–509. <https://doi.org/10.1109/TIT.2005.862083>
- Candès, E. J., & Wakin, M. B. (2008). An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2), 21–30. <https://doi.org/10.1109/msp.2007.914731>
- Carlisle, D. M., Falcone, J., Wolock, D. M., Meador, M. R., & Norris, R. H. (2010). Predicting the natural flow regime: Models for assessing hydrological alteration in streams. *River Research and Applications*, 26(2), 118–136. <https://doi.org/10.1002/RRA.1247>
- Chiang, S.-M., Tsay, T.-K., & Nix, S. J. (2002). Hydrologic regionalization of watersheds. I: Methodology development. *Journal of Water Resources Planning and Management*, 128(1), 3–11. [https://doi.org/10.1061/\(ASCE\)0733-9496\(2002\)128:1\(3\)](https://doi.org/10.1061/(ASCE)0733-9496(2002)128:1(3))
- Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen science as an ecological research tool: Challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics*, 41(1), 149–172. <https://doi.org/10.1146/annurev-ecolsys-102209-144636>
- Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52(4), 1289–1306. <https://doi.org/10.1109/TIT.2006.871582>

- Ebtehaj, A. M., Foufoula-Georgiou, E., Lerman, G., & Bras, R. L. (2015). Compressive Earth observatory: An insight from AIRS/AMSU retrievals. *Geophysical Research Letters*, 42(2), 362–369. <https://doi.org/10.1002/2014GL062711>
- Eltner, A., Sardemann, H., & Grundmann, J. (2020). Technical note: Flow velocity and discharge measurement in rivers using terrestrial and unmanned-aerial-vehicle imagery. *Hydrology and Earth System Sciences*, 24(3), 1429–1445. <https://doi.org/10.5194/HESS-24-1429-2020>
- Etter, S., Strobl, B., Seibert, J., & Ilja Van Meerveld, H. J. (2018). Value of uncertain streamflow observations for hydrological modelling. *Hydrology and Earth System Sciences*, 22(10), 5243–5257. <https://doi.org/10.5194/HESS-22-5243-2018>
- Gettelman, A., Geer, A. J., Forbes, R. M., Carmichael, G. R., Feingold, G., Posselt, D. J., et al. (2022). The future of Earth system prediction: Advances in model-data fusion. *Science Advances*, 8(14), 3488. <https://doi.org/10.1126/sciadv.abn3488>
- Gleason, C. J., & Durand, M. T. (2020). Remote sensing of river discharge: A review and a framing for the discipline. *Remote Sensing*, 12(7), 1107. <https://doi.org/10.3390/RS12071107>
- He, Y., Bárdossy, A., & Zehe, E. (2011). A review of regionalisation for continuous streamflow simulation. *Hydrology and Earth System Sciences*, 15(11), 3539–3553. <https://doi.org/10.5194/HESS-15-3539-2011>
- Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of predictions in ungauged basins (PUB)-a review. *Hydrological Sciences Journal*, 58(6), 1198–1255. <https://doi.org/10.1080/02626667.2013.803183>
- Jayaraman, B., & Al Mamun, S. M. A. (2020). On data-driven sparse sensing and linear estimation of fluid flows. *Sensors*, 20(13), 1–31. <https://doi.org/10.3390/s20133752>
- Katul, G. G., Porporato, A., Daly, E., Oishi, A. C., Kim, H. S., Stoy, P. C., et al. (2007). On the spectrum of soil moisture from hourly to interannual scales. *Water Resources Research*, 43(5), 1–10. <https://doi.org/10.1029/2006WR005356>
- Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., & Woods, R. A. (2020). A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments. *Water Resources Research*, 56(9), 1–23. <https://doi.org/10.1029/2019WR025975>
- Krause, P., Boyle, D. P., & Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, 5, 89–97. <https://doi.org/10.5194/ADGEO-5-89-2005>
- Li, X., Weller, D. E., & Jordan, T. E. (2010). Watershed model calibration using multi-objective optimization and multi-site averaging. *Journal of Hydrology*, 380(3–4), 277–288. <https://doi.org/10.1016/j.jhydrol.2009.11.003>
- Lustig, M., Donoho, D., Santos, J., & Pauly, J. (2008). Compressed sensing MRI. *Signal Processing* ..., March 2008, 25(2), 72–82. <https://doi.org/10.1109/msp.2007.914728>
- Mahdadi, M., le Moine, N., Ribstein, P., Mahdadi, M., le Moine, N., & Ribstein, P. (2021). How to build reach-averaged rating curves for remote sensing discharge estimation? The potential of periodic geometry hypotheses. *EGUGA*. EGU21-12476. <https://doi.org/10.5194/EGUSPHERE-EGU21-12476>
- Manohar, K., Brunton, B. W., Kutz, J. N., & Brunton, S. L. (2018). Data-driven sparse sensor placement for reconstruction: Demonstrating the benefits of exploiting known patterns. *IEEE Control Systems*, 38(3), 63–86. <https://doi.org/10.1109/MCS.2018.2810460>
- Manohar, K., Kutz, J. N., & Brunton, S. L. (2022). Optimal sensor and actuator selection using balanced model reduction. *IEEE Transactions on Automatic Control*, 67(4), 2108–2115. <https://doi.org/10.1109/TAC.2021.3082502>
- Mazzoleni, M., Verlaan, M., Alfonso, L., Monego, M., Norbiato, D., Ferri, M., & Solomatine, D. P. (2017). Can assimilation of crowdsourced data in hydrological modelling improve flood prediction? *Hydrology and Earth System Sciences*, 21(2), 839–861. <https://doi.org/10.5194/HESS-21-839-2017>
- 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](https://doi.org/10.1016/0022-1694(70)90255-6)
- Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydro-meteorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, 19(1), 209–223. <https://doi.org/10.5194/HESS-19-209-2015>
- Nolan, B. T., Fienen, M. N., & Lorenz, D. L. (2015). A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA. *Journal of Hydrology*, 531, 902–911. <https://doi.org/10.1016/J.JHYDROL.2015.10.025>
- Ohmer, M., Liesch, T., & Wunsch, A. (2022). Spatiotemporal optimization of groundwater monitoring networks using data-driven sparse sensing methods. *Hydrology and Earth System Sciences*, 26(15), 4033–4053. <https://doi.org/10.5194/HESS-26-4033-2022>
- Pan, F., Wang, C., & Xi, X. (2016). Constructing river stage-discharge rating curves using remotely sensed river cross-sectional inundation areas and river bathymetry. *Journal of Hydrology*, 540, 670–687. <https://doi.org/10.1016/j.jhydrol.2016.06.024>
- Parolari, A. J., Sizemore, J., & Katul, G. G. (2021). Multiscale legacy responses of soil gas concentrations to soil moisture and temperature fluctuations. *Journal of Geophysical Research: Biogeosciences*, 126(2), e2020JG005865. <https://doi.org/10.1029/2020JG005865>
- Patil, A., & Ramsankaran, R. A. A. J. (2017). Improving streamflow simulations and forecasting performance of SWAT model by assimilating remotely sensed soil moisture observations. *Journal of Hydrology*, 555, 683–696. <https://doi.org/10.1016/J.JHYDROL.2017.10.058>
- Pham, L. T., Luo, L., & Finley, A. (2021). Evaluation of random forests for short-term daily streamflow forecasting in rainfall- and snowmelt-driven watersheds. *Hydrology and Earth System Sciences*, 25(6), 2997–3015. <https://doi.org/10.5194/HESS-25-2997-2021>
- Pool, S., & Seibert, J. (2021). Gauging ungauged catchments—Active learning for the timing of point discharge observations in combination with continuous water level measurements. *Journal of Hydrology*, 598(January), 126448. <https://doi.org/10.1016/j.jhydrol.2021.126448>
- Pool, S., Viviroli, D., & Seibert, J. (2017). Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration? *Journal of Hydrology*, 554, 613–622. <https://doi.org/10.1016/j.jhydrol.2017.09.037>
- Pool, S., Viviroli, D., & Seibert, J. (2019). Value of a limited number of discharge observations for improving regionalization: A large-sample study across the United States. *Water Resources Research*, 55(1), 363–377. <https://doi.org/10.1029/2018WR023855>
- Razavi, T., & Coulibaly, P. (2013). Streamflow prediction in ungauged basins: Review of regionalization methods. *Journal of Hydrologic Engineering*, 18(8), 958–975. [https://doi.org/10.1061/\(asce\)he.1943-5584.0000690](https://doi.org/10.1061/(asce)he.1943-5584.0000690)
- Reza Md Towfiqul Islam, A., Chandra Pal, S., Chowdhuri, I., Salam, R., Saiful Islam, M., Mostafizur Rahman, M., et al. (2021). Application of novel framework approach for prediction of nitrate concentration susceptibility in coastal multi-aquifers, Bangladesh. *Science of the Total Environment*, 801. <https://doi.org/10.1016/j.scitotenv.2021.149811>
- Samaniego, L., Bárdossy, A., & Kumar, R. (2010). Streamflow prediction in ungauged catchments using copula-based dissimilarity measures. *Water Resources Research*, 46(2). <https://doi.org/10.1029/2008WR007695>
- Sanborn, S. C., & Bledsoe, B. P. (2006). Predicting streamflow regime metrics for ungauged streams in Colorado, Washington, and Oregon. *Journal of Hydrology*, 325(1–4), 241–261. <https://doi.org/10.1016/J.JHYDROL.2005.10.018>
- Tetzlaff, D., Seibert, J., McGuire, K. J., Laudon, H., Burns, D. A., Dunn, S. M., & Soulsby, C. (2009). How does landscape structure influence catchment transit time across different geomorphic provinces? *Hydrological Processes*, 23(6), 945–953. <https://doi.org/10.1002/HYP.7240>
- Wei, J., Wang, L., Liu, P., Chen, X., Li, W., & Zomaya, A. Y. (2017). Spatiotemporal fusion of MODIS and landsat-7 reflectance images via compressed sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12), 7126–7139. <https://doi.org/10.1109/TGRS.2017.2742529>

- Williams, D. A., Nelsen, B., Berrett, C., Williams, G. P., & Moon, T. K. (2018). A comparison of data imputation methods using Bayesian compressive sensing and Empirical Mode Decomposition for environmental temperature data. *Environmental Modelling & Software*, 102, 172–184. <https://doi.org/10.1016/j.envsoft.2018.01.012>
- Wu, X., Wang, Q., & Liu, M. (2014). In-situ soil moisture sensing: Measurement scheduling and estimation using sparse sampling. *ACM Transactions on Sensor Networks*, 11(2), 1–11. <https://doi.org/10.1145/2629439>
- Wyatt, B. M., Ochsner, T. E., Krueger, E. S., & Jones, E. T. (2020). In-situ soil moisture data improve seasonal streamflow forecast accuracy in rainfall-dominated watersheds. *Journal of Hydrology*, 590, 125404. <https://doi.org/10.1016/j.jhydrol.2020.125404>
- Yang, X., Magnusson, J., Huang, S., Beldring, S., & Xu, C. Y. (2020). Dependence of regionalization methods on the complexity of hydrological models in multiple climatic regions. *Journal of Hydrology*, 582, 124357. <https://doi.org/10.1016/j.jhydrol.2019.124357>
- Zhang, K., Bin Mamoon, W., Schwartz, E., & Parolari, A. J. (2023). Reconstruction of sparse stream flow and concentration time-series through compressed sensing. *Geophysical Research Letters*, 50(2), e2022GL101177. <https://doi.org/10.1029/2022GL101177>
- Zhang, K., Luhan, M., Brunner, M. I., & Parolari, A. J. (2023). Data for "Streamflow prediction in poorly gauged watersheds in the United States through data-driven sparse sensing. [Dataset]. <https://doi.org/10.4211/hs.49b0f3b0f6924b2d917b3659fb03926b>
- Zhang, L., Xing, M., Qiu, C. W., Li, J., Sheng, J., Li, Y., & Bao, Z. (2010). Resolution enhancement for inversed synthetic aperture radar imaging under low SNR via improved compressive sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 48(10), 3824–3838. <https://doi.org/10.1109/TGRS.2010.2048575>