`ForeSWE`: Forecasting Snow-Water Equivalent with an Uncertainty-Aware Attention Model

Krishu K Thapa, Supriya Savalkar, Bhupinderjeet Singh, Trong Nghia Hoang, Kirti Rajagopalan, Ananth Kalyanaraman

Washington State University, Pullman, WA.
{krishu.thapa, supriya.savalkar, bhupinderjeet.singh, trongnghia.hoang, kirtir, ananth}@wsu.edu

Abstract

Various complex water management decisions are made in snow-dominant watersheds with the knowledge of Snow-Water Equivalent (SWE)—a key measure widely used to estimate the water content of a snowpack. However, forecasting SWE is challenging because SWE is influenced by various factors including topography and an array of environmental conditions, and has therefore been observed to be spatio-temporally variable. Classical approaches to SWE forecasting have not adequately utilized these spatial/temporal correlations, nor do they provide uncertainty estimates—which can be of significant value to the decision maker. In this paper, we present `ForeSWE`, a new probabilistic spatio-temporal forecasting model that integrates deep learning and classical probabilistic techniques. The resulting model features a combination of an attention mechanism to integrate spatiotemporal features and interactions, alongside a Gaussian process module that provides principled quantification of prediction uncertainty. We evaluate the model on data from 512 Snow Telemetry (SNOTEL) stations in the Western US. The results show significant improvements in both forecasting accuracy and prediction interval compared to existing approaches. The results also serve to highlight the efficacy in uncertainty estimates between different approaches. Collectively, these findings have provided a platform for deployment and feedback by the water management community.

Introduction

Streamflow is vital for societal needs, including food production, irrigation, flood control, hydropower, and supporting endangered fish species. In snow-dominant watersheds of the Western U.S., 50–80% of annual streamflow originates from melting winter snowpack (Hunter, Tootle, and Piechota 2006; Li et al. 2017). Therefore, the state of the snowpack and the water it contains—known as *Snow Water Equivalent (SWE)*—is critical for determining the magnitude and timing of streamflow (Mankin et al. 2015; Harpold et al. 2017).

Information on the current state of SWE is widely used by local and federal water agencies in the United States, including reservoir operators and irrigation districts (U.S. Bureau of Reclamation, Research and Development Office 2024). The National Resources Conservation Service (NRCS) maintains a nationwide dashboard for monitoring SWE and related variables (U.S. Department of Agriculture, Natural Resources Conservation Service, National Water and Climate Center 2025). However, SWE forecast information, while critical, is not yet available in an operational context.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Accurate SWE forecasts are critical for both short- and long-term water management. Forecasts at multi-day, and multi-week timescales each serve important roles: multi-day forecasts help anticipate rapid snowmelt and potential flooding, enabling real-time interventions such as reservoir drawdowns, while multi-week forecasts of SWE and peak SWE inform subseasonal planning and allocation decisions across agricultural, ecological, and hydropower sectors (Pagano, Garen, and Sorooshian 2004; Huang et al. 2017; Stillinger et al. 2021). SWE also serves as a key input to streamflow forecasts (Mote et al. 2005) and improves sub-seasonal climate outlooks by capturing important land–atmosphere feedbacks (Diro and Lin 2020).

SWE forecasts can contribute to the growing body of climate-informed tools that offer direct potential for social good. Improvements in SWE forecasting can support more equitable and sustainable water use, reduce the societal impacts of extreme hydrologic events, and strengthen the resilience of communities and ecosystems that depend on predictable water availability.

Challenges. Real-time forecasting of SWE as the season progresses, is challenging due to variations in SWE patterns across space and time. These patterns are influenced by complex interactions between local weather and spatial attributes, as well as by the phase of SWE.

- **Temporal Variation.** For any location, the state of its snowpack at any point influences its SWE trend in the future. For example, a dry and cold snowpack has more liquid retention capacity and can act as buffer to reduce the risk of melt and flooding during storm events (Garvelmann, Pohl, and Weiler 2015). Meteorological variables (e.g., temperature, precipitation) vary temporally and determine the timing of snow accumulation and melt.
- **Spatial Variation.** The variation of SWE across locations results from different interrelated factors like orographic effects, elevation, terrain, wind, vegetation and radiation (Liston 2004). For instance, locations in higher

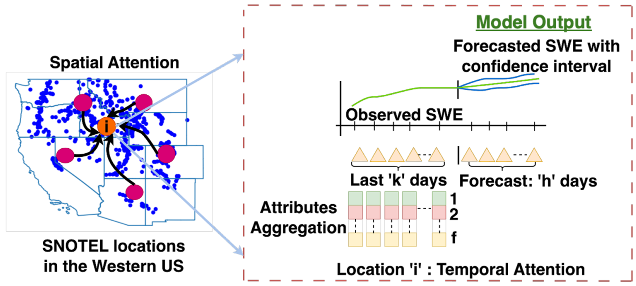


Figure 1: Spatiotemporal SWE forecasting with confidence interval for location i over a horizon h, using k days of historical observations with f attributes. Each blue dot in the map is a SNOTEL location.

elevation can have colder temperature and more snow accumulation; and southern-facing slopes can have faster snowmelt due to solar exposure. Additionally, locations at different ranges of proximity might exhibit spatial correlations (e.g., areas falling under same atmospheric river path).

- **Spatio-Temporal Interaction.** Spatiotemporal attributes of a location interact and influence SWE behavior (Garvelmann, Pohl, and Weiler 2015). For example, the relationship between surface air temperatures and precipitation phase (rain or snow) depends on relative humidity (Jennings et al. 2018). The precipitation phase can in turn impact snowmelt, creating a potential for floods (Musselman et al. 2018).
- **Environmental Uncertainty.** SWE trends are affected by the inherent stochastic nature of the process and unpredictable environmental anomalies. For example, an ongoing accumulation phase might unexpectedly shift to rapid melt due to untimely warming. Accounting for such uncertainty in the SWE forecasting is essential to optimize for planning decisions regarding resource deployment.

Contributions. In this paper, we present `ForeSWE`, a new uncertainty-aware attention model for SWE forecasting, that combines modern deep learning and classical probabilistic methods (see Figure 1). We focus on SWE forecasting at both daily and weekly timestep, as each serves distinct decision-making needs: daily forecasts are critical for emergency response scenarios such as flood risk management, while weekly forecasts inform agricultural planning, recreational activities (e.g., skiing season operations) and energy management (e.g., hydropower generation).

The paper makes the following contributions:

1. An attention-based deep learning (DL) model (Vaswani et al. 2017) parameterized with a new spatio-temporal attention module which is specifically designed to capture the correlation in space and time, as well as interaction between attributes in SWE forecasting context.
2. A probabilistic augmentation using a Gaussian process (GP) (Rasmussen and Williams 2006) as the prediction head, enabling spatio-temporal correlation learning and uncertainty quantification of SWE forecasts.
3. A standalone sparse raw GP implementation that represents a GP-only baseline for the SWE forecasting problem.
4. A thorough experimental evaluation that evaluates our proposed approaches against various spatial and/or temporal ML approaches. Results show that for daily forecasting `ForeSWE` and `Raw-GP` outperform all other approaches, while for weekly forecasting, `ForeSWE` is the most suited model as it delivers the best performance both by accuracy and by the quality of its uncertainty estimates.

The extended version of this article with the Appendix is available on ArXiV (Thapa et al. 2025).

Related Work

The current literature on SWE forecasting features both mechanistic and data-driven approaches. Mechanistic models utilize prior descriptive knowledge in physical and hydrological equations. However, our knowledge of these processes is limited leading to simplified models and large biases (Diro and Lin 2020). Alternatively, machine learning (ML) techniques can learn from historical data from a diverse collection of locations, and benefit where physical knowledge may be lacking; and be learnable in an incremental fashion as new data become available—e.g., past temporal observations of SWE, and additional spatial features such as remotely-sensed reflectances (Brodzik et al. 2016). Such data can be utilized to build SWE forecasting models.

Among the earlier data-driven works in SWE forecasting, (Sarhadi, Kelly, and Modarres 2014) uses statistical models like ARIMA (Box and Jenkins 1970) and SARIMA (Box and Jenkins 1976) with exogenous variables to predict daily (i.e. 30 days) and monthly (i.e. 6 months) SWE. However, all the locations under study are in low elevation (<2000ft) with a low annual max SWE (100 to 150mm), making the model too restrictive. Similarly, (Franz, Butcher, and Ajami 2010) combines twelve bio-physical models using a Bayesian Model Averaging (BMA) to forecast SWE with uncertainty quantification. Although the work shows competitive results, it has only been implemented on six locations and limited to a single day forecasting.

Deep Learning approaches have been explored more recently. (Cui, Anderson, and Bales 2023) couples the Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) and zonal bias correction data assimilation approach to predict the gridded SWE value at 1 km pixel for the next day or month. The assimilation approach depends on the presence of nearby observation sites, which makes it challenging due to the geographical sparsity in the observation layer. Recently, an attention-based model was implemented to predict SWE using spatial and temporal attention (Thapa et al. 2024) but it does not provide any uncertainty estimates. For a different application in climate modeling, a few approaches exist (Nguyen et al. 2023a,b). In comparison to these approaches, our model architecture is different—as it uses a form of spatio-temporal attention that captures attribute interaction (unlike a single learnable query vector in the *ClimaX model* (Nguyen et al. 2023a)), and is designed to provide uncertainty estimates.

Model Architecture

Our model design hinges on two core ideas. First, to incorporate spatio-temporal correlations as well as attribute interaction for SWE forecasting, we first model the problem as a sequence-to-sequence prediction problem, and present an adaptation of the self-attention mechanism (Vaswani et al. 2017). When used in real-time, this model will ingest the last k days of data for each location and will output the SWE values for the next h days (i.e., the forecast horizon), as shown in Figure 1. Secondly, to implement a principled quantification of prediction uncertainty, we augment the model with Gaussian process (GP) (Rasmussen and Williams 2006), which operates on the distribution of data points rather than sequences. For any given point, GPs—with appropriate kernel selection and adequate input features—capture correlation across all the training data points to make predictions with uncertainty quantification. However, using GPs directly can introduce a limitation in the number of dimensions. Therefore, in our model architecture, we train our self-attention model to capture spatio-temporal correlation in the data at a lower dimensional space, followed by using GP as an output head to give forecast with a prediction interval.

Figure 2(a) shows the overall architecture of our model, and Figure 2(b) shows the temporal attributes aggregation step of the model in detail.

Notation: $\mathbb{L}$ represents set of n SNOTEL station; $\mathbb{S}$ is set of SWE seasons or water years (Oct. 1 to Sep. 30); $\mathbb{A}$ is a set of f dynamic attributes attached to locations; k is input historical window; h is the forecasting horizon (in days or weeks); and m is day number in a water year.

Inputs: Each training example is a sequence of all locations, their daily features and historical attributes, and prompts. The input sequences for day t consist of:

- $\mathbf{Pr}_t$=$[\mathbf{pr}_t^i]_{i=1}^n$: a sequence of prompt vectors of n locations.
- $\mathbf{Sp}_t$=$[\mathbf{sp}_t^i]_{i=1}^n$: a sequence of spatial attribute vectors.
- $\mathbf{D}_{t,k}[1{:}n, 1{:}f, t-k{:}t] \in \mathbb{R}^{n\times f\times k}$: a collection of n locations with k historical observations of f attributes per day t. For location i, its attribute $a \in f$ is $\mathbf{D}_{t,k}^{i,a} = \mathbf{D}_{t,k}[i, a, t-k : t]$. Here, k can represent historical observations for different temporal resolutions as daily, weekly, monthly, and yearly.

Outputs: The output of the model is a sequence of SWE forecast values for the n locations, along with a prediction interval, for a forecast horizon of h days. For each location $i \in \mathbb{L}$ on day t, the forecast is a 2-D vector containing high (u), mean (p), and low (l) SWE respectively, given by $\mathbf{y}_{t:t+h}^i = [u_{t:t+h}^i, p_{t:t+h}^i, l_{t:t+h}^i] \in \mathbb{R}^{3\times h}$ which is referred to as the **prediction interval**. Finally, for all n locations on day t, the output sequence can be represented as $\mathbf{Y}_{t:t+h} \in \mathbb{R}^{n\times 3\times h}$. h can represent the forecasting horizon at different temporal resolutions such as daily and weekly points.

Location Embedding Block: This block is divided into two parts as shown in the left part of Figure 2(b).

Spatial Embedding Layer embeds *daily spatial features* (e.g., day number/length) on day t for each location, alongside key spatial features (e.g., lat/long, southness, elevation) to its corresponding embedding representation $\overline{\mathbf{Sp}}_t$. For embedding we used d_{model} = 1024 dimensions.

Prompt Embedding Layer provides prompts ($\overline{\mathbf{Pr}}_t$) that represent additional context for the location such as capture weather patterns, vegetation, and elevation range.

These two embeddings provide the location-specific embedding for day t: $\mathbf{L}_t = \overline{\mathbf{Sp}}_t + \overline{\mathbf{Pr}}_t$.

Temporal Attribute Aggregation Block: The temporal attribute aggregation for each location is obtained individually (Figure 2(b)). For location i on day t, given its k historical observation of each attribute as $\mathbf{D}_{t,k}^i = \mathbf{D}_{t,k}\ [i, 1 : f, t-k : t]$, each attribute's historical vector is passed to their corresponding attribute embedding layers, finally to obtain $\overline{\mathbf{D}}_{t,k}^i = \overline{\mathbf{D}}_{t,k}[i, 1 : f, 1 : d_{model}]$. Given location embedding vector $\boldsymbol{\ell}_t^i \in \mathbf{L}_t$ and its embedded attributes vectors $\overline{\mathbf{D}}_{t,k}^i$, we obtain the temporal attributes aggregated representation $\mathbf{x}_{t,k}^i = \mathbf{X}_{t,k}[i, 1 : d_{model}]$, where $\mathbf{X}_{t,k}$ includes representation for all the locations.

Window-wise Concatenation: From the multiple temporal attributes aggregation blocks, we obtain C different $\mathbf{X}_{t,k_c}$, with their corresponding window k_c. These representations of locations are concatenated to obtain $\mathbf{X}_t[1 : n, 1 : C \times d_{model}]$, where the vector dimension for each location is $C \times d_{model}$. For location i, its concatenated representation is obtained as $\mathbf{x}_t^i = \mathbf{X}_t[i, 1 : C \times d_{model}]$.

Modified Spatial Attention: Next, to capture spatial correlations among locations, the attribute aggregated vectors for each location ($\mathbf{X}_t$) are subjected to spatial attention ($\overline{\mathbf{X}}_t$)—as explained in (§ Technical Approach). The dimension of each representation is $C \times d_{model}$.

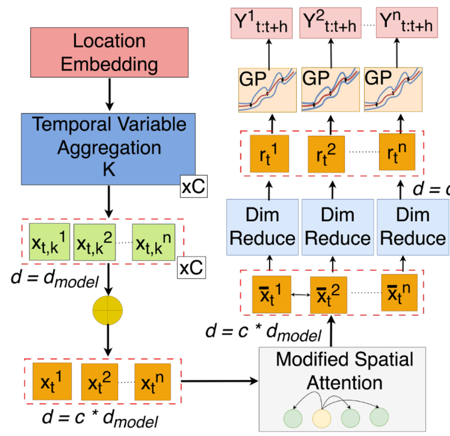
Gaussian Process (GP): Once all of the above blocks are trained using actual SWE, the spatio-temporal attention model is considered as the pre-trained model. To provide prediction uncertainty, we replace its prediction head with a GP model (Rasmussen and Williams 2006), which is learned using the dimension-reduced ($d_{GP}(8) \ll d_{model}$) input representation of the pre-trained model, denoted by $\mathbf{R}_t$, and its corresponding SWE output. Note that for any location, when we pass its learned spatio-temporal representation $\mathbf{r}_t^i \in \mathbf{R}_t$ to the GP model, we obtain SWE forecast for horizon h along with its prediction interval, represented by vector $\mathbf{y}_{t:t+h}^i$.

Technical Approach

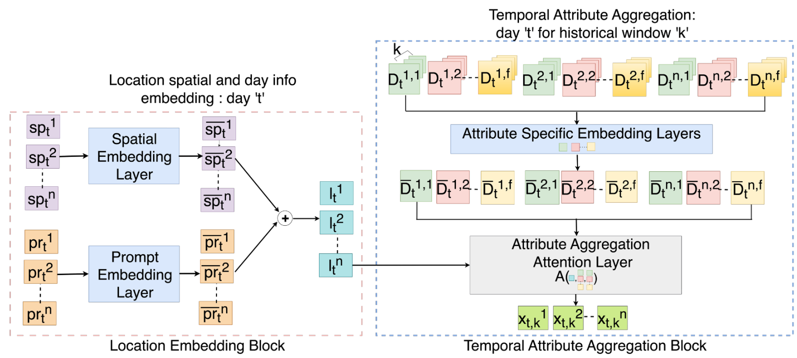
In what follows, we describe the key components of our model architecture.

Attributes Aggregation

Attribute aggregation is used to capture the attribute interactions at a certain time of a location. Our approach inspired from (Nguyen et al. 2023a) combines interactions of daily attributes into a vector representation. We implement the classical self-attention mechanism (Vaswani et al. 2017), which works with a sequence of tokens embedded into query



(a) Overall architecture



(b) Location embedding and attribute aggregation

Figure 2: `ForeSWE` model architecture (left) and its location embedding and temporal attribute aggregation components (right).

($\mathbf{Q}$), key ($\mathbf{K}$) and value ($\mathbf{V}$) vectors, as given by:

$$\text{Attention}\big(\mathbf{Q}, \mathbf{K}, \mathbf{V}\big) \;=\; \rho\left(\frac{\mathbf{Q}.\mathbf{K}^\top}{\sqrt{d_{\text{model}}}}\right)\mathbf{V}\;, \qquad (1)$$

where $\rho(.)$ represents the softmax function, and d_{model} represents the dimensions of all vectors. In our implementation, query comes from daily spatial attributes and prompts for each location i on day t, and is given by $\mathbf{q}^i_t = \mathbf{sp}^i_t + \mathbf{pr}^i_t$. Meanwhile, key and value come from its k historical collection of daily observations ($\mathbf{D}^i_{t,k}$). Therefore, the self-attention function is given by:

$$\mathbf{x}^i_{t,k} \;=\; \text{Attention}\left(\mathbf{q}^i_t.\mathbf{W}^Q_{tp}, \mathbf{D}^i_{t,k}.\mathbf{W}^K_{tp}, \mathbf{D}^i_{t,k}.\mathbf{W}^V_{tp}\right) \qquad (2)$$

$\mathbf{W}^Q_{tp}$, $\mathbf{W}^K_{tp}$ and $\mathbf{W}^V_{tp}$ are weight matrices for query, key, and value respectively. Each attribute has historical information associated with it on day t, enabling the equation above to capture the temporal dynamics between attributes for each location.

Modified Spatial Attention

The location representation is the combination of representations of their attributes aggregated temporal dynamics from C blocks, and is given by $\mathbf{X}_t$ for all n locations. It is then passed to the modified attention function $\overline{\text{Attention}}$ to capture spatial correlations across the locations,

$$\overline{\mathbf{X}}_t \;=\; \overline{\text{Attention}}\Big(\mathbf{X}_t.\mathbf{W}^Q_{sp}, \mathbf{X}_t.\mathbf{W}^K_{sp}, \mathbf{X}_t.\mathbf{W}^V_{sp}\Big) \qquad (3)$$

Here, $\bar{\mathbf{X}}_t$ is the encoded feature representation of all the n locations in day $t \in [1, m]$ of season $s \in \mathbb{S}$. $\mathbf{W}^Q_{sp}$, $\mathbf{W}^K_{sp}$ and $\mathbf{W}^V_{sp}$ are the weight matrices. This encoding captures spatiotemporal correlation information across and within locations.

The modified spatial function adds the Haversine distance (Inman 1835) and angularity between locations to derive the attention weights. The implementation details of Haversine distance and degree angularity can be found in the Appendix (Thapa et al. 2025). The rationale behind this is to add some bias of proximity to capture the spatial correlations between locations. However, the contributions of these additional elements in the attention weights are governed by learnable parameters, $\boldsymbol{w_H}$ and $\boldsymbol{w_\theta}$. Let $\mathbf{Q}_{sp} = \mathbf{X}_t.\mathbf{W}^Q_{sp}$; $\mathbf{K}_{sp} = \mathbf{X}_t.\mathbf{W}^K_{sp}$; $\mathbf{V}_{sp} = \mathbf{X}_t.\mathbf{W}^V_{sp}$ be the query, key and value vectors, respectively. The modified attention can be calculated as:

$$\overline{\text{Attention}} = \rho\left(\rho\left(\frac{\mathbf{Q}_{sp}.\mathbf{K}^\top_{sp}}{\sqrt{d_{model}}}\right) + \boldsymbol{w_H}\cdot\mathbf{d}_H + \boldsymbol{w_\theta}\cdot\boldsymbol{\theta}\right)\mathbf{V}_{sp} \quad (4)$$

where $\mathbf{d}_H$ and $\boldsymbol{\theta}$ represent 2-D vectors for distance and angularity between all location pairs respectively. The variable aggregation and the spatial attention will collectively compute the encoded representation of spatiotemporal correlations across the locations for a given day t, as ($\overline{\mathbf{X}}_t$). These encoded representations are dimensionally reduced to $d_{GP} \ll d_{model}$, transforming the sequence to $\mathbf{R}_t$, and used to train Gaussian process (Rasmussen and Williams 2006) model to forecast SWE and quantify uncertainty.

Probabilistic Prediction with Gaussian Process

To account for uncertainty in SWE forecasting, we replace the prediction head of the attention-based model in the previous sections with a Gaussian process (GP) regressor. This imposes a probabilistic prior on the prediction function $g(.)$ that maps from the pre-trained representation $\boldsymbol{r}^n_t$ at each spatio-temporal coordinate (t, n) to its corresponding SWE value $y^n_{t:t+h}$ (see Figure 6, Appendix).

To learn this prior, we remodel the pre-trained representation $\boldsymbol{r}^n_t$ as a tuple $\boldsymbol{z} = (\boldsymbol{r}, t)$ where $\boldsymbol{r} = \boldsymbol{r}^n_t$ and t corresponds to its time index. This will allow the GP prior to model and learn the temporal correlation separately. We parameterize it as a τ-component linear co-regionalized GP prior (Van der Wilk et al. 2020),

$$g(\boldsymbol{z}) \;\sim\; \text{GP}\Big(m(\boldsymbol{z};\gamma), \kappa\big(\boldsymbol{z}, \boldsymbol{z}'; \zeta_1, \zeta_2\big)\Big) \text{ with}$$

$$\kappa\big(\boldsymbol{z}, \boldsymbol{z}'; \zeta_1, \zeta_2\big) \;\triangleq\; \sum_{i=1}^{\tau} \zeta_2(t, i)\zeta_2(t', i)\cdot\kappa_i\big(\boldsymbol{r}, \boldsymbol{r}'; \zeta_{1,i}\big)\;, \quad (5)$$

where $\kappa_i(\boldsymbol{r}, \boldsymbol{r}'; \zeta_1)$ parameterizes the i-th component of the covariance between the overall spatio-temporal representation while the scalar product $\zeta_2(t,i)\zeta_2(t',i)$ parameterizes the i-th component of the correlation between time-steps t and t'. Here, the kernel component $\{\kappa_i\}_i$ are parameterized as RBF kernels with learnable parameters $\zeta_1 = \{\zeta_{1,i}\}_{i=1}^{\tau}$ while ζ_2 is a learnable matrix; and $m(\boldsymbol{z};\gamma)$ is the mean SWE function parameterized by γ. Detailed specifications of these parameterization are provided in Appendix (Thapa et al. 2025).

Remark. Although temporal information is already embedded in the spatio-temporal representation $\boldsymbol{r}_t^n$, our formulation retains an additional, separate temporal correlation component. Without this, temporal correlation would be modeled implicitly through the overall spatio-temporal similarity, effectively imposing a single, possibly averaged time scale. However, in the SWE context, temporal dependencies can evolve differently over short and long time horizons (Nijssen and Lettenmaier 2004). By explicitly modeling temporal correlation, we allow for a more flexible structure that can capture the correlation across multiple temporal scales.

Let $\boldsymbol{y} = [g(\boldsymbol{z})]_{\boldsymbol{z}} = [g(\boldsymbol{r}_t^n, t)]_{(t,n)}$ denote the corresponding column vector of SWE ground-truth values. Let $\boldsymbol{m}_\gamma = [m(\boldsymbol{z};\gamma)]_{\boldsymbol{z}} = [m(\boldsymbol{r}_t^n, t;\gamma)]_{(t,n)}$ denote the column vector of the parameterized mean function's outputs at $\boldsymbol{z} = (\boldsymbol{r}_t^n, t)$ for all observed spatio-temporal coordinates (t, n). Then,

$$\boldsymbol{y} \sim \mathbb{N}\Big(\boldsymbol{m}_\gamma; \boldsymbol{K}_\zeta\Big) \text{ where } \boldsymbol{K}_\zeta[\boldsymbol{z}, \boldsymbol{z}'] = \kappa(\boldsymbol{z}, \boldsymbol{z}'; \zeta_1, \zeta_2) \quad (6)$$

following the marginal property of GP. Having both the pre-trained input representation $\boldsymbol{z}$ and the corresponding SWE output $\boldsymbol{y}$, we can learn the parameters $(\gamma, \zeta = (\zeta_1, \zeta_2))$ via

$$(\gamma, \zeta) = \text{argmax} \ \log \mathbb{N}\Big(\boldsymbol{y}; \boldsymbol{m}_\gamma, \boldsymbol{K}_\zeta\Big) . \quad (7)$$

Given these GP parameters, the distribution over the true SWE value y_* at any spatio-temporal coordinate (t_*, n_*) with representation $\boldsymbol{z}_* = (\boldsymbol{r}_*, t_*)$ can be derived as the GP's predictive distribution (Rasmussen and Williams 2006),

$$y_* \sim \mathbb{N}\Big(\boldsymbol{\kappa}_*^\top \boldsymbol{K}_\zeta^{-1}\Big(\boldsymbol{y} - \boldsymbol{m}_\gamma\Big) + m(\boldsymbol{z}_*;\gamma), \boldsymbol{\sigma}_*^2\Big) , \quad (8)$$

where $\boldsymbol{\kappa}_* = [\kappa(\boldsymbol{z}_*, \boldsymbol{z}; \zeta_1, \zeta_2)]_{\boldsymbol{z}}$ denotes the column vector of covariance values between $\boldsymbol{z}_*$ and other observed spatio-temporal representations $\boldsymbol{z}$; and the predictive variance,

$$\boldsymbol{\sigma}_*^2 = \kappa\big(\boldsymbol{z}_*, \boldsymbol{z}_*; \zeta_1, \zeta_2\big) - \boldsymbol{\kappa}_*^\top \boldsymbol{K}_\zeta^{-1} \boldsymbol{\kappa}_* . \quad (9)$$

Eqs. (8) and (9) thus present an entire Gaussian distribution over our SWE prediction. This yields both the (mean) prediction,

$$p_{n_*}^{t_*} = \mathbb{E}[y_*] = \boldsymbol{\kappa}_*^\top \boldsymbol{K}_\zeta^{-1}\Big(\boldsymbol{y} - \boldsymbol{m}_\gamma\Big) + m(\boldsymbol{z}_*;\gamma) , \quad (10)$$

and its prediction variance σ_*^2, which can be used to provably compute any α-prediction interval $\boldsymbol{Y}_{t_*}^{n_*} = [\ell_{n_*}^{t_*}, u_{n_*}^{t_*}]$,

$$\ell_{n_*}^{t_*} = p_{n_*}^{t_*} - \Phi^{-1}(\alpha/2) \cdot \sigma_* , \quad (11)$$
$$u_{n_*}^{t_*} = p_{n_*}^{t_*} + \Phi^{-1}(\alpha/2) \cdot \sigma_* , \quad (12)$$

where $\Phi^{-1}(.)$ is the inverse cumulative function of an univariate normal and α is the confidence level. For example, in our experiment, we choose $\alpha = 0.95$. This means under our learned GP calibration (γ, ζ), the true SWE value is in $[\ell_{n_*}^{t_*}, u_{n_*}^{t_*}]$ with 95%. The cost of the above training and inference procedure is however cubic in the number of spatio-temporal training data points $\boldsymbol{z} = (\boldsymbol{r}, t)$. To sidestep this prohibitively expensive cost, we can adopt existing sparse approximations (Quiñonero-Candela and Rasmussen 2005; Titsias 2009; Hoang et al. 2014; Hoang, Hoang, and Low 2015, 2016, 2017; Hoang et al. 2019, 2020; Van der Wilk et al. 2020; Bui et al. 2024) of the above GP which scales the cost back to linear in the size of the training dataset.

Computational Complexity: A detailed complexity analysis is provided in the Appendix (Thapa et al. 2025). In a nutshell, the overall complexity of our model is governed by attribute aggregation, modified spatial attention, and Gaussian process blocks.

Implementation: Model implementation used Pytorch (v2.0.1) (LSTM and Attention models), GPyTorch (v1.12) (Gaussian process) packages. Data processing and visualization used multiple Python packages. This is an open-source project, and the code with data can be found in https://github.com/Krishuthapa/SWE-Forecasting.

Experimental Setup

Data Description: The set of static and dynamic features used along with their respective sources are listed below:

- **Static Features**: elevation, latitude, longitude (NRCS 2023); land cover (Yang et al. 2018); southness (NED 2014); prompt vectors generated from pre-trained language model (Reimers and Gurevych 2019) based on vegetation, weather type, etc.
- **Daily Features:** SWE (NRCS 2023), max/min temperature, precipitation accumulation, downward surface shortwave radiation, wind velocity, max/min relative humidity, and specific humidity (Abatzoglou 2013).
- **Daily Satellite Observations:** Passive microwave brightness temperature (19VE, 37VE, and their difference) (Brodzik et al. 2016)

The daily SNOTEL data, downloaded from (NRCS 2023), consists of 822 stations for 28 water years (1991-2019). We filtered out stations with more than 10% missing snow observation data in any given year. The resulting 512 stations comprised our main data set. Given our focus on SWE forecasting, we used daily data for ∼180 days starting on Dec 1, to cover the active SWE season. This resulted in a total of 2,580,480 (=512 × 28 × 180) ⟨location, year, day⟩ combinations. The satellite data for getting brightness reflectances, and Light Detection and Ranging (LiDAR) data for getting slope and elevation to calculate southness, following the approach discussed in (Thapa et al. 2024).

Evaluation Methodology: Out of the 28 years, we used 25 years (from 1994) for training and testing , splitting the 25 years into two sets: training (20 years), and testing (5 years). Here, we train and test our model on the same set

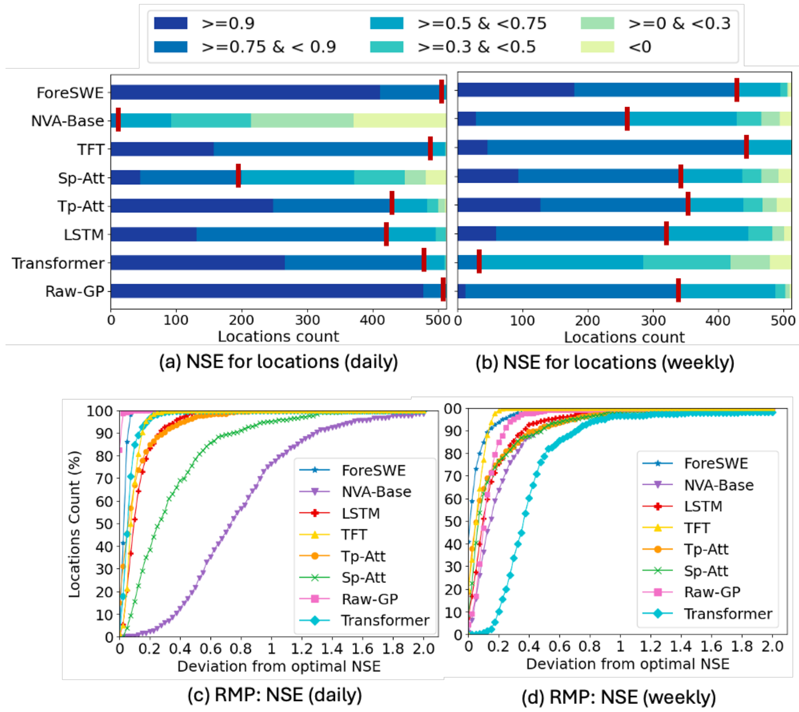


Figure 3: **Daily and Weekly forecasting** models comparison: (a) The distribution of locations across five NSE groups for all models. NSE is calculated for each location, with higher values (blue bars) indicating a better prediction. Locations to the left of the red line have NSE > 0.75. (b) Relative model performance (RMP) based on the NSE values. RMP chart: Each curve corresponds to a model; the closer and longer the line along the y-axis, the better the model performance. The X-axis shows the deviation of a model from the corresponding best performing model; Y-axis shows the fraction of locations with that deviation.

of locations. The data from (1991-1993) is separated as a buffer to include yearly historical data (3 years) for inputs starting in 1994. Test water years are 2015 through 2019—consecutive so that the data points belonging to these years do not appear anywhere during our training. These test years also cover a wide range in average SWE, from driest (2015) to wettest (2017). Additionally, the locations were binned into four "groups" based on their averaged peak SWE, from lowest (group 1) to highest (group 4)—as shown in Table 2 of Appendix.

In our experiments, we compared eight models in a design ablation study (as outlined in Table 5, Appendix):

- `LSTM`: Long Short-Term Memory model in autoregressive encoder-decoder format (Hochreiter and Schmidhuber 1997)
- `Raw-GP`: Sparse Gaussian process with multi-variate output implemented on raw input features (Rasmussen and Williams 2006)
- `Sp-Att` and `Tp-Att`: Spatial and temporal attention models, respectively, implemented using the attention mechanism presented in (Thapa et al. 2024). These models are not autoregressive—i.e., they generate the next h days of forecast from linear transformation of their encoded representations.
- `Transformer`: Temporal auto-regressive standard transformer (Vaswani et al. 2017)
- `TFT`: Temporal Fusion Transformer (Lim et al. 2021)
- `NVA-Base`: Non-Variable Aggregation is a trimmed down version of our proposed model without the variable aggregation and GP parts.
- **`ForeSWE`:** Our proposed uncertainty-aware attention model.

We evaluate models in two axes: forecasting accuracy and prediction interval. To compare models accuracy across location groups, we use Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe 1970), which compares the predicted (Y_p) versus actual (Y_a) over the entire prediction time (T) for any location $i \in \mathbb{L}$. Its value ranges from -∞ to 1, where the model performs best with its value closer to 1. Generally, in a long-horizon forecasting setup, **NSE $>$ 0.75** is preferred (Moriasi et al. 2007). A value below 0 means an averaged extrapolation would be better than the model prediction.

$$NSE_i = 1 - \frac{\sum_{t=1}^{T}(Y_{p,i}^t - Y_{a,i}^t)^2}{\sum_{t=1}^{T}(Y_{a,i}^t - \overline{Y}_{a,i}^t)^2} \tag{13}$$

Here, $\overline{Y}_{a,i}$ is the long-term mean actual SWE for location i. We also use relative bias that quantifies over-/under-predictions:

$$RB_i = \frac{\sum_{t=1}^{T}(Y_{p,i}^t - Y_{a,i}^t)}{\sum_{t=1}^{T} Y_{a,i}^t} \tag{14}$$

Additionally, to evaluate predictions with uncertainty quantification from `ForeSWE` and `Raw-GP`, we use metrics such as Negative Log Likelihood (NLL), Expected Calibration Error (ECE) and coverage percentage, further explained in the Appendix (Thapa et al. 2025).

As for the *forecasting horizon*, we trained and tested our model under two experimental settings: forecasting daily SWE for a 10-day horizon, and forecasting weekly SWE for 4-week horizon.

Experimental Results

This section presents and analyzes detailed evaluation of our approach in terms of performance and uncertainty quantification in the daily and weekly forecasting setups.

Model Comparison: We evaluate the model's performance using NSE values and relative bias (%). Figure 3 shows model comparison based on NSE for daily and weekly forecasting.

For *daily* forecasting, we observed that `ForeSWE` and `Raw-GP` models achieve the best NSE values with comparable performance. In particular, `ForeSWE` and `Raw-GP` achieve high NSE (> 0.75) in 99.2% and 99.6% of the locations, respectively (Figure 3a). Figure 3c shows the relative performance chart based on the NSE values comparing the different models against one another. Notably, temporal models perform better, indicating the prominence of temporal effect for near-term forecasting.

For *weekly* forecasting, `ForeSWE` achieves the best performance. Figure 3b shows that `ForeSWE` achieves NSE above 0.75 at over 427 locations while `Raw-GP` only achieves comparable NSE at only 340 locations. This performance advantage of `ForeSWE` over `Raw-GP` is also reflected in the relative model performance chart (Figure 3d). This shows that `ForeSWE` is the preferred model for the longer forecasting horizon (i.e., weekly), which makes it particularly useful in sub-seasonal scale decision making. Interestingly, the relative performance of `Tp-Att` is comparative or outperforms `Raw-GP` for the weekly horizon (Figure 3d). Furthermore, in weekly forecasting `Sp-Att` outperforms most of the temporal models such as `LSTM` and `Transformer`, underscoring the significance of spatial correlation in relatively long-term forecasting.

Furthermore, as observed from Fig. 3c and Fig. 3d, the performance of `Raw-GP` decreases substantially in the weekly forecasting setup while `ForeSWE` still maintains its top performance. This reveals a weakness of `Raw-GP` in spatio-temporal modeling. It operates in the raw feature space and over-smooths itself to the short-scale variation in observed data, over-emphasizing temporal correlations over spatial correlations. This helps `Raw-GP` generalize well in short-term daily forecasting but mislead it in long-term weekly forecasting where spatial correlation has a stronger impact on the snow pattern. In contrast, `ForeSWE` operates in the embedding space of a spatio-temporal transformer which were pre-trained to generate the most holistic spatio-temporal representation of data. This allows `ForeSWE` to preserve its best performance in both short-term and long-term forecasting, underscoring the importance of using deep spatio-temporal representation for both short-term and long-term snow forecasting.

Metric	Model	2015	2016	2017	2018	2019
NLL	`ForeSWE`	**6.94**	**7.48**	**7.48**	**6.66**	**6.49**
	`Raw-GP`	11.97	16.85	22.70	17.46	18.42
ECE	`ForeSWE`	**0.14**	**0.15**	**0.18**	**0.12**	**0.15**
	`Raw-GP`	0.362	0.38	0.41	0.37	0.40
Coverage	`ForeSWE`	**80.88**	**79.57**	**76.12**	**82.53**	**79.84**
	`Raw-GP`	58.71	56.62	53.89	57.15	54.54

Table 1: Reported qualities of uncertainty estimates produced by `ForeSWE` and `Raw-GP` over the years (detailed in Table 3, Appendix). The reported metrics include NLL (Negative Log Likelihood), ECE (Expected Calibration Error), and Coverage. Higher coverage and lower values for NLL and ECE indicate better uncertainty estimates.

Uncertainty in Forecasting: In addition to predictive performance, prediction uncertainty is also important for downstream decision making. To evaluate this aspect, we further evaluate and compare the prediction uncertainty of `ForeSWE` and `Raw-GP` over entire intervals of the test years. The models were compared using the NLL, ECE and coverage metrics (defined in Appendix) where it is desirable to have a high coverage percentage with low NLL and ECE values. In the *daily* forecasting setup, Table 1 shows that `ForeSWE` has better uncertainty estimates than `Raw-GP` across all metrics for all the test years, indicating that it is a more suitable tool to be integrated into downstream decision making processes. We also have similar observations in the *weekly* forecasting setting (see Table 4 in Appendix). Both methods also have lower coverage in the weekly setting compared to the daily setting which is expected since forecasting for longer horizons is generally associated with higher uncertainty. An example output for `ForeSWE` is shown in Figure 6 of Appendix.

To further understand how the `ForeSWE` forecast skill varies by the snow period, we analyzed the locations classified into four groups based on average SWE (group 1: low, to group 4: high; Table 2; Appendix). Figure 4(a) shows how `ForeSWE` model performs across the different months, locations, and forecasting horizons, under the daily forecasting setting. Figure 7 in Appendix shows for weekly forecasting. Our key observations are as follows:

- Forecasting accuracy is high in the active snow accumulation phase (Feb, Mar) across all location groups. March signifies an onset of the melting phase in low SWE locations (group 1), which accounts for a slight increase in error.
- April is when most locations reach their peak SWE with an accelerated melt phase and the forecast errors are higher for groups 1 and 2 (Figure 4(b)). However, groups 3 and 4, which have larger snowpacks, have a median relative bias close to 0%.
- In May, groups 3 and 4 show increasing relative bias with horizon, as most of the snow is melted (Figure 4(a)). Fur-

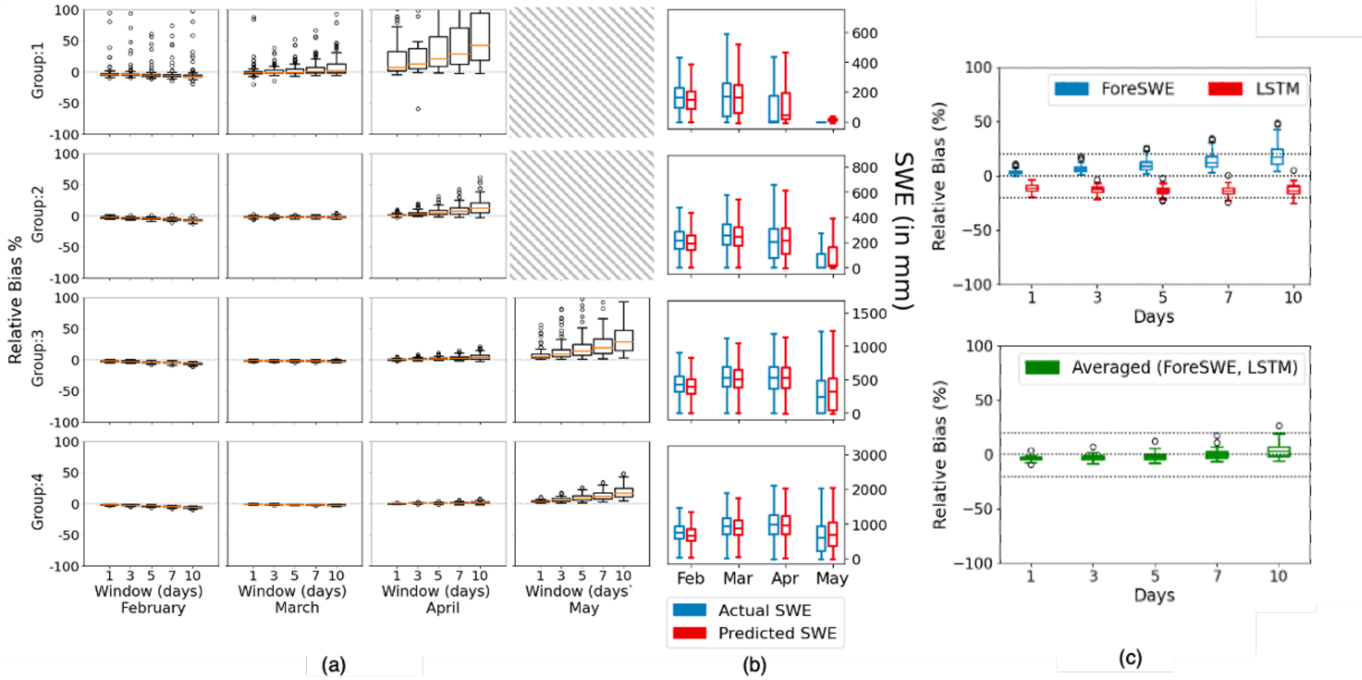


Figure 4: Daily Forecasting: (a) `ForeSWE` model's relative bias, by the different months and location groups (with forecasting horizons ranging from 1 to 10 days). Groups 1 and 2 are blocked for May as the snow has melted completely at these locations. (b) Actual and `ForeSWE` Predicted SWE availability in different groups across the active SWE months. (c) (upper) Relative bias of `ForeSWE` model against a temporal model (`LSTM`) for Group 4 in May over different forecasting windows. (lower) Relative bias with the ensembling of `ForeSWE` and `LSTM` for Group 4 in May. The dotted lines in the plot mark $\pm$ 20%.

thermore, we observed that a fully-temporal model such as `LSTM` consistently underpredicts in May, while our `ForeSWE` model overpredicts (Figure 4(c))—suggesting the suitability of an average ensemble of these two models during this low-snowpack month.

Road to Deployment

In this paper, we presented and demonstrated the effectiveness of a new uncertainty-aware attention model to forecast SWE in real-world settings. Our approach combines the strengths of the attention mechanism in exploiting spatial and temporal correlations, with the strengths of GP in uncertainty quantification.

To support deployment of this new model, our team is actively pursuing collaborative directions, building on existing partnerships with regional and state water management agencies. For example, we have engaged with a regional office of the U.S. Bureau of Reclamation (USBR) to outline the design of a dashboard that can inform their reservoir operations—a tool that is currently lacking. We plan to co-develop this dashboard with them to ensure it meets operational needs. Additionally, multiple regional agencies involved in water management coordinate through the regional River Forecast Center. Once a dashboard is implemented, we intend to broaden our engagement to include these additional partners. Through ongoing interactions with stakeholders, two critical requirements for deploying AI models have emerged: i) prediction explainability at the individual forecast level, by integrating interpretability tools (e.g., SHapley Additive exPlanations) and ablation studies, to provide actionable explanations for each prediction; and ii) transparent uncertainty quantification, which we have prioritized in the current work to bring us closer to operational readiness.

SWE is an important intermediate variable for streamflow forecasting, and with the uncertainty-aware SWE forecasting over different horizons, `ForeSWE` is capable of providing a well-calibrated input in real-time into streamflow forecasting models that USBR and its contemporaries internally uses for their reservoir management decision workflows.

Future directions: In addition to deployment efforts, we also plan to continue improving and extending the `ForeSWE` model—including expanding its applicability to previously unseen locations and incorporating feedback from end users as deployment progresses, enabling more robust and user-informed modeling. Feature-based ablation studies can further improve the interpretability of the model.

Acknowledgements

This research was supported by USDA NIFA award No. 2021-67021-35344 (AgAID AI Institute).

References

Abatzoglou, J. T. 2013. Development of gridded surface meteorological data for ecological applications and modeling. *International journal of climatology*, 33(1): 121–131.

Box, G. E. P.; and Jenkins, G. M. 1970. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Box, G. E. P.; and Jenkins, G. M. 1976. *Time Series Analysis: Forecasting and Control (2nd ed.)*. San Francisco: Holden-Day.

Brodzik, M.; Long, D.; Hardman, M.; Paget, A.; and Armstrong, R. 2016. MEaSUREs calibrated enhanced-resolution passive microwave daily EASE-grid 2.0 brightness temperature ESDR, version 1. *Digital Media*.

Bui, L. M.; Huu, T. T.; Dinh, D.; Nguyen, T. M.; and Hoang, T. N. 2024. Revisiting Kernel Attention with Correlated Gaussian Process Representation. In *The 40th Conference on Uncertainty in Artificial Intelligence*.

Cui, G.; Anderson, M.; and Bales, R. 2023. Mapping of snow water equivalent by a deep-learning model assimilating snow observations. *Journal of Hydrology*, 616: 128835.

Diro, G. T.; and Lin, H. 2020. Subseasonal Forecast Skill of Snow Water Equivalent and Its Link with Temperature in Selected SubX Models. *Weather and Forecasting*, 35(1): 273 – 284.

Franz, K. J.; Butcher, P.; and Ajami, N. K. 2010. Addressing snow model uncertainty for hydrologic prediction. *Advances in Water Resources*, 33(8): 820–832.

Garvelmann, J.; Pohl, S.; and Weiler, M. 2015. Spatio-temporal controls of snowmelt and runoff generation during rain-on-snow events in a mid-latitude mountain catchment. *Hydrological processes*, 29(17): 3649–3664. 12.02.03; LK 01.

Harpold, A. A.; Kaplan, M. L.; Klos, P. Z.; Link, T.; McNamara, J. P.; Rajagopal, S.; Schumer, R.; and Steele, C. M. 2017. Rain or snow: hydrologic processes, observations, prediction, and research needs. *Hydrology and Earth System Sciences*, 21(1): 1–22.

Hoang, Q. M.; Hoang, T. N.; and Low, K. H. 2017. A Generalized Stochastic Variational Bayesian Hyperparameter Learning Framework for Sparse Spectrum Gaussian Process Regression. In *Proc. AAAI*, 2007–2014.

Hoang, Q. M.; Hoang, T. N.; Pham, H.; and Woodruff, D. 2020. Revisiting the Sample Complexity of Sparse Spectrum Approximation of Gaussian Processes. In *Advances in Neural Information Processing Systems*.

Hoang, T. N.; Hoang, Q. M.; and Low, K. H. 2015. A Unifying Framework of Anytime Sparse Gaussian Process Regression Models with Stochastic Variational Inference for Big Data. In *Proc. ICML*, 569–578.

Hoang, T. N.; Hoang, Q. M.; and Low, K. H. 2016. A Distributed Variational Inference Framework for Unifying Parallel Sparse Gaussian Process Regression Models. In *Proc. ICML*, 382–391.

Hoang, T. N.; Hoang, Q. M.; Low, K. H.; and How, J. P. 2019. Collective Online Learning of Gaussian Processes in Massive Multi-Agent Systems. In *Proc. AAAI*.

Hoang, T. N.; Low, K. H.; Jaillet, P.; and Kankanhalli, M. 2014. Nonmyopic ϵ-Bayes-Optimal Active Learning of Gaussian Processes. In *Proc. ICML*, 739–747.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Huang, C.; Newman, A. J.; Clark, M. P.; Wood, A. W.; and Zheng, X. 2017. Evaluation of snow data assimilation using the ensemble Kalman filter for seasonal streamflow prediction in the western United States. *Hydrology and Earth System Sciences*, 21(1): 635–650.

Hunter, T.; Tootle, G.; and Piechota, T. 2006. Oceanic-atmospheric variability and western US snowfall. *Geophysical Research Letters*, 33(13).

Inman, J. 1835. *Navigation and Nautical Astronomy: For the Use of British Seamen*. London: W. Woodward.

Jennings, K.; Winchell, T.; Livneh, B.; and Molotch, N. 2018. Spatial variation of the rain–snow temperature threshold across the Northern Hemisphere. *Nature Communications*, 9.

Li, D.; Wrzesien, M. L.; Durand, M.; Adam, J.; and Lettenmaier, D. P. 2017. How much runoff originates as snow in the western United States, and how will that change in the future? *Geophysical Research Letters*, 44(12): 6163–6172.

Lim, B.; Arık, S. Ö.; Loeff, N.; and Pfister, T. 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*, 37(4): 1748–1764.

Liston, G. 2004. Representing Subgrid Snow Cover Heterogeneities in Regional and Global Models. *Journal of Climate - J CLIMATE*, 17.

Mankin, J. S.; Viviroli, D.; Singh, D.; Hoekstra, A. Y.; and Diffenbaugh, N. S. 2015. The potential for snow to supply human water demand in the present and future. *Environmental research letters*, 10(11): 114016.

Moriasi, D. N.; Arnold, J. G.; Liew, M. W. V.; Bingner, R. L.; Harmel, R. D.; and Veith, T. L. 2007. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Transactions of the ASABE*, 50(3): 885–900.

Mote, P. W.; Hamlet, A. F.; Clark, M. P.; and Lettenmaier, D. P. 2005. Declining mountain snowpack in western North America. *Bulletin of the American meteorological Society*, 86(1): 39–50.

Musselman, K.; Lehner, F.; Ikeda, K.; Clark, M.; Prein, A.; Liu, C.; Barlage, M.; and Rasmussen, R. 2018. Projected increases and shifts in rain-on-snow flood risk over western North America. *Nature Climate Change*, 8.

Nash, J. E.; and Sutcliffe, J. V. 1970. River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, 10(3): 282–290.

NED, U. 2014. United States Geological Survey, National Elevation Dataset (https://apps.nationalmap.gov/datasets/). Accessed: November 3, 2023.

Nguyen, T.; Brandstetter, J.; Kapoor, A.; Gupta, J. K.; and Grover, A. 2023a. ClimaX: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*.

Nguyen, T.; Shah, R.; Bansal, H.; Arcomano, T.; Maulik, R.; Kotamarthi, V.; Foster, I.; Madireddy, S.; and Grover, A. 2023b. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. *arXiv preprint arXiv:2312.03876*.

Nijssen, B.; and Lettenmaier, D. P. 2004. Effect of precipitation sampling error on simulated hydrological fluxes and states: Anticipating the Global Precipitation Measurement satellites. *Journal of geophysical research: atmospheres*, 109(D2).

NRCS. 2023. USDA NRCS NWCC Database for SNOTEL locations data. Accessed: 2023-08-03.

Pagano, T.; Garen, D.; and Sorooshian, S. 2004. Evaluation of official western US seasonal water supply outlooks, 1922–2002. *Journal of Hydrometeorology*, 5(5): 896–909.

Quiñonero-Candela, J.; and Rasmussen, C. E. 2005. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6: 1939–1959.

Rasmussen, C. E.; and Williams, C. K. I. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sarhadi, A.; Kelly, R.; and Modarres, R. 2014. Snow water equivalent time-series forecasting in Ontario, Canada, in link to large atmospheric circulations. *Hydrological Processes*, 28(16): 4640–4653.

Stillinger, T.; Costello, C.; Bales, R. C.; and Dozier, J. 2021. Reservoir Operators React to Uncertainty in Snowmelt Runoff Forecasts. *Journal of Water Resources Planning and Management*, 147(10): 06021010.

Thapa, K. K.; Savalkar, S.; Singh, B.; Hoang, T. N.; Rajagopalan, K.; and Kalyanaraman, A. 2025. ForeSWE: Forecasting Snow-Water Equivalent with an Uncertainty-Aware Attention Model. arXiv Preprint:2511.08856.

Thapa, K. K.; Singh, B.; Savalkar, S.; Fern, A.; Rajagopalan, K.; and Kalyanaraman, A. 2024. Attention-Based Models for Snow-Water Equivalent Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22969–22975.

Titsias, M. K. 2009. Variational learning of inducing variables in sparse Gaussian processes. In *Proc. AISTATS*.

U.S. Bureau of Reclamation, Research and Development Office. 2024. Snow Water Supply Forecasting Program. https://www.usbr.gov/research/snow/aboutsnow.html. Last updated August 20, 2024; accessed December 21, 2026.

U.S. Department of Agriculture, Natural Resources Conservation Service, National Water and Climate Center. 2025. Snow and Water Interactive Map. https://www.nrcs.usda.gov/resources/data-and-reports/snow-and-water-interactive-map. Accessed July 28, 2025; interactive mapping tool for snow water equivalent and related data :contentReference[oaicite:1]index=1.

Van der Wilk, M.; Dutordoir, V.; John, S.; Artemev, A.; Adam, V.; and Hensman, J. 2020. A framework for interdomain and multioutput Gaussian processes. *arXiv preprint arXiv:2003.01115*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yang, L.; Jin, S.; Danielson, P.; Homer, C.; Gass, L.; Bender, S. M.; Case, A.; Costello, C.; Dewitz, J.; Fry, J.; Funk, M.; Granneman, B.; Liknes, G. C.; Rigge, M.; and Xian, G. 2018. A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146: 108–123.