Prepared by/for:

Headquarters USACE

2023 General Investigation Project

Hydrologic Modeling using HEC-HMS and Machine Learning

U.S. Army Corps of Engineers

July 2024

CONTENTS

LIST OF TABLES

LIST OF FIGURES

SECTION 1
Introduction

Water distribution is paramount in the state of California, supporting nearly 40 million residents, powering the United States' most productive agricultural areas, and preserving a diverse array of freshwater species. Yet, the state faces critical challenges due to persistent droughts, extreme flooding events, and environmental degradation, highlighting challenges within its water management system [DWR, 2020; Hanak et al., 2011].

Much of the water available during the dry season comes from snowmelt or from scheduled releases from reservoirs located throughout the state. Accurate streamflow and snowpack forecasts are critical for water resources management, particularly considering climate change and variability and an overburdened infrastructure [Rhoades et al., 2018; Christian-Smith et al., 2015; Berg and Hall, 2018].

Hydrological forecasts are typically made using process-based (PB) or data-driven models (DD). PB models explicitly assume a representation of the important watershed processes in the model structure, are based on first principles (e.g. they conserve mass), but do not offer a complete representation of all processes. Examples of PB models for streamflow prediction include models of differing degrees of complexity, such as IHACRES, SAC-SMA, VIC, PRMS, and HEC-HMS. HEC-HMS [HEC, 2022] is the operational model used by the U.S. Army Corps of Engineers (USACE) for forecasting snowpack (i.e., snow water equivalent), snowmelt, and reservoir inflow.

In contrast, DD models implicitly derive the relationship between forcing and output variables directly from the data. DD models, particularly those based on Machine Learning (ML), and Deep Learning (DL), are increasingly showing great effectiveness in streamflow prediction, and have been tested over the CONUS, as well as OCONUS (e.g.: Kratzert et al., 2019, Lees et al., 2021).

Long-short-term-memory networks (LSTMs; Hochreiter & Schmidthuber, 1997), a type of recurrent DL neural network, have outperformed PB models in streamflow prediction, and are the state-of-the-art, as they are able to resolve long-term dependencies between forcing variables and predictands. However, their “black box” nature, and the uncertainty about their performance in a very different climate than that represented in training data, has slowed down their adoption in operational environments.

Recently, models that combine PB and DD approaches to insert process information into a DL model have been proposed [Willard et al., 2022]. Such models include LSTMs with dedicated architecture aimed at conserving mass [Hoedt et al, 2021], or with loss functions that include a penalty term for deviation from a water balance [Zhang et al., 2022]. These models are currently an active area of research.

A simple approach to create a physics-informed LSTM (PILSTM) is to use post-processing, where outputs of the PB model (e.g., internal state variables and predictions) are used to inform the LSTM model, in addition to the forcing variables used by both [Konapala et al., 2020; Nearing et al., 2020; Frame et al., 2021; Adeera et al, 2024]. There are many advantages to such an approach. For example, using a PB model to inform a LSTM can help apply physically realistic constraints, which is particularly important

in cases where the forecast domain is significantly different than the historical domain. At the same time, the ability of the LSTM to derive relationships from data that are not explicitly represented in the PB model, can help with bias-correcting the PB model and reflecting variable dependencies not represented in the PB model [Adera et al., 2024]. Because many of the input variables are a result of a PB model, a PILSTM is more interpretable than a typical DL model. PILSTMs have minimal impact on the operational workflow because post-processing a PB model has low computational demands.

In this case study, we compare the performance of these three approaches (PB, DD, and hybrid) to predict snow water equivalent (SWE, the amount of water available in the snow) and streamflow in a watershed of the Tule River that flows into the Shafer Dam reservoir. We test the models on both a historical time series (base case) as well as a resampled one (extreme case) to simulate a possible future climate with increasing extremes. Resampling is performed by identifying the wettest and driest years in the historical record and using them to evaluate the models, while the model training is performed on the remaining (average) years only (Table 3-4).

We find that all models exhibit good prediction skill with respect to SWE in the base case; the LSTM slightly outperforms the other models, but with minimal differences. In the extreme case the PILSTM outperforms the other models in the higher elevation snow-dominated basin, suggesting increased robustness under climate extremes. However, in the lower elevation basins with considerably less snow HEC-HMS gives the best results, suggesting that a PB model may be more robust when significant changes in phase are involved, and that a distributed model may be more appropriate when factors such as aspect, slope, and albedo also play an important role.

With respect to inflow prediction (a more difficult task for all models), a similar story emerges, but with more notable differences between the PILSTM, LSTM, and HEC-HMS models. The data-driven models, particularly the PILSTM, outperform HEC-HMS across most performance metrics. In the base case the performance of the LSTM and the PILSTM are comparable. Importantly, the PILSTM exhibits a marked improvement compared to the other approaches in the extreme climate experiments, in this case noticeably improving on the LSTM as well.

This report outlines the three methods used, the site and data used in this study, the experiments performed, and present our findings in detail. We conclude with a summary and recommendations for future investigations.

HEC staff, Matt Fleming and Natasha Sokolovskaya, developed the HEC-HMS model of the Tule River watershed. The effort included observed data preparation, and calibration/validation of the Tule River watershed HEC-HMS Model. U.C. Berkeley staff, Dino Bellugi, and students, Evan Roberts and Sumana Srivas, developed and calibrated the LSTM and PILSTM models, and performed analysis of HEC-HMS, LSTM, and PILSTM model results. Laurel Larsen is the principal investigator at U.C. Berkeley. Christopher Tennant, an employee of the Army Geospatial Center (USACE) and formerly in flood management in the Sacramento District, is the principal investigator at USACE and obtained funding for this project.

SECTION 2
Watershed Background

2.1 WATERSHED DESCRIPTION

The Tule River watershed upstream of Success Dam was used for this study. Success Dam (now commonly known as Schafer Dam) is located upstream of the town of Porterville, California in the central Sierra Nevada mountains. The watershed is approximately 390 square miles. Elevation ranges from 650 feet at the dam to over 8,000 feet along the watershed boundary. The figure below shows the watershed delineated into six subbasins. Two subbasins are highlighted. The MF_TuleR_S20 subbasin is a high elevation subbasin, the average elevation is 6286 feet, and the drainage area is 85.8 square miles. The TuleR_S10 subbasin is a low elevation subbasin, average elevation is 996 feet, and the drainage area is 32.9 square miles. As shown in Figure 2-1, a 2000-meter discretization was used for all subbasin elements.

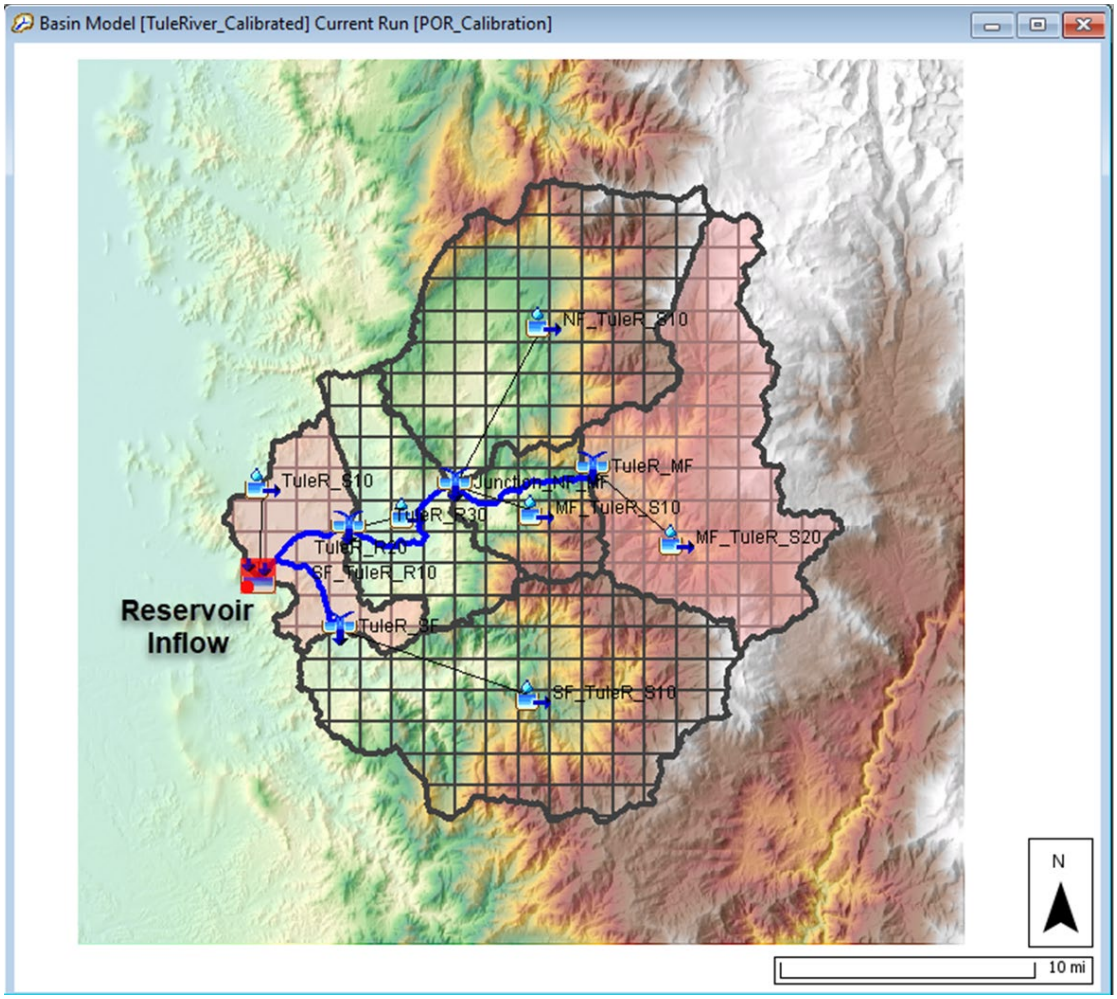


Figure 2-1 Shafer Dam Watershed.

2.2 DATA COMPILATION

This section describes the data collected for the study. The observed precipitation and temperature data were accepted as provided; additional quality control was not performed as part of this study.

2.3 PRECIPITATION DATA

Precipitation data used in the study was the Parameter-elevation Regressions on Independent Slopes Model (PRISM) dataset (http://www.prism.oregonstate.edu/). The continental-scale dataset was reprojected to a 2-kilometers grid size to match the HEC-HMS grid projection and resolution. Then, the precipitation grids were clipped to the boundary of the modeling domain. During model simulations, HEC-HMS disaggregated the gridded data from the original daily time step to the computational time step of 3-hours.

Analysis of Period of Record for Calibration (AORC) hourly precipitation data, provided from the NOAA office of Weather Prediction, was also gathered and reprojected to the 2-kilometer grid and used by the HEC-HMS model (https://hydrology.nws.noaa.gov/pub/AORC/V1.1/Documents/).

2.4 TEMPERATURE DATA

Temperature data that was used within this modeling effort was also originally generated using PRISM (http://www.prism.oregonstate.edu/). The daily average temperature dataset was used in the HEC-HMS model. During model simulation, HEC-HMS disaggregated the gridded temperature data from a daily time step to the computation time step of 3-hours.

USACE Sacramento District temperature gages were also used in the study (observational time-step was 1-hr). These data were imported into the HEC-HMS model and the HEC-HMS interpolation method was used to create temperature grids using elevations from the terrain grid, a user defined lapse rate, and the temperature data.

2.5 SNOW DATA

The University of Arizona SWE dataset was used as “observed” SWE. The gridded SWE data for water years 1982 - 2005 was downloaded from https://climate.arizona.edu/data/UA_SWE/. The University of Arizona SWE dataset was processed using the HEC-HMS Gridded Data Importer and Grid to Point tools. The subbasin average University of Arizona SWE time-series was added to the HEC-HMS as an "observed" SWE gage and linked to the appropriate subbasin elements.

2.6 STREAMFLOW

Daily average and 1-hour inflow into Schafer Dam was provided by the USACE Sacramento District. The flow data was computed using the observed reservoir stage (converted to storage), discharge from the reservoir, and the continuity equation. The 1-hour inflow dataset was not directly used when training the LSTM model because the computed inflow dataset was noisy, the flow time-series oscillated dramatically and did not represent natural conditions.

The Near Springville USACE gage is located upstream of Schafer Lake and includes 1-hour observations. The Near Springville gage was used to evaluate model performance when the hourly AORC and temperature gages were applied to the model. The Near Springville gage does not capture runoff from the entire Shafer Dam watershed.

2.7 SOFTWARE AND DOCUMENTATION

Table 2-1 provides a summary of the computer programs and versions used in the study.

Table 2-1 Computer Programs Utilized.

Program	Version	Capability	Developer
HEC-HMS	4.12	Snow and precipitation-runoff simulation	HEC
HEC-DSSVue	3.3	Plot, tabulate, edit, and manipulate data in HEC-DSS files	HEC
MATLAB	2024b	Read CSV-formatted data into table, train LSTM and PILSTM networks, output CSV-formatted results	The Mathworks Inc.
Jupyter Notebook	6.5.7	Read CSV-formatted results, produce plots and tables	Open-Source, packaged by Anaconda Inc.
Python	3.11.7	Scripting language used in Jupyter Notebook	Open-Source, packaged by Anaconda Inc.
Anaconda Navigator	2.6.0	Wrapper for Jupyter Notebook and Python scripting language	Anaconda Inc.

SECTION 3
HEC-HMS Model Development

An HEC-HMS model of the watershed upstream of Shafer Dam was developed for this study. Different modeling options were explored and compared to the LSTM model. State variables from the HEC-HMS model were also used as predictor variables for some of the LSTM models.

The period of record used for model simulations was October 1, 1981, through September 30, 2005. As described below, the period of record was divided into different calibration and validation periods.

3.1 INFILTRATION

Infiltration computations were executed using the Deficit and Constant and Soil Moisture Accounting Loss methods. Both loss models were used in conjunction with a canopy method, the canopy model removes moisture from the soil layer in the loss models. Method parameters were determined through model calibration. The HEC-HMS User's and Technical Reference manuals described method parameters and computational procedures.

3.2 EVAPOTRANSPIRATION

The Hamon evapotranspiration method was used to simulate evapotranspiration (ET) losses throughout the modeling domain. Within the Hamon method, ET losses are directly proportional to the daily average temperature and related to the location of interest and time of year. A modified, gridded version of the Hamon method was used to estimate potential ET losses using the previously mentioned daily average temperatures and a coefficient.

3.3 UNIT HYDROGRAPH TRANSFORM

The modified Clark (ModClark) unit hydrograph (UH) transform method was used to route excess precipitation to the subbasin outlet within each subbasin. This linear, quasi-distributed transform method uses a set of grid cells to represent travel times within a subbasin to the outlet point. As such, it explicitly accounts for variations in travel time from all areas within a subbasin using a time travel index for each grid cell. These grid cells were laid out using the Standard Hydrologic Grid (SHG) system with a 2- x 2-kilometer resolution.

3.4 BASEFLOW

The Linear Reservoir baseflow method was used to transform water infiltrated into interflow and baseflow and add these components to any direct runoff generated within each subbasin. For this modeling effort, the storage and movement of infiltrated water was simulated using three layers when coupled to the Deficit and Constant loss model and two layers when coupled to the Soil Moisture Accounting loss model. The layers are considered "linear" since the outflow at each time step of the simulation is a linear function of

the average storage during the time step. The modeler can control how water is distributed to the different groundwater layers using loss and baseflow parameters.

3.5 STREAMFLOW ROUTING

No stream routing was included in the model due to the larger simulation time steps and use of daily precipitation, temperature, and flow data to calibration and validate the model.

3.6 SNOWMELT

The Gridded Temperature Index and Gridded Radiation Temperature Index methods were used to simulate snowmelt processes within the HEC-HMS model. The HEC-HMS User's and Technical Reference manuals [HEC, 2022; HEC, 2000] describe method parameters and computational procedures.

3.7 CALIBRATION PARAMETERS AND APPROACH

Model performance was evaluated by comparing computed results against observed results. Model parameters were altered to minimize the differences between computed and observed SWE and inflow into Shafer Dam. Summary statistics were used to quantify model performance compared to observations. Statistics include Nash-Sutcliffe Efficiency (NSE) [Nash and Sutcliffe, 1970] and Percent Bias (PBIAS) [Yapo et al., 1996].

NSE measures the relative magnitude of the residual variance compared to the measured data variance. NSE ranges between -∞ and 1, where NSE = 1 is optimal. Values of NSE ≤ 0 indicate the mean observed value is a better predictor than the simulated value. NSE is computed using the equation listed in Section 4-3.

PBIAS measures the average tendency of the simulated data to be larger or smaller than the observed data. The optimal value for PBIAS is 0.0, with low absolute PBIAS indicating accurate model simulation. PBIAS is computed using the equation listed in Section 4-3

Summary statistic performance ratings are presented in Table 3-1. The ranges used for each of the performance ratings are based on values found in literature; however, the values used for the ranges were modified for this study due to the temporal scale used to simulate precipitation-runoff processes and the precipitation and temperature boundary condition data applied to the model.

Table 3-1 HEC-HMS Performance Ratings for Summary Statistics.

Performance Rating	NSE	PBIAS
Very Good	$0.75<NSE\leq 1.00$	$PBIAS< \pm 15$
Good	$0.55<NSE\leq 0.75$	$\pm 15\leq PBIAS<\pm 20$
Satisfactory	$0.40<NSE\leq 0.55$	$\pm 20\leq PBIAS<\pm 30$
Unsatisfactory	$NSE\leq 0.40$	$PBIAS\geq \pm 30$

3.8 HEC-HMS SWE CALIBRATION AND VALIDATION

As mentioned, two HEC-HMS snowmelt models were developed, the temperature index and radiation temperature index models. The first calibration and validation effort involved dividing the period of record into a calibration period, October 1981 – September 1997, and a validation period, October 1997 – September 2005. This is referred to as the base case calibration and validation periods. During model calibration, key snowmelt method parameters were adjusted within a reasonable range for the model to reproduce the observed data. Key snowmelt model parameters include the temperature discriminating liquid or frozen precipitation, the temperature initiating snowpack melt, and the melt rate antecedent temperature index relationship for the temperature index model and the rain threshold air temperature, snow threshold air temperature, and the melt factor for the radiation temperature index model. The parameter values determined through model calibration were then set for the validation simulation.

Figure 3-1 shows the simulated and observed snow water equivalent for the MF_TuleR_S20 subbasin for both the temperature index and radiation temperature index snowmelt methods for the calibration period. The HEC-HMS model results and observed snow water equivalent time-series were converted to monthly average values, and the average value for each month was computed to create the summary plot shown below. As shown, both temperature index and radiation temperature index HEC-HMS models successfully simulated the snowmelt accumulation and melt pattern for the MF_TuleR_S20 subbasin. All plots in this section are for a typical water year; therefore, month 1 is October.

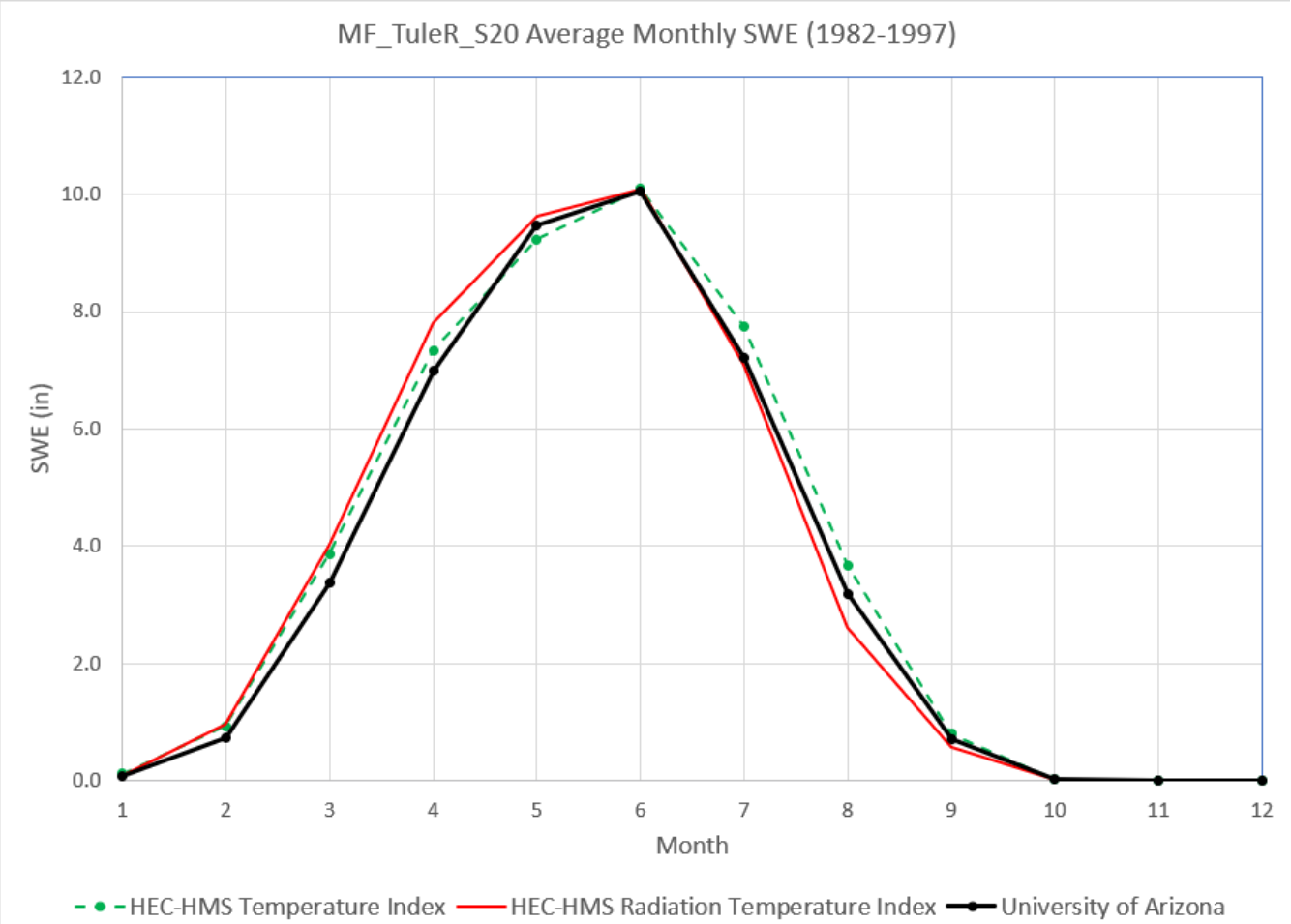


Figure 3-1 Monthly Average Snow Water Equivalent for the MF_TuleR_S20 Subbasin – Base Case Calibration Period.

Figure 3-2 shows the simulated and observed snow water equivalent for the MF_TuleR_S20 subbasin for both the temperature index and radiation temperature index snowmelt methods for the base case validation period. Results show that the calibrated model parameters can reproduce the snow accumulation and melt pattern for the validation period.

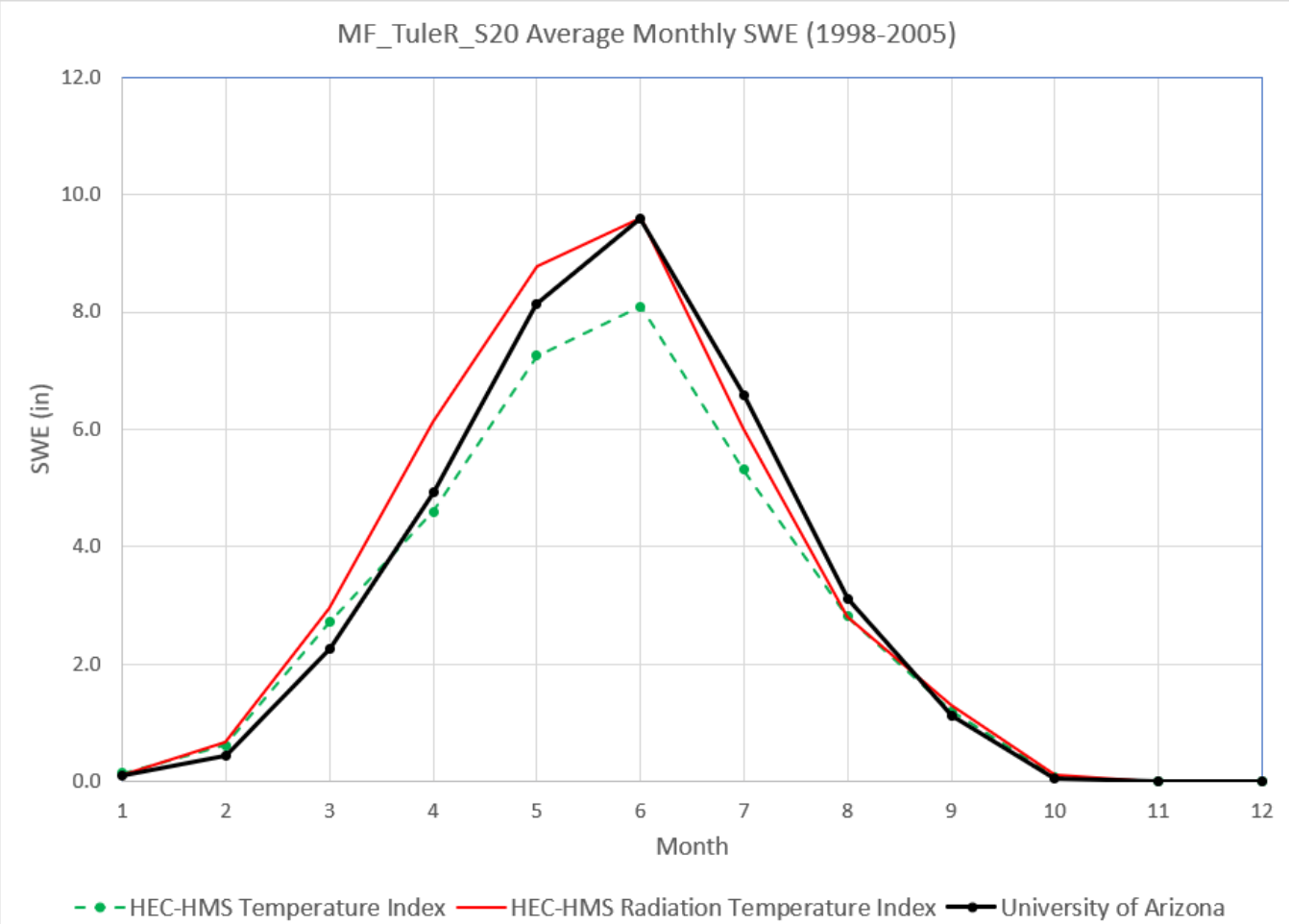


Figure 3-2 Monthly Average Snow Water Equivalent for the MF_TuleR_S20 Subbasin – Base Case Validation Period.

Table 3-2 contains performance metrics for the three highest elevation subbasins for the temperature index and radiation temperature index models. The radiation temperature index model performs slightly better in the base case calibration period.

Table 3-2 Monthly Average Snow Water Equivalent Performance Metrics for the Base Case Calibration Period.

	Temperature Index		Radiation Temperature Index	
Subbasin	**NSE**	**PBIAS**	**NSE**	**PBIAS**
MF_TuleR_S20	0.88	4.73	0.91	2.48
NF_TuleR_S10	0.91	2.48	0.91	3.71
MF_TuleR_S20	0.77	3.37	0.84	3.36

Table 3-3 contains performance metrics for the three highest elevation subbasins for the temperature index and radiation temperature index models. Both snow models perform well in the base case validation period.

Table 3-3 Monthly Average Snow Water Equivalent Performance Metrics for the Base Case Validation Period.

	Temperature Index		Radiation Temperature Index	
Subbasin	**NSE**	**PBIAS**	**NSE**	**PBIAS**
MF_TuleR_S20	0.95	-9.58	0.94	6.10
NF_TuleR_S10	0.93	-15.32	0.92	12.51
MF_TuleR_S20	0.92	-6.08	0.87	12.99

An additional round of calibration and validation was carried out where the four years with the largest and smallest measured SWE were set aside for validation, as shown in Table 3-4. This is referred to as the extreme case calibration and validation periods. The extreme case was used to see how well the HEC-HMS and LSTM models performed when applied to conditions not experienced during model calibration. Only the radiation temperature index model was used within this additional calibration and validation exercise.

Figures 3-3 and 3-4 show SWE extreme case calibration and validation results for the MF_TuleR_S20 subbasin. The NSE and PBIAS for the MF_TuleR_S20 subbasin are 0.88 and -3.11, respectively, for the calibration period and 0.93 and -4.12, respectively, for the validation period. Similar trends were seen for other subbasins in the model. HEC-HMS can simulate the accumulation and melting of snow in the watershed when the events used for model validation are larger and smaller than the ones used to calibrate the model.

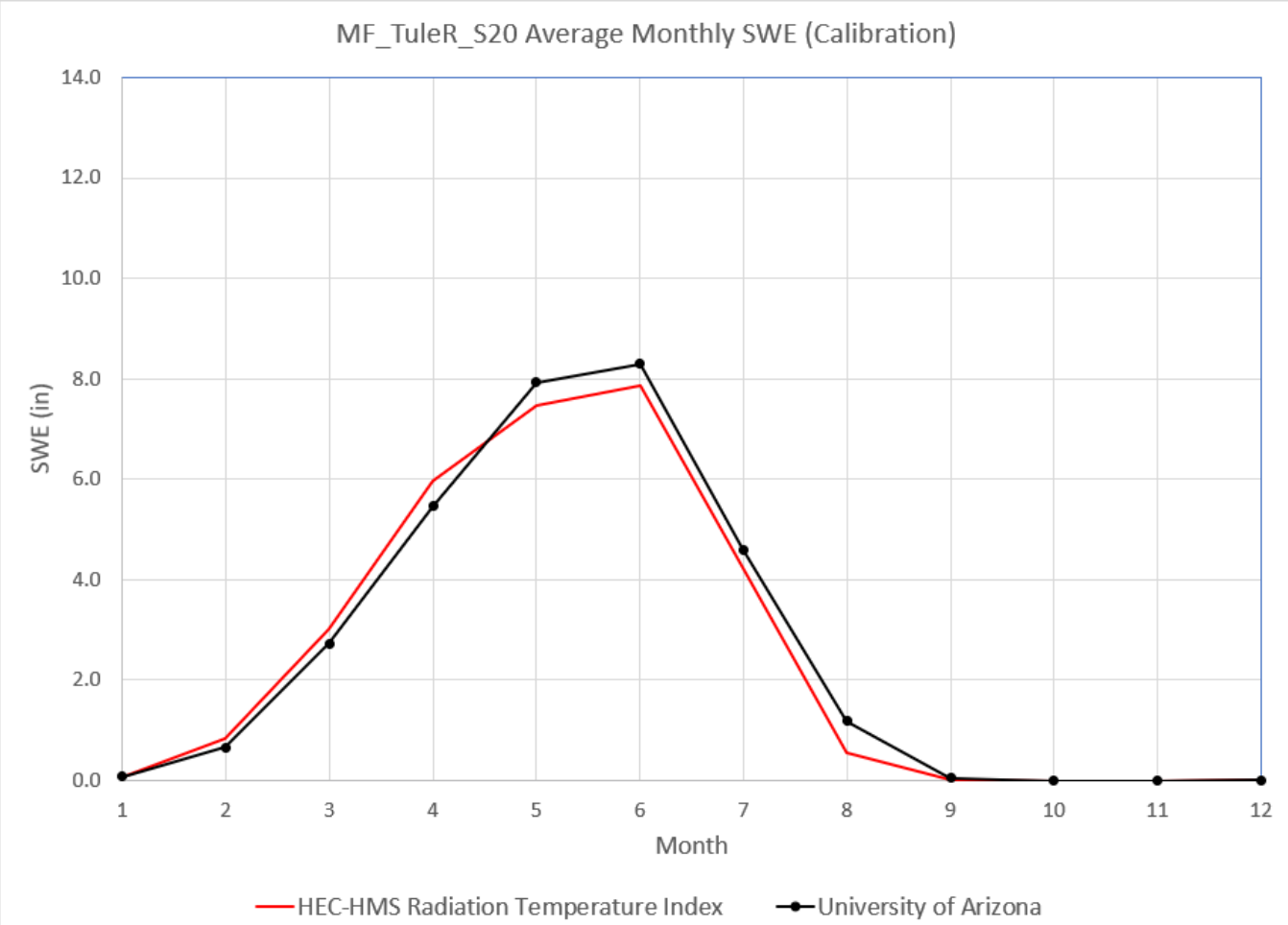


Figure 3-3 Monthly Average Snow Water Equivalent for the MF_TuleR_S20 Subbasin – Extreme Case Calibration Period.

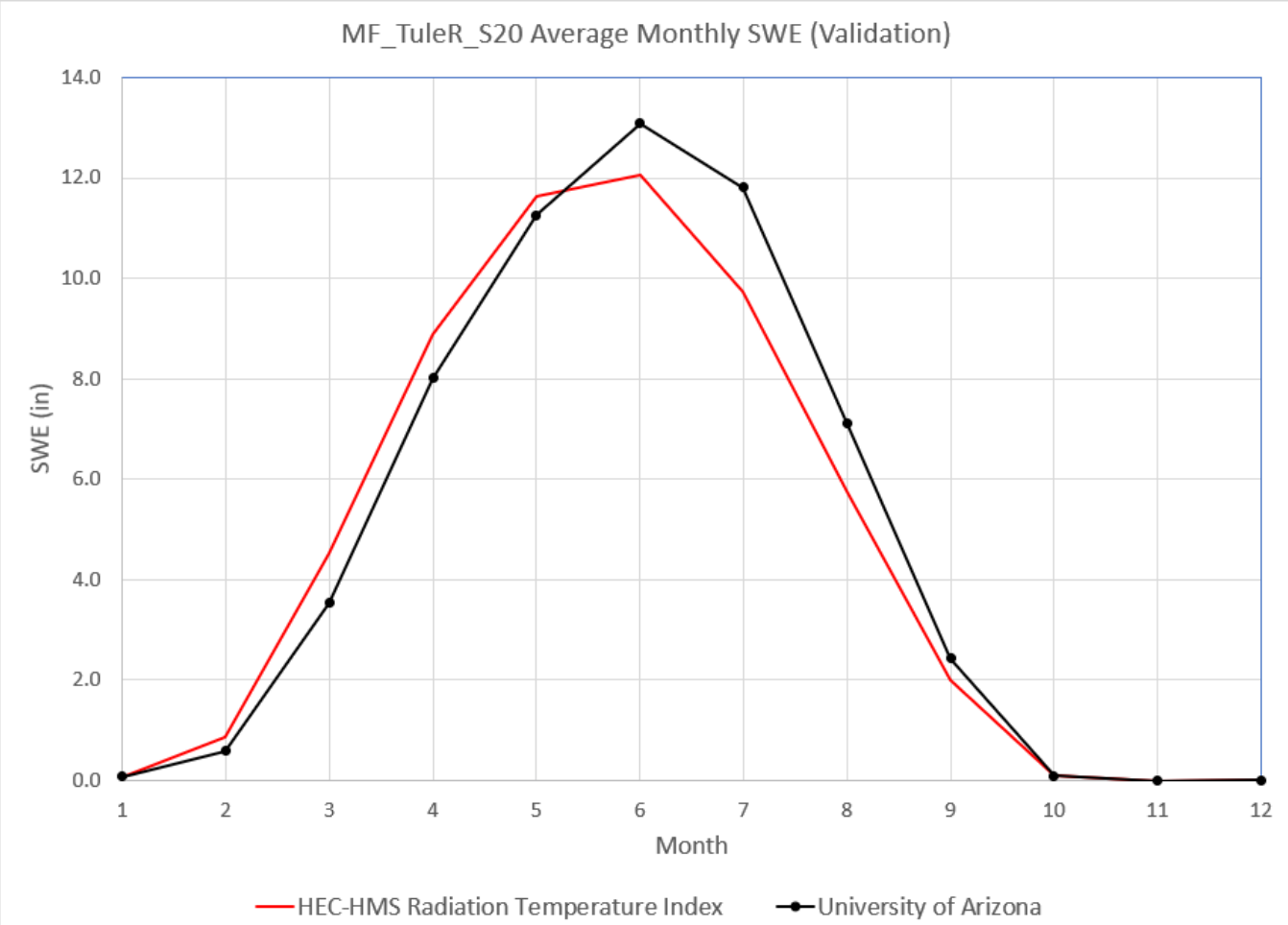


Figure 3-4 Monthly Average Snow Water Equivalent for the MF_TuleR_S20 Subbasin – Extreme Case Validation Period.

3.9 HEC-HMS FLOW CALIBRATION AND VALIDATION

The temperature index snowmelt model was selected for the flow modeling effort (due to run times being faster than the radiation temperature index model). As mentioned, both the deficit and constant and soil moisture accounting loss models were applied for simulating inflow into Shafer dam. Conceptually, the deficit and constant loss model is a one-layer model, and the soil moisture accounting model is a three-layer model. Both these loss models are linked to the canopy model, which is used to extract soil moisture, and baseflow model which is used to route saturated groundwater to the surface stream network.

Only the key parameter in the deficit and constant, soil moisture accounting, and linear reservoir baseflow models were adjusted during model calibration. Snow, canopy, unit hydrograph, and reach routing parameters were not adjusted during the flow model calibration. Parameters were adjusted within reasonable ranges. Key parameters for the deficit and constant model include the maximum deficit and the constant loss rate. When the Linear Reservoir baseflow model is linked to the deficit and constant

model, the key baseflow parameters include the fraction and storage coefficient. Key soil moisture accounting parameters included the maximum infiltration rate, soil storage, tension storage, groundwater storage, groundwater percolation rates, and groundwater storage coefficients. The parameter values determined through model calibration were then set for the validation simulations. Similarly to the snowmelt models, the base and extreme case calibration and validation periods were used, as shown in Table 3-4.

A note about simulation run times. The HEC-HMS models took approximately 60 seconds to run one 16-year simulation at a 3-hour simulation time step.

Figure 3-5 shows the simulated and observed flow for both the deficit and constant and soil moisture accounting loss models for the base case calibration period. The HEC-HMS model results and observed flow were converted to monthly average values, and the average value for each month was computed to create the summary plot. As shown, both deficit and constant and soil moisture accounting loss models capture the runoff response in the watershed.

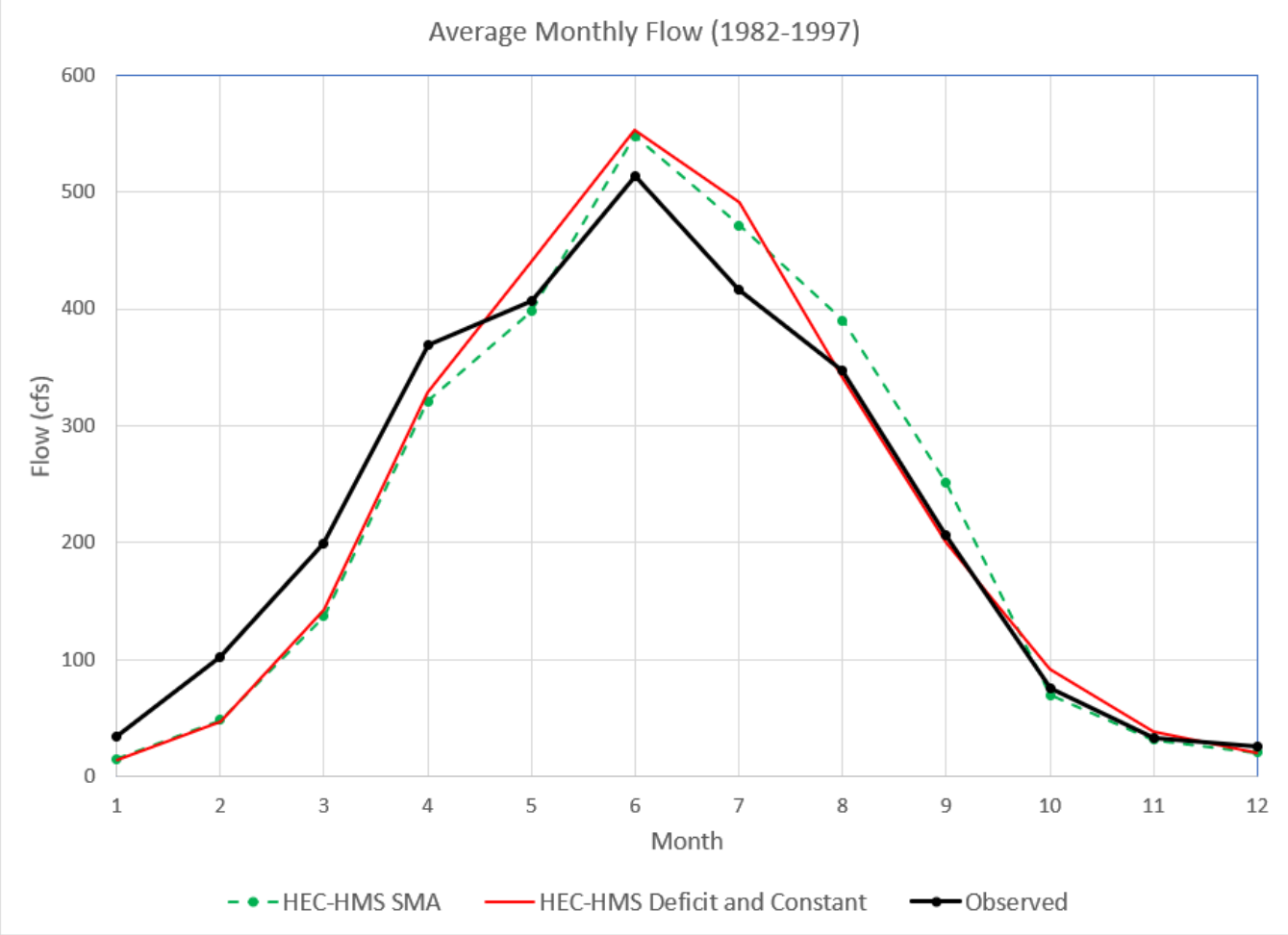


Figure 3-5 Monthly Average Inflow for the Base Case Calibration Period.

Figure 3-6 shows the simulated and observed flow for both the deficit and constant and soil moisture accounting loss models for the base case validation period. Results demonstrate that the calibrated model parameters can reproduce the runoff pattern for the validation period.

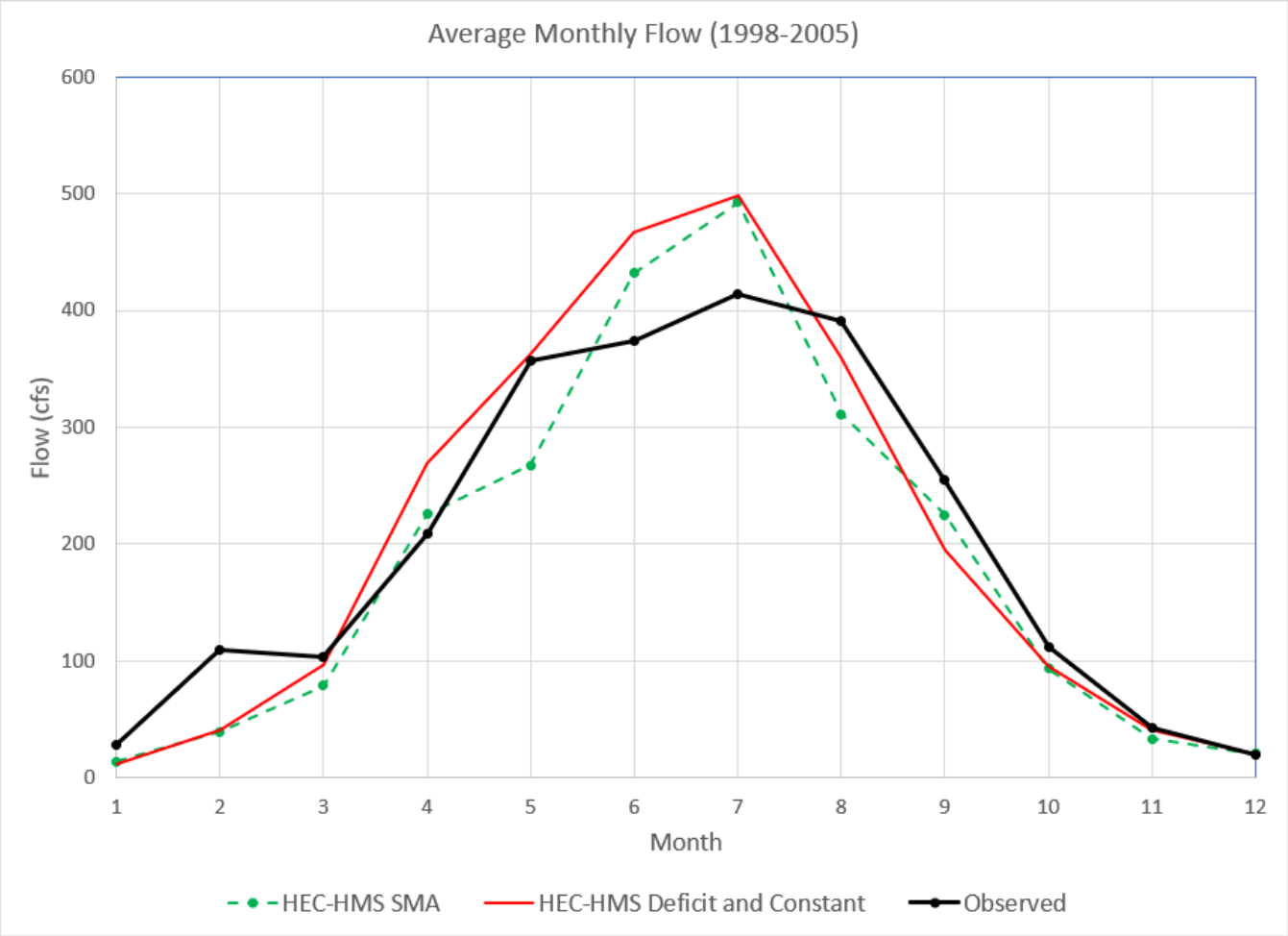


Figure 3-6 Monthly Average Inflow for the Base Case Validation Period.

Figure 3-7 shows the monthly average inflow into Shafer Dam for the entire period of record (base case), calibration, and validation periods, using both the deficit and constant and soil moisture account loss models.

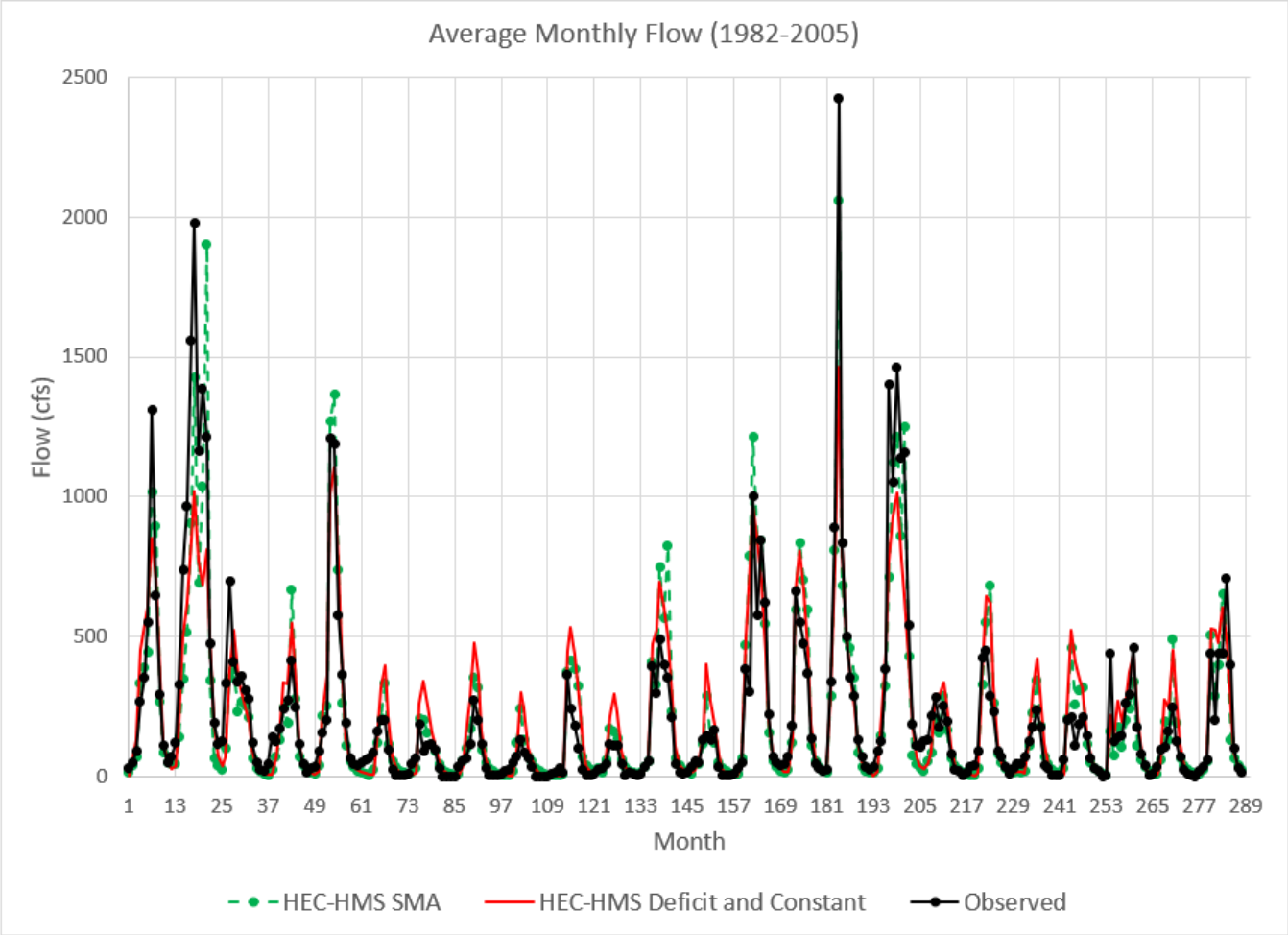


Figure 3-7 Monthly Average Inflow for the Period of Record (Base Case).

An additional round of calibration and validation was carried out where the four years with the largest and smallest measured average flow years were set aside for validation. This extreme test case was used to see how well the HEC-HMS and LSTM models performed when applied to conditions not experienced during model calibration. Only the soil moisture accounting loss model was used within this addition calibration and validation exercise. The water years used for calibration and validation are listed in Table 3-4.

Table 3-4 Water years used for calibrating and validating HEC-HMS SWE models in base case and extreme case scenarios.

Data split	Base Case	Extreme Case (SWE)	Extreme Case (Flow)
Calibration years (train)	1982 - 1997	1982, 1984-1989, 1991, 1994, 1996-1997, 1999-2001, 2004-2005	1982, 1984-1987, 1989, 1991, 1993, 1996, 1999-2005
Validation years (test)	1998 - 2005	1983, 1990, 1992-1993, 1995, 1998, 2002-2003	1983, 1988, 1990, 1992, 1994-1995, 1997-1998

Figures 3-8 and 3-9 show calibration and validation results for the extreme test case. HEC-HMS can simulate reservoir inflow in the watershed when the events used for model validation are larger and smaller than the ones used to calibrate the model.

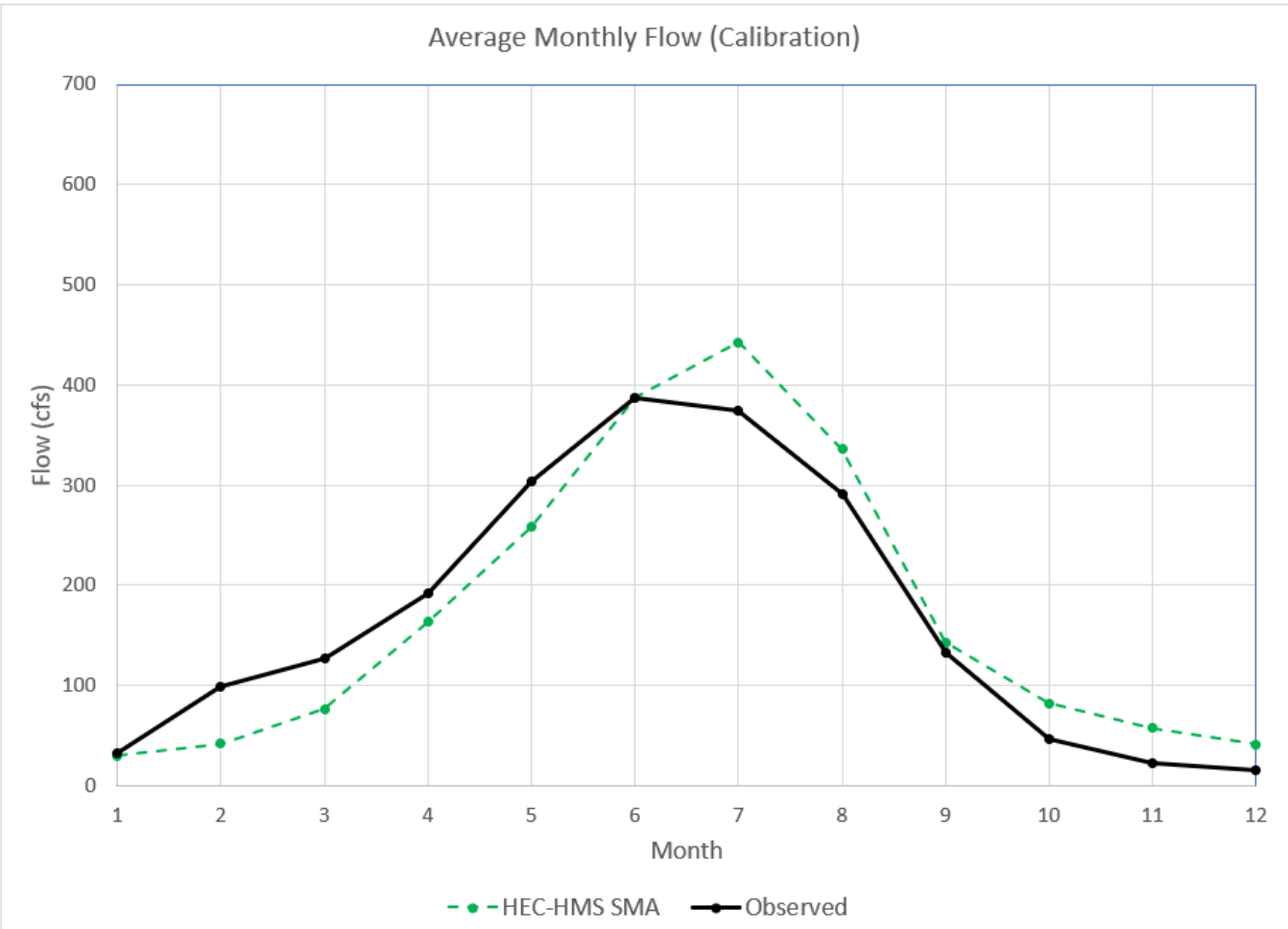


Figure 3-8 Monthly Average Inflow for the Calibration Period – Extreme Test Case.

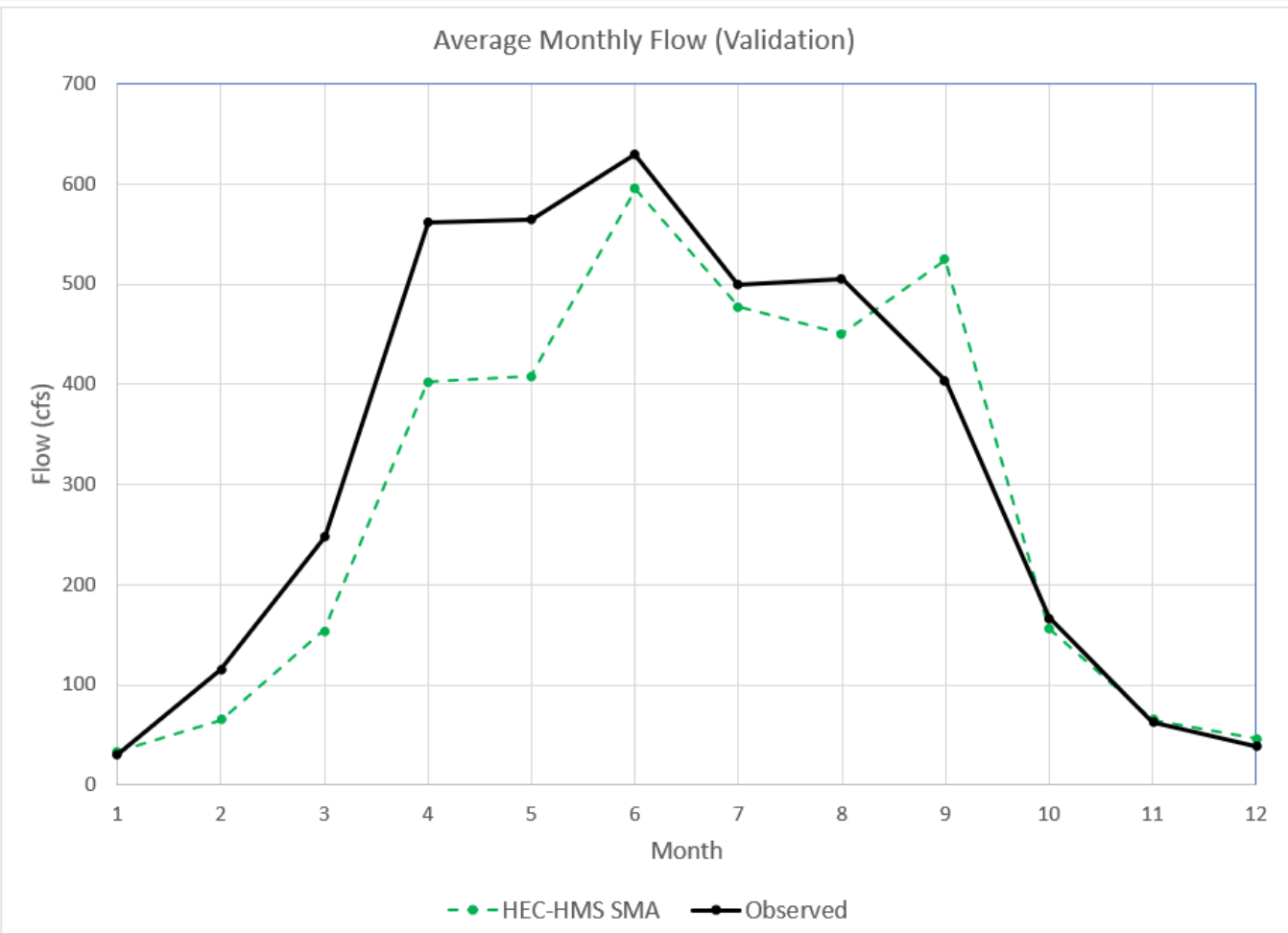


Figure 3-9 Monthly Average Inflow for the Validation Period – Extreme Test Case.

Table 3-5 contains performance metrics for the different loss models and calibration and validation periods. The soil moisture accounting loss model outperformed the deficit and constant loss model.

Table 3-5 Monthly Average Flow Performance Metrics.

	Calibration		Validation	
Loss Model	NSE Temperature Index	PBIAS Temperature Index	NSE Temperature Index	PBIAS Temperature Index
Deficit and Constant (base case)	0.78	-0.8	0.74	1.7
SMA (base case)	0.84	-1.0	0.82	-7.5
SMA (extreme case)	0.79	1.86	0.83	-11.7

One final model configuration was created using hourly precipitation and temperature data. Hourly AORC precipitation and temperature gage data (interpolated) were applied to an HEC-HMS configuration that included the temperature index snow model and soil moisture accounting loss model. The model was calibrated to water years 1996, 1998, and 2003 and validated to water years 1997 and 2005. One parameter set was determined for the calibration period and then applied to the validation period. The purpose of this additional modeling effort was to identify whether the HEC-HMS model could predict the hourly runoff response using hourly boundary condition data.

Hourly results are shown in the Figures 3-10 and 3-11. The NSE (computed from hourly results converted to daily average time-series) and PBIAS for the calibration period was 0.7 and -8.3, respectively. The NSE (computed from hourly results converted to daily average time-series) and PBIAS for the validation period was 0.79 and 2.4, respectively. Results demonstrate the HEC-HMS model can adequately predict the runoff response, daily flow magnitude, using hourly boundary conditions.

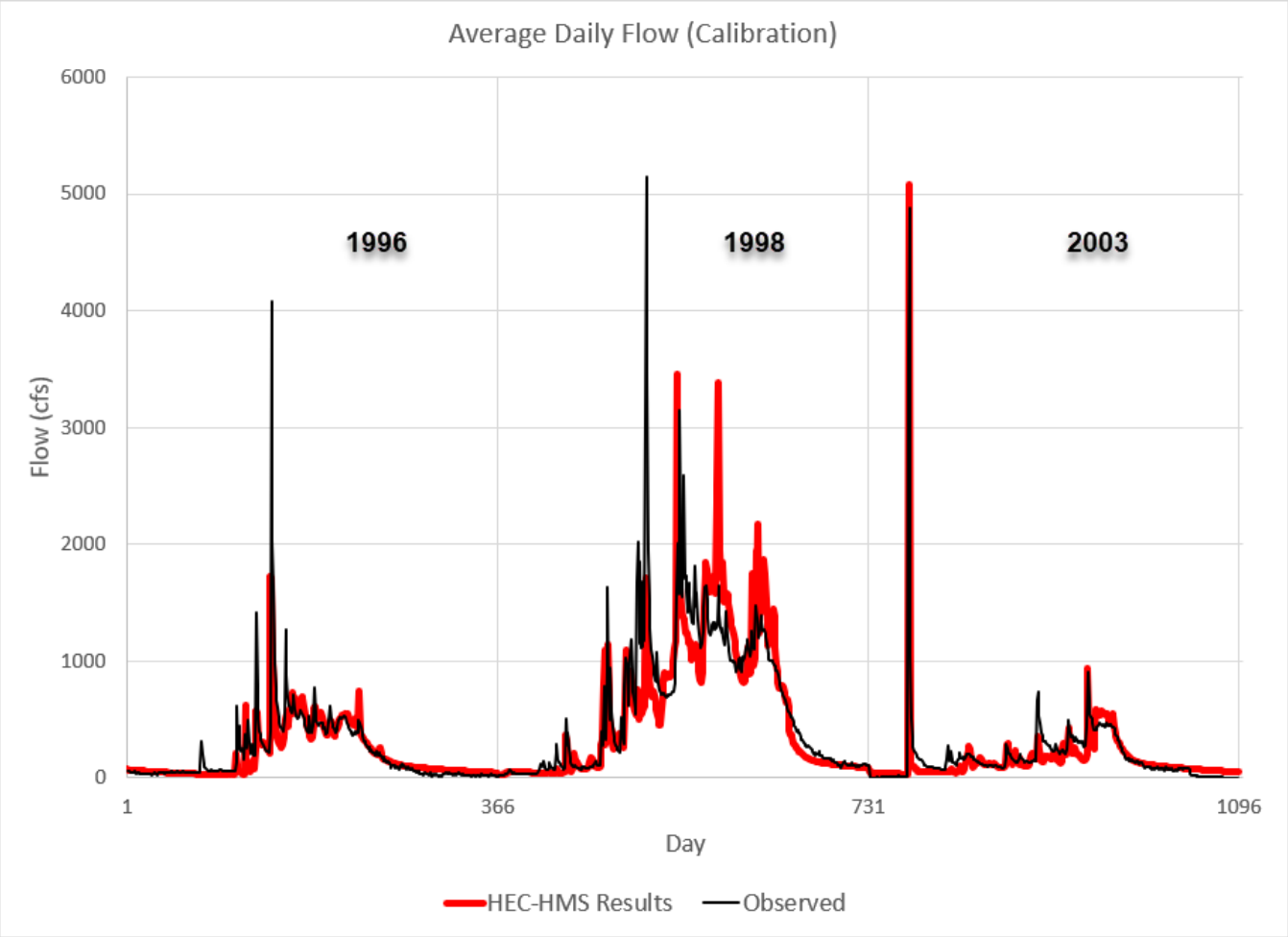


Figure 3-10 Hourly Flow Results aggregated to a daily average for the Calibration Period – 1996, 1998, and 2003.

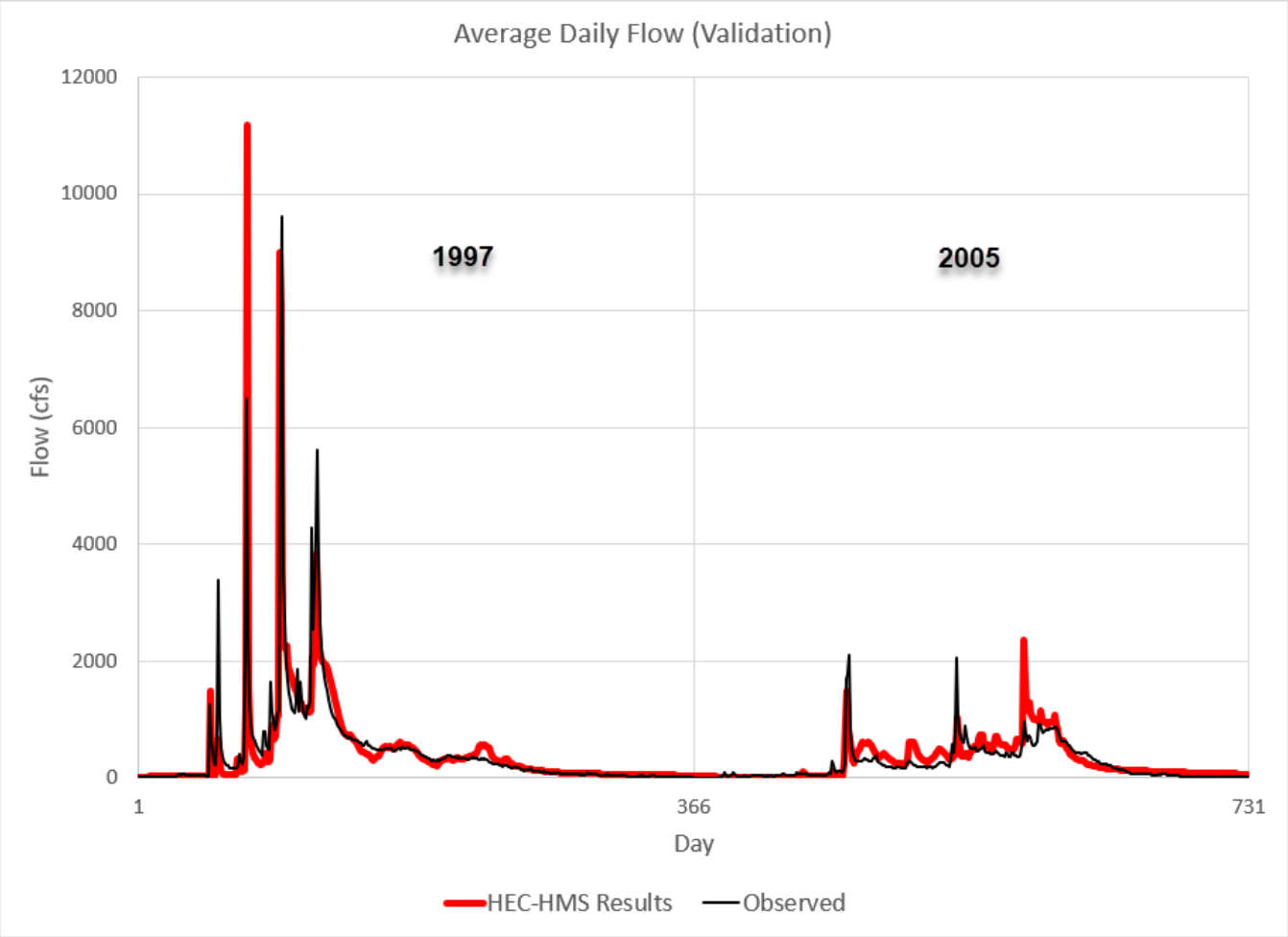


Figure 3-11 Hourly Flow Results aggregated to a daily average for the Validation Period – 1997 and 2005.

SECTION 4
Machine Learning Model Development

4.1 MODEL DESCRIPTION

The ML model adopted as the core of our DL and hybrid DL/PB approaches is an LSTM [Hochreiter & Schmidhuber, 1997]. LSTMs are a type of DL neural model with a recurrent neural network (RNN) architecture and are considered to represent the state of the art for data-driven hydrologic forecasting (e.g. Kratzert et al., [2019]; Nearing et al., [2021]; Lees et al., [2021]). They have dedicated memory cells that selectively store information over arbitrarily long time periods, enabling them to resolve long-term dependencies between input predictor variables. More specifically their structure consists of a cell state and three gates (input gate, forget gate, and output gate) which regulate the flow of information, where the cell state is the information "conveyor belt" that maintains long-term dependencies. The input, forget, and output gates control the extent to which new information is added to the cell state, what portion is instead discarded from the previous state, and what part of the current state should be used for the current prediction. They are trained using traditional backpropagation as their gated structure prevents issues like vanishing and exploding gradients that occur with traditional RNNs.

The following equations define the structure of the LSTM:

$$\boldsymbol{i}[t] = \sigma\left(\mathbf{W}_{\mathrm{i}}\boldsymbol{x}[t] + \mathbf{U}_{\mathrm{i}}\boldsymbol{h}[t-1] + \boldsymbol{b}_{\mathrm{i}}\right),$$
$$\boldsymbol{f}[t] = \sigma\left(\mathbf{W}_{\mathrm{f}}\boldsymbol{x}[t] + \mathbf{U}_{\mathrm{f}}\boldsymbol{h}[t-1] + \boldsymbol{b}_{\mathrm{f}}\right),$$
$$\boldsymbol{g}[t] = \tanh\left(\mathbf{W}_{\mathrm{g}}\boldsymbol{x}[t] + \mathbf{U}_{\mathrm{g}}\boldsymbol{h}[t-1] + \boldsymbol{b}_{\mathrm{g}}\right),$$
$$\boldsymbol{o}[t] = \sigma\left(\mathbf{W}_{\mathrm{o}}\boldsymbol{x}[t] + \mathbf{U}_{\mathrm{o}}\boldsymbol{h}[t-1] + \boldsymbol{b}_{\mathrm{o}}\right),$$
$$\boldsymbol{c}[t] = \boldsymbol{f}[t] \odot \boldsymbol{c}[t-1] + \boldsymbol{i}[t] \odot \boldsymbol{g}[t],$$
$$\boldsymbol{h}[t] = \boldsymbol{o}[t] \odot \tanh\left(\boldsymbol{c}[t]\right),$$

In the above equations, $\boldsymbol{x}$ is an input sequence $[\boldsymbol{x}[1],\ldots,\boldsymbol{x}[T]]$ with T time steps, with each element $\boldsymbol{x}[t]$ being a model inputs vector for time step t. The input, forget, and output gates are $\boldsymbol{i}[t]$, $\boldsymbol{f}[t]$, and $\boldsymbol{o}[t]$, respectively. The cell input is $\boldsymbol{g}[t]$, while $\boldsymbol{h}[t\text{-}1]$ and $\boldsymbol{c}[t\text{-}1]$ are the recurrent inputs and the cell state from the previous time step, respectively. $\boldsymbol{W}$, $\boldsymbol{U}$, and $\boldsymbol{b}$ are parameters for each gate (learned during training), with subscripts indicating which gate they are used for, $\sigma()$ is the sigmoid function, tanh() is the hyberbolic tangent function and $\odot$ is the element-wise multiplication.

LSTMs have many more parameters than the PB models, including internal weights and biases W, U, and b in the equations above), which are obtained during training. Additionally, LSTMs necessitate a set of hyperparameters defining the architecture of the system (e.g. number of layers, cells, training epochs, learning and dropout rates), which we determined by evaluating performance using an extended set of parameter combinations (grid-search, LaValle et al., [2004]). The hyper parameters chosen in this study are listed in Table SI-1.

The LSTM is trained using time sequences of predictor variables from the training dataset together with the respective daily observations. Parameters are then iteratively adjusted to minimize a loss function, in this case the Mean Squared Error between predictions and observation. Because of the stochasticity derived from the random initialization of weights and biases, as well as from the random dropout layer in which neurons are disconnected according to a specified dropout probability, each simulation consists of an ensemble of ten runs. Then, for each prediction we use the median value from the ensemble at each time step [Tennant et al., 2020].

We ran the LSTM-based models on the test dataset in standalone mode (i.e. using only the observational meteorological forcings) or as a PILSTM whereby the input variables are augmented with the HEC-HMS outputs (Figure 4-1). As mentioned earlier, the inclusion of HEC-HMS outputs provides the PILSTM additional physically based constraints. For SWE forecasts, a separate model is developed for each of the three sub-basins of interest (MF Tule S20, NF Tule S10, and SF Tule S10), while for inflow forecast a single model is developed for the entire basin.

As mentioned in Section 3, we perform two different test splits on the historical data described in Section 2. In the base case we train the models on the first part of the data and test on the remaining data. In a separate set of simulations, we chose as the test period the three driest and three wettest water years and trained the models on the rest of the historical data (Table 3-4). This re-sampling of the data is meant to simulate a future time period where extremes are increased with respect to the historical case.

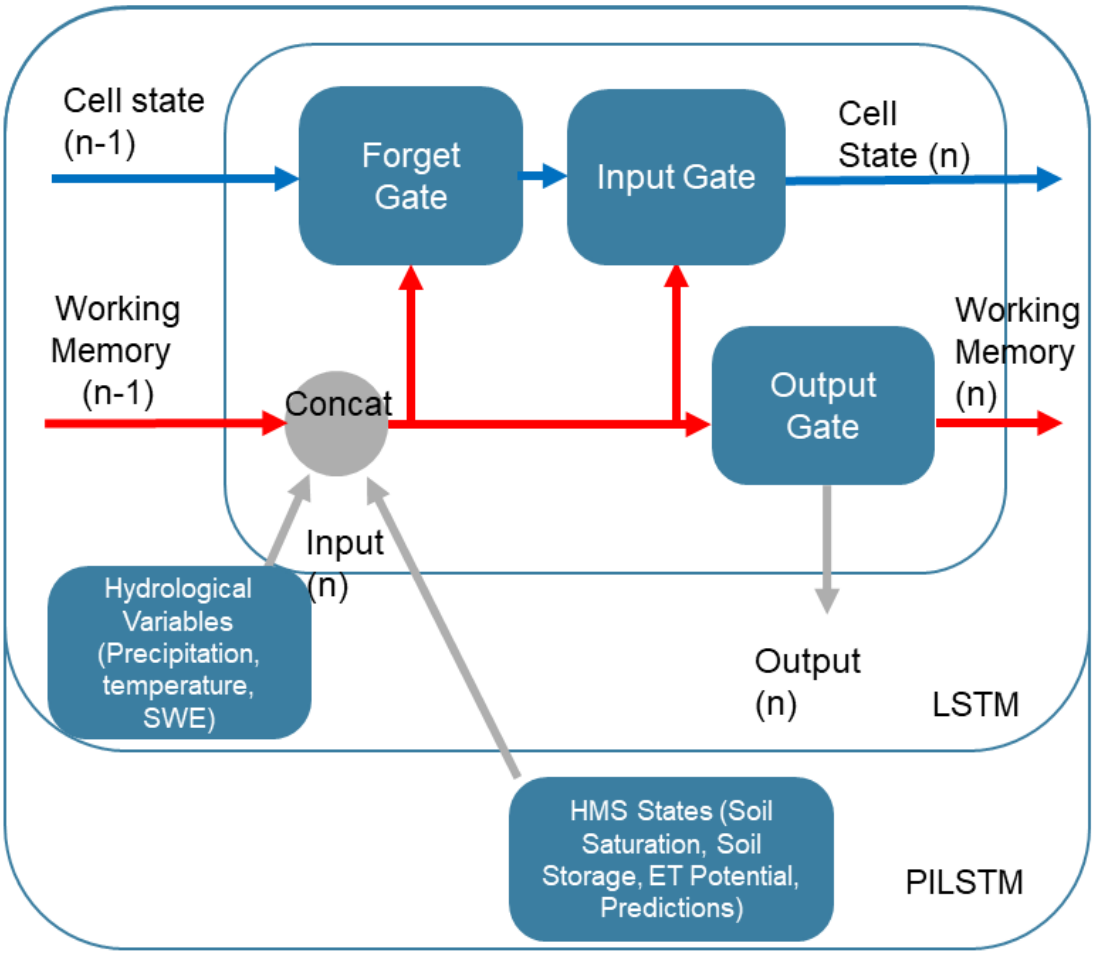


Figure 4-1 Schematic of LSTM and PILSTM architectures.

4.2 PREDICTOR VARIABLES

SWE predictor variables consist of meteorological forcings and HEC-HMS outputs. Table 4-1 provides a complete list of variables for both the LSTM and PILSTM. Note that each of the three sub-basins has its own set of values.

Table 4-1 SWE Variables.

Variables	Units	LSTM	PILSTM
SWE_OBSERVED	Inches	Predictand	Predictand
PRECIP_INC	Inches/day	✓	✓
TEMPERATURE_AVG	Degrees F	✓	✓
TEMPERATURE_MIN	Degrees F	✓	✓
TEMPERATURE_MAX	Degrees F	✓	✓
ET_POTENTIAL	Inches/day	×	✓
LW_RAD	Lang/min	×	✓
SW_RAD	Lang/min	×	✓
ALBEDO-EFFECTIVE	Dimensionless	×	✓
ATI	Degrees F	×	✓
HEAT_DEFICIT	Inches	×	✓
TEMPERATURE-RAD	Degrees F	×	✓
SWE_HMS	Inches	×	✓

Similarly, inflow predictor variables consist of meteorological forcings (including SWE observations) and HEC-HMS outputs. Table 4-2 provides a complete list of variables for both the LSTM and PILSTM. Note that in this case there is only one set of variables but those include observations and states for each sub-basin. The complete table is shown in Table SI-2.

All variables are standardized before model application by subtracting the mean of each time series and dividing by its standard deviation [Abramowitz and Stegun,1965]. The operation is inverted before data analysis and visualization.

Table 4-2 Inflow Variables.

Variables	Units	LSTM	PILSTM
ReservoirInflow_FLOW_Observed	CFS	Predictand	Predictand
ReservoirInflow_FLOW_HMS	CFS	×	✓
MF_TuleR_S20_ET-POTENTIAL	Inches/day	×	✓
MF_TuleR_S20_FLOW	CFS	×	✓
MF_TuleR_S20_PRECIP-INC	Inches/day	✓	✓
MF_TuleR_S20_SATURATION_FRACTION	Inches/Inches	×	✓
MF_TuleR_S20_STORAGE-GW-1	Inches	×	✓
MF_TuleR_S20_STORAGE-GW-2	Inches	×	✓
MF_TuleR_S20_STORAGE-SOIL	Inches	×	✓
MF_TuleR_S20_TEMPERATURE-AIR	Degrees F	✓	✓
MF_TuleR_S20_SWE-OBSERVED	Inches	✓	✓
MF_TuleR_S20_PRECIP-LWASS	Inches/day	×	✓

Note: equivalent variables for the other sub-basins are also used but are not listed here for brevity. The complete table is listed in the S.I (Table SI-2).

Training run times for a given set of hyperparameters are approximately 2 minutes per ensemble member over a 16-year time period at a daily time step while prediction run times over an 8-year period are less than 10 seconds per ensemble member over an 8-year time period. As with all DL models, the computationally expensive task is the one-time determination of hyperparameters for each model. In the case of a grid search, this implies training and testing over many combinations of hyperparameters (in our case in the order of 50 combinations, depending on the model). Because all these combinations are independent of each other, this can be done in parallel on a high-performance computer with linear scaling of run times with the number of available GPUs.

4.3 PERFORMANCE METRICS

We evaluate the performance of the models by computing and comparing a set of statistical metrics based on SWE levels or daily flows. These are evaluated over the different base and extreme time periods as well as for day-of-year averages and individual wet/dry years.

The following performance metrics are used to evaluate the models developed in this study [Nash & Sutcliffe, 1970; Yapo et al., 1996; Gupta et al., 2009]:

$$NSE = 1 - \left[\frac{\Sigma_{i=1}^{n}(Y_i^{obs} - Y_i^{sim})^2}{\Sigma_{i=1}^{n}(Y_i^{obs} - \underline{Y}^{obs})^2}\right]$$

$$PBIAS = \left[\frac{\Sigma_{i=1}^{n}(Y_i^{obs} - Y_i^{sim}) \times 100}{\Sigma_{i=1}^{n}(Y_i^{obs})}\right]$$

$$\alpha\ NSE = \frac{\sigma_{sim}}{\sigma_{obs}},$$

$$\beta\ NSE = \frac{\mu_{sim} - \mu_{obs}}{\mu_{obs}},$$

$$r = \frac{\sum_{i=1}^{N} (Q_{obs_i} - \mu_{obs})(Q_{sim_i} - \mu_{sim})}{\sigma_{sim}\sigma_{obs}}.$$

For inflows, we also adopt the following metrics based on the Flow Duration Curve (FDC) [Yilmaz et al., 2008]:

$$FHV_{bias} = \frac{\sum_{i=1}^{H} (Q_{sim_i} - Q_{obs_i})}{\sum_{i=1}^{H} Q_{obs_i}} * 100,$$

where h=1,..H are the flow indices for flows with exceedance probabilities lower than 0.02,

$$FMS_{bias} = \frac{[log(Q_{sim_{m1}}) - log(Q_{sim_{m2}})] - [log(Q_{obs_{m1}}) - log(Q_{obs_{m2}})]}{[log(Q_{obs_{m1}}) - log(Q_{obs_{m2}})]} * 100,$$

where m1 and m2 are the lowest and highest flow exceedance probabilities (0.2 and 0.7 respectively). A summary of all the metrics is provided in Table 5.3

$$FLV_{bias} = -1 * \frac{\sum_{l=1}^{L} [log(Q_{sim_i}) - log(Q_{sim_L})] - \sum_{l=1}^{L} [log(Q_{obs_i}) - log(Q_{obs_L})]}{\sum_{l=1}^{L} [log(Q_{obs_i}) - log(Q_{obs_L})]} * 100,$$

where l=1,..L are the flow indices for flows with exceedance probabilities greater than 0.7. A summary of the metrics used is reported in Table 4-3.

Table 4-3 Performance Metrics.

Metric	Description	Range	Ideal Value
NSE	Nash-Sutcliffe Efficiency: overall error in observed and simulated values	-∞ to 1	1
PBIAS	Percent bias: the average tendency of the simulated data to be larger or smaller than the observed data	-∞ to ∞	0
alpha-NSE	Daily variability: ratio between the standard deviation of the simulated and observed values	0 to ∞	1
beta-NSE	Daily bias: difference between the mean of the simulated and observed values, normalized by the standard deviation of the observed values	∞ to ∞	0
r	Pearson correlation coefficient: an indicator of the timing of values sequence	-1 to 1	1
FHV	FDC high-flow bias: defined for flow with exceedance probabilities less than 2%	-∞ to ∞	0
FMS	FDC mid-segment slope bias: defined for flow with exceedance probabilities between 20% and 70%	-∞ to ∞	0
FLV	FDC low-flow bias: defined for flow with exceedance probabilities greater than 70%	-∞ to ∞	0

SECTION 5
Results

5.1 SWE

For each sub-basin, we trained the HEC-HMS, LSTM and LSTM/HEC-HMS (referred to as PILSTM) models on water years 1982-1997 and applied them to water years 1998-2005. The corresponding predicted daily SWE time series are shown in Figure 5-1. All models showed great predictive skills with NSE values generally above 0.9, (Table 5-1), with minimal advantages for the models with a data-driven component. Similarly, the day-of-year average across the testing period shows no significant difference between the three models (Figure 5-2 and Table 5-1), with NSE values consistently above 0.9. As shown in Figures 5-3 through 5-5, a slight decrease in predictive performance can be seen during individual driest or wettest years in the lower elevation basins where NSE values can occasionally drop to the 0.3-0.4 range. In these cases, however, performance improves to the 0.7-0.8 range when the HEC-HMS outputs are combined with the LSTM inputs (PISLTM). (The HEC-HMS SWE results shown in this section are from simulations using the gridded Radiation Temperature Index snow model.)

In the extreme case experiments (i.e., very wet and very dry years) the results are more nuanced (Figures 5-6 through 5-8 show results for the MF Tule S20 subbasin). While overall performance is good for all models (NSE values ranging from 0.69 to 0.94 (Table 5-2)), there is a difference in performance between the higher and lower elevation basins. The PILSTM performs best in the MF Tule S20, which receives up to 40 inches of snow. In the NF and SF Tule S10 basins, where snow is always less than 8 inches, the HEC-HMS model performs best (Table 5-2). This is the case also when averaging over all the wettest or dry years, with the exception of the NF Tule S10 dry years, where the PILSTM is superior. This likely reflects the ability of the HEC-HMS's spatially distributed/process-based model to capture the variability in precipitation phase and the spatial variations in albedo (i.e., aspects with and without snow cover) that impacts snowpack distribution in a low elevation setting. We note that all models achieve their worst performance in the extreme case simulations during the driest year of the lowest elevation basins (Table SI-2 and Figure SI-8), with the LSTM and the PILSTM showing particularly poor performance, mostly due to excessive bias.

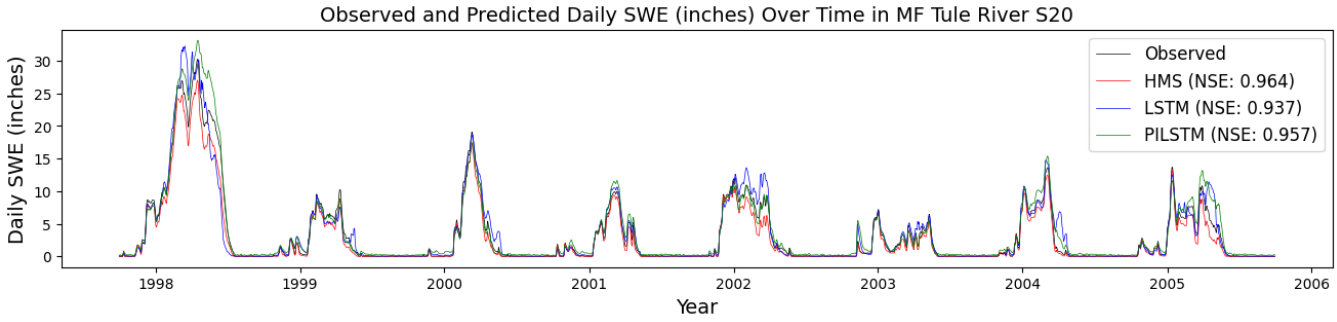


Figure 5-1 Base case test period time series of daily SWE (inches) for observed, HEC-HMS, LSTM, and PILSTM values in the MF Tule River S20 basin. The time series for all three sub-basins are shown in Figure SI-1.

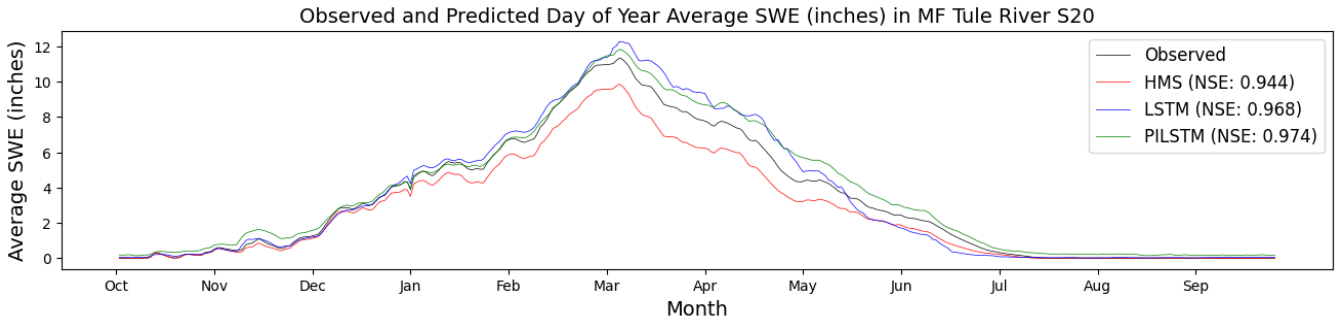


Figure 5-2 Base case test period day-of-year average time series of daily SWE (inches) for observed, HEC-HMS, LSTM, and PILSTM values in the MF Tule River S20 basin. The time series for all three sub-basins are shown in Figure SI-2.

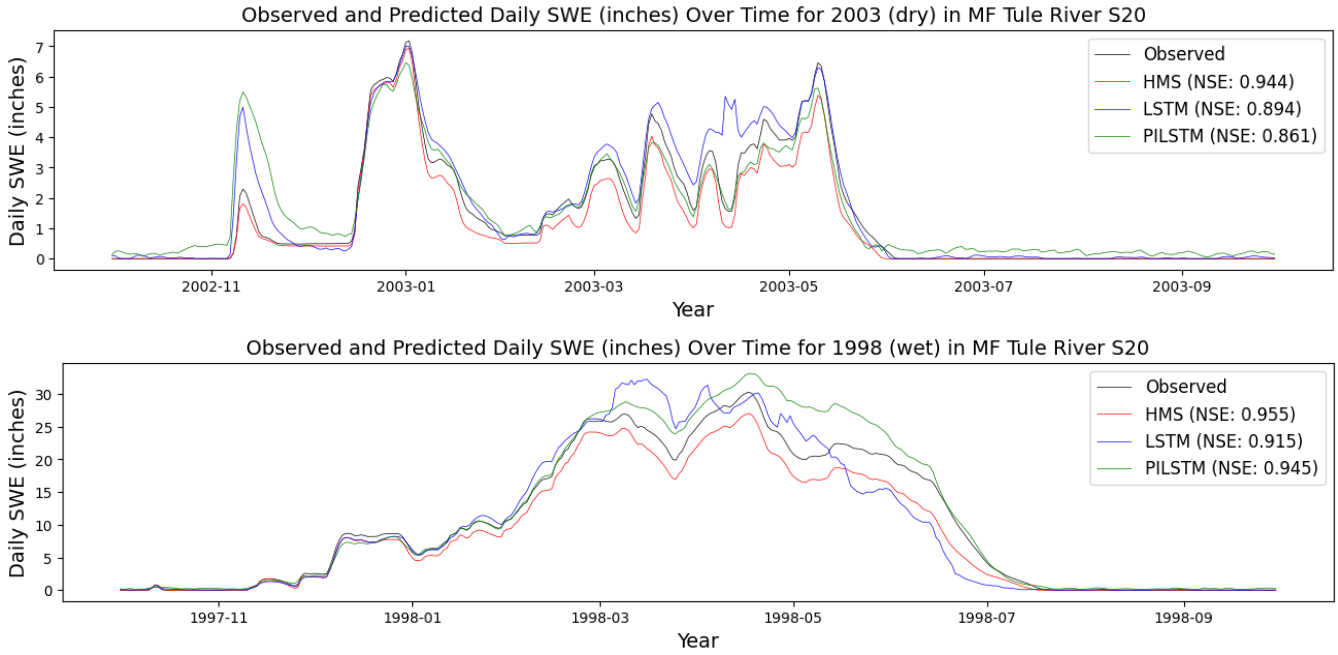


Figure 5-3 Time series of daily SWE (inches) for the wettest and driest years of record in the base case test period of observed, HEC-HMS, LSTM, PILSTM values in the MF Tule S20 sub-basin. The time series for the two wettest and driest years of record in the base case test period of observed, HEC-HMS, LSTM, PILSTM values in the MF Tule S20 sub-basin are shown in Figure SI-3.

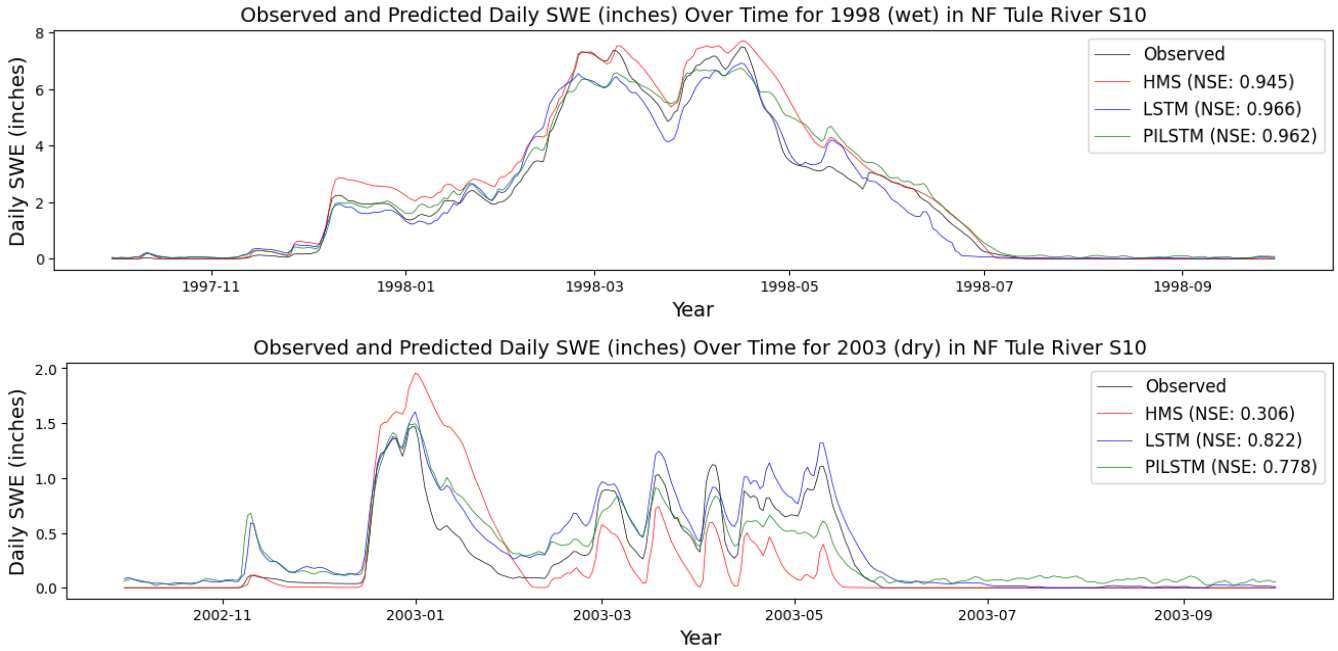


Figure 5-4 Time series of daily SWE (inches) for the wettest and driest years of record in the base case test period of observed, HEC-HMS, LSTM, PILSTM values in the NF Tule S10 sub-basin. The time series for the two wettest and driest years of record in the base case test period of observed, HEC-HMS, LSTM, PILSTM values in the NF Tule S10 sub-basin are shown in Figure SI-4.

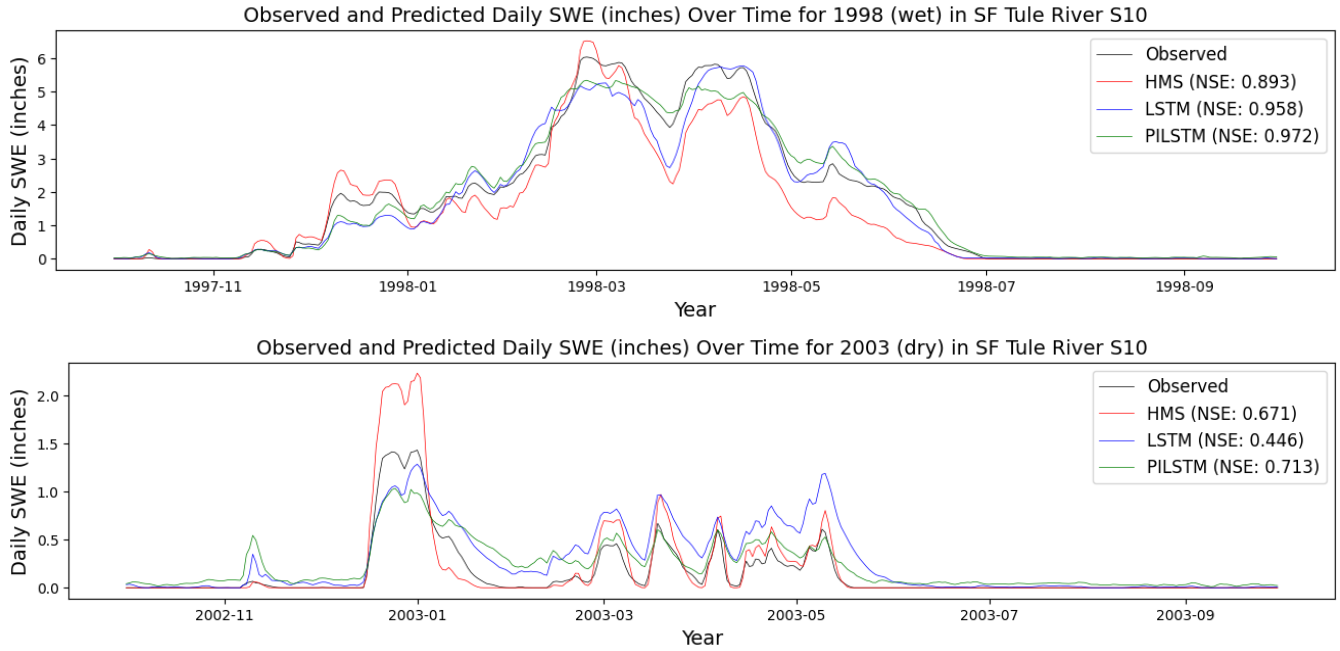


Figure 5-5 Time series of daily SWE (inches) for the wettest and driest years of record in the base case test period of observed, HEC-HMS, LSTM, PILSTM values in the SF Tule S10 sub-basin. The time series for the two wettest and driest years of record in the base case test period of observed, HEC-HMS, LSTM, PILSTM values in the SF Tule S10 sub-basin are shown in Figure SI-5.

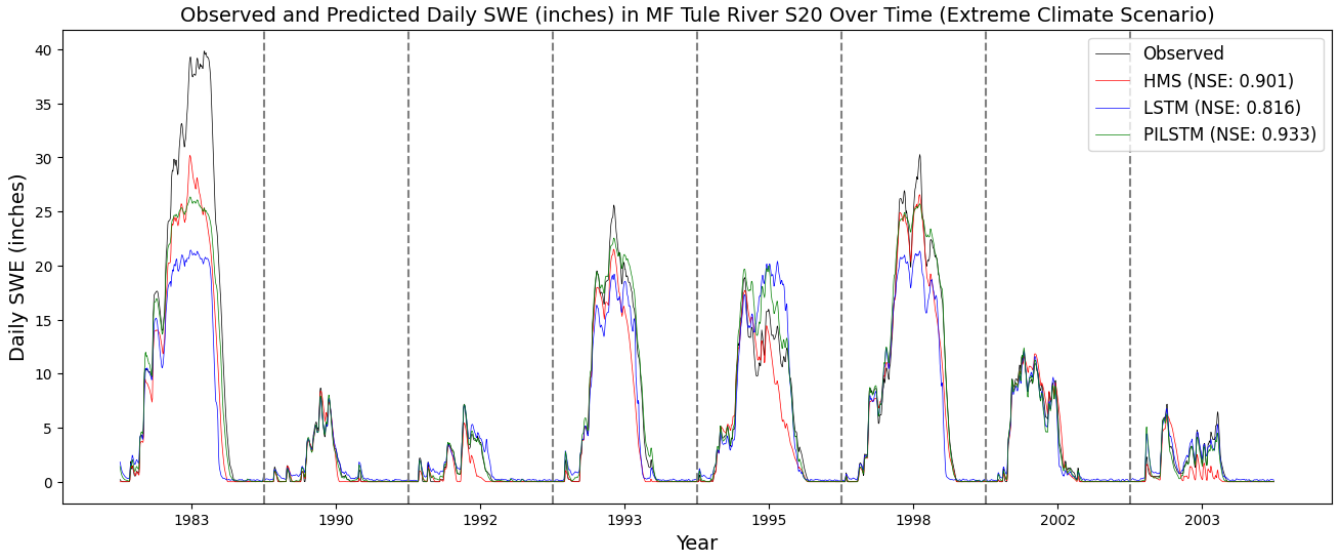


Figure 5-6 Extreme case test period time series of daily SWE (inches) for observed, HEC-HMS, LSTM, PILSTM (LSTM_HMS) values in the MF Tule River S20 basin. The equivalent time series for all three sub-basins are shown in Figure SI-6.

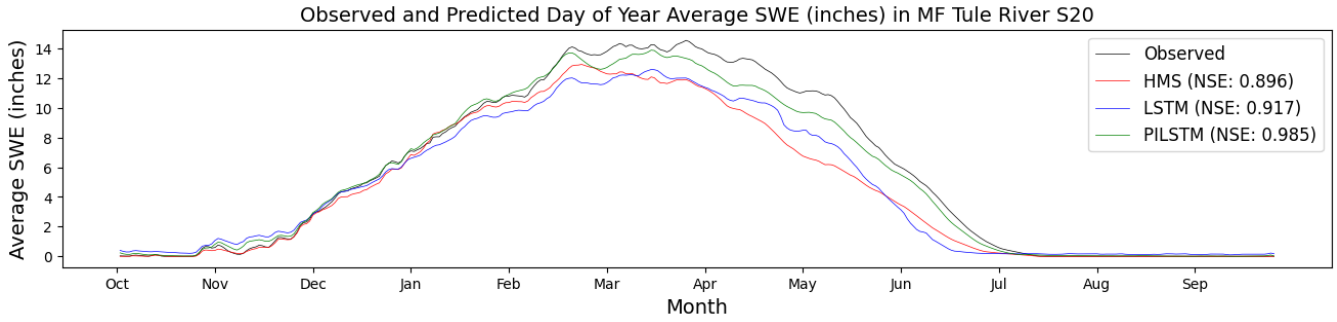


Figure 5-7 Extreme case test period day-of-year average time series of daily SWE (inches) for observed, HEC-HMS, LSTM, PILSTM (LSTM_HMS) values in the MF Tule River S20 basin. The equivalent time series for all three sub-basins are shown in Figure SI-7.

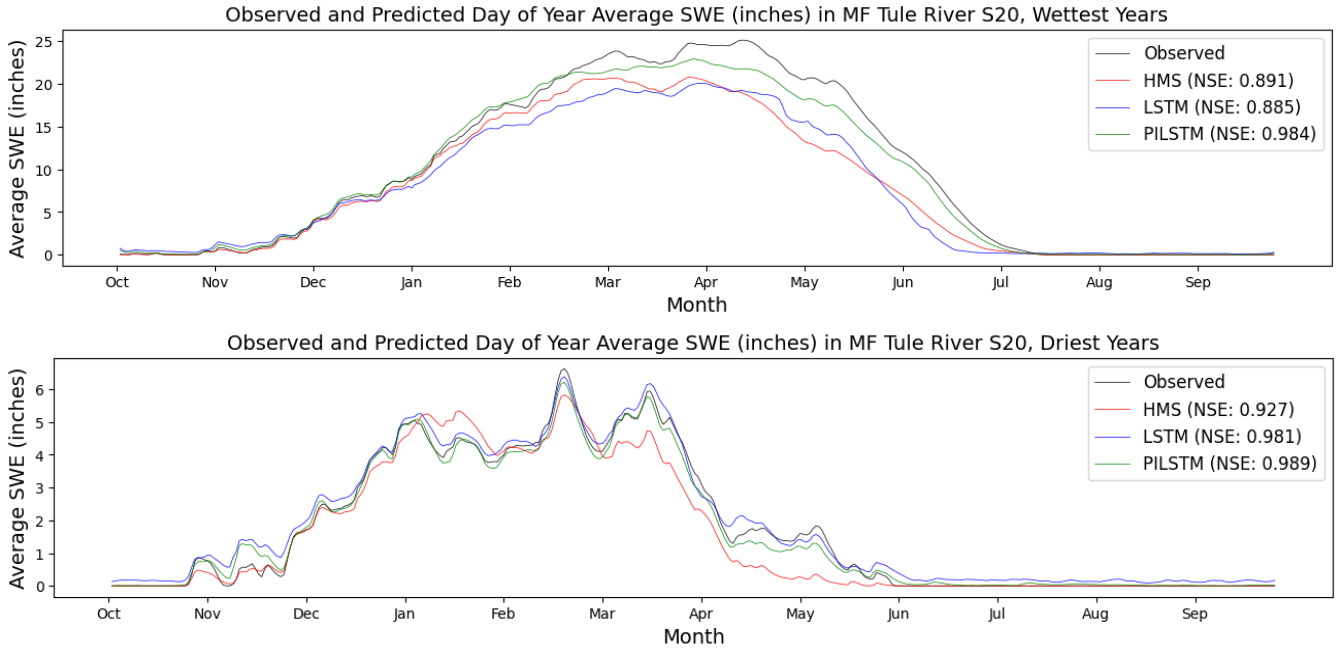


Figure 5-8 Day-of-year average time series of daily SWE (inches) for the wettest and driest years in the extreme case test period of observed, HEC-HMS, LSTM, PILSTM (LSTM_HMS) values in the MF Tule S20 sub-basin. The equivalent time series for all three sub-basins are shown in Figure SI-8.

Table 5-1 SWE performance metrics evaluated over the base case test period. Red values indicate NSE best performance, while maroon values indicate PBIAS best performance. The full set of metrics is shown in Table SI-3.

Sub-Basin	Time Period	HMS NSE	HMS PBIAS	LSTM NSE	LSTM PBIAS	PILSTM NSE	PILSTM PBIAS
MF Tule 20	Base	0.96	-17.15	0.94	6.26	0.96	12.35
NF Tule S10	Base	0.91	12.52	0.94	4.52	0.94	7.88
SF Tule S10	Base	0.88	-5.94	0.93	-1.18	0.93	6.83
MF Tule 20	Base, DOY avg.	0.94	-17.15	0.97	6.27	0.97	12.37
NF Tule S10	Base, DOY avg.	0.95	12.51	0.96	4.53	0.97	7.90
SF Tule S10	Base, DOY avg.	0.95	-5.97	0.97	-1.17	0.98	6.85
MF Tule 20	Base, wettest year 1	0.96	-14.88	0.92	-1.72	0.94	11.67
MF TuleS20	Base, wettest year 2	0.91	-21.56	0.91	13.39	0.99	0.66
MF Tule 20	Base, driest year 1	0.98	-11.96	0.98	12.00	0.93	31.26
MF Tule 20	Base, driest year 2	0.94	-19.58	0.89	16.26	0.86	11.27
NF Tule S10	Base, wettest year 1	0.95	16.45	0.97	-4.21	0.96	7.38
NF Tule S10	Base, wettest year 2	0.94	-11.48	0.93	-10.31	0.94	-9.32
NF Tule S10	Base, driest year 1	0.87	10.64	0.87	35.08	0.94	13.63
NF Tule S10	Base, driest year 2	0.31	-13.17	0.82	42.40	0.78	23.52
SF Tule S10	Base, wettest year 1	0.89	-18.68	0.96	-5.86	0.97	0.45
SF Tule S10	Base, wettest year 2	0.59	-7.27	0.86	0.66	0.83	27.02
SF Tule S10	Base, driest year 1	0.78	8.07	0.96	6.40	0.86	43.80
SF Tule S10	Base, driest year 2	0.67	28.88	0.45	86.00	0.71	53.96

Table 5-2 SWE performance metrics evaluated over the extreme test period. Red values indicate NSE best performance, while maroon values indicate PBIAS best performance. The full set of metrics is shown in Table SI-4.

Sub-Basin	Time Period	HMS NSE	HMS PBIAS	LSTM NSE	LSTM PBIAS	PILSTM NSE	PILSTM PBIAS
MF Tule S20	Extreme	0.90	-18.94	0.82	-16.65	0.93	-5.13
NF Tule S10	Extreme	0.94	-1.01	0.64	-33.85	0.82	-17.69
SF Tule S10	Extreme	0.90	-9.25	0.59	-40.60	0.69	-31.48
MF Tule S20	Extreme, DOY avg.	0.90	-18.95	0.92	-16.64	0.99	-5.13
NF Tule S10	Extreme, DOY avg.	0.98	-1.02	0.73	-33.86	0.89	-17.69
SF Tule S10	Extreme, DOY avg.	0.95	-9.27	0.68	-40.59	0.79	-31.48
MF Tule S20	Extreme, wettest years	0.89	-20.05	0.88	-21.48	0.98	-5.54
MF Tule S20	Extreme, driest years	0.93	-12.85	0.98	10.11	0.99	-2.87
NF Tule S10	Extreme, wettest years	0.98	-4.99	0.56	-44.84	0.82	-25.58
NF Tule S10	Extreme, driest years	0.67	28.83	0.66	48.68	0.79	41.59
SF Tule S10	Extreme, wettest years	0.93	-14.79	0.48	-51.17	0.63	-41.39
SF Tule S10	Extreme, driest years	0.76	30.83	0.59	36.39	0.55	40.68

5.2 DAILY AVERAGE INFLOW

Reservoir inflow predictions are a more difficult task than SWE predictions with NSE values in the 0.7-0.9 range for the base case scenario (Table 5-3). Notably, there is a significant increase in performance between the PISLTM and the HEC-HMS models, suggesting that post-processing a processed-based model with an LSTM can be extremely beneficial to predictive skill. In particular, the LSTM and PILSTM models are better able to capture the runoff response earlier in the wet season (Oct - Jan) where HEC-HMS is struggling (Figure 5-11). This is likely a limitation with HEC-HMS and how it models groundwater (or how the watershed becomes saturated after long periods with no precipitation). Additionally, this could be a limitation of using daily precipitation to simulate inflow magnitudes. Figures 3-10 and 3-11 show that HEC-HMS obtains better performance predicting inflow magnitude when hourly precipitation is used. (The HEC-HMS results shown in this section are from simulations using the Soil Moisture Accounting (SMA) loss model and gridded Temperature Index snow model.)

As with SWE predictions, in the extreme case scenario the PILSTM offers a significant improvement compared to the other models, highlighting the potential of our hybrid approach. NSE values increase from 0.6 (HEC-HMS), to 0.7 (LSTM) to 0.8 (PILSTM), as shown in Table 5-4. These trends are particularly noticeable in the wettest and driest years of record (Fig. 5-14) where the differences are even more significant. This suggests that while a purely data-driven model can implicitly better capture relationships between the forcing variables, the introduction of physical constraints consistently improves out-of-sample performance.

Figure 5-9 shows simulated and observed daily flow for the base case test period. The NSE metrics in the legend highlight the LSTM and PILSTM models perform better than the HEC-HMS model.

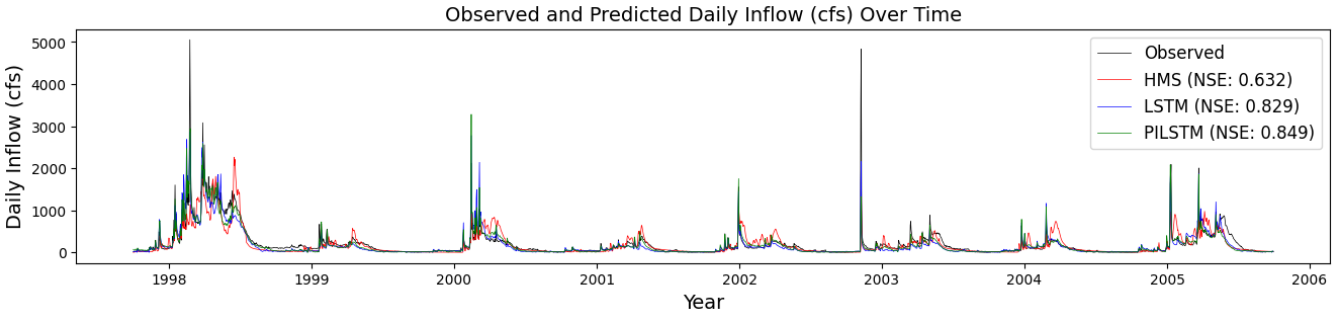


Figure 5-9 Base case test period time series of daily inflow (CFS) for observed, HEC-HMS, LSTM, PILSTM values at Shafer Dam Reservoir.

Figure 5-10 shows the average flow for each day of the year for the base case test period. All models can predict the seasonal runoff response pattern in the watershed that includes snow accumulation in winter months, snowmelt in spring and early summer, and baseflow recession in summer and fall before precipitation begins again. The NSE performance metrics identify improved model performance from the PILSTM model.

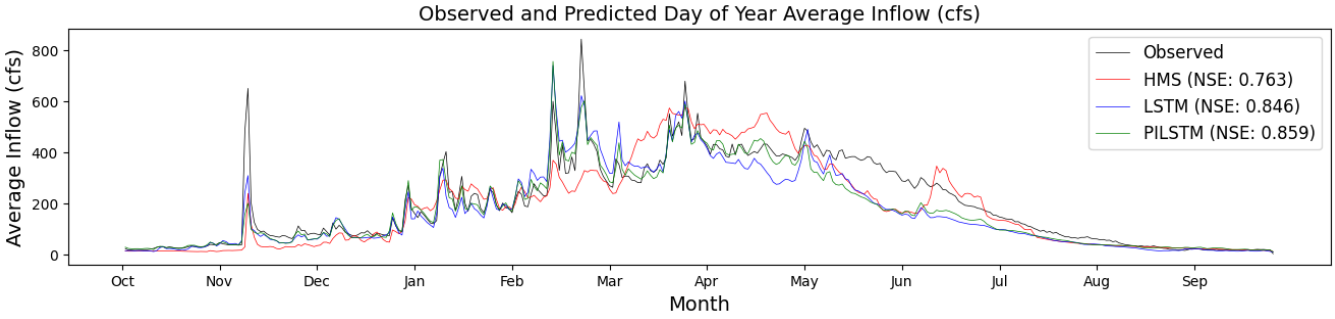


Figure 5-10 Base case test period day-of-year average of daily inflow (CFS) for observed, HEC-HMS, LSTM, PILSTM values at Shafer Dam Reservoir.

Figure 5-11 shows simulated daily flow for two water years (wettest and driest) in the base case test period. Results show LSTM and PILSTM models have improved model performance over the HEC-HMS model when simulating the watershed becoming saturated after a prolonged period with no precipitation, simulating the largest flow events, and simulating dryer water years. Also refer to Figure SI-9 for the two wettest and two driest years for the same scenario.

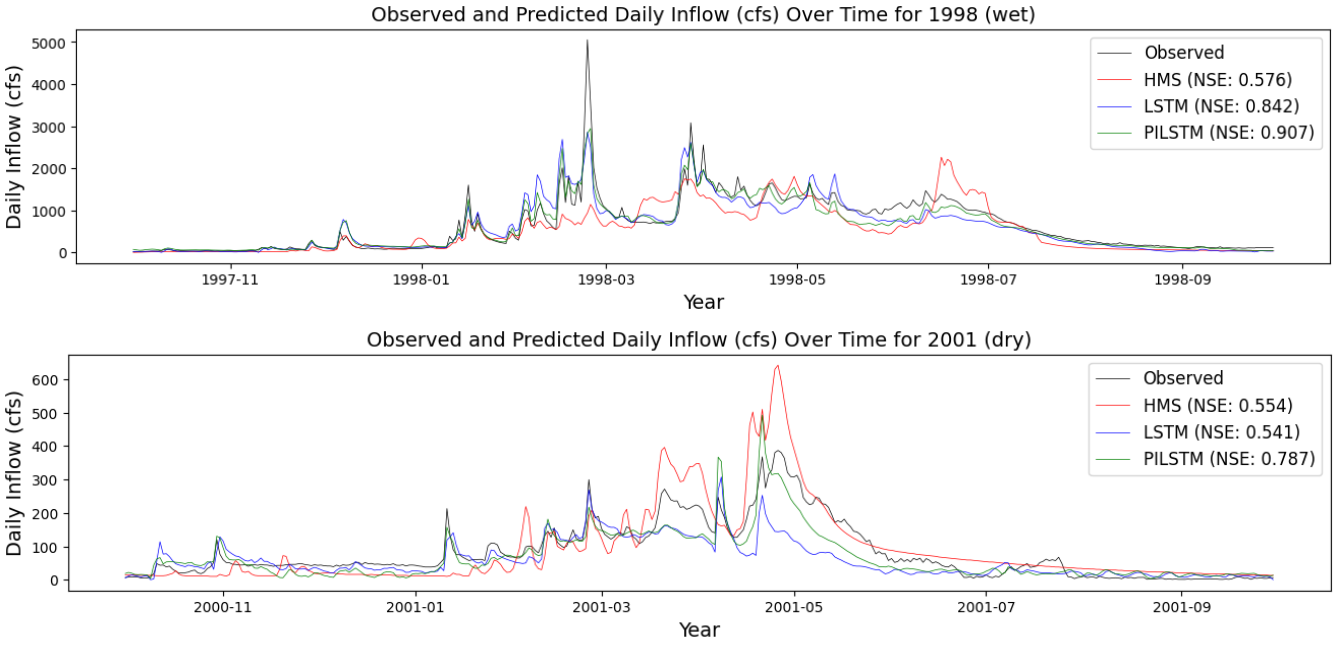


Figure 5-11 Base case test period time series of daily inflow (CFS) in the wettest and driest years for observed, HEC-HMS, LSTM, PILSTM values at Shafer Dam Reservoir.

Figure 5-12 shows simulated and observed daily flow for the extreme case test period. The NSE metrics in the legend highlight the LSTM and PILSTM models perform better than the HEC-HMS model.

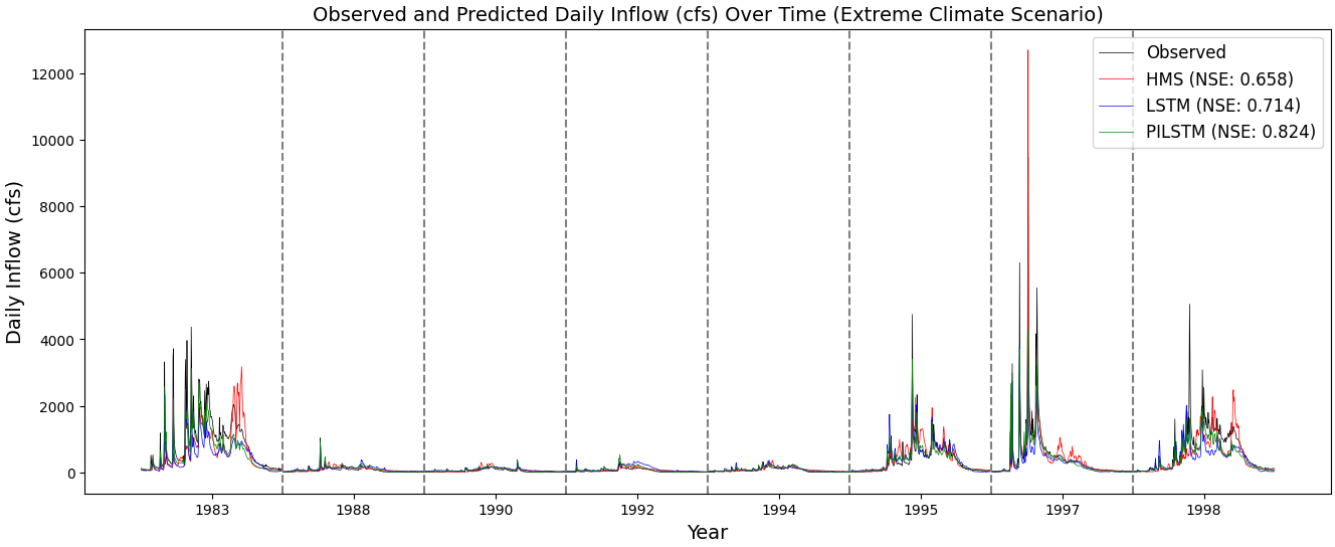


Figure 5-12 Extreme case test period time series of daily inflow (CFS) for observed, HEC-HMS, LSTM, PILSTM (LSTM_HMS) values at Shafer Dam Reservoir.

Figure 5-13 shows the average flow for each day of the year for the extreme case test period. All models can predict the seasonal runoff response pattern in the watershed that includes snow accumulation in winter months, snowmelt in spring and early summer, and baseflow recession in summer and fall before precipitation begins again. The NSE performance metrics identify a similar level of performance for the HEC-HMS and LSTM models and improved performance by the PILSTM model.

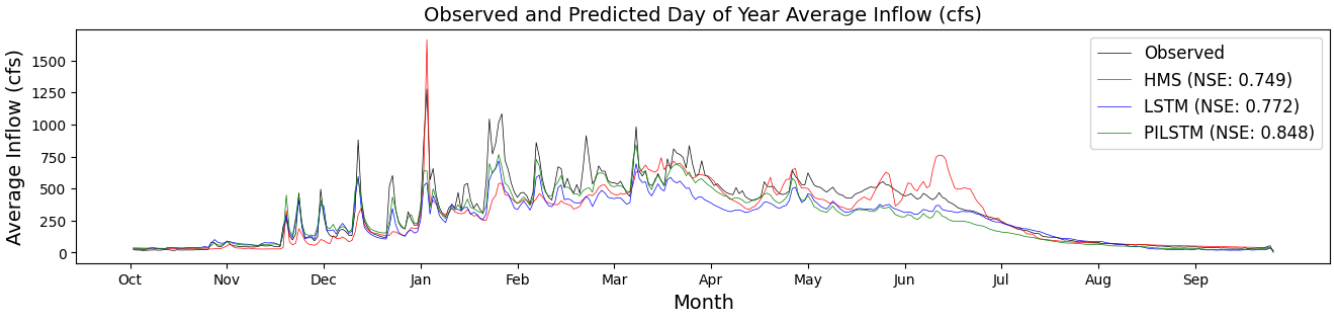


Figure 5-13 Extreme case test period day-of-year average of daily inflow (CFS) for observed, HEC-HMS, LSTM, PILSTM (LSTM_HMS) values at Shafer Dam Reservoir.

Figure 5-14 shows simulated daily flow for the day-of-year averaged four wet and four dry years of the eight water years in the extreme case test period. Results show LSTM and PILSTM models have improved model performance over the HEC-HMS model when simulating the watershed becoming saturated after a prolonged period with no precipitation and simulating the largest flow events. It is challenging when configuring and calibrating process-based models to identify one parameter set that performs well for both large flows and baseflow conditions experienced when the watershed is dryer than average. Calibration processes, tools, and metrics tend to focus on those parameters that have the largest impact on the model's response. Both wet and dry day-of-year results show the HEC-HMS results are better than the LSTM results; however, the PILSTM model has the best performance for these two periods.

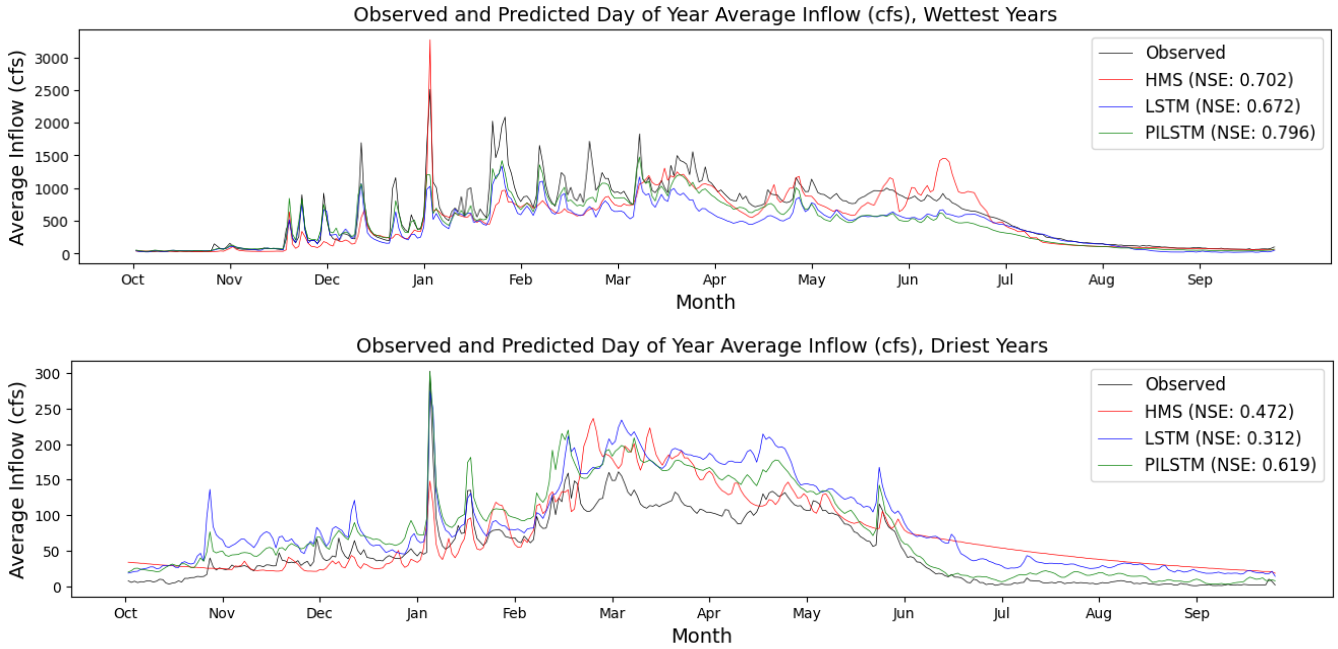


Figure 5-14 Extreme case test period day-of-year averaged time series of daily inflow (CFS) in the wettest and driest years for observed, HEC-HMS, LSTM, PILSTM (LSTM_HMS) values at Shafer Dam Reservoir.

Table 5-3 Inflow performance metrics evaluated over the base case test period. Red values indicate NSE best performance, while maroon values indicate PBIAS best performance. The full set of metrics is shown in Table SI-5.

Time Period	HMS NSE	HMS PBIAS	LSTM NSE	LSTM PBIAS	PILSTM NSE	PILSTM PBIAS
base	0.63	-7.35	0.83	-13.25	0.85	-11.27
base, DOY avg.	0.76	-7.36	0.85	-13.25	0.86	-11.28
base, wettest year 1	0.58	-18.50	0.84	-7.35	0.91	-6.24
base, wettest year 2	0.50	-6.90	0.77	-4.75	0.76	-16.53
base, driest year 1	0.55	11.21	0.54	-26.74	0.79	-17.49
base, driest year 2	-0.28	41.55	0.74	20.18	0.81	16.50

Table 5-4 Inflow performance metrics evaluated over the extreme case test period. Red values indicate NSE best performance, while maroon values indicate PBIAS best performance. The full set of metrics is shown in Table SI-6.

Time Period	HMS NSE	HMS PBIAS	LSTM NSE	LSTM PBIAS	PILSTM NSE	PILSTM PBIAS
Extreme	0.66	-11.57	0.71	-21.91	0.82	-15.77
Extreme, DOY avg.	0.75	-11.55	0.77	-21.90	0.85	-15.78
Extreme, wettest years	0.70	-16.07	0.67	-29.83	0.80	-21.42
Extreme, driest years	0.47	37.55	0.31	64.50	0.62	45.64

SECTION 6
Conclusions and Recommendations for Future Investigations

In this case study, we selected a basin in the Tule river watershed, where a long record of meteorological, snow water equivalent (SWE), and reservoir inflow exists, and compared the performance of a purely process-based model (HEC-HMS), with that of a purely data driven model (LSTM), and that of a hybrid physics informed data driven model (PILSTM) that incorporated output variables from HEC-HMS to provide additional physically based constraints.

We used these three approaches (PB, DD, and hybrid) to predict SWE and streamflow into the Shafer Dam reservoir. We tested the models on the historical time series (base case); then constructed a resampled time series by identifying the wettest and driest years in the historical record and using them to evaluate the models, while the model training is performed on the remaining (average) years only (extreme case).

With respect to SWE, we found that all models exhibit good prediction skill in the base case. In the extreme case the PILSTM outperforms the other models in the higher elevation snow-dominated basin, suggesting increased robustness under climate extremes. However, this is not the case in the lower elevation basins with considerably less snow, where HEC-HMS gives the best results, suggesting that a PB model may be more robust when significant changes in phase are involved, and that a distributed model may be more appropriate when albedo, slope and aspect also play an important role.

Inflow prediction is a slightly more challenging task for all models, but a similar story emerges, though with more notable differences between the PILSTM, LSTM, and HEC-HMS models. The data-driven models, particularly the PILSTM, outperform HEC-HMS across performance metrics where flow magnitude is measured. While in the base case the performance of the LSTM and the PILSTM are comparable, in the extreme climate experiments the PILSTM exhibits a marked improvement compared to the other approaches, noticeably improving on the LSTM as well. Noted improvements by the LSTM and PILSTM models included better prediction of runoff during the fall and early winter months as the watershed transitioned from dry to saturated conditions, better prediction of runoff for higher flow events, and better prediction of runoff for the driest watersheds.

Next steps include testing this approach on other basins, with the goal of developing a single model for all the California reservoirs with similar meteorology and hydrology. This would require extensive data collection, and the development of static attributes (e.g., characteristic climatology, physiography, land use, and land type, etc.) to distinguish individual basins (e.g. Kratzert et al., 2019). The single model approach has been shown to benefit streamflow prediction over the CONUS, due to training over a larger dataset spanning a variety of characteristics [Nearing et al., 2021].

We also plan to explore the usefulness of incorporating hourly data to obtain higher temporal resolution predictions and better capture sub-daily events. In this study we have shown how HEC-HMS can achieve

better performance when run at the hourly scale and aggregated at the daily scale, suggesting that the PILSTM model could benefit as well from considering sub-daily timescales.

A suite of different software platforms was used in this case study. We plan to make the code base simpler and more uniform to simplify operation.

Finally, we wish to test our approach in a true forecast mode, training the model on both observed and forecasted data such as the National Weather Service 7-day forecast which provides a suite of meteorological data (e.g. precipitation, temperature, pressure, humidity, etc.). A challenge could be obtaining archived forecast data to be used for model training, but such a model would provide a true test of operational functionality and provide a basis for comparison with current practices.

We note that the HEC-HMS model was manually calibrated in an iterative process to improve model performance. The HEC-HMS team is in the process of building automated calibration tools that will allow much more comprehensive and efficient calibration focusing on a range of performance metrics. Similarly, the U.C. Berkeley team is developing an automated calibration procedure on a parallel high-performance computer that will allow the testing of a much vaster number of hyper-parameter combinations, resulting in superior model training. The analysis should be performed again to verify the level of performance gained by post processing HEC-HMS results using an LSTM model.

Nevertheless, this case study suggests that post-processing a PB model such as HEC-HMS with an LSTM is an efficient and effective way to obtain improved predictive skills, particularly when considering a climate change scenario where the model is applied to events larger and smaller than those used to calibrate the model.

SECTION 7
Supplemental Information

Table SI-1 LSTM and PILSTM Hyper-Parameters.

Basin	Model Type	Scenario	Sequence Length	Layers	Hidden Units	Dropout Probability	Training Epochs	Learning Rate
MF Tule S20 SWE	LSTM	Base	120 days	1	128	0.35	64	0.001
NF Tule S10 SWE	LSTM	Base	120 days	1	64	0.35	16	0.001
SF Tule S10 SWE	LSTM	Base	120 days	1	64	0.35	32	0.001
MF Tule S20 SWE	LSTM	Extreme	120 days	1	16	0.35	32	0.001
NF Tule S10 SWE	LSTM	Extreme	90 days	1	16	0.5	32	0.001
SF Tule S10 SWE	LSTM	Extreme	120 days	1	32	0.5	32	0.001
Success Dam Inflow	LSTM	Base	60 days	1	64	0.35	32	0.001
Success Dam Inflow	LSTM	Extreme	60 days	1	64	0.35	32	0.001
MF Tule S20 SWE	PILSTM	Base	120 days	1	16	0.35	32	0.001
NF Tule S10 SWE	PILSTM	Base	90 days	1	16	0.5	32	0.001
SF Tule S10 SWE	PILSTM	Base	120 days	1	32	0.5	32	0.001
MF Tule S20 SWE	PILSTM	Extreme	120 days	1	32	0.35	32	0.001
NF Tule S10 SWE	PILSTM	Extreme	60 days	1	32	0.5	16	0.001
SF Tule S10 SWE	PILSTM	Extreme	120 days	1	32	0.5	32	0.001
Success Dam Inflow	PILSTM	Base	90 days	1	64	0.5	32	0.001
Success Dam Inflow	PILSTM	Extreme	90 days	1	64	0.5	32	0.001

Table SI-2 Inflow Variables.

Variables	Units	LSTM	PILSTM
ReservoirInflow_FLOW_Observed	CFS	Predictand	Predictand
ReservoirInflow_FLOW_HMS	CFS	×	✓
MF_TuleR_S20_ET-POTENTIAL	Inches/day	×	✓
MF_TuleR_S20_FLOW	CFS	✓	✓
MF_TuleR_S20_PRECIP-INC	Inches/day	✓	✓
MF_TuleR_S20_SATURATION_FRACTION	Inches/Inches	×	✓
MF_TuleR_S20_STORAGE-GW-1	Inches	×	✓
MF_TuleR_S20_STORAGE-GW-2	Inches	×	✓
MF_TuleR_S20_STORAGE-SOIL	Inches	×	✓
MF_TuleR_S20_TEMPERATURE-AIR	Degrees F	✓	✓
MF_TuleR_S20_SWE-OBSERVED	Inches	✓	✓
MF_TuleR_S20_PRECIP-LWASS	Inches/day	×	✓
NF_TuleR_S10_ET-POTENTIAL	Inches/day	×	✓
NF_TuleR_S10_FLOW	CFS	✓	✓
NF_TuleR_S10_PRECIP-INC	Inches/day	✓	✓
NF_TuleR_S10_SATURATION_FRACTION	Inches/Inches	×	✓
NF_TuleR_S10_STORAGE-GW-1	Inches	×	✓
NF_TuleR_S10_STORAGE-GW-2	Inches	×	✓
NF_TuleR_S10_STORAGE-SOIL	Inches	×	✓
NF_TuleR_S10_TEMPERATURE-AIR	Degrees F	✓	✓
NF_TuleR_S10_SWE-OBSERVED	Inches	✓	✓
NF_TuleR_S10_PRECIP-LWASS	Inches/day	×	✓
SF_TuleR_S10_ET-POTENTIAL	Inches/day	×	✓
SF_TuleR_S10_FLOW	CFS	✓	✓
SF_TuleR_S10_PRECIP-INC	Inches/day	✓	✓
SF_TuleR_S10_SATURATION_FRACTION	Inches/Inches	×	✓
SF_TuleR_S10_STORAGE-GW-1	Inches	×	✓
SF_TuleR_S10_STORAGE-GW-2	Inches	×	✓
SF_TuleR_S10_STORAGE-SOIL	Inches	×	✓
SF_TuleR_S10_TEMPERATURE-AIR	Degrees F	✓	✓
SF_TuleR_S10_SWE-OBSERVED	Inches	✓	✓
SF_TuleR_S10_PRECIP-LWASS	Inches/day	×	✓

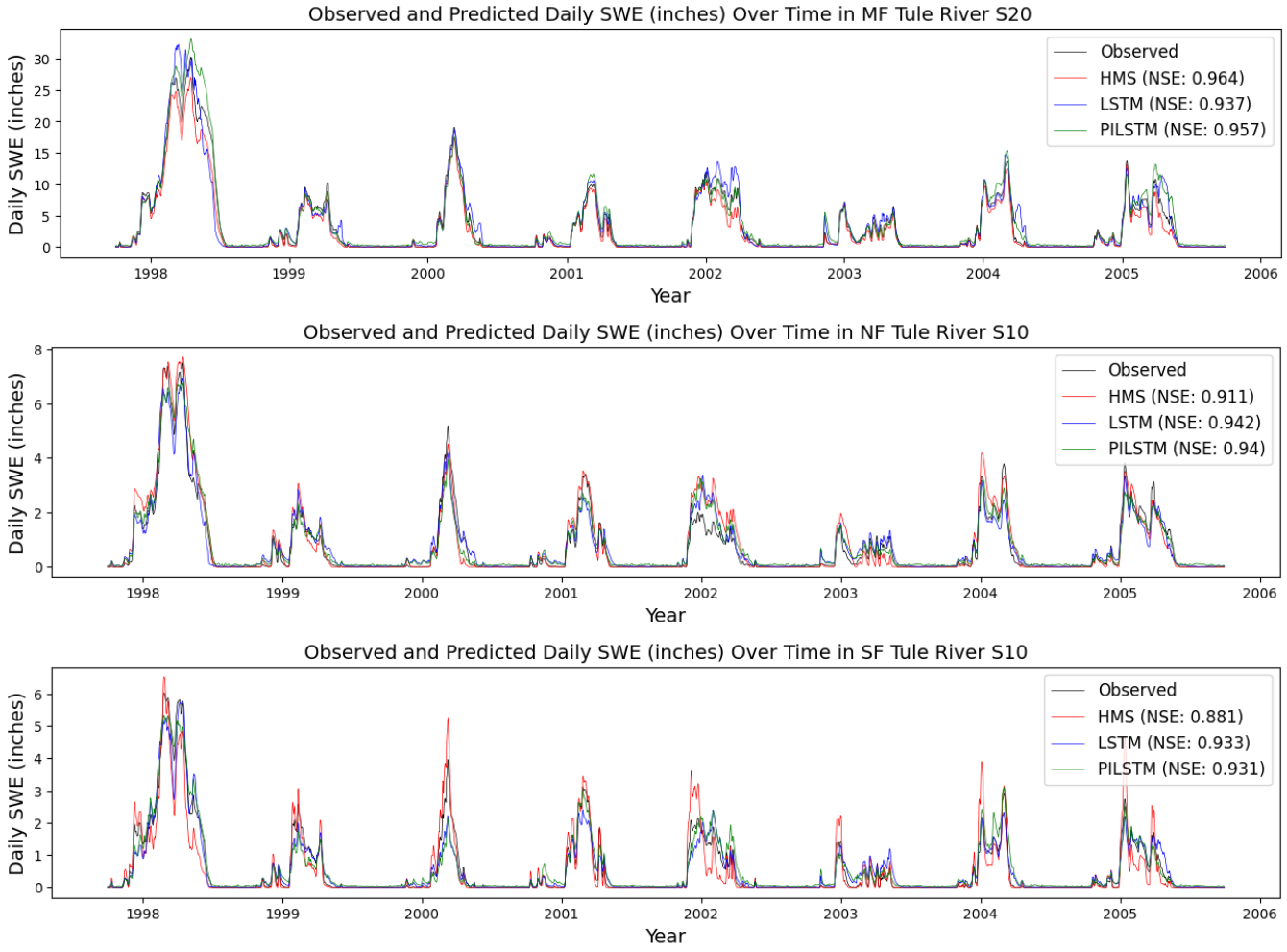


Figure SI-1 Base case test period time series of daily SWE (inches) for observed, HEC-HMS, LSTM, and PILSTM values in the three sub-basins.

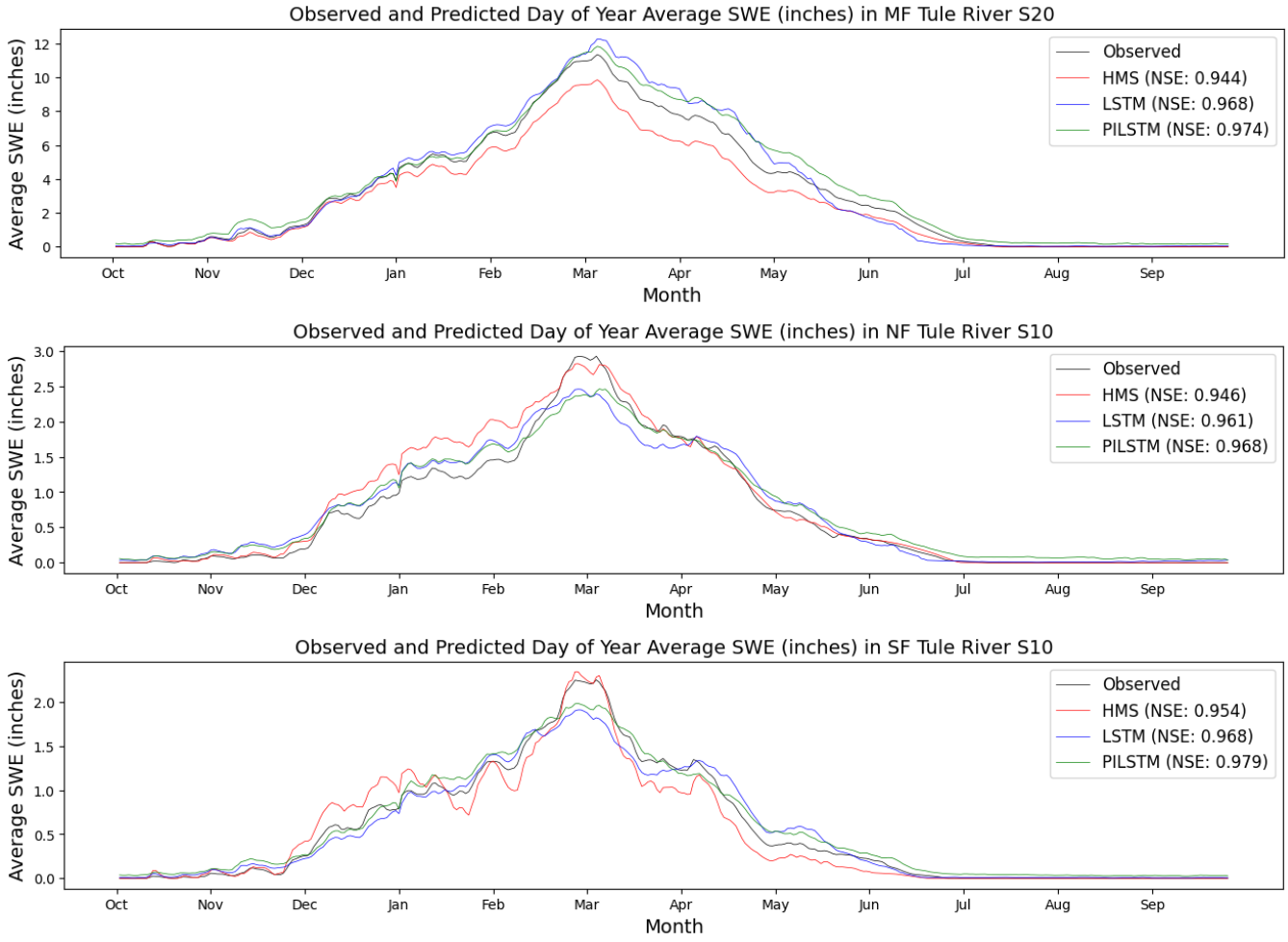


Figure SI-2 Base case test period day-of-year average time series of daily SWE (inches) for observed, HEC-HMS, LSTM, and PILSTM values in the three sub-basins.

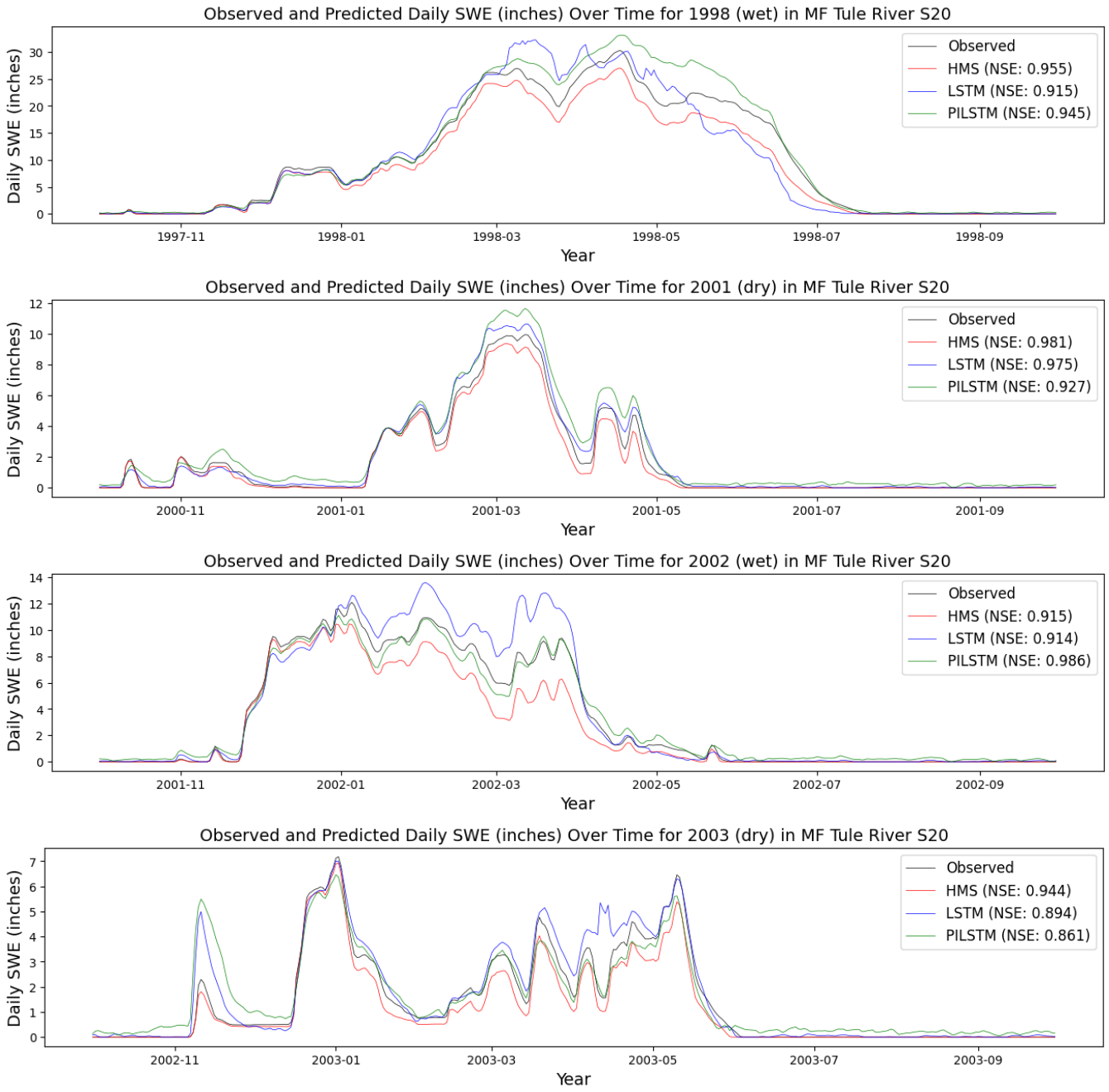


Figure SI-3 Time series of daily SWE (inches) for the two wettest and driest years of record in the base case test period of observed, HEC-HMS, LSTM, PILSTM values in the MF Tule S20 sub-basin.

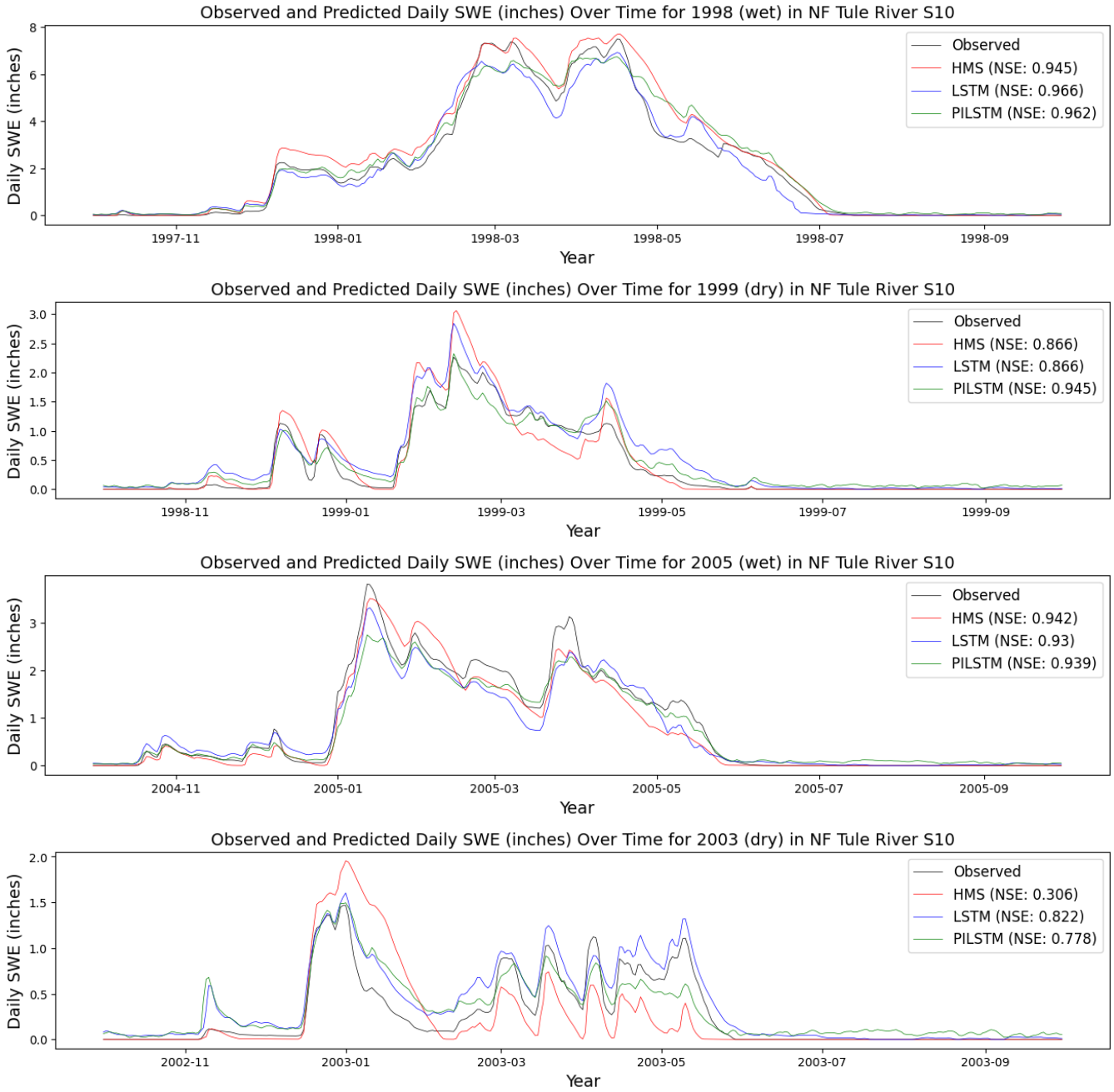


Figure SI-4 Time series of daily SWE (inches) for the two wettest and driest years of record in the base case test period of observed, HEC-HMS, LSTM, PILSTM (LSTM_HMS) values in the NF Tule S10 sub-basin.

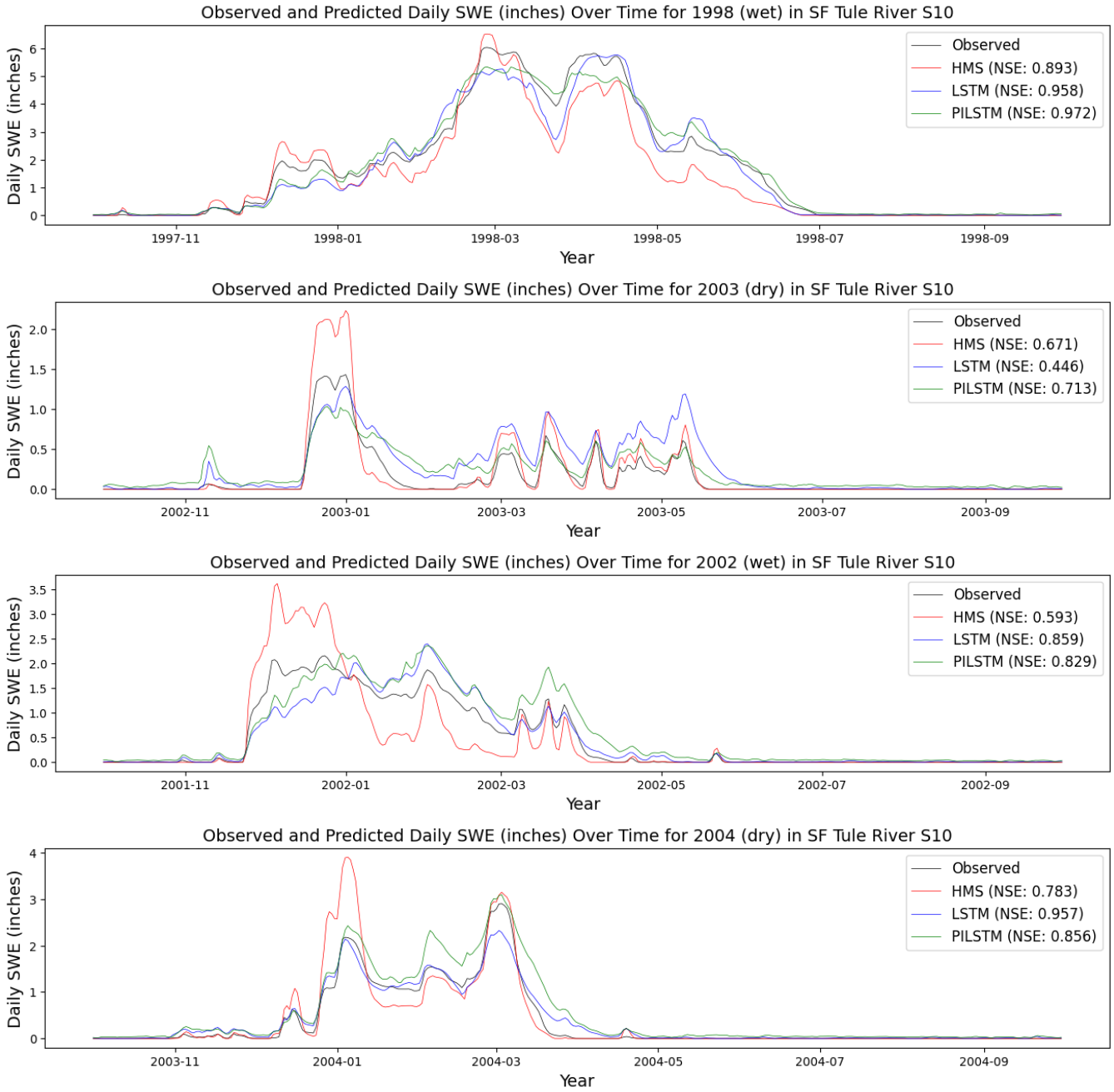
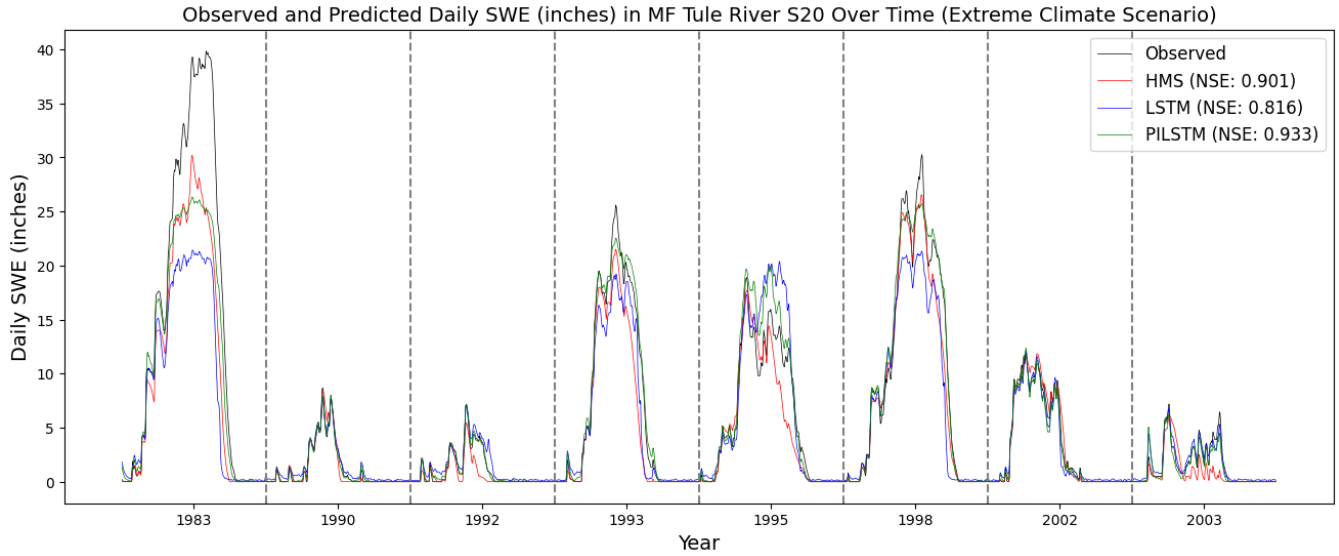
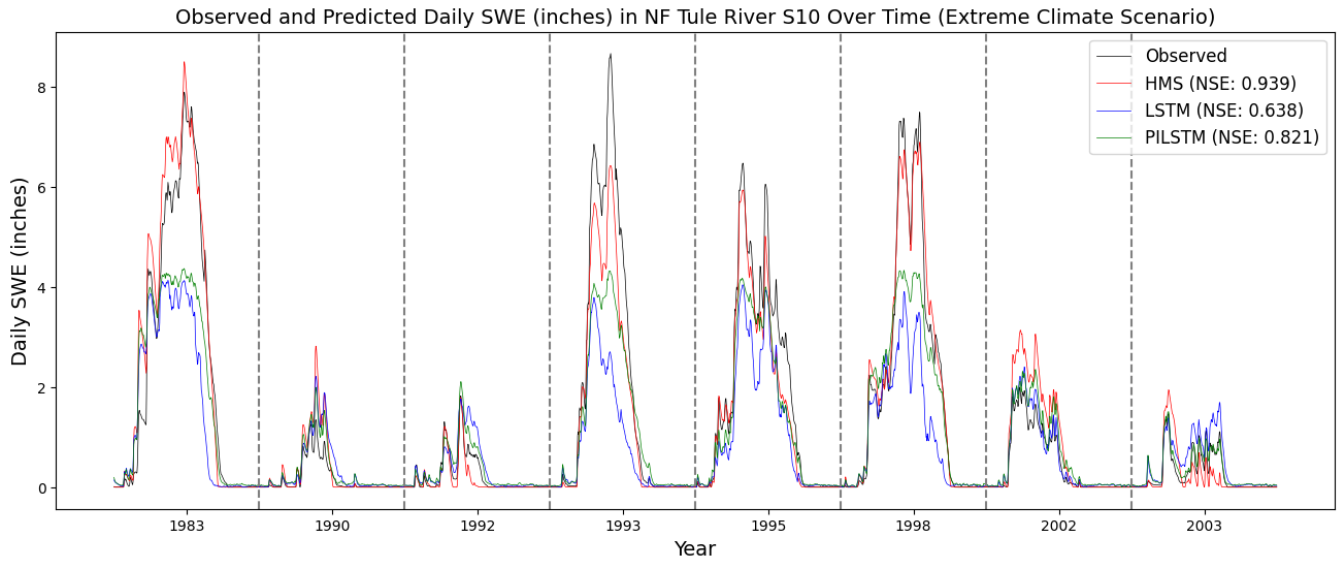


Figure SI-5 Time series of daily SWE (inches) for the two wettest and driest years of record in the base case test period of observed, HEC-HMS, LSTM, PILSTM (LSTM_HMS) values in the SF Tule S10 sub-basin.


Observed and Predicted Daily SWE (inches) in MF Tule River S20 Over Time (Extreme Climate Scenario)
Observed
HMS (NSE: 0.901)
LSTM (NSE: 0.816)
PILSTM (NSE: 0.933)
Daily SWE (inches)
Year
1983
1990
1992
1993
1995
1998
2002
2003


Observed and Predicted Daily SWE (inches) in NF Tule River S10 Over Time (Extreme Climate Scenario)
Observed
HMS (NSE: 0.939)
LSTM (NSE: 0.638)
PILSTM (NSE: 0.821)
Daily SWE (inches)
Year
1983
1990
1992
1993
1995
1998
2002
2003

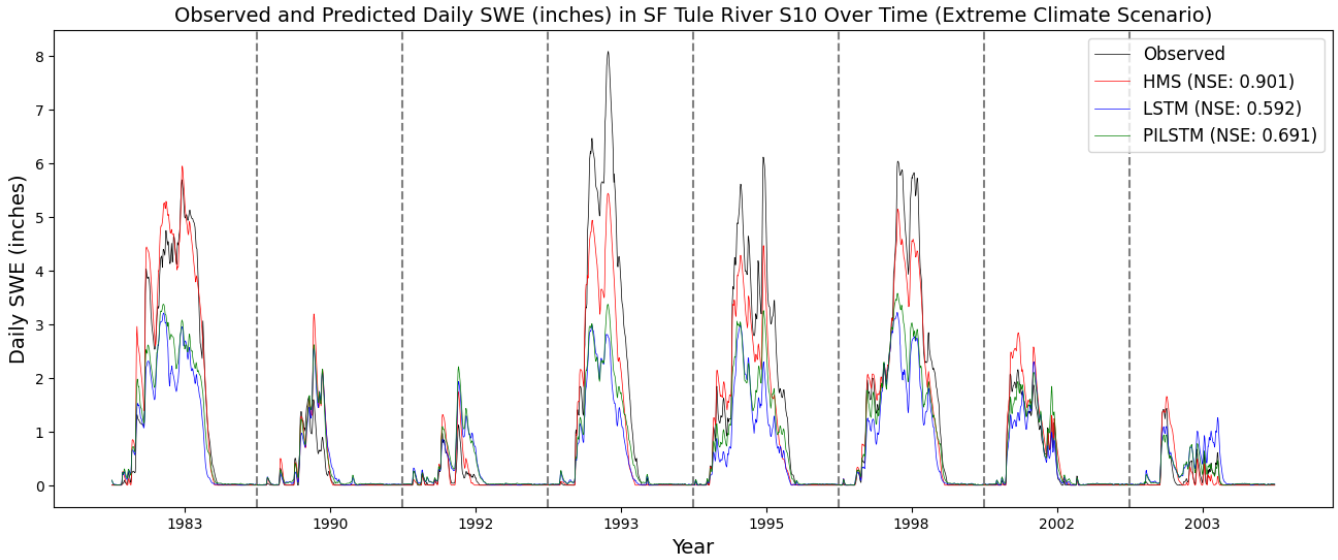


Figure SI-6 Extreme case test period time series of daily SWE (inches) for observed, HEC-HMS, LSTM, PILSTM values in the three sub-basins.

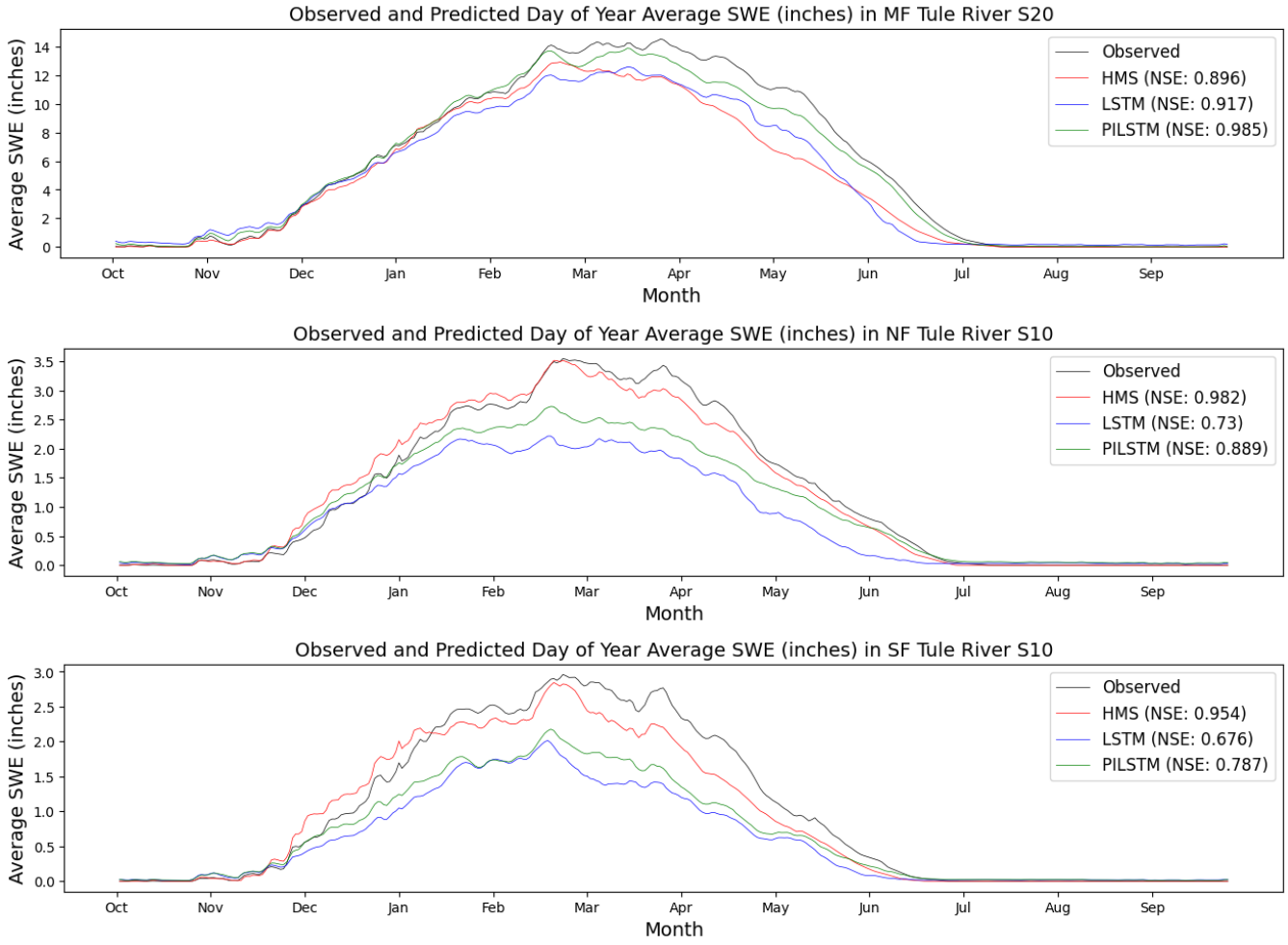
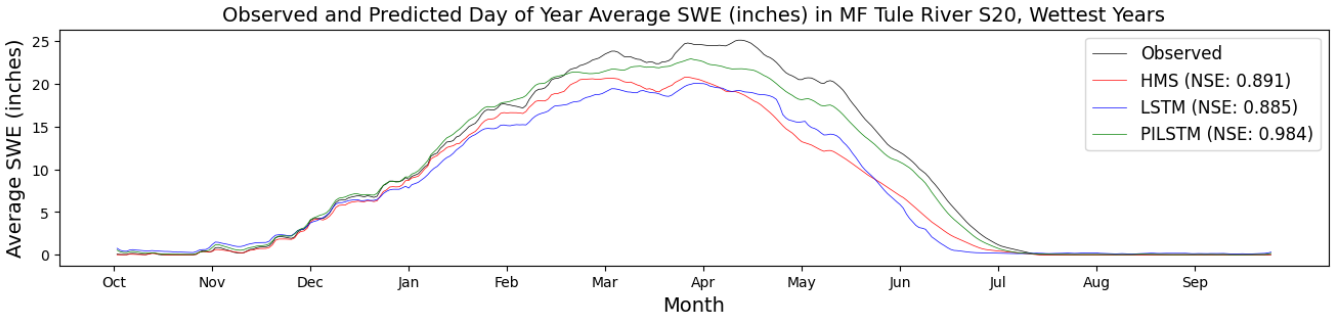
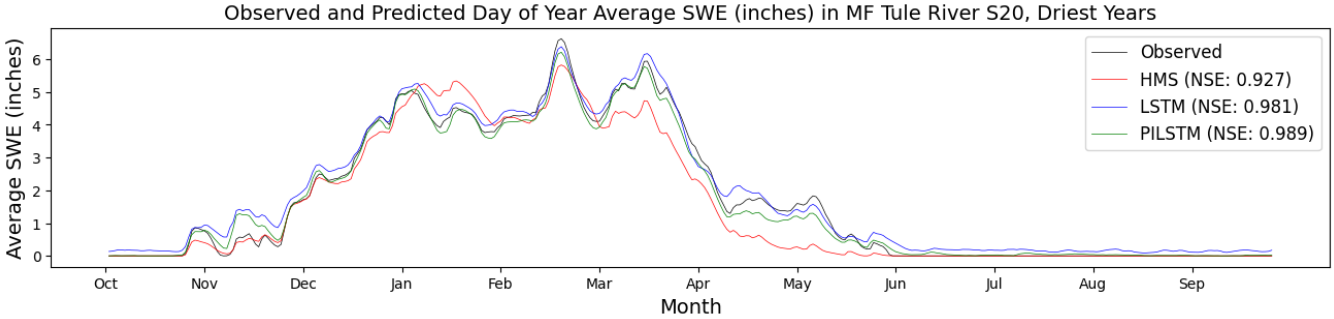
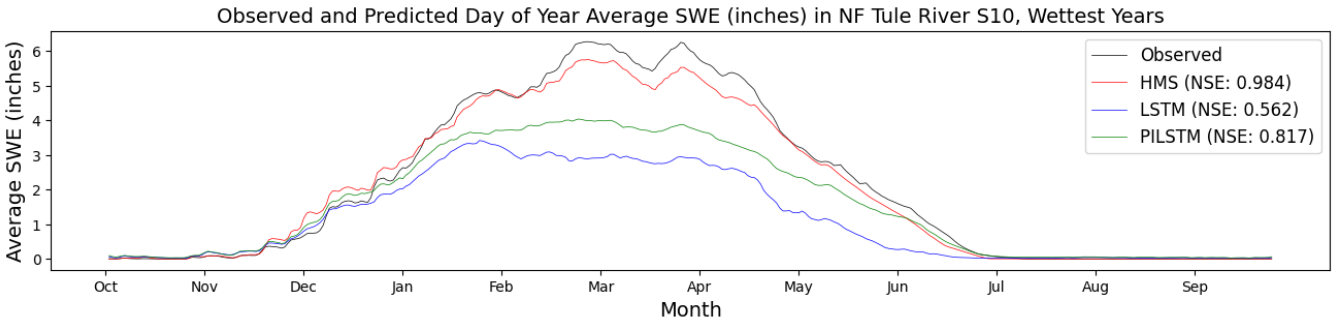
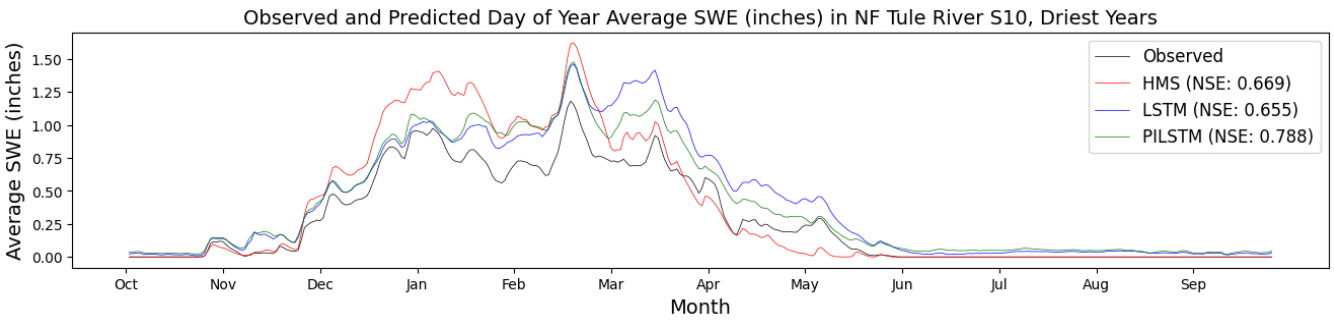


Figure SI-7 Extreme case test period day-of-year average time series of daily SWE (inches) for observed, HEC-HMS, LSTM, PILSTM values in the three sub-basins.


Observed and Predicted Day of Year Average SWE (inches) in MF Tule River S20, Wettest Years
Average SWE (inches)
Month
Observed
HMS (NSE: 0.891)
LSTM (NSE: 0.885)
PILSTM (NSE: 0.984)
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep


Observed and Predicted Day of Year Average SWE (inches) in MF Tule River S20, Driest Years
Average SWE (inches)
Month
Observed
HMS (NSE: 0.927)
LSTM (NSE: 0.981)
PILSTM (NSE: 0.989)
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep


Observed and Predicted Day of Year Average SWE (inches) in NF Tule River S10, Wettest Years
Average SWE (inches)
Month
Observed
HMS (NSE: 0.984)
LSTM (NSE: 0.562)
PILSTM (NSE: 0.817)
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep


Observed and Predicted Day of Year Average SWE (inches) in NF Tule River S10, Driest Years
Average SWE (inches)
Month
Observed
HMS (NSE: 0.669)
LSTM (NSE: 0.655)
PILSTM (NSE: 0.788)
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep

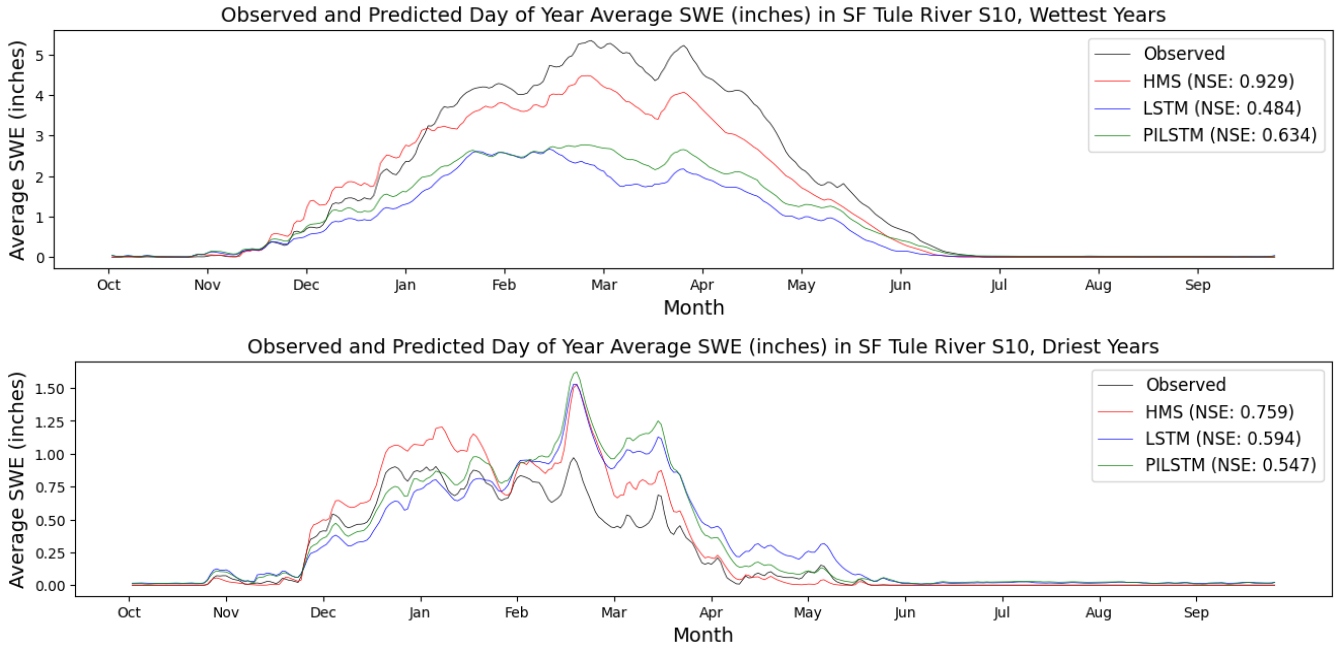


Figure SI-8 Time series of day-of-year averaged daily SWE (inches) for the wettest and driest years in the extreme case test period of observed, HEC-HMS, LSTM, PILSTM values in the three sub-basins.

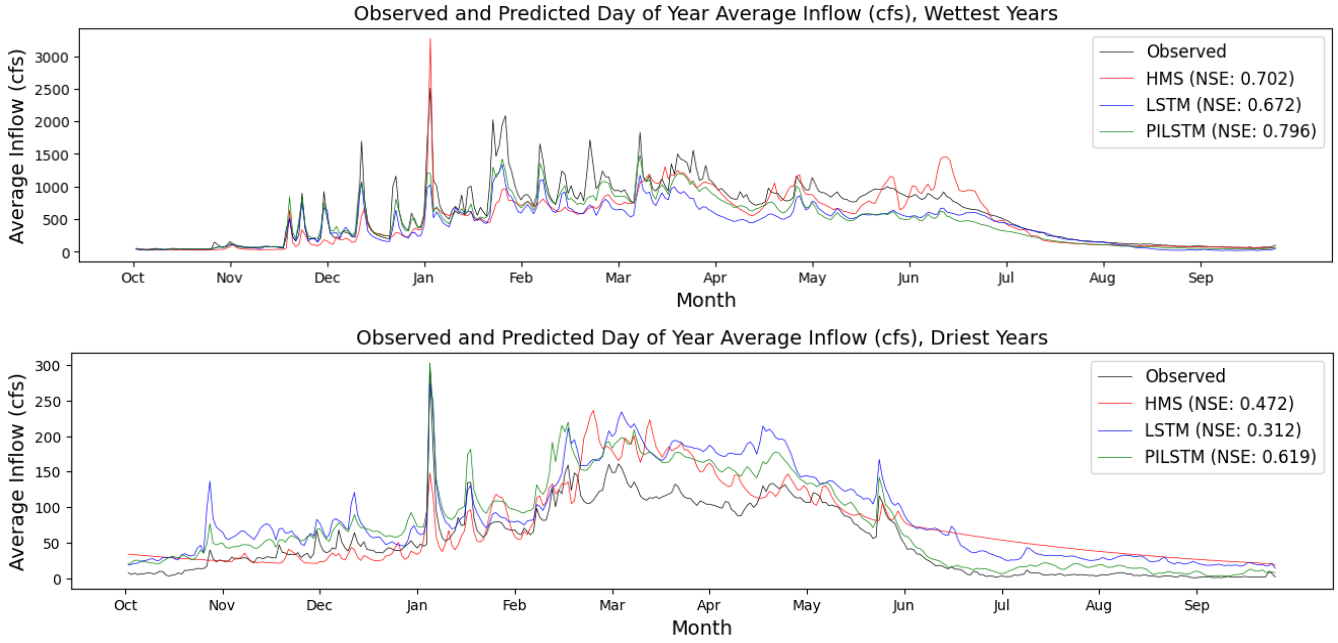


Figure SI-9 Base case test period time series of day-of-year averaged daily inflow (CFS) for the wet and dry years for observed, HEC-HMS, LSTM, PILSTM values at Shafer Dam Reservoir.

Table SI-3 SWE performance metrics evaluated over the base case test period. Red values indicate overall best performance (NSE), while maroon value indicate best performance for specific metrics.

Sub-Basin	Time Period	HMS NSE	HMS PBIAS	HMS alpha-NSE	HMS beta-NSE	HMS r	LSTM NSE	LSTM PBIAS	LSTM alpha-NSE	LSTM beta-NSE	LSTM r	PILSTM NSE	PILSTM PBIAS	PILSTM alpha-NSE	PILSTM beta-NSE	PILSTM r
MF Tule S20	Base	0.96	-17.15	0.87	-0.10	0.99	0.94	6.26	1.05	0.04	0.97	0.96	12.35	1.09	0.07	0.99
NF Tule S10	Base	0.91	12.52	1.11	0.07	0.97	0.94	4.52	0.92	0.03	0.97	0.94	7.88	0.96	0.04	0.97
SF Tule S10	Base	0.88	-5.94	0.99	-0.03	0.94	0.93	-1.18	0.91	-0.01	0.97	0.93	6.83	0.94	0.04	0.97
MF Tule S20	Base, DOY avg.	0.94	-17.15	0.85	-0.16	1.00	0.97	6.27	1.12	0.06	0.99	0.97	12.37	1.04	0.12	1.00
NF Tule S10	Base, DOY avg.	0.95	12.51	1.07	0.11	0.98	0.96	4.53	0.93	0.04	0.98	0.97	7.90	0.91	0.07	0.99
SF Tule S10	Base, DOY avg.	0.95	-5.97	0.97	-0.05	0.98	0.97	-1.17	0.92	-0.01	0.99	0.98	6.85	0.95	0.06	0.99
MF Tule S20	Base, wettest year 1	0.96	-14.88	0.88	-0.15	1.00	0.92	-1.72	1.09	-0.02	0.96	0.94	11.67	1.15	0.12	0.99
MF Tule S20	Base, wettest year 2	0.91	-21.56	0.83	-0.17	0.98	0.91	13.39	1.16	0.11	0.98	0.99	0.66	0.92	0.01	1.00
MF Tule S20	Base, driest year 1	0.98	-11.96	0.92	-0.07	1.00	0.98	12.00	1.09	0.08	0.99	0.93	31.26	1.14	0.20	0.99
MF Tule S20	Base, driest year 2	0.94	-19.58	0.87	-0.15	0.99	0.89	16.26	1.09	0.13	0.96	0.86	11.27	0.91	0.09	0.93
NF Tule S10	Base, wettest year 1	0.95	16.45	1.09	0.14	0.99	0.97	-4.21	0.95	-0.04	0.98	0.96	7.38	0.98	0.06	0.98
NF Tule S10	Base, wettest year 2	0.94	-11.48	0.97	-0.09	0.97	0.93	-10.31	0.85	-0.08	0.98	0.94	-9.32	0.83	-0.08	0.99
NF Tule S10	Base, driest year 1	0.87	10.64	1.17	0.07	0.96	0.87	35.08	1.12	0.23	0.97	0.94	13.63	0.91	0.09	0.98
NF Tule S10	Base, driest year 2	0.31	-13.17	1.20	-0.09	0.73	0.82	42.40	1.11	0.30	0.97	0.78	23.52	0.91	0.17	0.90
SF Tule S10	Base, wettest year 1	0.89	-18.68	0.88	-0.16	0.96	0.96	-5.86	0.96	-0.05	0.98	0.97	0.45	0.96	0.00	0.99
SF Tule S10	Base, wettest year 2	0.59	-7.27	1.26	-0.05	0.87	0.86	0.66	0.97	0.00	0.93	0.83	27.02	1.09	0.19	0.94
SF Tule S10	Base, driest year 1	0.78	8.07	1.21	0.04	0.93	0.96	6.40	0.90	0.03	0.98	0.86	43.80	1.17	0.23	0.97
SF Tule S10	Base, driest year 2	0.67	28.88	1.48	0.14	0.97	0.45	86.00	1.13	0.42	0.84	0.71	53.96	0.80	0.26	0.89

Table SI-4 SWE performance metrics evaluated over the extreme case test period. Red values indicate overall best performance (NSE), while maroon values indicate best performance for specific metrics.

Sub-Basin	Time Period	HMS NSE	HMS PBIAS	HMS alpha-NSE	HMS beta-NSE	HMS r	LSTM NSE	LSTM PBIAS	LSTM alpha-NSE	LSTM beta-NSE	LSTM r	PILSTM NSE	PILSTM PBIAS	PILSTM alpha-NSE	PILSTM beta-NSE	PILSTM r
MF Tule S20	Extreme	0.90	-18.94	0.81	-0.12	0.97	0.82	-16.65	0.74	-0.11	0.93	0.93	-5.13	0.88	-0.03	0.97
NF Tule S10	Extreme	0.94	-1.01	0.96	-0.01	0.97	0.64	-33.85	0.55	-0.20	0.89	0.82	-17.69	0.67	-0.11	0.96
SF Tule S10	Extreme	0.90	-9.25	0.86	-0.05	0.96	0.59	-40.60	0.49	-0.24	0.90	0.69	-31.48	0.56	-0.18	0.93
MF Tule S20	Extreme, DOY avg.	0.90	-18.95	0.85	-0.20	0.97	0.92	-16.64	0.84	-0.17	0.98	0.99	-5.13	0.94	-0.05	1.00
NF Tule S10	Extreme, DOY avg.	0.98	-1.02	0.97	-0.01	0.99	0.73	-33.86	0.65	-0.32	0.96	0.89	-17.69	0.75	-0.17	0.99
SF Tule S10	Extreme, DOY avg.	0.95	-9.27	0.91	-0.08	0.98	0.68	-40.59	0.59	-0.37	0.98	0.79	-31.48	0.66	-0.29	0.99
MF Tule S20	Extreme, wettest years	0.89	-20.05	0.83	-0.20	0.98	0.88	-21.48	0.80	-0.22	0.98	0.98	-5.54	0.93	-0.06	1.00
MF Tule S20	Extreme, driest years	0.93	-12.85	0.97	-0.11	0.97	0.98	10.11	0.98	0.09	0.99	0.99	-2.87	0.96	-0.02	1.00
NF Tule S10	Extreme, wettest years	0.98	-4.99	0.93	-0.05	1.00	0.56	-44.84	0.54	-0.42	0.95	0.82	-25.58	0.67	-0.24	0.99
NF Tule S10	Extreme, driest years	0.67	28.83	1.44	0.24	0.97	0.66	48.68	1.31	0.41	0.97	0.79	41.59	1.24	0.35	0.99
SF Tule S10	Extreme, wettest years	0.93	-14.79	0.83	-0.13	0.99	0.48	-51.17	0.48	-0.46	0.97	0.63	-41.39	0.54	-0.37	0.99
SF Tule S10	Extreme, driest years	0.76	30.83	1.37	0.23	0.98	0.59	36.39	1.22	0.27	0.88	0.55	40.68	1.34	0.30	0.91

Table SI-5 Inflow performance metrics evaluated over the base case test period. Red values indicate overall best performance (NSE), while maroon value indicate best performance.

Time Period	HMS NSE	HMS PBIAS	HMS alpha-NSE	HMS beta-NSE	HMSr	HMS FHV	HMS FMS	HMS FLV	LSTM NSE	LSTM PBIAS	LSTM alpha-NSE	LSTM beta-NSE	LSTM r	LSTM FHV	LSTM FMS	LSTM FLV	PILSTM NSE	PILSTM PBIAS	PILSTM alpha-NSE	PILSTM beta-NSE	PILSTM r	PILSTM FHV	PILSTM FMS	PILSTM FLV
base	0.63	-7.35	0.84	-0.04	0.80	-17.98	28.17	-376.50	0.83	-13.25	0.93	-0.08	0.91	-4.60	12.88	-377.82	0.85	-11.27	0.90	-0.07	0.92	-7.45	19.83	-377.23
base, DOY avg.	0.76	-7.36	1.03	-0.09	0.89	-13.02	32.10	72.50	0.85	-13.25	0.95	-0.16	0.93	-6.56	12.33	62.20	0.86	-11.28	0.93	-0.14	0.94	-9.23	7.35	52.96
base, wettest year 1	0.58	-18.50	0.84	-0.18	0.78	-34.16	11.97	-1324.00	0.84	-7.35	0.92	-0.07	0.92	-17.62	-6.18	-1342.85	0.91	-6.24	0.89	-0.06	0.96	-18.45	-11.29	38.35
base, wettest year 2	0.50	-6.90	0.85	-0.06	0.72	-39.39	10.45	53.78	0.77	-4.75	0.95	-0.04	0.88	-4.51	-0.57	23.53	0.76	-16.53	0.87	-0.14	0.88	-4.20	-2.00	-513.66
base, driest year 1	0.55	11.21	1.42	0.11	0.91	57.01	54.02	82.74	0.54	-26.74	0.64	-0.26	0.80	-32.62	25.25	-796.74	0.79	-17.49	0.87	-0.17	0.90	3.82	40.06	24.83
base, driest year 2	-0.28	41.55	1.66	0.33	0.78	47.80	-7.32	34.06	0.74	20.18	1.31	0.16	0.95	38.86	-6.64	-58.35	0.81	16.50	1.23	0.13	0.95	29.91	-5.03	-1490.24

Table SI-6 Inflow performance metrics evaluated over the extreme case test period. Red values indicate overall best performance (NSE), while maroon value indicate best performance.

Time Period	HMS NSE	HMS PBIAS	HMS alpha-NSE	HMS beta-NSE	HMS r	HMS FHV	HMS FMS	HMS FLV	LSTM NSE	LSTM PBIAS	LSTM alpha-NSE	LSTM beta-NSE	LSTM r	LSTM FHV	LSTM FMS	LSTM FLV	PILSTM NSE	PILSTM PBIAS	PILSTM alpha-NSE	PILSTM beta-NSE	PILSTM r	PILSTM FHV	PILSTM FMS	PILSTM FLV
Extreme	0.66	-11.57	0.84	-0.06	0.81	-22.75	-10.25	91.96	0.71	-21.91	0.60	-0.12	0.91	-39.78	-13.80	-30.47	0.82	-15.77	0.71	-0.08	0.94	-28.31	-13.74	-23.62
Extreme, DOY avg.	0.75	-11.55	0.91	-0.15	0.88	-15.03	6.31	72.29	0.77	-21.90	0.69	-0.28	0.96	-37.75	-21.94	66.55	0.85	-15.78	0.81	-0.20	0.95	-28.75	0.60	46.80
Extreme, wettest years	0.70	-16.07	0.89	-0.20	0.86	-16.40	15.92	32.33	0.67	-29.83	0.64	-0.37	0.95	-40.85	-19.38	-0.86	0.80	-21.42	0.76	-0.27	0.95	-32.89	3.89	39.58
Extreme, driest years	0.47	37.55	1.10	0.41	0.84	19.84	-47.20	74.65	0.31	64.50	1.27	0.70	0.95	28.82	-40.39	56.19	0.62	45.64	1.29	0.50	0.98	24.67	-24.61	1.61

SECTION 8
References

Abramowitz, M. and Stegun, I.A., eds. (1965). Handbook of Mathematical Functions. New York, NY: Dover.

Adera, S., Bellugi, D. Dhakal, A., Larsen, L., 2024, Streamflow prediction at the intersection of physics and machine learning: a case study of two Mediterranean-climate watersheds, *WRR.*

Berg, N. & Hall, A. Anthropogenic warming impacts on California snowpack during drought. Geophys. Res. Lett. 44, 2511–2518 (2017).

California Department of Water Resources (DWR)**.** (2020). California Water Plan Update 2018.

Christian-Smith, J., Levy, M. C., & Gleick, P. H. (2015). Maladaptation to drought: a case report from California, USA. *Sustainability Science*, *10*(3), 491–501.

Frame, J. M., Kratzert, F., Raney II, A., Rahman, M., Salas, F. R., & Nearing, G. S. (2021). Post-Processing the National Water Model with Long Short-Term Memory Networks for Streamflow Predictions and Model Diagnostics. *JAWRA Journal of the American Water Resources Association*, *57*(6), 885–905. https://doi.org/10.1111/1752-1688.12964.

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

Hanak, E. et al. Managing California's water: From conflict to reconciliation. (Public Policy Instit. Of CA, 2011).

HEC, 2000, U.S. Army Corps of Engineers, Hydrologic Engineering Center. 2000. *HEC-HMS Hydrologic Modeling System,* Technical Reference Manual, CPD-74B. Hydrologic Engineering Center, Davis, CA.

HEC, 2022, U.S. Army Corps of Engineers, Hydrologic Engineering Center. 2012. *HEC-HMS Hydrologic Modeling System,* User's Manual, Version 4.0, CPD-74A. Hydrologic Engineering Center, Davis, CA.

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

Hoedt, P. J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G. S., ... & Klambauer, G. (2021, July). Mc-lstm: Mass-conserving lstm. In *International conference on machine learning* (pp. 4275-4286). PMLR.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Benchmarking a catchment-aware long short-term memory network (LSTM) for large-scale hydrological modeling. *Hydrol. Earth Syst. Sci. Discuss*, *2019*, 1-32.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089-5110.

Konapala, G., Kao, S.-C., Painter, S. L., & Lu, D. (2020). Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. *Environmental Research Letters*, *15*(10), 104022. https://doi.org/10.1088/1748-9326/aba927.

LaValle, S. M., Branicky, M. S., & Lindemann, S. R. (2004). On the relationship between classical grid search and probabilistic roadmaps. *The International Journal of Robotics Research*, *23*(7–8), 673–692.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., & Dadson, S. J. (2021). Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. *Hydrology and Earth System Sciences*, *25*(10), 5517-5534.

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.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What role does hydrological science play in the age of machine learning?, Water Resources Research, 57, e2020WR028 091, 2021.

Nearing, G., Sampson, A. K., Kratzert, F., & Frame, J. (2020). Post-Processing a Conceptual Rainfall-Runoff Model with an LSTM. Retrieved from https://eartharxiv.org/repository/view/122/.

Rhoades, A. M., Jones, A. D. & Ullrich, P. A. The Changing Character of the California Sierra Nevada as a Natural Reservoir. Geophys. Res. Lett. 45, 13,008-13,019 (2018).

Tennant, C., Larsen, L., Bellugi, D., Moges, E., Zhang, L., & Ma, H. (2020). The utility of information flow in formulating discharge forecast models: A case study from an arid snow-dominated catchment. *Water Resources Research*, *56*(8), e2019WR024908.

Yapo, P.O.; Gupta, H.V.; Sorooshian S. (1996). Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data. Journal of Hydrology. v181 i1-4. 23–48. doi:10.1016/0022-1694(95)02918-4.

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research*, *44*(9), W09417. https://doi.org/10.1029/2007WR006716.

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2022). Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. *ACM Computing Surveys*. https://doi.org/10.1145/3514228.

Zhang, L., Bellugi, D. G., Li, S., Kamat, A., Kadi, J., Moges, E., ... & Larsen, L. (2022, December). A physics-informed machine learning model for streamflow prediction. In *AGU Fall Meeting Abstracts* (Vol. 2022, pp. H31E-01).