Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

A MACHINE LEARNING FRAMEWORK TO EXPLAIN COMPLEX GEOSPATIAL SIMULATIONS: A CLIMATE CHANGE CASE STUDY

Tanvir Ferdousi
Abhijin Adiga
Mandy Wilson
S. S. Ravi
Anil Vullikanti
Madhav V. Marathe
Samarth Swarup

University of Virginia
1001 N. Emmet St.
Charlottesville, VA 22903, USA

Mingliang Liu
Kirti Rajagopalan
Jennifer Adam

Washington State University
1815 NE Wilson Rd
Pullman, WA 99164, USA

2023 Winter Simulation Conference (WSC) | 979-8-3503-6966-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/WSC60868.2023.10408406

ABSTRACT

The explainability of large and complex simulation models is an open problem. We present a framework to analyze such models by processing multidimensional data through a pipeline of target variable computation, clustering, supervised classification, and feature importance analysis. As a use case, the well-known large-scale hydrology and crop systems simulator VIC-CropSyst is utilized to evaluate how climate change may affect water availability in Washington, United States. We study how snowmelt varies with climate variables (temperature, precipitation) to identify different response characteristics. Based on these characteristics, spatial units are clustered into six distinct classes. A random forest classifier is used with Shapley values to rank static soil and land parameters that help detect each class. The results also include an analysis of risk across different classes to identify areas vulnerable to climate change. This paper demonstrates the usefulness of the proposed framework in providing explainability for large and complex simulations.

1 INTRODUCTION

We utilize model simulations to validate theories, optimize performance, detect potential failures, and predict scenarios where empirical data is unavailable (Maria 1997). The term *explainability* is commonly associated with black-box modeling approaches, such as machine learning (Burkart and Huber 2021). However, mechanistic simulation-based models with complex interactions are commonplace and produce massive quantities of data, posing challenges to out-of-the-box explainability and interpretability of the output data. Such systems can therefore benefit from explainability as a form of model validation.

An important family of mechanistic simulation systems that fit the above criteria consists of global circulation, land surface, and crop growth models. These models capture many processes that occur in the environment, including the energy cycle, the water cycle, soil composition, and plant growth (Sepúlveda et al. 2022; Siad et al. 2019). Most of these models use spatially gridded time series data and produce similarly structured outputs that pertain to water, soil, environment, and plant properties. Such output data sets can quickly get overwhelming to analyze, interpret, and aggregate for obtaining useful insights. One such situation arises with the coupled agro-hydrological model, VIC-CropSyst (Malek et al. 2017), which captures hydrology and plant processes over large regions. It operates at the cell level of a gridded area to produce time series data on multiple variables for each cell. Consequently, post-processing and aggregation

become unwieldy when simulations are performed over large geographical regions with thousands of cells where each result variable is available for each step in space and time.

We present a framework to analyze large-scale geospatial datasets produced by mechanistic simulations used in civil, agricultural, and environmental engineering. The framework combines supervised and unsupervised machine learning techniques to better *explain* the outcomes as a function of the underlying models and initial conditions. Our pipeline includes parameter scanning, clustering of model response, training of machine learning surrogates, and feature importance analysis. As a compelling use case, we perform a climate change study for selected regions of Washington using the VIC-CropSyst simulator. Specifically, we study the transition of the water cycle from snow- to rain-dominated regimes with changing climate variables, which affects agricultural water availability.

1.1 Our Contributions

Explainability Framework for Geospatial Models: This paper presents a general framework architecture for analyzing large spatial models. This framework consists of several components that run simulation models, perform input parameter scans, post-process output data to generate response surfaces for quantities of interest, detect clusters in space based on response surfaces, fit machine learning classifiers using static input data to predict cluster memberships, and compute Shapley values to rank those static parameters.

Analysis of VIC-CropSyst as Case Study: We demonstrate the proposed framework using VIC-CropSyst, running it for a selected set of watersheds in Washington. First, a parameter scan is performed by varying temperatures and precipitation to study the model's response to climate change. This step involves multiple simulation runs for all possible temperature and precipitation delta value combinations. Then, to quantify the effect of climate change on snow and water dynamics, we compute the ratio of snow-derived runoff to total runoff and generate response surfaces for the two scanned parameters.

Evaluation of Factors Influencing Model Dynamics: We perform k-means clustering of the gridded cells in Washington based on the response surfaces to identify six distinct classes. Next, a random forest classifier is fitted with static input parameters to optimize class identification. Then, using Shapley values (Lundberg and Lee 2017), we compute feature importance to characterize and rank how these parameters determine the climate change response for every class. Finally, we discuss the implications of climate change and identify which classes are vulnerable with respect to water.

Paper Organization: Section 2 compiles a literature review on several aspects of this paper, including the climate issue we address, agro-hydro models, AI, sensitivity analysis, and model explainability. We present the high-level framework and our case study implementation using VIC-CropSyst for the climate problem in Section 3. The case study results are presented in Section 4, and we provide some concluding remarks in Section 5.

2 BACKGROUND

Climate Issues: In the US Pacific Northwest, snow accumulates in regions with high elevations during winter. During warmer seasons, melting snow generates runoff, which flows through streams and groundwater. The snowpack is critical for water storage, and the snowmelt timing coincides with most irrigation and other water demands in downstream areas. This process is crucial as it provides water at the right time for agricultural uses in snow-dominant watersheds, including parts of Washington. However, it is at risk, and it was estimated that the volume of the snowpack has declined by 21% since 1915 (Mote et al. 2018). Climate change may further deteriorate snowpack storage, and snow may no longer dominate hydrological processes in many regions (Klos et al. 2014). Certain areas may shift to rainwater-dominated under a warming climate scenario. Land surface models calibrated using historical data with stationarity assumptions of the underlying hydrological processes may not perform well under such transitions (Karimi et al. 2022). It is not only essential to evaluate how snowmelt changes with climate but also useful to characterize how static soil and land properties drive these dynamic behaviors.

Quantification of Snow: Studies have evaluated the contribution of snow in generating runoff (Doesken and Judson 1997) and how it may change in the future (Huning and AghaKouchak 2020). Climate change may shift the seasonality to early spring and reduce the amount of snow-derived runoff in later months, effectively causing water scarcity in summer when demand is high (Barnett et al. 2005). The ratio of snow-derived runoff to total runoff, denoted by $f_{Q,snow}$, is a commonly used metric to indicate the relative contribution of snow, and it has been estimated before by various techniques (Li et al. 2017). Projected future increases in temperature caused by greenhouse emissions will potentially change this ratio along with the snowmelt timing (Qin et al. 2020). It has been estimated that snow historically contributed to 53% of the total runoff in the Western United States, which will fall to 39.5% and 30.4% under the RCP4.5 and RCP8.5 scenarios (Li et al. 2017). The regions where runoff is dominated by snow are also likely to change in response to climate. However, every region may not respond similarly to climate change. Our study aims to identify these differences and analyze them. In particular, we perform a sensitivity analysis of $f_{Q,snow}$ with respect to changing temperature and precipitation. We then relate those response dynamics to static soil and land properties corresponding to various regions.

Hydrology and Plant Growth Models: Hydrological models can simulate processes such as evapotranspiration, snow accumulation, surface runoff, baseflow, etc., to estimate hydrological fluxes (Sepúlveda et al. 2022). These models take as input land/soil characteristics, atmospheric variables, and vegetation (Siad et al. 2019). Some examples of hydrological models include SWAT, JULES, and VIC (Arnold et al. 1998; Best et al. 2011; Liang et al. 1994). On the other hand, crop growth models connect crop yields to water availability, soil features, land and water management policies, climate data, and weather variables (Siad et al. 2019). Crop models such as EPIC, ALMANAC, and CROPSYST (Williams 1990; Kiniry et al. 1992; Stöckle et al. 1994; Stöckle et al. 2003) have been widely used for various applications. Hydrological models are often coupled with crop models to capture water and plant processes better. Depending on the implementation, models can be loosely coupled through parameter exchanges or integrated at the code level (Siad et al. 2019). VIC-CropSyst is a framework with tight coupling at the source code level between the VIC model for hydrology and the CROPSYST model for crop growth and phenology (Malek et al. 2017).

Use of Artificial Intelligence (AI): AI has been used in climate and hydrological systems modeling to achieve various objectives. Convolutional neural networks (CNN) have been used to predict precipitation and how climate change affects extreme events (Davenport and Diffenbaugh 2021). Support vector machine (SVM) classifiers have been used to predict atmospheric rivers by training them on large climate model simulator outputs (Muszynski et al. 2019). Watt-Meyer et al. (2021) used a random forest for bias correction of a global circulation model (GCM). Unsupervised machine learning techniques such as clustering and segmentation have been used to monitor changes in land surface water availability (Chen et al. 2015).

Parameter Scanning: Parameter scanning, or sensitivity analysis, is a common technique for evaluating model behavior. It can be used for various purposes, including model similarity identification, parameter importance, identification of highly sensitive regions in the parameter space, reduction of factors, and uncertainty assessments (Razavi and Gupta 2015). Sensitivity analysis (SA) for VIC commonly reports over-parameterization in many studies. Sepúlveda et al. (2022) found the model sensitive to only 28% of the parameters used in their analyses. Karimi et al. (2022) performed uncertainty quantification of crop yield for VIC-CropSyst under changing climate scenarios. Work by Gou et al. (2020) performed sensitivity analyses for the parameter calibration of the VIC model over China. Temperature and precipitation were varied in a single step for future climate scenarios to observe the changes in streamflow by Nijssen et al. (2001). They found that snow-dominated basins at mid to high elevations underwent the largest changes with increased streamflow in spring. In contrast, transitional snow basins showed a decrease in peak streamflow in spring, but an increase in winter.

Model Explainability: Explainability can often become essential to model simulations. There are different reasons why explaining a model might be required. Some examples are incompleteness in the problem formalization (Doshi-Velez and Kim 2017), a mismatch between model objectives and user requirements (Lipton 2018), and model failures with certain data points (Kabra et al. 2015). Explainability

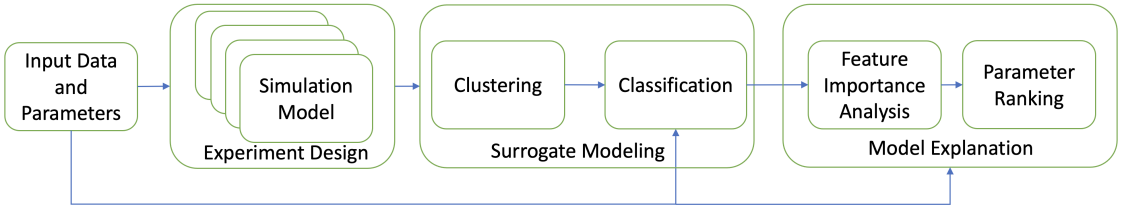


Figure 1: The general framework for explainability and parameter ranking of complex simulation models.

and interpretability are often used interchangeably, and both can contribute to model validation. Many techniques have been developed to explain models, including Lime (Ribeiro et al. 2016), LRP (Bach et al. 2015), SHAP (Lundberg and Lee 2017), etc. In this work, we use SHAP (Shapley Additive Explanations), which is a game theoretic approach where a *feature* is interpreted as a player in a cooperative game, and a *prediction* is interpreted as the payout. The objective is to distribute the payout fairly to the players.

3 METHODOLOGY

3.1 The Framework

A high-level diagram of the framework is shown in Figure 1. The main idea is to combine the surrogate modeling of simulations with recent advances in explaining machine learning models to allow the explanation of complex simulations. We develop an experiment design to generate the appropriate simulation data for training a surrogate model. There are multiple ways to train surrogate models. The most general and classic method is response surface learning, where a regression model is trained to try to predict the exact output of a target variable from the simulation at every point in the parameter space. However, it can be hard to get good performance at this task, and often, domain experts are more interested in the ranges of simulation outputs that are qualitatively different. We propose to cluster the simulation outputs first and then train the surrogate model to predict which cluster the simulation output will fall into for a given parameter setting. This turns the surrogate modeling problem into a simpler classification problem while preserving the qualitative variability in the output. Once the classifier has been satisfactorily trained, we pass it to a model explanation/interpretation method, as described below. In what follows, we work through this framework in detail for the specific example of VIC-CropSyst.

3.2 Definitions

Given Washington's Columbia River Basin (CRB) study area, we want to rank static soil and land properties that drive snowmelt response behavior when certain forcing inputs are changed from historically observed values. The two forcing variables used in this study are temperature (T) and precipitation (P). We use the ratio of the snow-derived runoff to the total runoff as the metric (Li et al. 2017) for evaluating snowmelt response behavior. This metric, denoted by $f_{Q,snow}$, is defined as follows:

$$f_{Q,snow} = \Sigma_t \, Q_{snow} / \Sigma_t \, Q \, . \tag{1}$$

Here, Q_{snow} is defined as the runoff originating from snow, and Q is defined as the total runoff (from snow and rain). The runoffs are computed from the output fluxes of the VIC-CropSyst simulator using a snowmelt tracker algorithm (Li et al. 2017) for each timestep. The summations are done over time to compute the long-term accumulation of runoffs. This computation is done for each grid cell. We define a grid cell to be *snow-dominated* if $f_{Q,snow} > 0.5$, and *rain-dominated* if it is ≤ 0.5.

Let $x_i(t)$ be the value of a forcing variable i at time step t, where i can indicate any of the variables: precipitation (P), minimum temperature (T_{min}), or maximum temperature (T_{max}). The forcing variables involving temperatures are changed as follows.

$$x'_i(t) = x_i(t) + \Delta x_T;\ \text{for } i \in T_{max}, T_{min}, \tag{2}$$

where Δx_T is the constant (across time steps) temperature amount that shifts the mean value of the time series. The input precipitation is modified as follows.

$$x'_i(t) = x_i(t) \times (1 + \Delta x_P);\ \text{for } i \in P, \tag{3}$$

where Δx_P is the constant (across time steps) fractional change in precipitation that modifies the original values of the time series.

The two change variables, namely Δx_T and Δx_P, take a range of values from T_{range} and P_{range}, respectively. For each data point, the value of $f_{Q,snow}$ is computed. This results in a matrix for each grid cell ($c \in [1:N]$); we denote this matrix by $F(c)$. Each element f_{jk} of this matrix is the value of $f_{Q,snow}$ for $\Delta x_T = T_{range}(j)$ and $\Delta x_P = P_{range}(k)$. Here, N is the total number of cells when the target region is gridded at some predefined spatial resolution.

3.3 Run Configuration

The simulations were run on Rivanna, a high-performance computing (HPC) system maintained by the research computing division of the University of Virginia. Our framework is implemented in Python with SLURM scripts to facilitate the parameter scan, simulation run, monitoring, and post-processing. The simulations are run on daily timesteps (24 hours) from 1970 to 2015.

Study Area: The study area consists of 57 watersheds in Washington. The entire region is gridded and divided into $N = 4,834$ cells at $1/16^\circ$ spatial resolution. These cells have a mean area of 33.4 km^2.

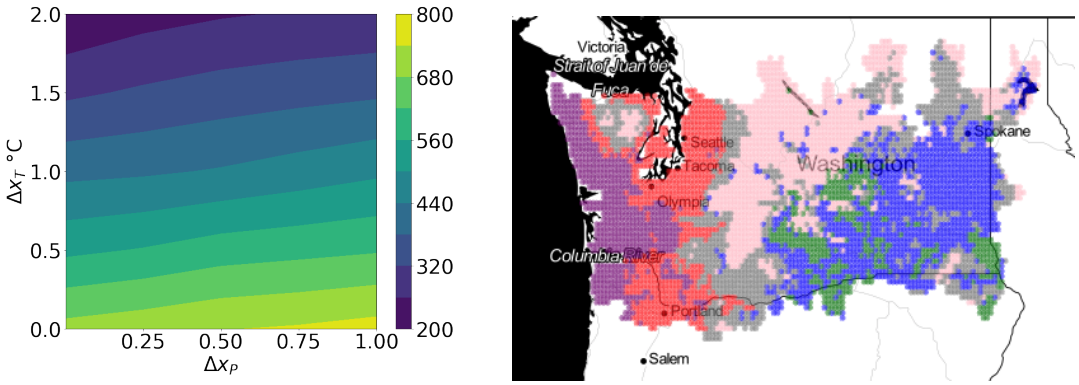
Input Parameters: The VIC-CropSyst simulator takes as inputs layered soil parameters, vegetation, irrigation, etc. In addition to temporally static variables, VIC-CropSyst takes time series of atmospheric forcing data as input, including air temperature and precipitation for each cell. We used the gridMET dataset (Abatzoglou 2013) that contains high-resolution spatial data for the contiguous United States. We vary three forcing parameters: minimum temperature (T_{min}), maximum temperature (T_{max}), and precipitation (P) as per equations (2) and (3). We set $T_{range} = [0, 0.5, 1.0, 1.5, 2.0]$ (in °C units) and $P_{range} = [0, 0.25, 0.50, 0.75, 1.0]$ (unitless). This results in a parametric search space of a 5×5 matrix.

3.4 Post Processing

Snowmelt Tracking: The cell-wise hydrological flux outputs are processed by a snowmelt tracking algorithm (Li et al. 2017), which computes $f_{Q,snow}$ as defined in equation (1). The tracking algorithm aggregates the historical contributions of snow and rain for each cell over a period of 30 years (from October 1986 to September 2015). This results in a single value of $f_{Q,snow}$ for each cell per simulation run (parameter combination). Eventually, we get a 5×5 matrix of $F(c)$ for each cell $c \in [1:N]$ where $N = 4,834$.

Clustering: Using the K-means clustering algorithm with the Euclidean distance metric, we categorize the cells into M distinct classes based on their $F(c)$ matrices. Here, M is the optimum number of classes for the given data. M is determined using the *knee method* on variance reduction plot.

Feature Selection: The static soil and land parameters of VIC-CropSyst are further used here to explain the clusters of $F(c)$. Each cell has 15 vectors (17 soil layers) and 12 scalars, resulting in 267 parameters. A complete list can be found here in the official VIC documentation (UW-Hydro 2023). Exploratory data analysis (EDA) was performed on them. During the primary inspection, 40 parameters were dropped because they were nearly constant (standard deviation $< 1 \times 10^{-8}$) across the grid cells. Temperature and precipitation parameters (2) were dropped as they were tied to scanned variables. The remaining 225 parameters were evaluated using pairwise correlation analysis. A semantic approach was taken to keep meaningful parameters and drop the ones highly correlated to the preferred ones. Parameters were dropped if found to correlate with a Pearson correlation coefficient > 0.5. After an iterative approach, 212 parameters were dropped, leaving 13 for the following stages.



(a) (b)

Figure 2: (a) A contour plot depicting the number of snow-dominated grid cells with varying temperatures and precipitation. The results are summarized for 4,834 total grid cells. The color bar on the right indicates snow-dominated grid cell counts. (b) The grid cells are clustered into 6 classes highlighted by distinct colors: Blue (1), Red (2), Green (3), Grey (4), Purple (5), and Pink (6). Each grid cell is depicted as a point on the map with a color corresponding to its class.

Classification: The 13 soil parameters were used as features and class memberships of cells as targets to train a random forest classifier. A random split of 80:20 was performed on 4,834 samples for training and testing. Bootstrap samples were used with the Gini impurity metric to build 50 trees. The classifier was regularized by requiring a minimum of 30 samples to form a leaf node and allowing a max tree depth of 9. The hyperparameters were tuned to optimize accuracy without overfitting.

Feature Importance: To identify and rank which features contribute more in predicting the correct class for a grid cell, the Python SHAP package was used (Lundberg et al. 2020).

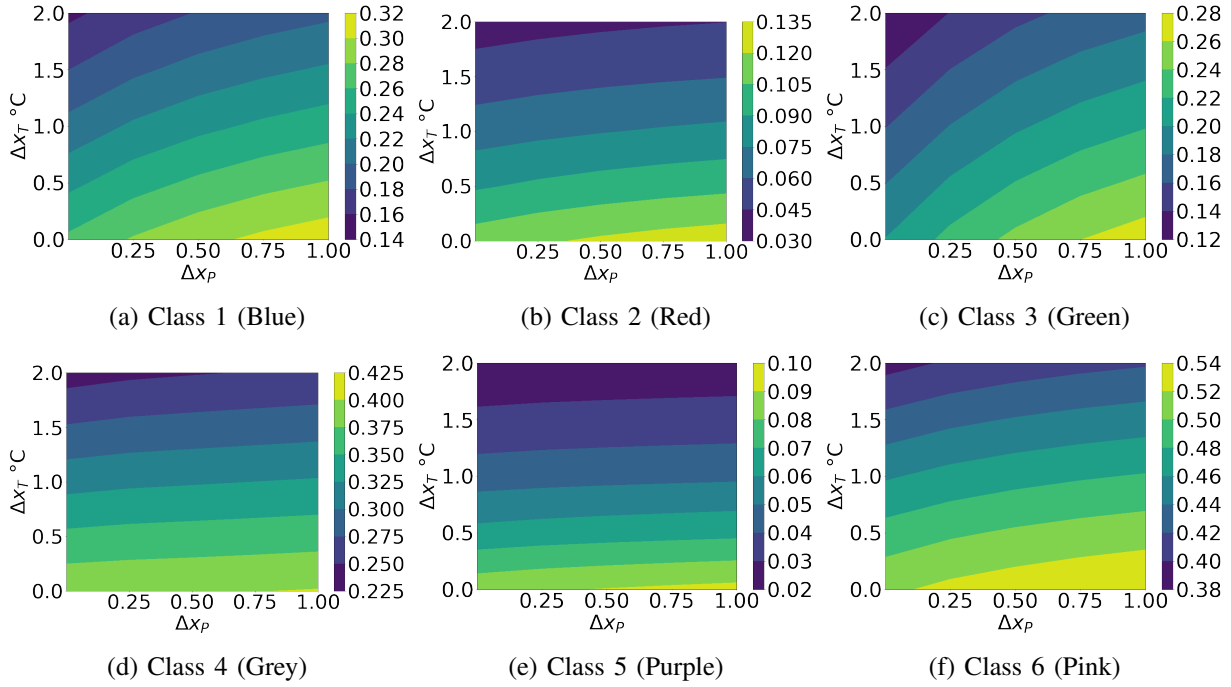
4 RESULTS

A summary of the results from the snowmelt tracking step is depicted as a contour plot in Figure 2a. It shows the total number of snow-dominated cells in the target region (the rest are rain-dominated) based on aggregate simulation data of 30 years (1986-2015). While the changes in temperature and precipitation are not directly comparable, the contour lines of Figure 2a indicate that $f_{Q,snow}$ is more sensitive to temperature than precipitation, and the two parameters have opposite effects.

4.1 Clustering of Cells

The optimum number of classes (M) was determined to be 6, using the *knee method* on the variance reduction plot. Hence, the cells are categorized with class ids ranging from 1 to 6. In Figure 2b, the cells belonging to the 6 classes are color-coded as follows, Blue (1), Red (2), Green (3), Grey (4), Purple (5), and Pink (6). The number of cells belonging to classes 1-6 are 1107, 685, 392, 1004, 560, and 1086, respectively. At the baseline ($\Delta x_T = 0°\text{C}$ and $\Delta x_P = 0$) scenario, four out of the six classes have one or more snow-dominated cells, with classes 1, 3, 4, and 6 having 2, 1, 110, and 587 such cells, respectively.

To understand more about how each class behaves in response to changes in temperature and precipitation, the mean response surfaces are plotted as contours for the 6 classes in Figure 3. A common trait across all classes is that the sensitivity of $f_{Q,snow}$ to temperature (Δx_T) is significantly higher compared to precipitation (Δx_P). Classes 1 (Blue) and 3 (Green) are closely matched in response dynamics and mean $f_{Q,snow}$. These two classes are mostly adjacent in Figure 2b. The Green cells are the most sensitive to changing precipitation



(a) Class 1 (Blue) (b) Class 2 (Red) (c) Class 3 (Green)

(d) Class 4 (Grey) (e) Class 5 (Purple) (f) Class 6 (Pink)

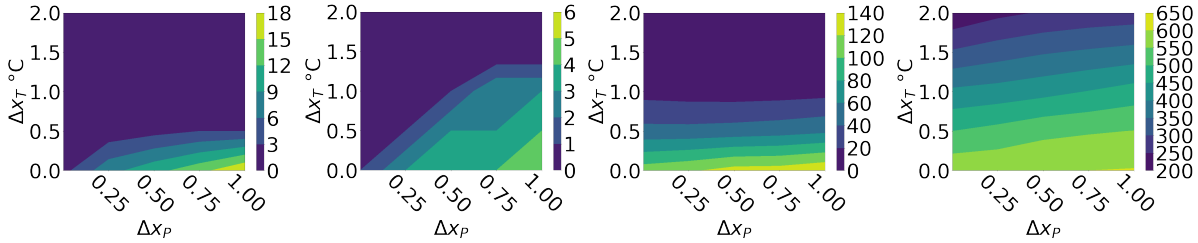
Figure 3: The mean of response curves ($F(c)$) for the cells belonging to each of the 6 classes. Each data point denotes the mean value of $f_{Q,snow}$ in a class for the given $\Delta x_T, \Delta x_P$ combination. Please note that the range of values (and corresponding color bars) are different.

among the entire bunch. Classes 4 and 5 (Grey and Purple) appear to have very low sensitivity to precipitation. The Purple cells have some of the smallest values of $f_{Q,snow}$, compared to others. As is evident from the map (Figure 2b), class 5 (Purple) consists of near-sea-level cells closer to the Pacific coast. Classes 4 (Grey) and 6 (Pink), on the other hand, have much higher snow-rain ratios. Class 6 (Pink) cells are comparatively more sensitive to precipitation than class 4 (Grey) cells. Class 2 (Red) has a closer association with class 5 (Purple), where snow-derived runoff quickly diminishes with increased temperature. While elevation data were not used in clustering, the results indicate that $f_{Q,snow}$ is strongly affected by elevation.

To identify the regions vulnerable to climate change and at high risk of losing snow dominance, we plot the number of *snow-dominated* cells in each class in response to temperature and precipitation changes in Figure 4. Only 4 classes out of the 6 had snow-dominated cells in the parameter search space. Class 1 (Blue) showed clear sensitivity to increased precipitation by adding 16 snow-dominated cells when precipitation is doubled. However, it would lose most of those if the temperature is increased by 0.5°C. The neighboring class 3 (Green) cells have fewer snow-dominated ones for elevated precipitation. However, some of these cells are more resilient to temperature changes and may sustain up to 1°C increase. These two classes of cells mostly lie in the Columbia Basin area. Classes 4 (Grey) and 6 (Pink) have significantly more snow-dominated cells. These cells are in and around the proximity of the Cascade Mountains. These classes see a sharp decline in snow-dominated cells with increased temperature. In class 4 (Grey), 88% of cells lose snow dominance if the temperature increases by a degree (°C). In class 6 (Pink), 30.2% and 64.6% of cells lose snow dominance if the temperature is increased by 1°C and 2°C, respectively.

4.2 Prediction of Clusters Using Random Forest Classification

A random forest (RF) classifier was fitted to static soil data, using the membership of one of the 6 classes as the target. The trained model achieved 75% overall accuracy over test data samples. RF classifier



(a) Class 1 (Blue) (b) Class 3 (Green) (c) Class 4 (Grey) (d) Class 6 (Pink)

Figure 4: Contour plots depicting the number of snow-dominated grid cells with varying temperature and precipitation for each class. Plots for the Red (2) and Purple (5) classes are not given here, as those regions are entirely rain-dominated in the parameter search space.

Table 1: Summary of Random Forest Classification of grid cell clustering. N is the number of test data samples.

Class	Precision	Recall	F_1 Score	N
1 (Blue)	0.73	0.74	0.74	223
2 (Red)	0.83	0.80	0.82	141
3 (Green)	0.61	0.56	0.59	73
4 (Grey)	0.65	0.72	0.68	208
5 (Purple)	0.94	0.80	0.86	111
6 (Pink)	0.76	0.78	0.77	211

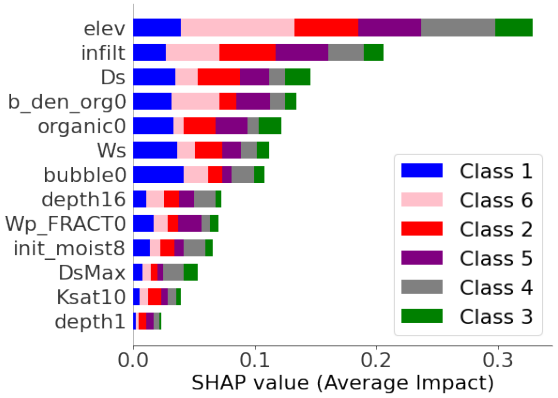


Figure 5: Shapley values indicating the relative importance of soil parameters.

performance is summarized in Table 1. Accuracy varied noticeably across classes. The classifier had the highest F_1 score for class 5 (Purple) cells (Pacific coast region) and the lowest F_1 score for class 3 (Green) cells (basin area).

4.3 Ranking of Static Parameters

The 13 selected features used to train the RF classifier are ranked using Shapley values in Figure 5. The summary plot of Figure 5 illustrates the mean impact and stacks feature contributions across different clusters to obtain an overall impact of a feature for all grid cells. The greater the overall impact, the higher a feature ranks. The most impactful feature is the elevation (*elev*), which contributes considerably to identifying class 6 (Pink) cells (forming around the Cascade, Olympic, and Rocky Mountains). The infiltration curve parameter ($infilt$) is the second most prominent, and its contribution is evenly distributed across most classes. It is a calibrated parameter, and higher values produce more runoff in the model. The next parameter is D_s, the fraction of max velocity where non-linear baseflow occurs.

To better understand how these parameters affect classification, we plot Shapley values for all 6 classes separately in Figure 6. Class 1 (Blue) was primarily identified by bubbling pressure (*bubble0*), elevation (*elev*), and the organic fraction of the soil layer (*organic0*). Relatively smaller values for *bubble0* and *organic0* are correlated with being classified as 1 (Blue). Elevation data is quite mixed, though. Despite that, the classification is generally good (F_1 score of 0.74). Class 2 (Red) is marked by low *elev*, high $infilt$, and low D_s. Clear trends for these top 3 parameters result in good classification performance (F_1 score of

Table 2: Summary of the 6 classes identified by our framework.

Class	**Region**	**Elevation (m)** *mean [Q_1 Q_3]*	**Baseline**	**Temperature Sensitivity**	**Risk**
1 (Blue)	Columbia Basin	568 [405 729]	2	1	Low
2 (Red)	Puget Lowland	271 [84 409]	0	n/a	None
3 (Green)	Columbia Basin	430 [244 491]	1	1	Low
4 (Grey)	Mixed	824 [621 1034]	110	0.88	High
5 (Purple)	Coastal	237 [117 319]	0	n/a	None
6 (Pink)	Mountains	1074 [723 1403]	587	0.30	High

0.82). Class 3 (Green) is marked by low *elev*, *organic0*, and high D_s. However, for many cells, lower values of organic fraction relate to being in class 1 (Blue), which complicates classification performance, scoring poorly at 0.59 (F_1 score). Class 4 (Grey) also has poor precision, which can be attributed to mixed values of the top contributing parameter (mixed elevation values on the right side of the vertical line with SHAP impact 0 in Figure 6d). Class 4 (Grey) cells are adjacent to class 6 (Pink) but with moderate elevations. Class 5 (Purple) cells show clear trends toward low *elev* and high $infilt$. Such clarity in the top 2 features results in excellent predictive performance (F_1 score of 0.86). Class 6 (Pink) cells lie in higher elevations as indicated by the top feature in Figure 6f.

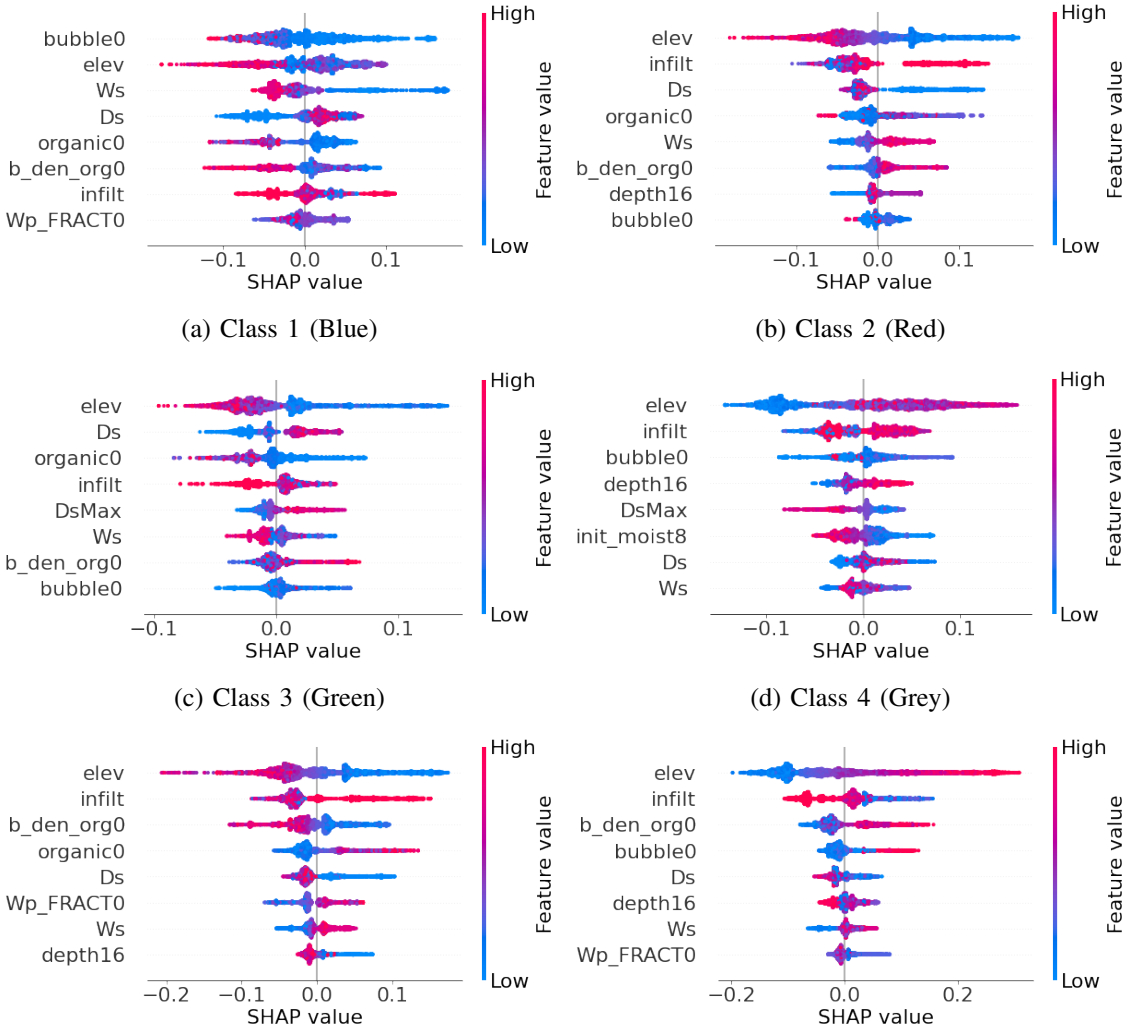
A summary of the 6 classes is shown in Table 2. The Elevation column contains the mean elevation with 1st (Q1) and 3rd (Q3) quartile values. The Baseline column shows the count of snow-dominated cells at $\Delta x_T = 0^\circ$C and $\Delta x_P = 0$. The Temperature Sensitivity column shows the fraction of snow-dominated cells that would transition to rain-dominated when $\Delta x_T = 1^\circ$C over baseline. Each class is associated with a risk level when applicable based on the total number of cells (hence the overall area) that undergo this transition. Classes 4 (Grey) and 6 (Pink) are marked as high-risk, as they see many cells transitioning to rain-dominated in response to climate change. This translates to a loss of snowpack storage.

5 DISCUSSION

We described a framework to analyze large-scale geospatial simulation data. We performed a climate change study using the VIC-CropSyst simulator to demonstrate its effectiveness. Our analyses characterized the gridded cells of Washington into 6 distinct classes based on their snowmelt dynamics. A machine learning surrogate was trained to detect such classifications satisfactorily without performing computationally expensive simulations. We also ranked important factors that determine those dynamic behaviors using Shapley values. On top of these, we identified risks associated with climate change by evaluating and comparing snowmelt behaviors across classes. While this paper demonstrates one use case of simulation data analysis, the framework can be applied to other models with large spatial/temporal datasets and complex interactions where interpretability is limited. It is well suited for generalizing across frameworks because it uses machine learning to relate outputs with inputs and does not look at the model's internals. Nevertheless, there are limitations when it comes to generalizing our approach. Experiment design would require domain-specific knowledge, and some system components must be adapted to model-specific input and output data formats. As a validation, our conclusions are consistent with other works performing extensive parameter sensitivity analyses on similar models. A theoretical validation is out of the scope of this paper and is left for future work. Future expansions can also focus on adaptation to multiple domains, model comparisons, and creating ensembles of interpretable results by combining models.

ACKNOWLEDGMENTS

We thank the reviewers and editors for their valuable suggestions. This material is based upon work supported by the AI Research Institutes program supported by NSF and USDA-NIFA under the AI Institute: Agricultural AI for Transforming Workforce and Decision Support (AgAID) award No. 2021-67021-35344 and NSF



(a) Class 1 (Blue)

(b) Class 2 (Red)

(c) Class 3 (Green)

(d) Class 4 (Grey)

(e) Class 5 (Purple)

(f) Class 6 (Pink)

Figure 6: Density scatter plots of SHAP values of the top 8 features for each cluster. Each data point indicates a sample of a feature with color indicating its relative magnitude. The sum of the absolute SHAP value magnitudes sorts the features. Note when the data points don't fit on a line, they are piled up to indicate density.

award No. OAC-1916805 (CINES). Opinions, findings, and conclusions are those of the author(s) and do not necessarily reflect the view of funding entities.

REFERENCES

Abatzoglou, J. T. 2013. "Development of Gridded Surface Meteorological Data for Ecological Applications and Modelling". *International Journal of Climatology* 33(1):121–131.

Arnold, J. G., R. Srinivasan, R. S. Muttiah, and J. R. Williams. 1998. "Large Area Hydrologic Modeling and Assessment Part I: Model Development 1". *JAWRA Journal of the American Water Resources Association* 34(1):73–89.

Bach, S., A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. 2015. "On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation". *PLOS ONE* 10(7):e0130140.

Barnett, T. P., J. C. Adam, and D. P. Lettenmaier. 2005. "Potential Impacts of a Warming Climate on Water Availability in Snow-Dominated Regions". *Nature* 438(7066):303–309.

Best, M. J., M. Pryor, D. Clark, G. G. Rooney, R. Essery, C. Ménard, J. Edwards, M. Hendry, A. Porson, N. Gedney et al. 2011. "The Joint UK Land Environment Simulator (JULES), Model Description–Part 1: Energy and Water Fluxes". *Geoscientific Model Development* 4(3):677–699.

Burkart, N., and M. F. Huber. 2021. "A Survey on the Explainability of Supervised Machine Learning". *Journal of Artificial Intelligence Research* 70:245–317.

Chen, X. C., A. Khandelwal, S. Shi, J. H. Faghmous, S. Boriah, and V. Kumar. 2015. "Unsupervised Method for Water Surface Extent Monitoring using Remote Sensing Data". In *Machine Learning and Data Mining Approaches to Climate Science: Proceedings of the 4th International Workshop on Climate Informatics*, 51–58. Springer.

Davenport, F. V., and N. S. Diffenbaugh. 2021. "Using Machine Learning to Analyze Physical Causes of Climate Change: A Case Study of US Midwest Extreme Precipitation". *Geophysical Research Letters* 48(15):e2021GL093787.

Doesken, N. J., and A. Judson. 1997. *The Snow Booklet: A Guide to the Science, Climatology, and Measurement of Snow in the United States*. Colorado State University Publications & Printing.

Doshi-Velez, Finale and Kim, Been 2017. "Towards a Rigorous Science of Interpretable Machine Learning". arXiv preprint arXiv:1702.08608, 13 pages.

Gou, J., C. Miao, Q. Duan, Q. Tang, Z. Di, W. Liao, J. Wu, and R. Zhou. 2020. "Sensitivity Analysis-based Automatic Parameter Calibration of the VIC Model for Streamflow Simulations over China". *Water Resources Research* 56(1):e2019WR025968.

Huning, L. S., and A. AghaKouchak. 2020. "Global Snow Drought Hot Spots and Characteristics". *Proceedings of the National Academy of Sciences* 117(33):19753–19759.

Kabra, M., A. Robie, and K. Branson. 2015. "Understanding Classifier Errors by Examining Influential Neighbors". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3917–3925. IEEE.

Karimi, T., P. Reed, K. Malek, and J. Adam. 2022. "Diagnostic Framework for Evaluating How Parametric Uncertainty Influences Agro-Hydrologic Model Projections of Crop Yields Under Climate Change". *Water Resources Research* 58(6):e2021WR031249.

Kiniry, J. R., J. Williams, P. W. Gassman, and P. Debaeke. 1992. "A General, Process-Oriented Model for Two Competing Plant Species". *Transactions of the ASAE* 35(3):801–810.

Klos, P. Z., T. E. Link, and J. T. Abatzoglou. 2014. "Extent of the Rain-Snow Transition Zone in the Western US under Historic and Projected Climate". *Geophysical Research Letters* 41(13):4560–4568.

Li, D., M. L. Wrzesien, M. Durand, J. Adam, and D. P. Lettenmaier. 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.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges. 1994. "A Simple Hydrologically based Model of Land Surface Water and Energy Fluxes for General Circulation Models". *Journal of Geophysical Research: Atmospheres* 99(D7):14415–14428.

Lipton, Z. C. 2018. "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is both Important and Slippery". *Queue* 16(3):31–57.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. 2020. "From Local Explanations to Global Understanding with Explainable AI for Trees". *Nature Machine Intelligence* 2(1):56–67.

Lundberg, S. M., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions". In *Advances in Neural Information Processing Systems*, Volume 30, 4765–4774. NIPS.

Malek, K., C. Stöckle, K. Chinnayakanahalli, R. Nelson, M. Liu, K. Rajagopalan, M. Barik, and J. C. Adam. 2017. "VIC–CropSyst-v2: A Regional-scale Modeling Platform to Simulate the Nexus of Climate, Hydrology, Cropping Systems, and Human Decisions". *Geoscientific Model Development* 10(8):3059–3084.

Maria, A. 1997. "Introduction to Modeling and Simulation". In *Proceedings of the 1997 Winter Simulation Conference*, edited by S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson, 7–13. Atlanta, Georgia, USA: IEEE.

Mote, P. W., S. Li, D. P. Lettenmaier, M. Xiao, and R. Engel. 2018. "Dramatic Declines in Snowpack in the Western US". *Npj Climate and Atmospheric Science* 1(1):1–6.

Muszynski, G., K. Kashinath, V. Kurlin, M. Wehner et al. 2019. "Topological Data Analysis and Machine Learning for Recognizing Atmospheric River Patterns in Large Climate Datasets". *Geoscientific Model Development* 12(2):613–628.

Nijssen, B., G. M. O'Donnell, A. F. Hamlet, and D. P. Lettenmaier. 2001. "Hydrologic Sensitivity of Global Rivers to Climate Change". *Climatic Change* 50:143–175.

Qin, Y., J. T. Abatzoglou, S. Siebert, L. S. Huning, A. AghaKouchak, J. S. Mankin, C. Hong, D. Tong, S. J. Davis, and N. D. Mueller. 2020. "Agricultural Risks from Changing Snowmelt". *Nature Climate Change* 10(5):459–465.

Razavi, S., and H. V. Gupta. 2015. "What Do We Mean by Sensitivity Analysis? The Need for Comprehensive Characterization of "Global" Sensitivity in Earth and Environmental Systems Models". *Water Resources Research* 51(5):3070–3092.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. ""Why Should I Trust You?" Explaining the Predictions of Any Classifier". In *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 1135–1144. ACM.

Sepúlveda, U. M., P. A. Mendoza, N. Mizukami, and A. J. Newman. 2022. "Revisiting Parameter Sensitivities in the Variable Infiltration Capacity Model Across a Hydroclimatic Gradient". *Hydrology and Earth System Sciences* 26(13):3419–3445.

Siad, S. M., V. Iacobellis, P. Zdruli, A. Gioia, I. Stavi, and G. Hoogenboom. 2019. "A Review of Coupled Hydrologic and Crop Growth Models". *Agricultural Water Management* 224:105746.

Stöckle, C. O., M. Donatelli, and R. Nelson. 2003. "CropSyst, A Cropping Systems Simulation Model". *European Journal of Agronomy* 18(3-4):289–307.

Stöckle, C. O., S. A. Martin, and G. S. Campbell. 1994. "CropSyst, A Cropping Systems Simulation Model: Water/Nitrogen Budgets and Crop Yield". *Agricultural Systems* 46(3):335–359.

UW-Hydro 2023. "Variable Infiltration Capacity (VIC) Macroscale Hydrologic Model". https://vic.readthedocs.io/.

Watt-Meyer, O., N. D. Brenowitz, S. K. Clark, B. Henn, A. Kwa, J. McGibbon, W. A. Perkins, and C. S. Bretherton. 2021. "Correcting Weather and Climate Models by Machine Learning Nudged Historical Simulations". *Geophysical Research Letters* 48(15):e2021GL092555.

Williams, J. R. 1990. "The Erosion-Productivity Impact Calculator (EPIC) Model: A Case History". *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 329(1255):421–428.

AUTHOR BIOGRAPHIES

TANVIR FERDOUSI is a Postdoctoral Research Associate at Biocomplexity Institute, University of Virginia. His email address is tanvir@virginia.edu.

MINGLIANG LIU is an Assistant Research Professor in the Department of Civil and Environmental Engineering at Washington State University. His email address is mingliang.liu@wsu.edu.

KIRTI RAJAGOPALAN is an Assistant Professor in the Department of Biological Systems Engineering at Washington State University. Her email address is kirtir@wsu.edu.

JENNIFER ADAM is a Distinguished Professor in the Department of Civil and Environmental Engineering at Washington State University. Her email address is jcadam@wsu.edu.

ABHIJIN ADIGA is a Research Associate Professor at Biocomplexity Institute, University of Virginia. His email address is aa5ts@virginia.edu.

MANDY WILSON is a Senior Scientist at Biocomplexity Institute, University of Virginia. Her email address is alw4ey@virginia.edu.

S. S. RAVI is a Research Professor at Biocomplexity Institute, University of Virginia. His email address is ssr6nh@virginia.edu.

ANIL VULLIKANTI is a Professor of the Computer Science Department and the Biocomplexity Institute at the University of Virginia. His email address is asv9v@virginia.edu.

MADHAV V. MARATHE is a Professor of the Computer Science Department and a Division Director of the Biocomplexity Institute at the University of Virginia. His email address is mvm7hz@virginia.edu.

SAMARTH SWARUP is a Research Associate Professor at Biocomplexity Institute, University of Virginia. His email address is swarup@virginia.edu.