MONITORING AND MODELING ECOHYDROLOGICAL PROCESSES IN VEGETATED WATERSHEDS

By

Leo Triet Pham

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Forestry – Master of Science

2022

ABSTRACT

MONITORING AND MODELING ECOHYDROLOGICAL PROCESSES IN VEGETATED WATERSHEDS

By

Leo Triet Pham

Ecohydrology links ecological and hydrological processes and considers interactions between water resources and ecosystems. Modeling tools are not only important for studying the mechanisms of ecological patterns and processes but also for assessing the effects of environmental change on hydrological and ecological processes, providing insights and solutions to issues in water management. This thesis explores various data-driven approaches to monitor and model these processes at 95 watersheds in western USA using a combination of seasonal and annual climate, hydrometric, and remotely sensed vegetation data. In one analysis, we show that a trend in earlier peak in spring vegetation activity may be a linked to reduced runoff availability during drought years compared to non-drought years. We also provide evidence that an increase drought severity is consistent with a decrease in runoff ratio in forested catchments through regression analysis, supporting the hypothesis that the relationship among water-balance components may shift during hydrological drought events. In another analysis, we show that the type and amount of vegetation coverage, among other catchment characteristics, can affect the accuracy of data-driven runoff models. These results suggest that a better understanding of the ecohydrologic processes and characteristics is vital to development of effective long-term strategies to improve the resilience of watersheds.

ACKNOWLEDGEMENTS

I would like to thank my advisers, Dr. Andrew Finley and Dr. Lifeng Luo, for their academic guidance and professional support. I'd also like to acknowledge my committee members Dr. Asia Dowtin for her invaluable comments and criticisms of my thesis work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

CHAPTER 1

ECOHYDROLOGICAL IMPLICATIONS OF DROUGHTS IN SIERRA NEVADA WATERSHEDS

1.1 Introduction

Drought is considered as a sustained period of less water compared to normal conditions of a region. Depending on the severity and duration, these periods of water deficit can have important implications on human activities and natural systems. In forested catchments, the relationships among drought, streamflow dyanmics, groundwater recharge and vegetation response are not straightforward because forest ecosystems have a large capacity to regulate precipitation partitioning through the evapotranspiration process (Sun et al., 2011) and are able to withstand certain disturbances through various physiological, morphological, and behavioral adaptations [Lytle and Poff, 2004]. As more severe and widespread droughts are projected for the 21st century [Dai, 2013], examining these interactions and feedbacks can provides a valuable ecohydrological context for understanding and evaluating the impacts of drought on water resources and management practices [Rodriguez-Iturbe, 2000, Brauman et al., 2007].

In the Mediterranean climate regions such as the Sierra Nevada, California, a substantial fraction of precipitation during the wet winter (November-April) provides the water supply to recharge reservoirs, replenishing groundwater, and build snowpack that provides water storage for the dry summer season (June-August). The recent multi-year droughts in California have been characterized by both large precipitation deficits and abnormally high temperatures during both wet and dry seasons [Swain, 2015, Luo et al., 2017]. The combined effect of warming temperature and variable precipitation due to climate change is expected to intensify drought and prolong periods of water stress [Diffenbaugh et al., 2015, Barnett et al., 2005]. Most studies agree that decreases in mean annual flow, earlier snowmelt runoff,

and higher evaporative demand are expected [Medellín-Azuara et al., 2008, Vicuña et al., 2008]. Because of the heterogeneity in topographic and physiographic features, vegetation types and structure, and their interaction with local climate, individual watersheds within this region will likely have different sensitivity where hydrologic responses to drought can be either mitigated or exacerbated by forest vegetation depending upon vegetation water use [Vose et al., 2016].

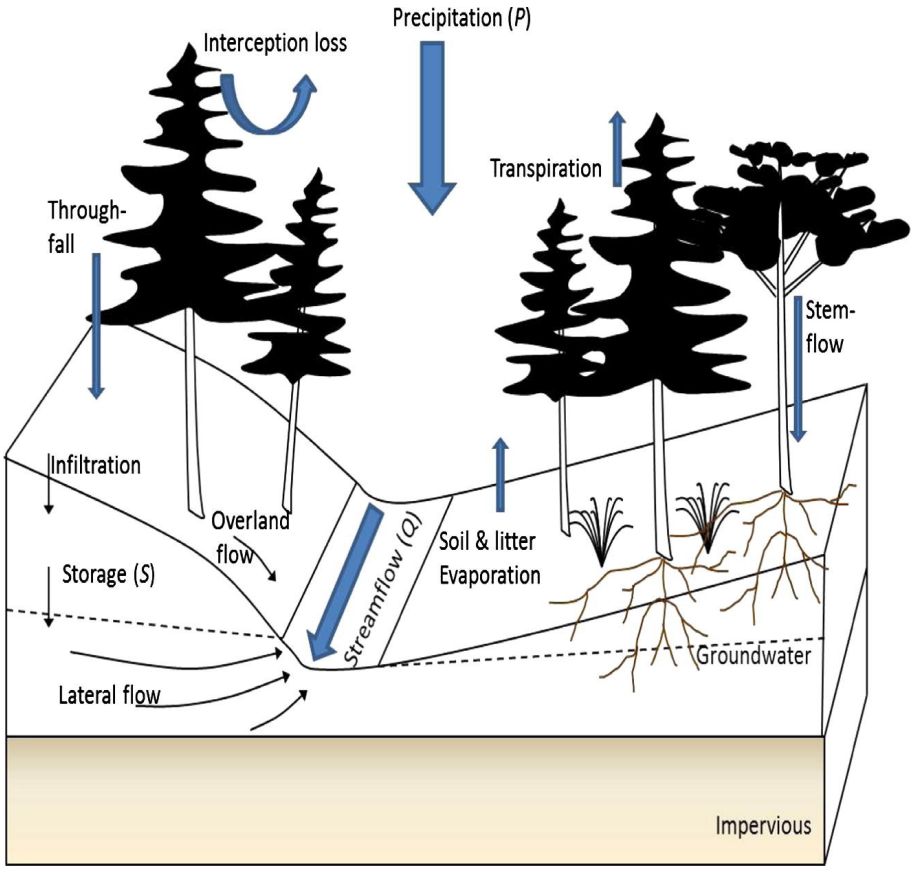
It is well established that forests grow where the water supply is sufficient to support perennial woody vegetation, as evapotranspiration (the sum of interception, transpiration and soil evaporation) is much greater in forest systems compared to other vegetation types [Frank and Inouye, 1994]. As drought is directly related to the balance between soil water supply from incoming precipitation and tree water demand, the simple water balance equation provides an intuitive framework to assess the relationships among different pools and partitioning of precipitation under different conditions.

At annual and longer time scales, the water balance for a catchment can be written as:

$$P = ET + Q + \Delta S \tag{1.1}$$

where P is precipitation, ET is evapotranspiration, Q is surface runoff measured as streamflow, and ΔS represents the change in subsurface storage within the watershed (Fig. 1.1). The sensitivity of streamflow (Q) to drought can be framed by examining how much evaporative and transpiration losses occur relative to the total precipitation. During drought conditions, reduction in P, increase in evaporative and water demand in trees can contribute to reduction to Q [D'Amato et al., 2013, Vicuna and Dracup, 2007]. However, the effect of these factors can vary greatly in time and space depending on the intensity, frequency, and type of precipitation input and the timing of water deficit. It is expected that in mountainous snowmelt-dominated regions less water is lost to evaporation and potential ET has little effect on the total annual runoff as the precipitation is delivered efficiently in forms of meltwater in large pulses [Wolock and McCabe, 1999]. Barnett et al. [2005] argued that the increased soil moisture earlier in the season due to melt happens when potential evaporation

Figure 1.1 Partitioning of water at catchment scale (Figure adapted from Vose et al. [2016]).



(dominated by net radiation) is low, thus attenuating the effects of ET changes to runoff production in these regions. However, more recent works show that warming temperature condition result in accelerating mountain vegetation growth and ET [Goulden and Bales, 2014], and longer growing season in lower elevations [Hunsaker et al., 2012], which in turn lead to runoff vulnerability as a greater proportion of snowmelt is converted to ET. Bales et al. [2018] summarized the mechanisms in which droughts can shift the P-Q relationship in mountainous catchments which include both priority of partitioning of precipitation into ET vs. discharge during drought, warmer than normal conditions creating higher vegetation evaporative demand, and spatial heterogeneity across the watershed and sources of P. These mechanisms highlight the relative importance of ET and vegetation dynamics in catchment response to drought. Thus, integrating vegetation's response to drought can benefit our

understanding of the partitioning of water and drought impacts.

Most studies that highlight the impact of droughts in California and Sierra Nevada have been conducted at global and regional scales [Mann and Gleick, 2015, Diffenbaugh et al., 2015]). However, there is still a dearth in little research conducted at the catchment scale. Moreover, Saft et al. [2015] found that local catchment properties such as mean slope and percentage of woody cover can play a role in changes in P-Q relationship induced by drought. In this study, we investigate of the drought conditions on the on water resources and vegetation dynamics in the 9 selected Sierra Nevada catchments. We are particularly interested in testing the hypothesis that the annual relationship between P and Q shifts during drought conditions compared to non-drought. The study consists of an empirical analysis of seasonal and annual hydrometric, and remotely sensed vegetation index and ET data supported by spatial information on catchment characteristics. As the hydrological processes in mountains are thus highly sensitive to changes in climate particularly drought [Beaulieu et al., 2016], understanding processes that control the water balance in mountainous regions is crucial and relevant for water resource management and can improve future hydrological model predictions.

1.2 Methods and data

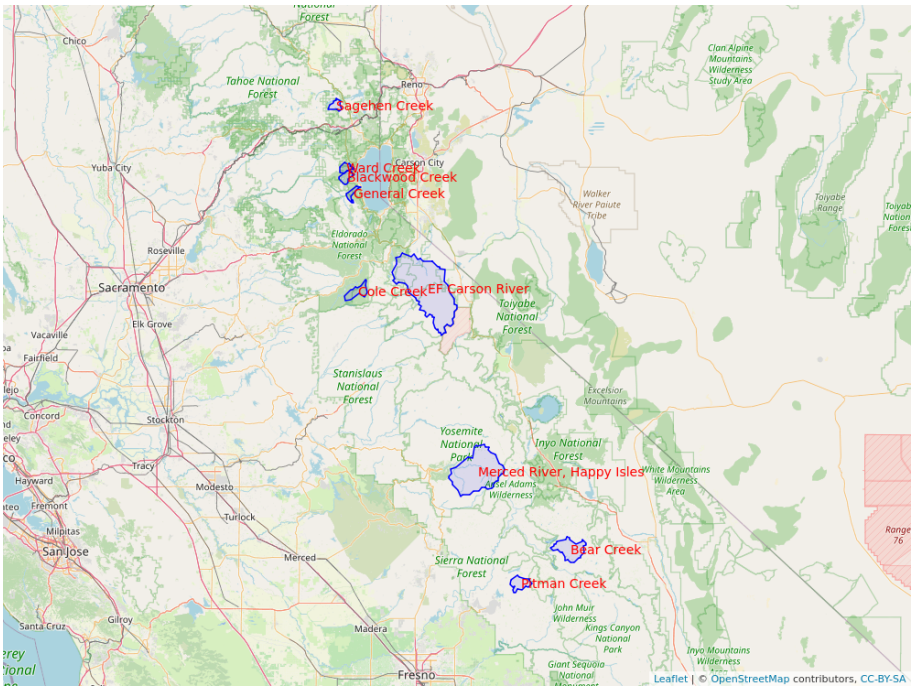
1.2.1 Study area

California's Sierra Nevada mountain range runs north-south, separating California's Central Valley from the Basin and Range province to the east. Due to the rain-shadowed effect, the eastern side receives less annual precipitation and is drier compared to western side. The southern part of the Sierra Nevada is generally higher, with elevations greater than 4,000 m at the crest, while the northern part is generally less than 3,000 m at peak elevations.

Table 1.1 Summary statistics of selected hydroclimatic, topographic, and vegetation characteristics for the 9 watersheds under study.

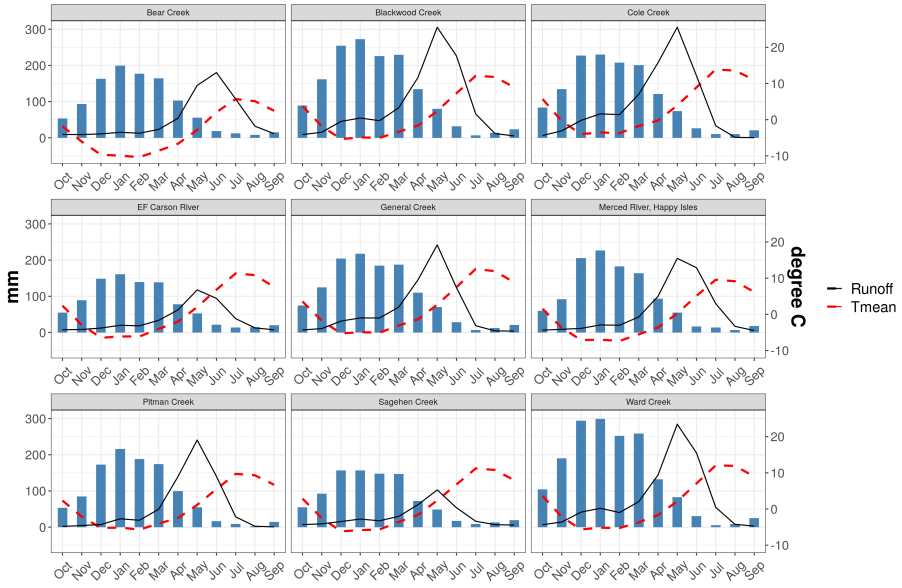
No.	STAID	Name	Area (km^2)	Mean elevation (m)	Annual P (mm)	Annual T ($°C$)	Runoff ratio	Percentage of forest	Major forest type
1	10336645	General Creek	19.59	2185.5	1201	5.76	0.55	78.38	Mixed conifer
2	10336660	Blackwood Creek	29.79	2216.39	1486	5.78	0.64	70.15	Red fir
3	10336676	Ward Creek	24.71	2218.42	1549	5.71	0.51	77.39	Mixed conifer
4	10343500	Sagehen Creek	27.60	2159.80	976	5.2	0.33	88.28	Mixed conifer
5	11237500	Pitman Creek	59.82	2437	1175	5.65	0.54	86.99	Mixed conifer
6	11230500	Bear Creek	135.53	3244.58	1149	0.43	0.56	26.4	Lodgepole pine
7	11315000	Cole Creek	54.01	2261.55	1410	7.10	0.74	55.16	Mixed conifer
8	10308200	EF Carson River	716.42	2410.54	980	4.82	0.43	41.20	Lodgepole pine
9	11266500	Merced River	467.98	2746.07	1198	6.08	0.58	42.02	Mixed conifer

Figure 1.2 Spatial map of nine watersheds in the study.



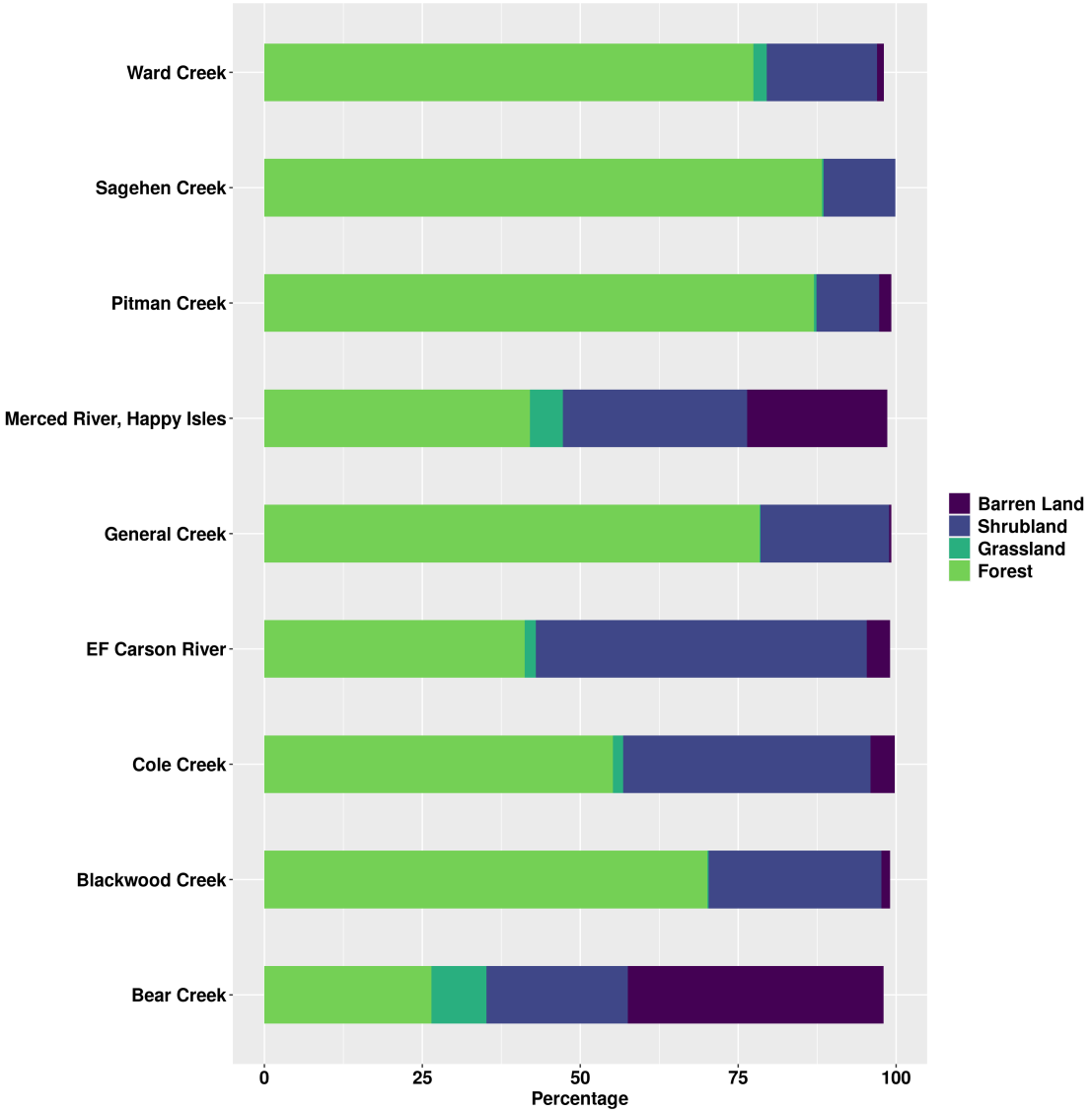
In this study, we consider 9 small and medium-sized mountainous watersheds that represent the variability in topography, vegetation coverage, and elevation gradients of the Sierra Nevada (Fig. 1.2). These include Bear Creek, Blackwood Creek, Cole Creek, East Folk (EF) Carson River, General Creek, Merced River at Happy Isles, Pitman Creek, Sagehen Creek, and Ward Creek. From north to south, Sagehen Creek, Ward Creek, Blackwood Creek, and General Creek are tributaries of the Truckee River systems and are small catchments located on the eastern slope of Sierra Nevada. Cole Creek is part of the North Fork Mokelumne River that flows west and contributes to Salt Springs Reservoir [Silverman, 2010]. EF Carson River and Merced River basins are larger catchments with drainage areas of approximately 700 km^2 and 500 km^2 respectively where the Carson River drains the rain-shadowed eastern slope of and the Merced River drains the wetter, western slopes [Dettinger et al., 2004]. The majority of the streamflow comes from the snowpack in these watersheds. Bear Creek and Pitman Creek are in the more south-central region of the Sierra Nevada and part of the San Joaquin River system. Their drainage areas are in the range 19.6 km^2-716.4 km^2 and ele-

Figure 1.3 Annual hydrographs for nine watersheds. Values reflect monthly average over the 30-year period between 1989 and 2018.



vation of 2159.8-3244.56 m. These watersheds have been referenced to have least-disturbed hydrologic condition and are free from current human influences like dams, diversions, and major land-use changes [Falcone, 2011]. They have been included in previous hydroclimatic studies that examined the long-term changes in discharge [Stewart et al., 2005, Peterson et al., 2005, Yang et al., 2018, Krogh et al., 2020] and forest and ecological monitoring [Podolak et al., 2015, Stevens et al., 2016, Loheide et al., 2009]. The 30-year climate normal (1980–2010) from PRISM indicate mean annual precipitation in the range of 980 - 1550 mm. Monthly precipitation, temperature, and runoff for individual watersheds are shown in Fig. 1.3. These watersheds have varying vegetation coverage and types with a combination of woody plants, shrubs, and meadows (Fig. 1.4, Table 1.1). Dominant tree species include mixed conifers such as lodgepole pine (*Pinus contorta*), Ponderosa pine (*Pinus ponderosa*), Jeffrey pine (*Pinus jeffreyi*), Douglas fir (*Pseudotsuga menziesii*), red fir (*Abies magnifica*), and incense cedar (*Calocedrus decurrens*) [North, 2012]. Chaparral and montane shrubs are also common at lower elevation foothills. Soils are either granitic or volcanic, which have origins from glaciers deposits.

Figure 1.4 Land-use classification for the 9 watersheds based on National Land Cover Database 2006 Classification.



1.2.2 Data

Monthly precipitation and temperature observations were obtained from the AN81m PRISM dataset [Daly et al., 2008]. This gridded dataset has a resolution of 4-km, covers the entire continental US from January 1981 to present, and is continuously updated every 6 months. Catchment-average time series were constructed by computing the arithmetic mean for precipitation and temperature values of all grid points that fall within watershed polygon. 31 years of record between 1988 and 2018 were considered in the study. These time series were later used to compute Standardized Precipitation Evapotranspiration Index (SPEI) values (described in Sect. 1.2.3).

Streamflow data was obtained through the USGS National Water Information System (NWIS) (`https://waterdata.usgs.gov/nwis/sw`) for the 9 watersheds. The observation records are mostly complete with less than 5 days of missing data in a given year. Daily discharge were aggregated to obtain monthly and annual values for the water years in the period 1989-2018. These values were then normalized by catchment area.

Two vegetation indices are currently produced from the the Moderate Resolution Imaging Spectrometer (MODIS) sensor, Enhanced Vegetation Index (EVI) and Normalized Difference Vegetation Index (NDVI). For our study, we chose EVI over the more commonly used NDVI because EVI has been reported to be less sensitive to soil and atmospheric effects than NDVI and remains sensitive to increases in vegetation density beyond where NDVI becomes saturated [Huete et al., 2002, Waring et al., 2006]. EVI values were obtained from MODIS at 500-m spatial resolution and 16-day compositing period from the MOD13Q1 dataset [Didan, 2015]. We focus on the vegetation activity during the growing season (May-Sep) as these are times when demand for water is high. To ensure quality of the pixels and reduce the possible bias in the resulting EVI values, we removed cloud and snow contaminated pixels before computing watershed average. As vegetation EVI typically are in the range of 0-1, we removed pixel with values outside of this range to minimize the possibility of added noise in the data. All remote sensing images were processed using Google Earth Engine. Due to the

availability of the data, 18 years of EVI between 2001 and 2018 water years are used in our study. We also excluded the month of April in our analysis due to large number of pixels are still covered with snow. 16-day values were linearly interpolated into daily time series and subsequently aggregated to obtain monthly EVI value for the months of May-Sep for each watershed.

Total ET was retrieved from MOD16A2 Version 6 [Mu et al., 2013] at 8-day temporal and 500-m pixel resolutions for each watershed for water years between 2001 and 2018. A water year begins on October 1st of the previous year and ends on September 30th. MODIS ET was processed in a similar procedure to EVI data. The algorithm used for the MOD16A2 data product collection is based on the logic of the Penman-Monteith equation, which includes inputs of daily meteorological reanalysis data along with MODIS remotely sensed data products such as vegetation property dynamics, albedo, and land cover. Cloud-contaminated pixels were included from our analysis. 8-day ET time series at each watershed was linearly interpolated to obtain monthly values.

1.2.3 Quantifying droughts using standardized precipitation evapotranspiration index (SPEI)

SPEI was first proposed by Vicente-Serrano et al. [2010] as an improved drought index of SPI and has since been widely used in many studies to capture drought propogations and reconstructions [Allen et al., 2011, Li et al., 2012, Labudová et al., 2017, Manzano et al., 2019]. An advantage of SPEI over SPI is that it accounts for the effect of temperature in the drought development and climate water balance, defined as the difference between precipitation and potential evapotranspiration (PET), and therefore provides a more reliable measure of drought severity than only considering precipitation [Beguería et al., 2014]. This difference is used as the input in the computation of SPEI. Thornthwaite equation [Thornthwaite, 1948] is used for estimation of PET. Alternatively, Penman–Monteith method (PM) can be used to estimate PET but often requires more extensive data (solar radiation, temperature,

wind speed and relative humidity) and more parameters, and long-term records of these variables are not always available. Previous studies have demonstrated that the two methods often yield comparable results. As the log-logistic distribution has been shown to provide better fit than other distributions [Vicente-Serrano et al., 2010, Beguería et al., 2014], we adopted this method to obtain SPEI series in standardized z units. We used R package `SPEI` [Beguería et al., 2013] to compute SPEI. Watersheds are considered to experience drought when n-month SPEI is less than -0.5 [Yang et al., 2016]. The lower the SPEI value, the more severe the drought condition (Table 1.2).

SPEI values can be calculated for different time scales using the cumulative water balance over the previous n-months. To understand the timescale at which SPEI affects vegetation activity, we calculated 1-, 3-, 6-, and 12-month SPEI for each watershed and obtained the Pearson's correlation between monthly EVI and each of these SPEI time series for the growing season months (May-Sep) in the period 2001-2018. We consider a significance threshold of $\alpha < 0.01$. The annual SPEI is the 12-month SPEI value for the month of September and represents accumulated water balance at the end of the water year [Feng et al., 2020]. Drought develops gradually and water deficit can accumulate over a long period of time, making it difficult to pinpoint drought and quantify their duration, magnitude in time and space with a single variable or metric [Mishra and Singh, 2011, Vicente-Serrano et al., 2010]. In order to access the impact of drought on the catchment's annual water partitioning, we characterized a year as a drought year if its annual SPEI was less than -0.5.

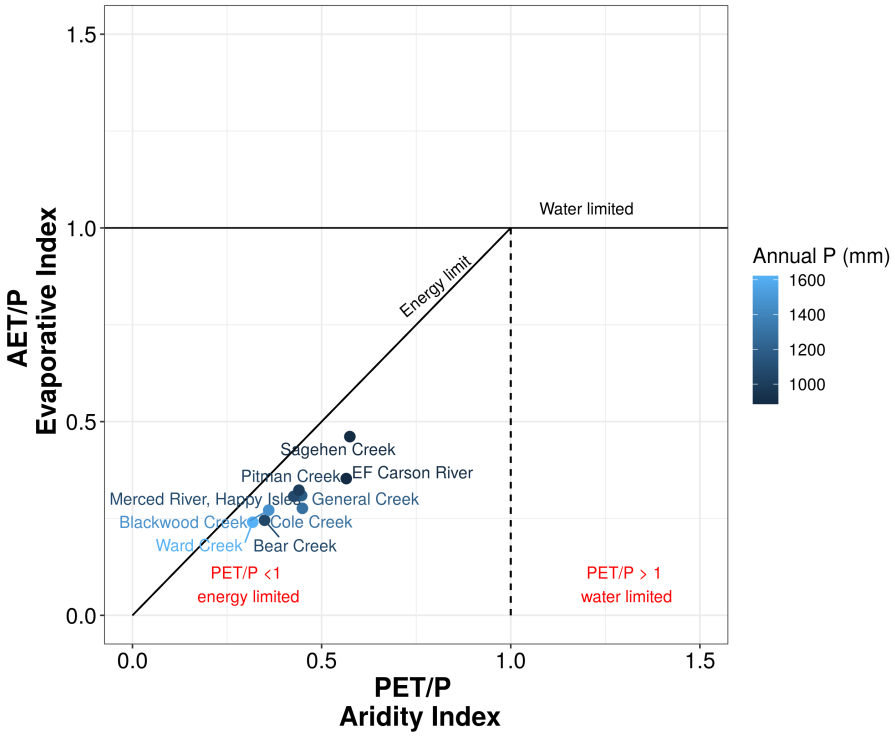
Table 1.2 Drought level classifications based on SPEI [Yao et al., 2018].

SPEI values	Categories
SPEI > -0.5	No drought
$-1.0 <$ SPEI ≤ -0.5	Mild drought
$-1.5 <$ SPEI ≤ -1.0	Moderate drought
$2.0 <$ SPEI ≤ -1.5	Severe drought
SPEI $\leq$ -2.0	Extreme drought

1.2.4 Assessing drought impacts of catchment water balances using the Budyko framework

The Budyko framework, which relates the dependence of actual evapotranspiration on energy availability represented by the potential evaporation and water availability represented by the precipitation, has been successfully used to understand and predict the climatic and landscape controls on long-term water balance [Budyko, 1974]. Recent studies have shown success in extending the framework to investigate between climate, vegetation in the hydrologic cycle [Donohue et al., 2007, Li et al., 2012] and study interannual variability in water partitioning at individual catchments and [Carmona et al., 2014, Koster and Suarez, 1999, Yang et al., 2007, Cheng et al., 2011]. Figure 1.5 shows the relationship of water pools in the 9 catchments plotted on the hypothetical Budyko framework.

Figure 1.5 Long-term annual water balance represented by theoretical Budyko framework.



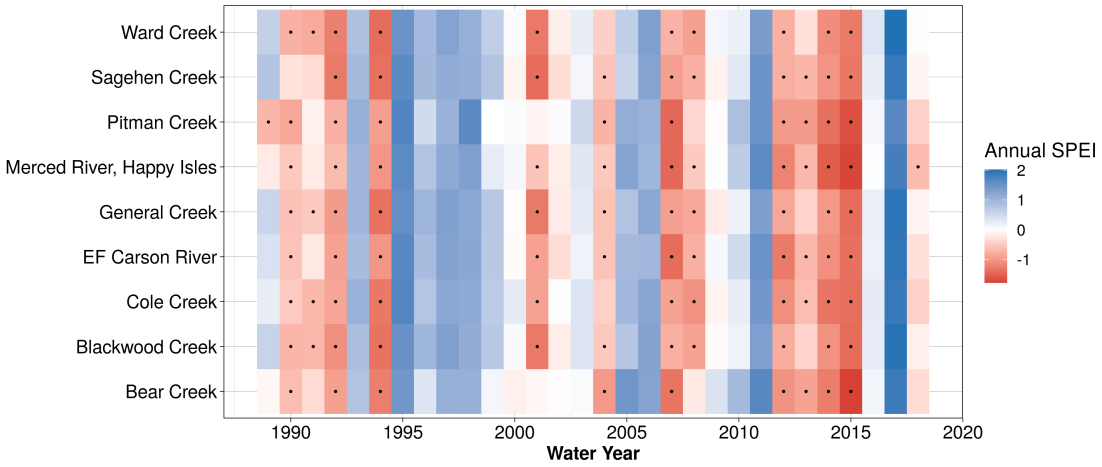
1.3 Results and discussion

1.3.1 Characterizing annual drought using 12-month SPEI

As shown in Fig. 1.6, the 12-month SPEI time series capture the development of the major multi-year drought periods in California in the last 40 years including the 1988–1992, 2007–2009, and more recently 2012–2015 droughts [He et al., 2017] across the watersheds. The severity of the 2014/15 drought, which broke many historical records [Funk et al., 2014], is also well reflected with SPEI reaching well below -2. At the annual scale, we identified between 9 and 12 drought years at individual watersheds (Fig. 1.3, Table 1.3).

Figure 1.6 12-month SPEI time series for 30 years between 1989 and 2018.

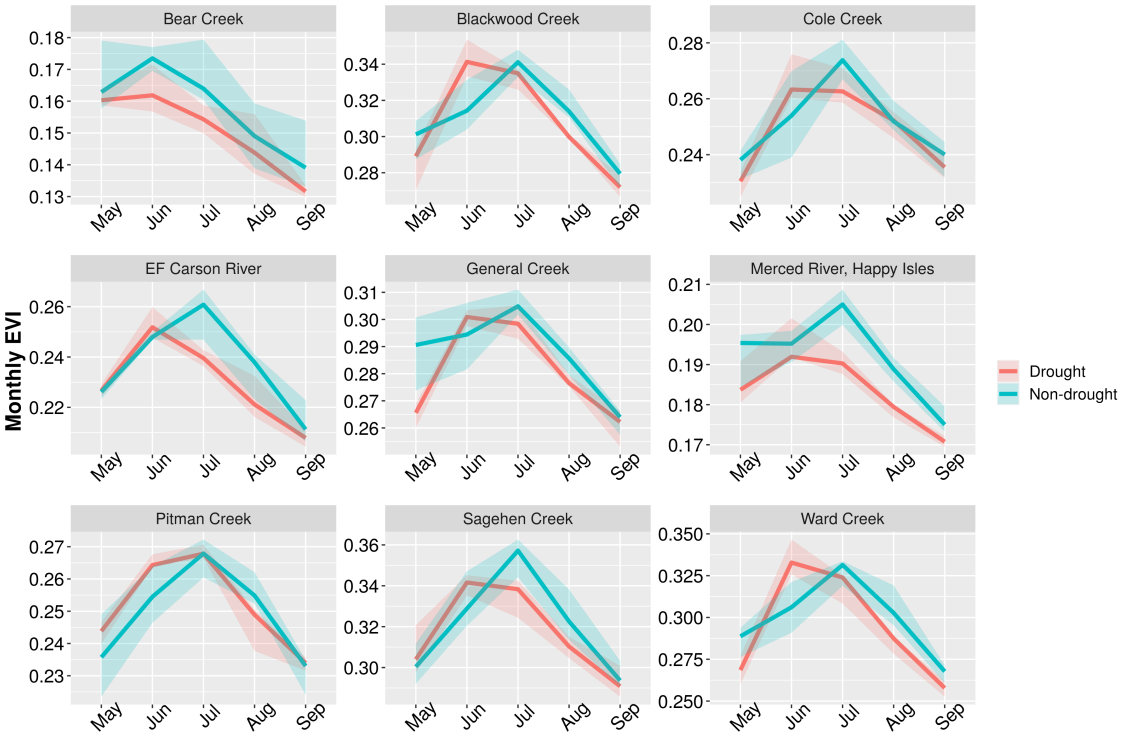
Figure 1.7 Annual SPEI for the period 1989-2018. Annual SPEI is calculated as the cumulative water balance for the 12-month period in the water year. Drought years with annual SPEI < -0.5 are indicated by black dots.



1.3.2 Trends in vegetation response to drought at annual timescale

Differences in geography and vegetation composition can affect vegetation's response to droughts (Fig. 1.8). There is a general pattern of decrease in vegetation activities during drought years in Bear Creek, EF Carson River, and Merced River, which locate in the south-central part of the Sierra Nevada. We also observe an apparent trend in earlier vegetation greening and peak in EVI among 8 out of 9 watersheds (except for Bear Creek) in drought years. As transpiration in conifer forests in the Sierra Nevada is broadly temperature-limited in winter and water-limited in late summer [Royce and Barbour, 2001], earlier onset of snowmelt during drought years likely provides soil moisture and warmer air temperature conditions can potentially drive photosynthesis and support vegetation growth at the beginning of the growing season. Late into the summer months (Jul-Aug), the melt water supply diminishes and drought stress likely resulted in lesser vegetation activity indicated by lower EVI across all 9 watersheds. This trade-off in longer growing yet lower mountain forest productivity under drought-induced snowpack reduction may be a scenario in future climate [Knowles et al., 2018, Goulden and Bales, 2014]. The summer droughts therefore will likely have detrimental effects on summer ecosystems in this region.

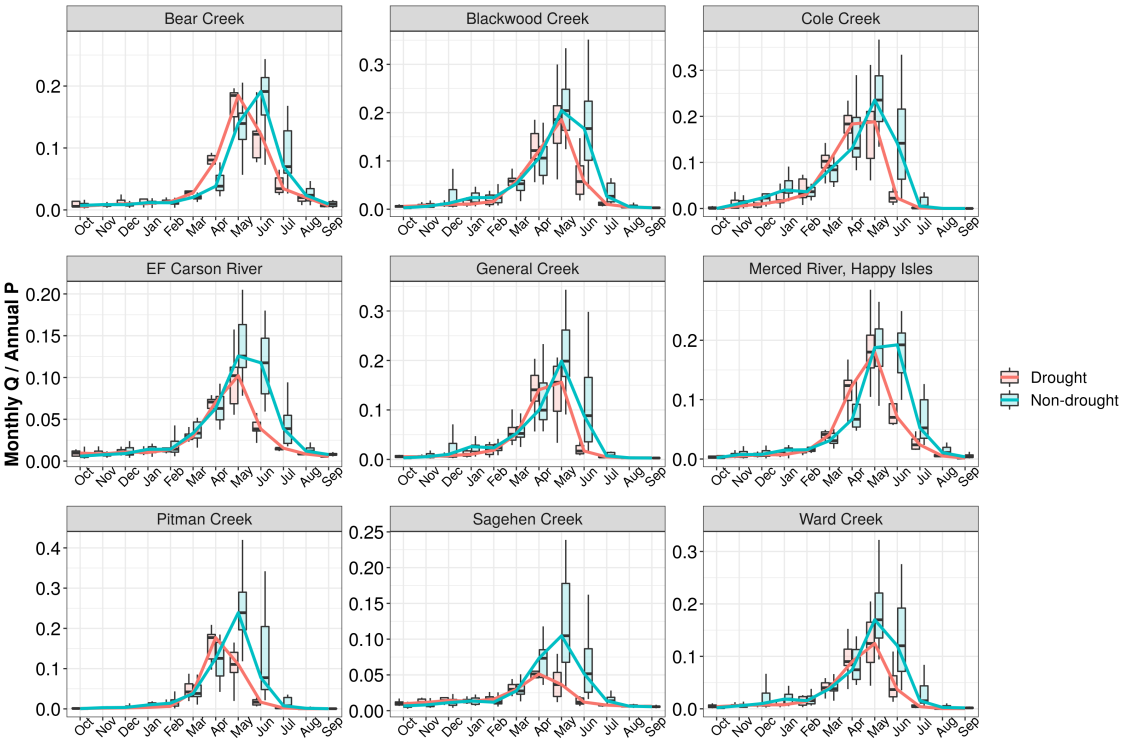
Figure 1.8 Growing season EVI (May-Sep) during drought and non-drought conditions for water years between 2001 and 2018. Drought years were identified using the annual SPEI value. Solid lines indicate the median values and shaded areas represent the Q_1 and Q_3 of the monthly values.



1.3.3 Impact of drought on precipitation-runoff relationship and catchment water balance

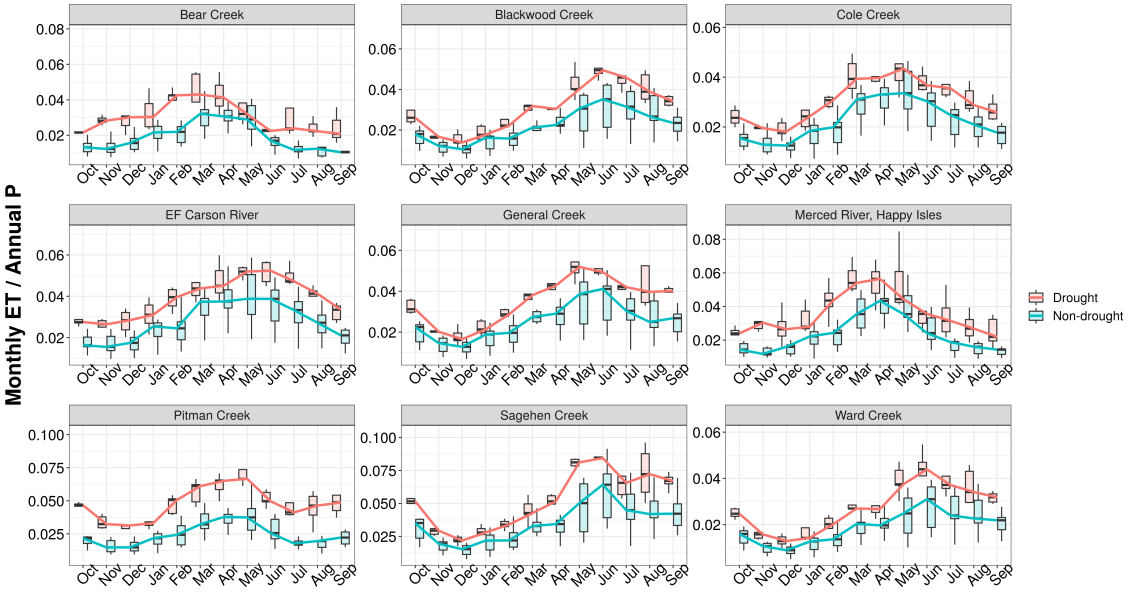
We observe a shift in precipitation-ruff relationship at 9 watersheds (Table 1.3) where there is a decrease in mean runoff ratio in drought years compared to non-drought years. The magnitude of the shift varies among the watersheds where the runoff ratio of Bear Creek is relatively resilient to droughts compared to the other watersheds. This is our first line of evidence that supports our hypothesis that the P-Q relationship changes under drought conditions. Monthly runoff (normalized by annual precipitation) shows a shift in earlier peak flow timing in a number of watersheds (Bear Creek, Pitman Creek, and Sagehen Creek) and highlights the impacts of droughts on summer water availability (Fig. 1.9). Noticeably, there is significantly less runoff in the months of May and June in drought years. This reduction

Figure 1.9 Fraction of monthly runoff to annual precipitation for the period 2001-2018 for non-drought and drought years.



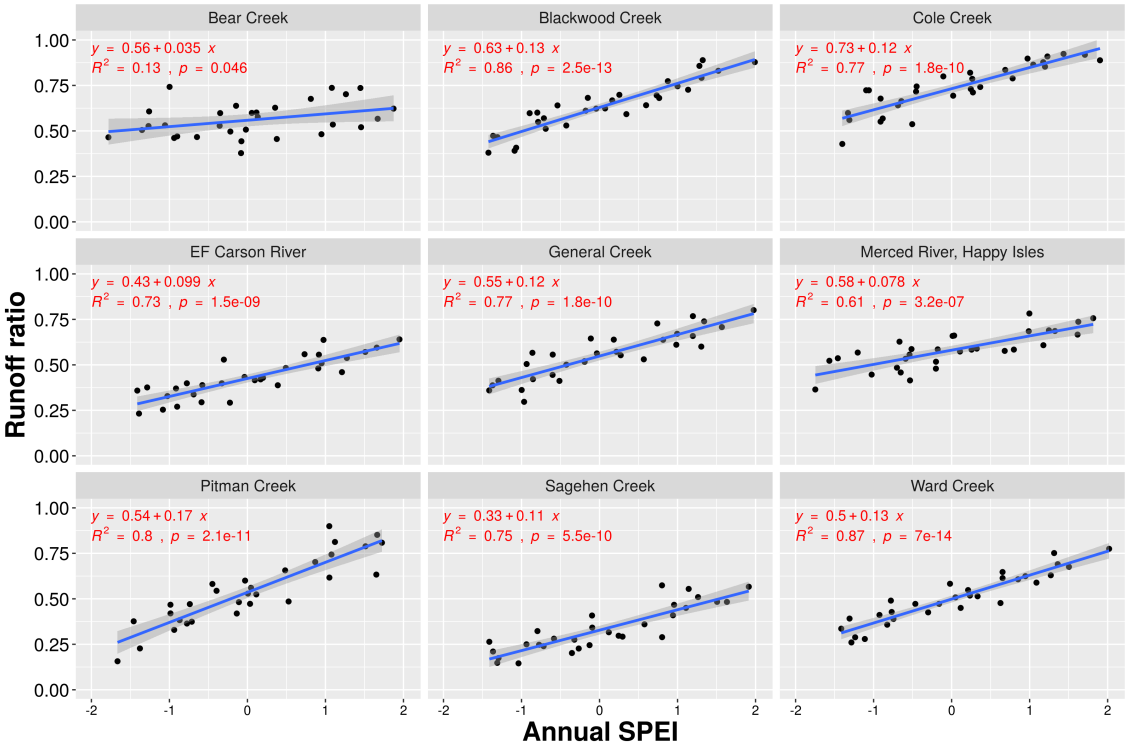
have important implications on human and ecosystem water needs as it occurs during the beginning of the peak water demand and summer growing season.

Figure 1.10 Fraction of monthly ET to annual precipitation in the period 2001-2018 for non-drought and drought years.



While elevated temperatures associated with drought conditions can result in drier, hotter atmospheric conditions favorable to drive ET, climate model simulations suggests that the early spring greening can result in soil moisture deficits and decrease surface runoff persisting well into the summer months [Lian et al., 2020]. It is evident that there is an increase in ET as a fraction of available P during drought years in the watersheds under study (Fig. 1.10). We also see that the large decrease in runoff in the month of June coincides with the peaks in EVI observed in Fig. 1.8 at Blackwood Creek, Ward Creek, and EF Carson River. Because runoff is influenced not only by ET, but by many other factors, it is difficult to conclude whether the observed reduction in Q/P fraction during droughts can be entirely attributed to earlier vegetation activity and water uptake. Previous studies have suggested that growing season in Sierra Nevada is restricted to a brief window in late spring and early summer when air temperatures are warm enough for photosynthesis and melt-supplied soil moisture remains plentiful [Dettinger et al., 2004, Goulden and Bales, 2014]. It makes sense that during droughts, there is higher priority allocation of water to replenish the soil moisture deficit and thus shifting the fraction of local P partitioned to Q.

Figure 1.11 Annual runoff ratio plotted against annual SPEI for 1989-2019. Best-fit line was obtained using simple linear regression.



We further observe this trend in Fig. 1.11 where there is a significant positive linear relationship between annual SPEI and annual runoff ratio in all watersheds ($p < 0.05$). In other words, drought severity may have a direct effect on the magnitude of the partitioning shift. However, such relationship varies among the watershed where annual SPEI explains 87% of variations in runoff ratio in Ward Creek but only 13% in Bear Creek. This highlights the fact that sensitivity of catchment function to drought likely differs among watersheds and is an interplay of both climate and catchment properties [Renner et al., 2012, Veettil et al., 2018], which, in our study, possibly includes vegetation cover and elevation among others (Appendix A.2, Fig. A.2).Possibly, the generation of high-elevation runoff such as that in Bear Creek, which is more resilient to increases in PET due to overall lower temperatures, can help mitigate runoff losses [Goulden et al., 2012]. Given the orographic effect of the Sierra Nevada, watersheds at high elevations may also be less susceptible to decreases in precipitation. These are consistent with previously studies conducted in the Sierra Nevada

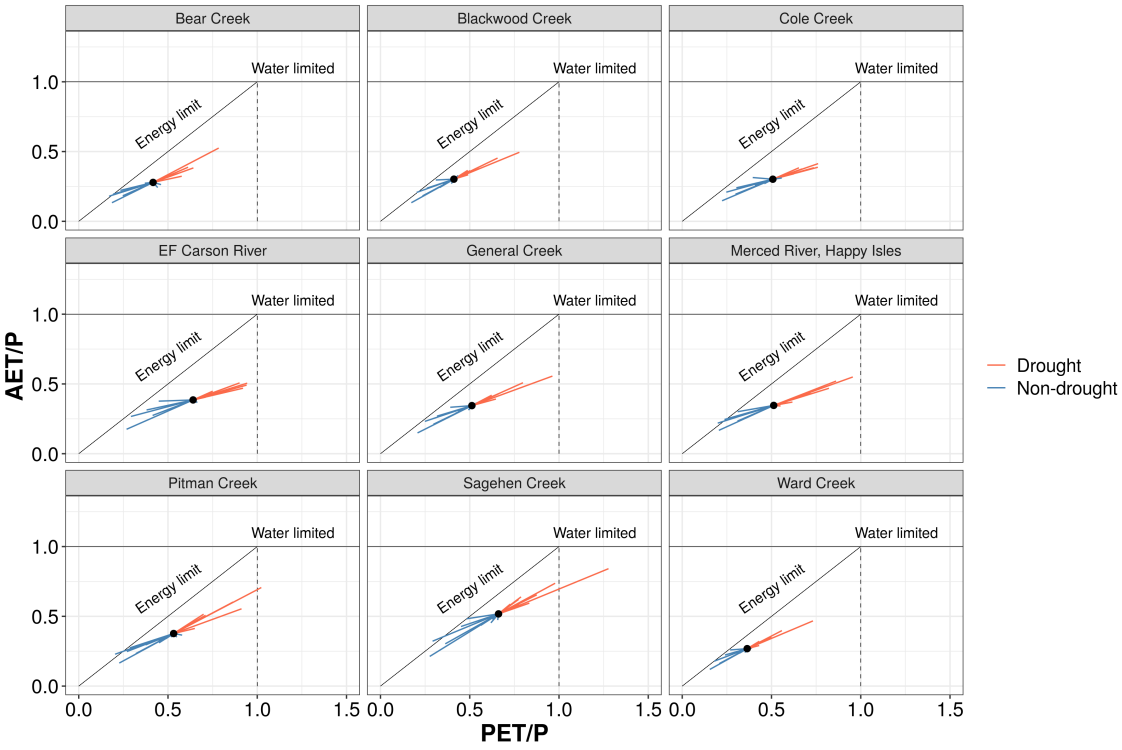
[Avanzi et al., 2020]. Prior studies, including Saft et al. [2015], Potter et al. [2011], and Avanzi et al. [2020], used linear regression-based approaches to identify factors associated with drought-induced changes to the precipitation–runoff relationship in Australia and California. A better and more robust understanding and prediction on the impacts of drought on P-Q relationship will benefit from consideration both only meteorological characteristics and catchment variables.

1.3.4 Catchment balance under the Budyko framework

It is expected that under natural climate variability, individual catchments can move in all directions through Budyko "space" [Van der Velde et al., 2014]. In Fig. 1.12, in 8 watersheds exept Bear Creek, there is a tendency of the catchment in Budyko space moving towards the theoretical water (horizontal) and energy (vertical) limits during drought years. In these years, the larger evaporative demand and increase in forest ET (higher AET/P) associated with warmer temperature (higher PET/P) allows for a higher fraction of water turned into latent heat. On the other hand, forests may have reacted to increased temperature by increasing their ET. The combined effect likely at the expense of runoff (lower Q/P). While Loarie et al. [2009] pointed out that the temperature increase associated with climate change is relatively slow in mountainous biomes including subtropical coniferous forests such as the Sierra Nevada, we suspect that the combined effect of lower precipitation and higher temperature associated with drought conditions may drive the eco-hydrology change in this region.

The Budyko framework, however, considers allocation of water relative to the aridity index, a combination of two major water balance drivers (PET and precipitation), rather than precipitation alone.

Figure 1.12 Interannual variability of catchment water balance under the Budyko framework for water years in period 2001-2018. Black dot is the catchment centroid and indicates 18-year average.



The Budyko framework governs available water partitioning by physical behavior under limit conditions (when the aridity index is zero, all water goes to runoff; when the aridity index is one, all water goes to ET) and allows for the possibility that even expected and predictable water balance changes during drought may be nonlinear and that some shifts observed in other studies may be the result of factors that are not captured in a two-dimensional precipitation– runoff plane [Maurer et al., 2022]. The Budyko framework can be leveraged to model more predictable regime versus less predictable partitioning shifts during droughts. Further research is needed to analyze a more comprehensive set of feedback mechanisms and compare the Budyko framework to other nonlinear approaches.

1.3.5 Limitations

We acknowledge that there are uncertainties associated with MODIS ET estimates and sources of error have been linked to LAI and meterological data quality, sensor calibration, and atmospheric corrections [Demarty et al., 2007, Mu et al., 2011]. While the ground data from the eddy covariance flux towers provide the best ET estimates, these are not spatially consistent and scaling from tower to watershed scale poses as a challenge due to the heterogeneous landscape, particularly among larger basins such as Merced River and EF Carson River in our study. Evaluation results of MOD16 ET over the conterminous United States using point and gridded FLUXNET and water balance ET by Velpuri et al. [2013] indicate that MOD16 ET products effectively reproduced basin scale ET response (up to 25% uncertainty) compared to CONUS-wide point-based ET response (up to 50–60% uncertainty), illustrating the reliability of MODIS ET products for basin-scale ET estimation.

1.4 Conclusions

In this study, we jointly explored the impacts of drought on the ecohydrological processes at 9 small and medium-sized watersheds in the Sierra Nevada. We found a general trend in earlier peak in vegetation activity during drought years compared to non-drought years watersheds. A similar trend is observed in peak runoff timing. This is likely due to warmer temperature. Significant, positive linear relationship between annual SPEI and runoff ratio suggests that drought conditions may affect runoff generation processes and cause changes in the P-Q relationship. Catchment properties such as elevation and vegetation cover can affect streamflow elasticity or catchment resilience to disturbance.

As climate change is expected to increase the odds of worsening drought in many parts of the United States, our study shows that both ecological and hydrological processes in the Sierra Nevada watersheds may be vulnerable to drought conditions. Future hydrologic modeling research could identify threshold responses in watersheds to changes in precipitation deficit and temperature associated with drought, which can provide relevant insights

for water and natural resource management.

Table 1.3 Mean runoff ratio for drought and non-drought years for water years 1987-2018.

Watersheds	Non-drought		Drought		Percentage difference
	Mean runoff ratio	Number of years	Mean runoff ratio	Number of years	
Bear Creek	0.58	21	0.50	9	-13 %
Blackwood Creek	0.70	19	0.51	11	-27 %
Cole Creek	0.82	19	0.60	11	-26 %
EF Carson River	0.48	19	0.34	11	-29 %
General Creek	0.64	19	0.41	11	-36 %
Merced River	0.63	18	0.53	12	-16 %
Pitman Creek	0.61	20	0.38	10	-37 %
Sagehen Creek	0.38	20	0.24	10	-36 %
Ward Creek	0.59	20	0.37	10	-37 %

CHAPTER 2

EVALUATION OF RANDOM FORESTS FOR SHORT-TERM DAILY STREAMFLOW FORECASTING IN RAINFALL AND SNOWMELT-DRIVEN WATERSHEDS

2.1 Introduction

Nearly all aspects of water resource management, risk assessment, and early-warning systems for floods rely on accurate streamflow forecast. Yet streamflow forecasting remains a challenging task due to the dynamic nature of runoff in response to spatial and temporal variability in rainfall and catchment characteristics. Therefore, development of skillful and robust streamflow models is an active area of study in hydrology and related engineering disciplines.

While physical models remain a common and powerful tool for predicting streamflow, machine learning (ML) models are gaining popularity due to some of their unique qualities and potential advantages. Compared with the often labor-intensive and computationally expensive task of parameterizing in physical model [Tolson and Shoemaker, 2007, Boyle et al., 2000], ML models are data-driven and can identify patterns in the input-output relationship without explicit knowledge of the physical processes and onerous computational demand. To make up for their limited ability to provide interpretation of the underlying mechanisms, ML models often require less calibration data than physical models, have demonstrated high accuracy in their predictive performance, are computationally efficient, and can be used in real-time forecasting [Adamowski, 2008, Mosavi et al., 2018]. ML models are particularly useful when accurate prediction is the central inferential goal [Dibike and Solomatine, 2001], whereas conceptual rainfall-runoff model can provide a better understanding of hydrologic phenomena and catchment yields and responses [Sitterson et al., 2018]. Artificial neural networks (ANN), neuro-fuzzy (a combination of ANNs and fuzzy logic), support vector machine (SVM), and decision trees (DT) are reported to be among the most popular and effective

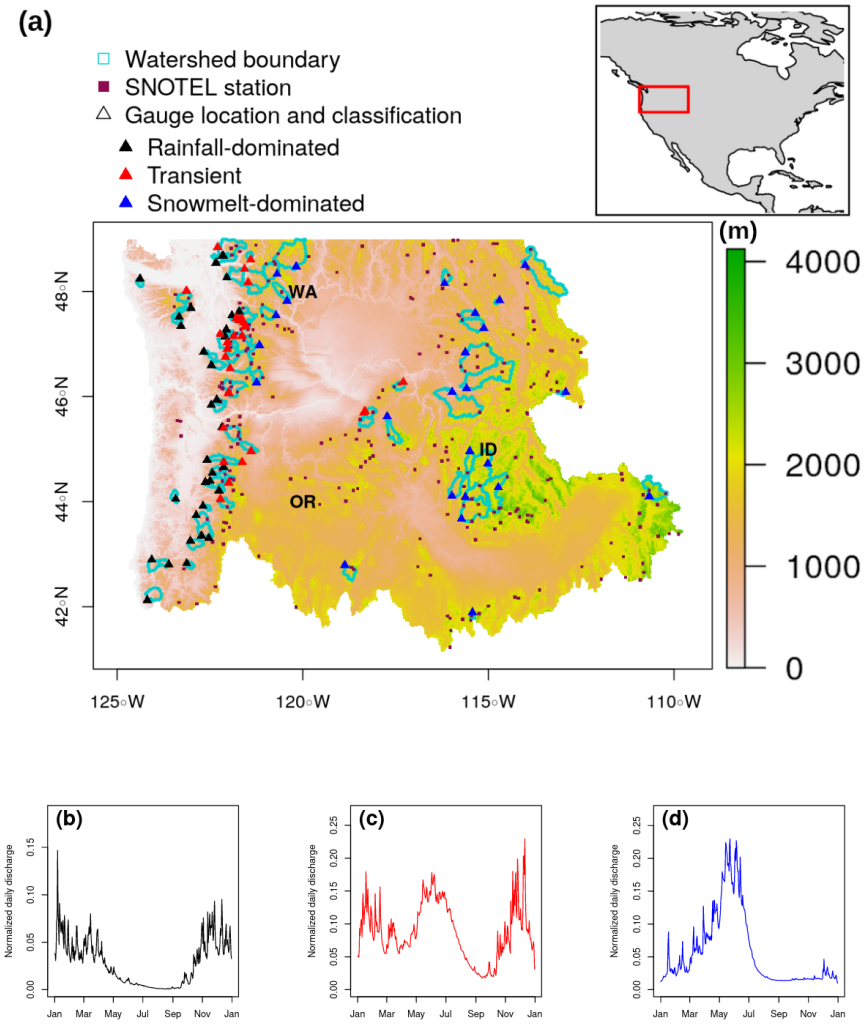
for both short-term and long-term flood forecast [Mosavi et al., 2018]. For example, Dawson et al. [2006] provided flood risk estimation at ungauged sites using ANN at catchments across the United Kingdom. Rasouli et al. [2012] predicted streamflow at lead times of 1-7 days with local observations and climate indices using three ML methods: Bayesian neural network (BNN), SVM, and Gaussian process (GP). They found BNN outperformed multiple linear regression (MLR) as well as the other two ML models. Their study also found models trained using climate indices yielded improved longer lead time forecasts (e.g., 5–7 days). Tongal and Booij [2018] forecasted daily streamflow in four rivers in the United States with SVR, ANN, and RF coupled with a baseflow separation method (i.e., separating the two different components of streamflow into baseflow and surface flow). Obringer and Nateghi [2018] compared eight parametric, semi-parametric, and non-parametric ML algorithms to forecast urban reservoir levels in Atlanta, Georgia. Their results showed random forests (RF) yielded the most accurate forecasts.

Despite the promising results reported in existing literature, most ML streamflow forecast applications are limited to watersheds where rainfall is the major contributor. In many settings, particularly non-arid mountainous regions in Western USA, a combination of rainfall and spring snowmelt can drive streamflow [Johnstone, 2011, Knowles et al., 2007]. The amount of snow accumulation and its contribution to discharge also vary among the watersheds [Knowles et al., 2006]. Both watershed-scale hydrologic and statistical models have been used to assess the current and future stream hydrology and associated flood risks [Salathé Jr et al., 2014, Wenger et al., 2010, Tohver et al., 2014, Pagano et al., 2009]. Safeeq et al. [2014] simulated streamflows in 217 watersheds at annual and seasonal time scales using the Variable Infiltration Capacity (VIC) model at 1/16° and 1/20° spatial resolutions. The study found that the model was able to capture the hydrologic behavior of the studied watersheds with a reasonable accuracy. Yet the authors recommend careful site-specific model calibration, using not only streamflow but also snow water equivalent (SWE) data, would be expected to improve model performance and reduce model bias. Pagano et al. [2009] applied

Z-score regression to daily SWE from Snow Telemetry (SNOTEL) stations and year-to-date precipitation data to predict seasonal streamflow volume in unregulated streams in Western US. The authors reported the skill of these forecasts is comparable to the official published outlooks. A natural question is whether ML models can produce comparable performance in these watersheds where streamflow contributions come from a mixture of snowmelt and rainfall, as well as where snowmelt dominates sources. Considering the prominent role of snowpack in water management and contribution of rapid snowmelt in flood events, such question is worth exploring. To this end, we evaluate the potential of RF in making short-term streamflow forecast at 1-day lead time across 86 watersheds in the Pacific Northwest Hydrologic Region (Fig. 2.1). The U.S. Geological Survey [2020] defines this region as hydrologic region 17 or HUC 17. HUC-17 consists of sub-basins and watersheds of the Columbia River that span varying hydrologic regimes. The selected watersheds have long-term record of unregulated streamflow and different streamflow contributions of rainfall and snowmelt. Drainage basin factors such as topography, vegetation, and soil can affect the response time and mechanisms of runoff [Dingman, 2015]. Few studies attempted to account for or report these effects on models' performance. Without such consideration, it is difficult to determine if a data-driven model can be generalized to watersheds not included in the given study. Therefore, our objectives are (1) to examine and compare the performance of RF in a number of watersheds across hydrologic regimes and (2) to explore the role of catchment characteristics in model performance that are overlooked in previous studies.

In practice, RF can be trained to forecast streamflow at various timescales, depending on the input variables provided. Rasouli et al. [2012] forecasted streamflow at 1-7 day lead times using three ML models and data from combinations of climate indices and local meteo-hydrologic observations. The authors concluded that models with local observations as predictors were generally best at shorter lead times while models with local observations plus climate indices were best at longer lead times of 5–7 days. Also, the skillfulness of all three models decreased with increasing lead times. In our study, we focused on 1-day lead

Figure 2.1 (a) Elevation (m) shading map showing the Pacific Northwest Hydrologic Unit, 86 selected stream gauges (triangles), and their drainage area (cyan delineation lines), and SNOTEL stations (brown squares). Examples of annual hydrographs of (b) rainfall-dominated, (c) transient, and (d) snowmelt-dominated watersheds. Figures (b-d) are based on 2009-2018 daily flow data at three sites 12043300 (48.2° N, 124.4° W), 12048000 (48° N, 123.1° W), and 10396000 (42.7° N, 118.9° W).

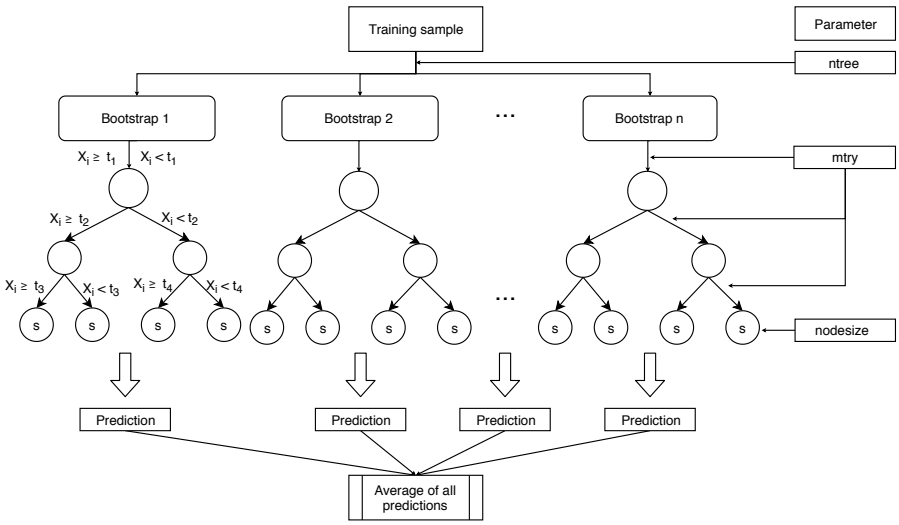


time forecasting and therefore did not include long-term climate information. At longer lead times, changes in weather conditions would likely exert much greater control on runoff and the performance of the model.

We select RF to forecast streamflow for two reasons. First, RF has been referenced to deliver high performance in short-term streamflow forecasts [Mosavi et al., 2018, Papacharalampous and Tyralis, 2018, Li et al., 2019, Shortridge et al., 2016], making it a good candidate for our study. Second, RF allows for some level of interpretability. This is delivered through two measures of predictive contribution of variables: mean decrease in accuracy (MDA) and mean decrease in node impuritiy (MDI). These two measures have been widely used as means for variable selection in classification and regression studies in bioinformatics [Chen and Ishwaran, 2012], remote sensing classification [Pal, 2005], and flood hazard risk assessment [Wang et al., 2015]. The interpretability of a ML model, however, can be a controversial subject and remains an active area of study [Ribeiro et al., 2016, Carvalho et al., 2019]. Both model-agnostic,such as permutation-based feature importance [Breiman, 2001], and model-specific, such as gini-based for RF [Breiman et al., 1984] and gradient-based for ANNs [Shrikumar et al., 2017], interpretation methods can provide useful insights into how the ML models make their predictions. While the referred interpretability does not directly translate to interpretation of the physical processes, it can provide insight into relationships among predictors and streamflow response.

The remainder of the paper is arranged as follows. Section 2.2 provides a brief introduction to RF, relevant parameters and selected evaluation criteria. Section 2.23 describes the study area, datasets, and predictor selection. Results and discussion are given in Section 2.4 along with limitations and recommendation for future research. A summary and indication of future work are is provided in Section 2.5.

Figure 2.2 Structure of a RF and relevant parameters.



2.2 Methods and data

2.2.1 Random forests

Proposed by Breiman [2001], RF is a supervised, non-parametric algorithm within the decision tree family that comprises an ensemble of decorrelated trees to yield prediction for classification and regression tasks. Non-parametric methods such as RF do not assume any particular family for the distribution of the data [Altman and Bland, 1999]. Since a single decision tree can produce high variance and is prone to noise [James et al., 2013], RF addresses this limitation by generating multiple trees where each tree is built on a bootstrapped sample of the training data (Fig. 2.2, Algorithm 2.1). Each time a binary split is made in a tree (also known as split node), a random subset of predictors (without replacement) from the full set of predictor variables is considered (Fig. 2.2). One predictor from these candidates is used to make the split where the expected sum variances of the response variable in the two resulting nodes is minimized (Algorithm 2.1, Step 3). The randomization process in generating the subset of the features prevents one or more particularly strong predictor from getting repeatedly chosen at each split, resulting in highly correlated trees [Breiman,

2001]. After all the trees are grown, the forests make prediction on a new data point by having all trees run through the predictors. In the end, the forests cast a majority vote on a label class for classification task or produce a value for regression task by averaging all predictions. Breiman [2001] provided full details on RF and its merit. The `randomForest` package in R developed by Liaw et al. [2002] was used for model training and validation in our study. The step-by-step of building a regression RF follows:

Algorithm 2.1 Building a regression RF

Step 1: n bootstrap samples are drawn from training set, each has the same size as the training sample. This is also known as `ntree` or number of trees in the forest.
Step 2: At each binary node split, a subset of `mtry` predictors, X_i, is randomly selected from p predictor space, Ω_p, that results in $X_i \in \Omega_p$ for {i $\in$ 1,..., `mtry`}, `mtry` $< p$.
Step 3: The single best combination of predictor X_i among X predictor variables and threshold t is selected to split the observations, y_j, into binary regions $R_1 = \{\ y_j | X_i < t\ \}$ and $R_2 = \{\ y_j | X_i \geq t\ \}$ that minimize:

$$\sum_{j:y_j \in R_1} (y_j - \hat{y}_{R_1})^2 + \sum_{j:y_j \in R_2} (y_j - \hat{y}_{R_2})^2 \tag{2.1}$$

where $\hat{y}_{R_1}$ is the mean of observations in R_1 and $\hat{y}_{R_2}$ is the mean of observations in R_2.
Step 4: Repeat step 2-3 until all terminal region contains less than `nodesize` observations.

Due to sampling with replacement, some observations may not be selected during the bootstrap. These are referred to as out-of-bag or OOB and used to estimate the error of the tree on unseen data. It has been estimated that approximately 37% of samples constitute OOB data [Huang and Boutros, 2016]. An average OOB error is calculated for each subsequently added tree to provide an estimate of the performance gain. The OOB error can be particularly sensitive to the number of random predictors used at each split `mtry` and number of trees `ntree` [Huang and Boutros, 2016]. Generally, the predictive performance improves (or OOB error decreases) as `ntree` increases. However, recent research has shown that depending on the dataset, there is a limit for number of trees where additional growing does not improve performance [Oshiro et al., 2012]. It has been advised that `mtry` is set to no larger than 1/3 of total number of predictors for optimal regression prediction [Liaw

et al., 2002], which is also the default value in `randomForest` function in R and widely adopted in literature. Nevertheless, Huang and Boutros [2016] found that this value is dataset-dependent and could be tuned to improve the performance of RF. Bernard et al. [2009] argued that the number of relevant predictors highly influences optimal `mtry` value. In this study, we select the optimal `mtry` using an exhaustive search strategy, in which all possible values of `mtry` are considered, using R package `Caret` [Kuhn et al., 2008]. While all considered parameters might have an effect on the performance of RF, we chose to focus on two parameters, `ntree` and `mtry`, for a number of reasons. The main reason is that these two parameters were originally introduced by Breiman [2001] in the development of RF algorithm. Second, `ntree` in a forest is a parameter that is tunable but not optimized and should be set sufficiently high [Oshiro et al., 2012, Probst et al., 2019] for RF to achieve good performance. It has been theoretically proven that more trees are always better [Probst et al., 2019]. In other words, optimal ntree value can go to infinity. The reduction in error, however, becomes negligible after a sufficiently large number of trees. Furthermore, empirical results provided in previous works suggest that `mtry` is the most influential out of parameters in RF [Bernard et al., 2009, Van Rijn and Hutter, 2018, Probst et al., 2019]. Figure 2.2 illustrates the step-by-step operating principle of growing RF and its the relevant parameters.

2.2.2 Variable importance in random forests

In addition to assessing a model's overall predictive ability, there is also interest in understanding the contribution of each predictor variable to model performance. There are two built-in measures for assessing variable importance in RF: mean decrease in accuracy (MDA) and mean decrease in node impurity (MDI). Both were developed by Breiman [Breiman et al., 1984, Breiman, 2001]. After all trees are grown, OOB data during training is used to compute the first measure. At each tree, the mean squared error (MSE) between predicted and observed is calculated. Then the values of each of the p predictors are randomly permuted with other predictor variables held constant. The difference between the previous and new

MSE is averaged over all trees. This is considered the predictor variable's MDA [Liaw et al., 2002] and values are reported in percent difference in MSE. The procedure is repeated for each predictor variable. Given that there is a strong association between a predictor and response variable, breaking such bond would potentially result in large error in the prediction (i.e., large MDA). MDA value can be negative where a predictor has no predictive power and adds noise to the model. Strobl et al. [2007], however, expressed caution that permutation-based measures such as MDA could show a bias towards correlated predictor variables by overestimating their importance, particularly in high-dimensional data sets.

The second method, MDI, measures the each time a predictor is selected to make a split during training. It is based on the principle that a binary split only occurs when residual errors (or impurity) of two descendent nodes are less than that of their parent node. The MDI of a predictor is the sum of all gains across all trees divided by the number of trees. Because the scale of MDI depends on values of response variable, raw MDI provides little interpretation. Following Wang et al. [2015], we computed relative MDI for each variable, which in our case is calculated by dividing each predictor variable's MDI by the sum of MDI from all predictors at each watershed. When scaled by 100, this relative MDI is a percentage and can be interpreted as the relative contribution of each predictor to the total reduction in node impurities. In the case where a predictor makes no contribution during the splitting, the relative MDI would be effectively zero. For both measures, the larger the value, the more important the predictor.

2.2.3 Benchmark models

FWe benchmark the performance of RF during the validation period against multiple linear regression (MLR) and simple naïve models using the calculated Pearson correlation coefficient (r) between forecasted and observed values for each model. In naïve model, we assume "minimal-information" scenario and the best estimate of the streamflow from the next day is the observed value from current day [Gupta et al., 1999]. Its r, in this case, is the 1-day

autocorrelation coefficient in the time series and measures of the strength of persistence. We train and verify MLR model using same data sets and predictors supplied to RF model.

2.2.4 Performance evaluation criteria

There exist different model performance criteria and each provides unique insights on the correspondence between forecasted and observed streamflow values. While r and its square, namely coefficient of determination (R^2), are often used, Legates and McCabe Jr [1999] discussed the limitation of these two measures where they were reported to be especially oversensitive to extreme values or outliers. The authors recommended that absolute error measures (i.e., root mean squared error or mean absolute error) and goodness-of-fit measure, such as the Nash-Sutcliffe efficiency (NSE), could provide more reliable and conservative assessment of the models. Kling-Gupta efficiency (KGE) is a relatively new metric that was developed based on a decomposition of NSE [Gupta et al., 2009]. This goodness-of-fit measure is gaining popularity as a benchmark metric for hydrologic models by addressing several shortcomings diagnosed with NSE. For these reasons, we selected the following four criteria to evaluate RF performance: R^2, RMSE, MAE, and KGE. These criteria cover various aspects of model's performance and also provide intuitive interpretation as explained in the remainder of this section.

R^2 can be interpreted as the proportion of the variance in the observed values that can be explained by the model. Values are in the range between 0 and 1 where 1 indicates the model is able to explain all variation in the observed dataset.

$$R^2 = \left(\frac{\sum_{i=1}^{N} (\hat{y}_i - \overline{\hat{y}})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N} (\hat{y}_i - \overline{\hat{y}})^2} \sqrt{\sum_{i=1}^{N} (y_i - \overline{y})^2}} \right)^2 \tag{2.2}$$

where N is total number of the observations during the validation period, $\hat{y}_i$ and y_i are the forecasted and observed values at day i respectively.

MAE provides an average magnitude of the errors in the model's predictions without considering the direction (underestimation or overestimation).

$$MAE = \frac{\sum_{i=1}^{N} |\hat{y}_i - y_i|}{N} \tag{2.3}$$

RMSE is the standard deviation of the residuals between the predictions and observations. It is more sensitive to larger error due to the squared operation. Both MAE and RMSE scores range between 0 and ∞ where a score of 0 indicates a perfect match between predicted and observed data. The standardization in streamflow measurements (described in Sect. 3) allows comparison of MAE and RMSE across gauges.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}} \tag{2.4}$$

KGE metric ranges between negative infinity and 1. While there currently is not a definitive KGE scale, Knoben et al. [2019] showed KGE values in the range between 0.41 and 1 indicate the model improves upon the mean flow benchmark, which assumes the predicted streamflow values equal to the mean of all observations. KGE value of 1 suggests the model can perfectly reproduce observations. KGE is calculated as follows:

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \tag{2.5}$$

where r is the Pearson correlation coefficient, α is a measure of relative variability in the forecasted and observed values, and β represents the bias:

$$\alpha = \frac{\sigma_{\hat{y}}}{\sigma_y} \quad \text{and} \quad \beta = \frac{\mu_{\hat{y}}}{\mu_y} \tag{2.6}$$

where $\sigma_{\hat{y}}$ is the standard deviation in observations, σ_y is the standard deviation in forecasted values, $\mu_{\hat{y}}$ is the forecasted mean, and μ_y is observation mean.

In hydrological forecast, one might be interested in the ability of the model to capture more extreme events rather than the overall performance. This is particularly relevant

in flood risk assessment and flood forecasting where floods are associated with discharge exceeding a high percentile (typically $\geq 90^{\text{th}}$)[Cayan et al., 1999]. The definition of "extreme" depends on the objective of the study. Here, we adopt the peak-over-threshold method. For the validation period, we calculated the 90^{th}, 95^{th}, and 99^{th} percentile streamflow values at each watershed. These are considered thresholds. If an observed daily streamflow exceeded this threshold, it would be considered an extreme event. We measure the ability of RF to capture these events using two additional criteria: probability of detection (POD) and false alarm rate (FAR). The calculation followed as in [Karran et al., 2013].

$$POD = \frac{P(\hat{y}_i > \omega | y_i > \omega)}{P(y_i > \omega)} \tag{2.7}$$

and

$$FA = \frac{P(\hat{y}_i > \omega | y_i < \omega)}{P(y_i < \omega)} \tag{2.8}$$

where ω is a specified threshold.

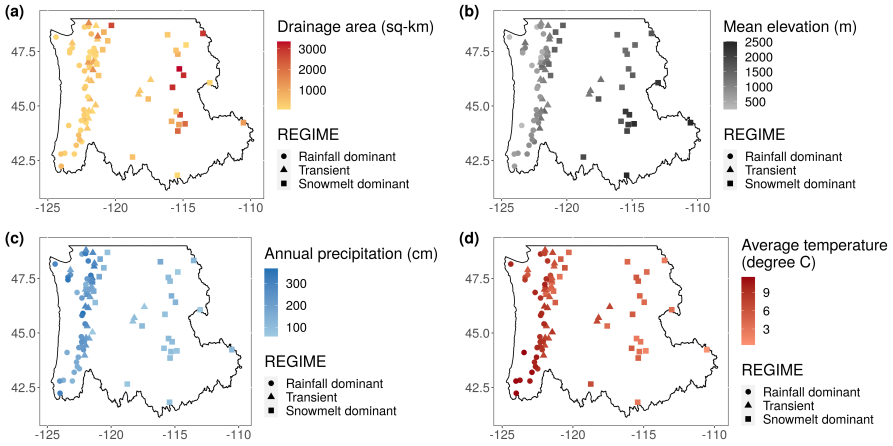
2.2.5 Study area: Pacific Northwest Hydrologic Region

In this study, we focus on watersheds in the Pacific Northwest Hydrologic Region (Fig. 2.1). This region covers an area of 836,517 km^2 and encompasses all of Washington, six other states, and British Columbia, Canada. For the purpose of maintaining consistency in monitoring protocol and data, we only consider watersheds on the US territory. The Columbia River and its tributaries make up the majority of the drainage area, traveling more than 2000 km with an extensive network of more than 100 hydroelectric dams and reservoirs have been built along these river channels. Hydropower in the Columbia River Basin supplies approximately 70 percent of Pacific Northwest energy [Payne et al., 2004]. Flood control is also an important aspect of reservoir operation in this region.

The north-south running Cascade Mountain Range divides the region into eastern and western parts and strongly influence the regional climate. The windward (west) side of the mountain receives an ample amount of winter precipitation compared to the leeward (east)

side. When temperature falls near freezing point, precipitation comes in the form of snow and provides water storage for dry summer months. Summers tend to be cool and comparatively dry. East of the Cascades, summer rainfall result from rapidly built thunderstorm and convective events that can produce flash floods [Mass, 2015]. For this region, proximity to the ocean creates a more moderate climate with a narrower seasonal temperature range compared to the inland areas, particularly in the winter. Spatial trends and variations in annual mean temperature, total precipitation, drainage area, and elevation of the watersheds are shown in Fig. 2.3.

Figure 2.3 Gauge locations with color gradient indicating variations in (a) drainage area (km^2), watershed mean elevation (m), (c) annual precipitation (cm), and (d) annual mean temperature (°C).



2.2.6 Data

The following section describes the different sources of data used in the study.

2.2.6.1 Streamflow

Our analysis uses streamflow data available through the USGS National Water Information System (NWIS) (https://waterdata.usgs.gov/nwis/sw). From NWIS, we selected daily streamflow time series for gauges using the following criteria: 1) continuous operation dur-

ing the 10-year period between 2009 and 2018, 2) have less than 10 percent of missing data, and 3) positioned in watersheds with "natural" flow that is minimally interrupted by anthropogenic intervention. The third criterion was met using the GAGES-II: Geospatial Attributes of gauges for Evaluating Streamflow dataset [Falcone, 2011] classification to identify watersheds with least-disturbed hydrologic condition and represented natural flow. We performed additional screening by computing correlation coefficient between the respective gauge and mean basin streamflow and removed those with a correlation of less than 0.5. We also excluded small creeks with drainage area less than 50 km^2. In total, 86 watersheds were selected (Fig. 2.1).

Following methodology proposed in [Wenger et al., 2010], the watersheds were further grouped into three classes of hydrologic regimes based on the timing of center-of-annual flow, which is defined as the date at which half of the total annual flow volume is exceeded. The annual flow calculations follow a water-year calendar that begins October 1^{st} and ends September 30^{th}. These three hydrologic regimes include: "early" streams with flow time < 150 (27 February), "late" streams with flow time > 200 (18 April), and "intermediate" streams with flow time between 150 and 200. These hydrologic regimes correspond to rainfall-dominated, snowmelt-dominated, and transient or transitional (mixture of rain and snowmelt) hydrographs, respectively. While this particular classification and its variants have been used in various studies related to water resources in this region [Mantua et al., 2009, Elsner et al., 2010, Vano et al., 2015], we adopted this partition in our study for two reasons. First, as Regonda et al. [2005] pointed out, the classification provides a summary of information about type and timing of precipitation, timing of snowmelt, and the contribution of these hydro-climatic variables to streamflow. This helps us assess model performance in consideration of sources of runoff. Second, the classification provides a basis to generalize the results to other watersheds that are not part of the study.

On average, records at these watersheds have less than 3 percent missing data during the 2009–2018 period. The drainage area of the watersheds range between 51 km^2 and 3355

km^2, and the mean elevation range from 239 m and 2509 m, estimated from 30-m resolution digital elevation model.

2.2.6.2 Precipitation

Daily precipitation observations were obtained from the AN81d PRISM dataset [Di Luzio et al., 2008]. This gridded dataset has a resolution of 4 km, covers the entire continental US from January 1981 to present, and is continuously updated every 6 months. Best estimate gridded value is derived by using all the available data from numbers of station networks ingested by the PRISM Climate Group. A combination of climatologically aided interpolation (CAI) and radar interpolation were used in developing PRISM dataset. In our study, watershed daily precipitation time series were constructed by computing the arithmetic mean for precipitation values of all grid points that fall within the given watershed.

2.2.7 Snow water equivalent and temperatures

SWE is defined as the depth of water that would be obtained if a column of snow were completely melted [Pan et al., 2003]. Daily SWE data were retrieved from 201 SNOTEL stations in HUC 17. These stations are part of the network of over 800 sites located in remote, high-elevation mountain watersheds in the western U.S. The elevation of these stations are in the range of 128 m and 3142 m. At SNOTEL sites, SWE is measured by a snow pillow—a pressure sensitive pad that weighs the snowpack and records the reading via a pressure transducer. As the temperature shift is the primary trigger for snowmelt, daily maximum temperature (TMAX) and minimum temperature (TMIN) from SNOTEL sensors were also retrieved and included as predictors for streamflow. The obtained data reflected the last measurement recorded for the respective day at each site. We only supplied the last measurement from SNOTEL stations because not all predictors have sub-daily values. The dataset is mostly complete, with 99.6 %, 99.6 %, and 99.9 % of the observations available for three variables TMAX, TMIN, and SWE respectively. Because of the sparse coverage

of SNOTEL sites, daily average values were calculated at USGS basin level (6-digit Hydrological Unit), similar to the currently reported snow observations from National Water and Climate Center (www.wcc.nrcs.usda.gov/snow/snow_map.html), and subsequently applied to the watersheds located in that basin. There is a total of 15 basins, each contains a number of SNOTEL stations in the range between 6 and 30 (Table S2 in the Supplement). It is noted the *in situ* data from these of stations cannot capture the spatial variability of snow accumulation and computing an area-averaged snowpack value from observations remains a challenging task [Mote et al., 2018]. The SNOTEL averages therefore represent first-order estimates of snow coverage and temperature conditions.

2.2.7.1 Predictor selection

Future daily mean streamflow (Q_{t+1}) is the response variable in our study. We attempt to explain the variability in Q_{t+1} using eight relevant predictors from the three datasets (Table 2.1). Selection of predictors is based on thorough review of the literature from previous studies and our understanding of the hydrology of this region. Specifically, precipitation (P_t) is intuitively a driver of streamflow. SWE_t provides storage information on the amount of accumulated snow available for runoff and is influenced by changes in temperature ($TMAX_t$ and $TMIN_t$). Given that there is high temporal correlation in daily temperatures, TMIN and TMAX data can provide useful signal to our streamflow forecast. Previous day streamflow (Q_t) is particularly important due to high degree of persistence that exist in the time series. A hydrological year consists of 73 pentads where each comprises of five consecutive days and observation for each day is indexed with a pentad value between 1 and 73. Data preprocessing showed moderate to strong non-linear temporal correlation between daily streamflow and the pentad at each gauge. We also derived two variables: sum of 3-day precipitation ($P3_t$) and snowmelt (SD_t) from available data. Inclusion of 3-day precipitation was to account for large winter storms that can last for several days, which often result in surges in streamflow. SD_t was calculated as the difference between SWE at day t and $t-1$. A positive value of

Table 2.1 List of predictors.

No.	Predictors	Index	Unit	Source
1	Streamflow at day t	Q_t	$m^3\ s^{-1}$	USGS
2	Precipitation	P_t	mm	PRISM
3	Sum of 3-day precipitation	$P3_t$	mm	Derived from PRISM
4	Snow water equivalent	SWE_t	mm	SNOTEL
5	Maximum temperature	$TMAX_t$	°C	SNOTEL
6	Minimum temperature	$TMIN_t$	°C	SNOTEL
7	Snowmelt (SW_t - SW_{t-1})	SD_t	mm	Derived from SNOTEL
8	Pentad	PEN_t	-	-

SD_t indicates snow accumulation and negative value indicates melt.

Soil moisture is also a relevant variable in streamflow modeling as it controls the partition between infiltration and runoff of precipitation [Aubert et al., 2003]. However, soil moisture data is often limited and incomplete, especially at daily interval and therefore not included in this study. The data were divided into two sets: training consisting of seven years 2009–2015 and a validation set of three years 2016–2018. We standardized training and validation data at each gauge using min-max scaling. First, we computed the min and max values from training data sets for each of the predictor and response variables at each watershed. These min and max values were then used to standardize both training and validation data sets. The training data, which were used to compute min-max values for standardization, therefore have values between 0 and 1. A flowchart representing the input-output model using RF is shown in Fig. 2.4.

2.3 Results and discussion

2.3.1 Parameter tuning

As we mentioned in Sect. 2, error rate in RF can be sensitive to two parameters: the number of trees `ntree` and number of randomly selected predictors available for splitting at each node `mtry`. We tested RF on training data sets of 30 randomly chosen watersheds and observed that the reduction in out-of-bag MAE error is negligible after 2000 trees. We then

Figure 2.4 Flowchart showing the input-output model using RF.

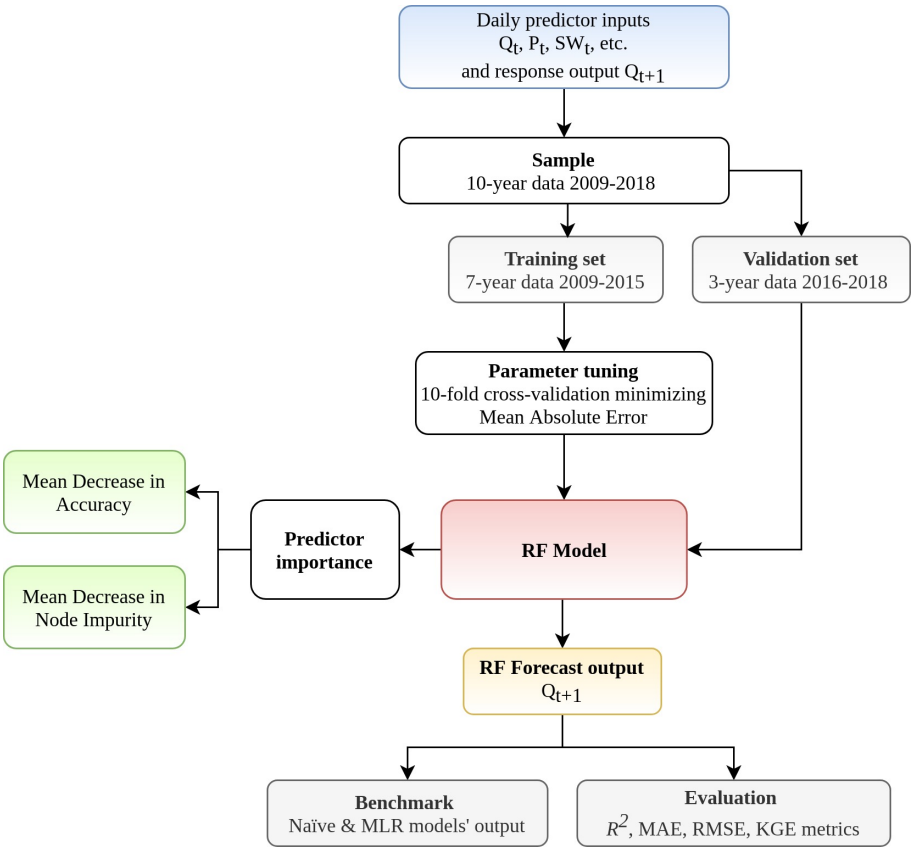


Figure 2.5 Out-of-bag mean absolute error plotted against `mtry` during optimal parameter search at Carbon River Watershed (USGS site 12094000).

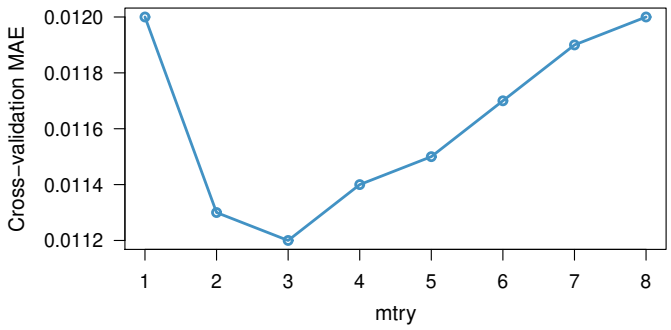


Table 2.2 The optimized parameter `mtry` using exhaustive-search strategy (`mtry` = {1, 2, 6, 7, 8} were considered but not found as the optimal value at any gauge).

`mtry`	Number of gauges	Median MAE
3	29	0.0127
4	44	0.0116
5	13	0.0079

set `ntree`=2000 for all 86 watersheds. `mtry`, on the other hand, was tuned empirically using a combination of exhaustive search approach and cross-validation.

The goal of tuning is to select the `mtry` parameter value that would optimize the performance of the model. The candidates were evaluated based on their OOB mean absolute error (MAE). At each watershed, eight possible candidate values of `mtry` (1-8) were analyzed by 3 repetitions of 10-fold cross validation from the train data set. Averaging the MAE of repetitions of the cross-validation procedure can provide more reliable results as the variance of the estimation is reduced [Seibold et al., 2018]. To illustrate, in Fig. 2.5, lowest cross-validation MAE is obtained at `mtry` = 3 at Carbon River Watershed (USGS Site 12094000). The results of tuning for all gauges (Table 2.2) show that the optimal `mtry` values are {3, 4, 5} with median MAE of 0.0127, 0.0116, and 0.0079 respectively. The optimal `mtry` at each gauge was then used in both training and validating the model. Because the number of predictors in our study is relatively small, computation burden of the exhaustive search was manageable. As the number of candidate grows, a random search strategy [Probst et al., 2019], in which values are drawn randomly from a specified space, can be more computationally efficient.

2.3.2 Benchmark RF against MLR and naïve models

Figure 2.6 shows the distributions of Pearson correlation coefficient (r) between forecasted and observed values obtained from the three models: RF, naïve, and MLR. Non-parametric, two-sample Wilcoxon rank-sum significance tests [Wilcoxon et al., 1970], which are used to assess whether the values obtained between two separate groups are systematically differ-

Figure 2.6 Boxplots for Pearson correlation coefficient between forecasted and observed values for three models: RF, naïve, and MLR across three flow regimes. Two-sample Wilcoxon rank-sum significance tests are performed and p-value (in black) are included for each pair of models.

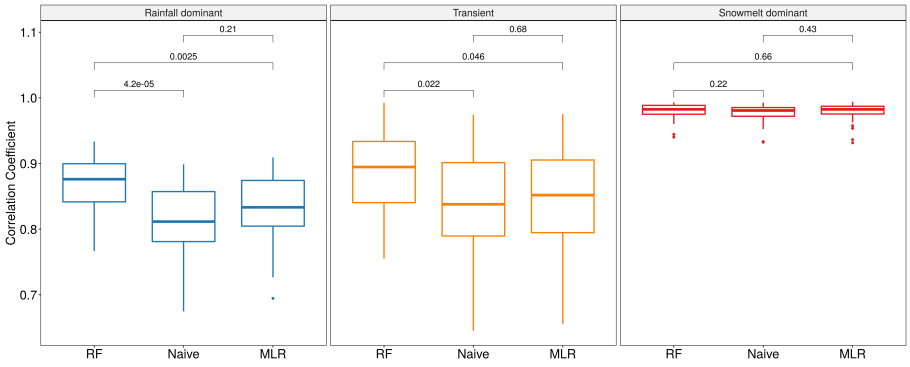
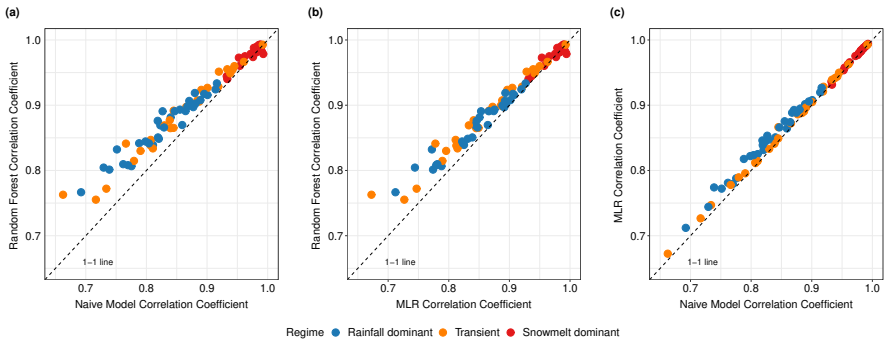


Figure 2.7 Pairwise scatter plots of Pearson correlation coefficient between forecasted and observed values among watersheds for (a) RF vs. naïve model, (b) RF vs MLR, and (c) MLR vs. naïve model. Each dot represents a watershed (n=86).



ent from one another, suggest that the pair-wise differences in r values between RF and the other two models are statistically significant ($p < 0.05$) in two flow regimes. RF is observed to outperform both naïve and MLR models in rainfall-driven and transient watersheds. Among snowmelt-driven watersheds, the three models yield similar correlation coefficients ($p > 0.05$). In Fig. 2.7a, we observe most points lie on the left of the 1-to-1 line, suggesting that RF outperforms naïve model at most individual watersheds in rainfall-driven and transient regimes. We also discern that large improvement, defined as the positive difference in r values between RF and naïve model, tends to occur with lower persistence (lower r values from the naïve model). This suggests that application of RF would be most

benefiting at watersheds where next-day streamflow is less dependent on the condition of the current day. Among snowmelt-driven watersheds, the data points lie on the 1-to-1 line, indicating that the three models show marginal difference in r values. As Mittermaier [2008] pointed out, the choice of reference can affect the perceived performance of the forecast system. Our pair-wise comparisons highlight the fact that evaluating data-driven models should be performed in consideration of the autocorrelation structure in the data [Hwang et al., 2012]. Without accounting for persistence, it would be inadequate to conclude that RF gives better performance in snowmelt-driven watersheds. Nevertheless, we observe RF outperformed MLR in all rainfall-dominated and transitional watersheds and 19 out of 25 snowmelt-dominated watersheds. The median r values for RF in the three groups are (0.88, 0.89, 0.98) compared to (0.85, 0.87, 0.98) for MLR. This may reflect RF's better ability to capture non-linear relationship between streamflow and other variables.

2.3.3 Evaluation of RF overall performance

We next evaluated the overall performance of RF across three flow regimes using four criteria: R^2, KGE, MAE, and RMSE (Table 2.3, Fig. 2.8). Here, we observe a similar trend in R^2, KGE, MAE, and RMSE scores compared to r-value trend in Fig. 6, where RF performs better in snowmelt-dominated than in rainfall-dominated (higher R^2 and KGE, lower MAE and RMSE). Snowmelt-dominated watersheds have the smallest range of R^2 values across the three groups. This may suggest that there is less variability in flow behaviors at individual gauges in this group and is consistent with the observed data where the hydrographs of snowmelt-driven watersheds tend to be less flashy compared to rainfall-driven watersheds. Not surprisingly, transitional group has the largest spread in R^2 values as watersheds in this group share characteristics from the other two groups.

Because RMSE is more sensitive to larger errors compared to MAE, the difference between the two scores represents the extent in which outliers are present in error values [Legates and McCabe Jr, 1999]. In rainfall-driven and transient groups, the shape of the

Figure 2.8 Streamflow daily forecast scores computed over the validation period for RF model in four metrics: R-squared, KGE, MAE, and RMSE.

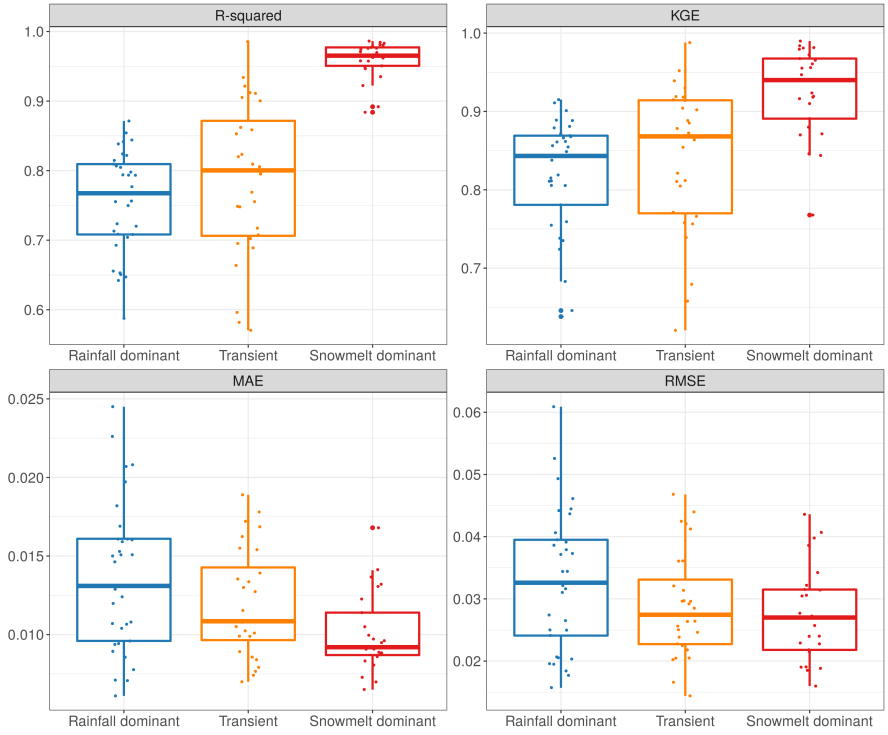


Table 2.3 Descriptive statistics of the four criteria used to evaluate the overall performance of RF: R^2, KGE, MAE, and RMSE.

Metric	Flow regime	Min	Q1	Median	Q3	Max
R^2	Rainfall-dominated	0.59	0.71	0.77	0.81	0.87
	Transient	0.57	0.71	0.80	0.87	0.99
	Snowmelt-dominated	0.88	0.95	0.97	0.98	0.99
KGE	Rainfall-dominated	0.64	0.78	0.84	0.87	0.92
	Transient	0.62	0.77	0.86	0.91	0.99
	Snowmelt-dominated	0.77	0.89	0.94	0.97	0.99
MAE	Rainfall-dominated	0.0061	0.0096	0.0131	0.0161	0.0245
	Transient	0.0070	0.0097	0.0109	0.0143	0.0189
	Snowmelt-dominated	0.0065	0.0087	0.0092	0.0114	0.0168
RMSE	Rainfall-dominated	0.0157	0.0241	0.0326	0.0395	0.0609
	Transient	0.0144	0.0227	0.0275	0.0331	0.0468
	Snowmelt-dominated	0.0160	0.0218	0.0270	0.0315	0.0436

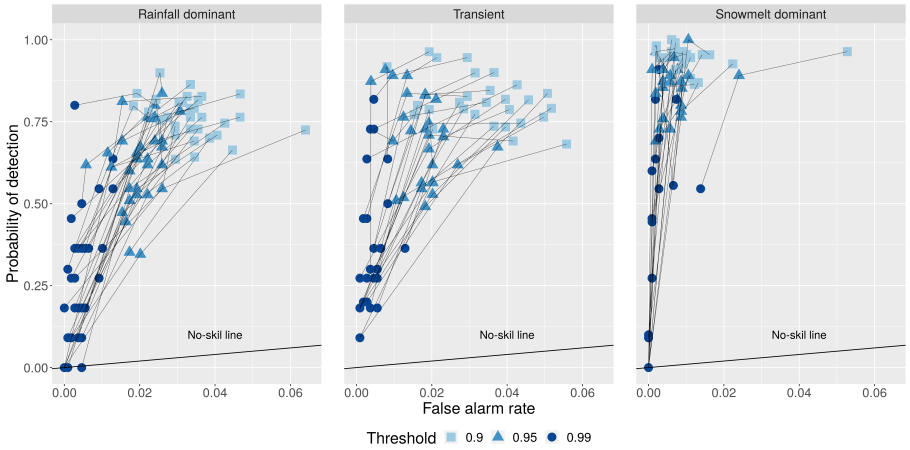
boxplot distributions remain fairly consistent between the two error scores, suggesting that distribution of large errors is similar to that of mean errors in these watersheds (Fig. 2.8). The MAE scores are heavily skewed towards 0 while RMSE scores are more evenly spread among snowmelt-driven watersheds. In snowmelt-driven watersheds, we observe a noticeably wider interquartile range (difference between first quartile and third quartile) in RMSE plot compared to MAE plot. This indicates that RF can still be susceptible to underestimation or overestimation in watersheds where the mean error is relatively low.

In Table 2.3, KGE scores are reported in a range of 0.64–0.99 for all watersheds. The median values for each flow regime are 0.84, 0.87, and 0.94. As observed mean flow is used in the calculation of KGE, Knoben et al. [2019] suggested that a KGE score greater than -0.41 indicates a hydrologic model improves upon the forecast with mean flow, independent of the basin. Therefore, RF can be seen to give satisfactory performance at all watersheds in our study. Our results are comparable to findings in [Tongal and Booij, 2018] where authors compare the performance of RF, SVM, and ANN to simulate daily discharge with baseflow separation at four rivers in California and Washington. Although authors did not classify these basins, it can be inferred that three of the rivers were rainfall-driven and one was snowmelt-driven. RF model in their study produced KGE scores of 0.41, 0.81, and 0.92 for the rainfall-driven water basins (without baseflow separation). However, our KGE scores for snowmelt-fed watersheds (with a median of 0.94) are higher compared to the reported 0.55 in their study.

2.3.4 RF performance on extreme streamflows

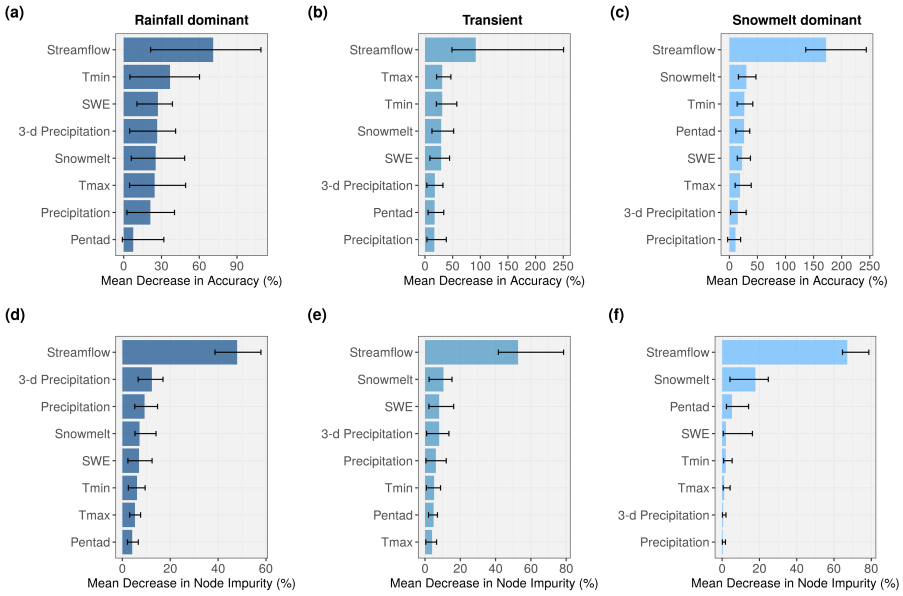
We also examine the model's capacity to forecast extreme events because of their potential high impact and associated flood risks in this region. Ability of RF to correctly detect extreme flows exceeding 90^{th}, 95^{th}, and 99^{th} percentile thresholds (defined as the POD) for each watershed are plotted against the FAR in Fig. 2.9. A threshold point falling below the no-skill line indicates the model yields higher FAR than POD and is considered to have no

Figure 2.9 The probability of detection (POD) plotted against the false alarm rate (FAR) for three extreme thresholds: 90^{th}, 95^{th}, and 99^{th} percentiles. Thin black line connects values from the same watershed. (Vertical axis) Number of times RF *correctly* forecasted events that exceeded the threshold divided by the total number of exceedance. (Horizontal axis) Number of times RF *incorrectly* forecasted events that exceeded the threshold divided by the total number of non-exceedance. It is noted that the scales of the horizontal and vertical axes are not 1-to-1 in the plotted partial receiver operating characteristic (ROC) curve.



predictive power for that threshold. RF becomes expectedly less skilful in its forecasts with increase in magnitude of the events. The model tends to perform better among snowmelt-dominated watersheds (higher POD, lower FAR) compared to those in transient and rainfall-driven groups. At the 95^{th} threshold, RF can forecast correctly at least 50 percent of the extreme events (POD ¿ 0.5) at most watersheds. At the 99^{th} threshold, the difference in RF's ability to forecast extreme streamflow among the three flow regimes becomes less obvious. In snowmelt-driven watersheds, 8 out of 25 have POD > 0.5, 9 have POD between 0.01 and 0.5, and 8 have a POD of 0. While few studies have examined complex diurnal hydrologic responses in high-elevation catchments [Graham et al., 2013], our particular result suggests large surges in streamflow sustained by spring and early summer snowmelt can be difficult to predict, even at 1-day lead time, and is an ongoing research subject [Ralph et al., 2014, Cho and Jacobs, 2020]. In our study, we observe high POD is accompanied by low FAR for the same threshold. This may suggest that RF is skillful in its forecasts of extreme events.

Figure 2.10 Barplots show importance of predictor variables using (a-c) MDA and (d-f) MDI criteria. Length of the blue bars indicates the median value across the watersheds for each flow regime and the thin black bar represents the full range of the values.



2.3.5 Analysis of variable importance

Variable importance is a useful feature in both understanding the underlying process of current model and generating insights for selection of variable in future studies [Louppe et al., 2013]. RF quantifies variable importance through two measures: MDA and MDI (Fig. 2.10). In both measures, the higher value indicates variable contributes more to the model accuracy. Intuitively, streamflow from previous day is shown to be the most importance variable due to persistence. This is reflected across three flow regimes and two measures. We also observe the sum of 3-day precipitation tends to have more predictive power than than 1-day precipitation. Maximum temperature and minimum temperature share similar contribution where minimum temperature tends to receive slightly higher scores. Among snowmelt-dominated watersheds (Fig. 2.10c and 2.10f), we anticipate snow indices (SD_t and SWE_t) contribute more in the prediction than precipitation and this is also reflected. Surprisingly, pentad comes third and fourth in MDI and MDA respectively. This supports the long-term snowpack memory of daily streamflow [Zheng et al., 2018] and can be useful in real-

time prediction. Precipitation does not seem to have significant contribution to the model's accuracy among the snowmelt-dominated watersheds. Although PRISM precipitation data includes both rainfall and snowfall, it is likely that the majority of fallen precipitation in these high-altitude watersheds is stored as snow on the surface and does not immediately contribute to runoff. Li et al. [2017] estimated that 37 % of the precipitation falls as snow in western US, yet snowmelt is responsible for 70 % of the total runoff in mountainous areas. It is still very surprising to observe such low contribution of precipitation variable to RF model accuracy. Nevertheless, we observe general agreement between the two measures in ranking of the variables in snowmelt-driven group.

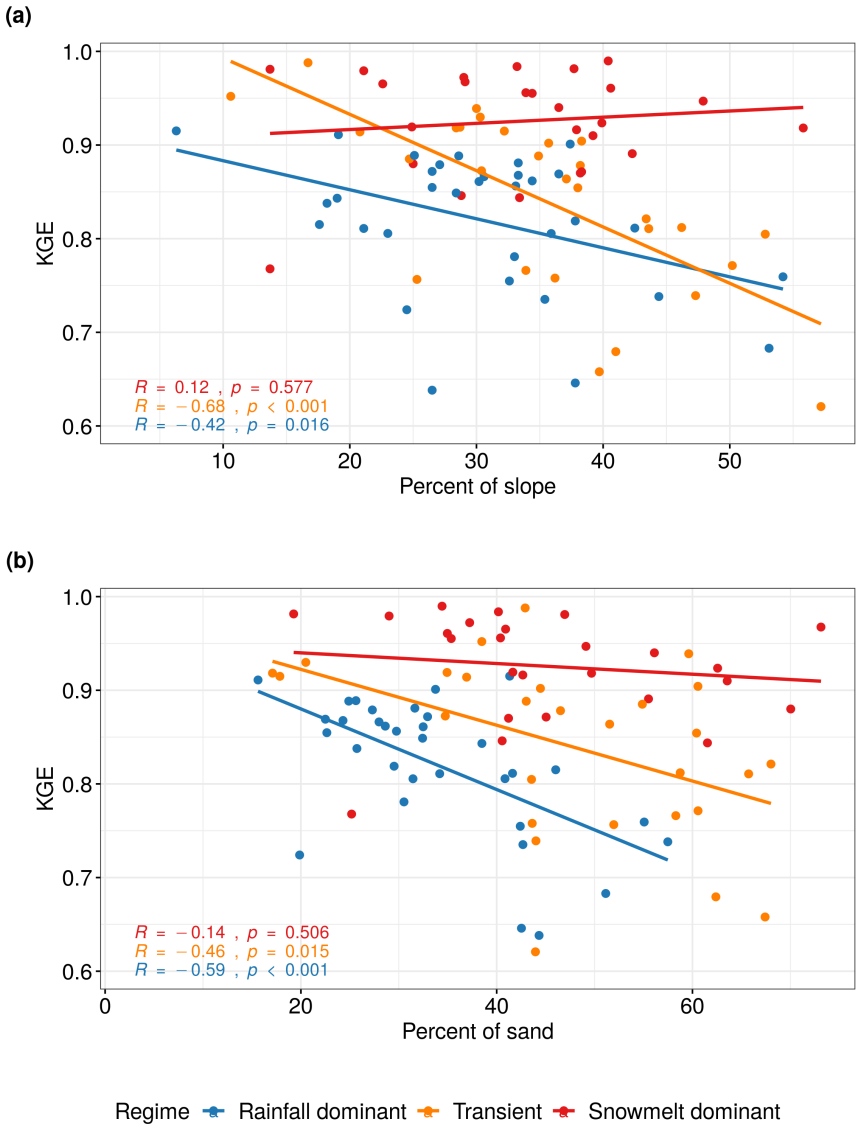
In transient and rainfall-dominated groups, there is noticeable disagreement between the two criteria. Precipitation (P_t) and 3-day precipitation ($P3_t$) tend to rank lower in MDA measure (Fig. 2.10a and 2.10b) compared to MDI (Fig. 2.10d and 2.10e). Specifically, in rainfall-dominated group, 3-day precipitation and precipitation are placed 2^{nd} and 3^{rd} based on median MDI compared to 4^{th} and 7^{th} in MDA. Maximum and minimum temperatures, on the other hand, tend to be more important in MDA calculation compared to in MDI. In Shortridge et al. [2016], RF model was used to predict streamflow at five rain-fed rivers in Ethiopia. Similarly calculated MDA in that study suggested precipitation was less important (7.71 %) than temperature (12.74 %). Linear model in the same study, however, considered the coefficient for precipitation to be significant ($p << 0.01$) while temperature coefficient was not ($p = 0.08$). In Obringer and Nateghi [2018], the authors predicted daily reservoir levels in three reservoirs in Indiana, Texas, and Atlanta using RF and other ML techniques. Precipitation was reported as the least important variable and ranked behind dew point temperature and humidity. Inspecting the probability density functions of our predictors, we suspect that for variables that are heavily skewed and zero-inflated (e.g., precipitation), permutation-based MDA may underestimate their importance compared to those that are more normally distributed such as maximum and minimum temperatures. In our precipitation data (both training and validation), at least 30 percent of the daily observations are

zeros across the watersheds. There is a high likelihood that the day with zero precipitation ends up with the same value during the shuffling process, thus potentially affecting the randomness created to compute MDA. While we did not perform additional simulation to further confirm whether MDA and MDI measures are sensitive to highly-skewed and zero-inflated variables, this can be a topic of future research. Strobl et al. [2007], however, showed RF variable importance measures can be unreliable in situations where predictor variables vary in their scale of measurement. It is noted that the scale of measurement does not only refer to the numeric range but also the nature of the data (e.g., ordinal vs. continuous). Among our 8 predictors in our study, pentad is considered an ordinal variable. Also, the scales of measurement of precipitation and temperature variables are slightly different. Precipitation is a flux variable and comprises discrete and continuous components in that if it does not rain the amount of rainfall is discrete whereas if it rains the amount is continuous. Temperature is a state variable and always continuous. Temperature predictors receiving higher MDA can also be due to identified bias where permutation-based importance measures overestimates the true contribution of correlated variables [Gregorutti et al., 2017]. In our study, temperature variables tend to have more correlation with other predictors than do the two precipitation variables. This is likely because temperature controls both the form of precipitation (snowfall vs. rainfall) and the timing of snowmelt. There is also an ongoing discussion regarding the stability of both measures, in which the two variable importance measures can yield noticeably different rankings, in simulated datasets [Calle and Urrea, 2010, Nicodemus, 2011, Ishwaran and Lu, 2019]. Although results from MDI make more sense in our case, we suggest RF users to exert caution when interpreting outputs from these two measures.

2.3.6 Effects of watershed characteristics on model performance

To explore the role of catchment characteristics such as geology, topography, and land cover on the performance of RF model, we perform Pearson correlation test between the KGE

Figure 2.11 KGE scores plotted against (a) the average percent of slope and (b) the average percent of sand in soil at each watershed. Best-fit lines were determined using simple linear regression. Pearson correlation coefficients were computed with associated significance.



scores and selected basin physical characteristics for each flow regime. These watershed characteristics were compiled as part of GAGES-II dataset using national data sources including US National Land Cover Database (NLCD) 2006 version, 100 m-resolution National Elevation Dataset (NED), and Digital General Soil Map of the United States (STATSGO2) (Table S1 in the Supplement). The results are shown in Table 2.4. There is a strong negative correlation ($p < 0.05$) between KGE scores and watershed slopes among rainfall-dominated and transient watersheds (Fig. 2.11a). As steeper hillslope often associates with faster surface and subsurface water movement during event-flow runoff, this can result in shorter response time. We observe a similar trend between KGE scores and percent of sand in the soil (Fig. 2.11b) where the RF performs worse in watersheds with higher hydraulic conductivity (i.e., higher sand content). This could be a result of rapid subsurface flow from soil profile enabled by soil macropores in mountainous forested area [Srivastava et al., 2017], where subsurface flow is the predominant mechanism. Without a quantification of the partition of discharge into surface flow and subsurface flow at individual watersheds, it is difficult to determine the relative importance of subsurface runoff mechanisms in regulating streamflow and how that may have affected the RF performance. The findings, however, suggest RF performance can deteriorate at watersheds with quick-response runoff when supplied with 1-day delayed observation data.

It appears that stream density and the amount of vegetation cover may also affect the performance of RF. Specifically, an increase in the amount of evergreen forest seems to improve the RF model among the snowmelt-dominanted watersheds but not the other two regimes. Aspect eastness, drainage area, and basin compactness are not determining factors to variability in the KGE scores. We also explored the impact of land-use and land-cover, which can be represented by the extent of impervious cover in each watershed. However, because we only selected unregulated watersheds that experienced minimal human disruption during the initial screening, most watersheds have very little impervious cover (less than 5%). It is noted that these selected characteristics are not meant to be exhaustive, but

Table 2.4 Pearson correlation coefficient between KGE scores and selected basin physical characteristics. Bolded value indicates the relationship is significant at 5 percent or 1 percent level.

Watershed characteristics	Hydrologic regime		
	Rainfall dominant	Transient	Snowmelt dominant
Slope	**-0.42**	**-0.68**	0.12
Aspect eastness	-0.02	0.12	-0.12
Drainage area	0.14	-0.12	0.11
Basin compactness	0.09	-0.12	-0.16
Stream density	-0.10	0.29	-0.27
Percent of sand	**-0.59**	**-0.46**	-0.14
Percent of evergreen forested area	–0.13	0.31	**0.41**

rather representative of various types of factors that could help explain the variability in model performance. Furthermore, an alternative approach to Pearson's correlation is to use ANOVA to test for marginal significance of each catchment variable to KGE while accounting for their interaction. Because our objective is not to make inference on KGE based on these variables and ANOVA analysis can be complicated to interpret, we choose to compute correlation coefficient.

2.3.7 Limitations and future research

There are some notable limitations in our study as well as RF in general. The classification of watersheds into three flow regimes was based on the timing of the climatological mean of the annual flow volume, which can fluctuate from year to year. This is particularly true for the watersheds in the transient group where streamflow is contributed by a mixture of runoff from winter rainfall and springtime snowmelt and the inter-annual variability is tremendous in both magnitude and timing [Lundquist et al., 2009]. Therefore, the membership of the classified watersheds from this group can vary. In fact, Mantua et al. [2009] discussed the future shift of transient runoff watersheds towards rainfall-dominated in Washington State. Because we trained RF using the same input variables for all watersheds regardless of flow regimes and calculated performance criteria separately, the classification does not alter the

results at individual watershed.

In the study, we used estimated precipitation from PRISM, which is an interpolation product and combines data from various rain gauges from multiple networks. Despite possible introduced errors and uncertainty, we believe the use of spatially distributed product better represents the areal estimation of precipitation over the watershed than a single rain gauge measurement. In real-time forecast, this would be not be feasible due to the added time to compile and process such data. Similarly, we provided RF model with a basin-average SWE from SNOTEL stations as an estimate of snowpack condition. Using a more spatially consistent SWE data such as the Snow Data Assimilation System [Pan et al., 2003] product would potentially improve model accuracy. As our results indicate that RF can produce reasonable forecasts, potential future research could explore the sensitivity of the model using satellite derived snow products a station data and even include $t + 1$ precipitation forecast as a predictor in the model.

An inherent limitation of RF is the lack of direct uncertainty quantification in prediction. In our case, the forecasted streamflow using RF does not yield a standard error comparable to that provided by traditional regression model, and hence no way to provide probabilistic confidence intervals on predictions. Methods to estimate confidence intervals have been proposed by Wager et al. [2014], Mentch and Hooker [2016], and Coulston et al. [2016], but they are not widely applied. For future work, computation of confidence interval in RF prediction will be useful in addressing and understanding uncertainty.

2.4 Conclusions

Accurate streamflow forecast has extensive applications across disciplines from water resources and planning to engineering design. In this study, we assessed the ability of RF to make daily streamflow forecasts at 86 watersheds in the Pacific Northwest Hydrologic Region. Key results are summarized below:

- Based on the KGE scores (ranging from 0.62 to 0.99), we show that RF is capable of

producing skilfull forecasts across all watersheds.

- RF performs better in snowmelt-dominated watersheds, which can be attributed to stronger persistence in the streamflow time series. The largest improvements in forecast compared to naïve model are found among rainfall-dominated watersheds.

- The two approaches for measuring predictor importance yield noticeably different results. We recommend interpretation of the these two measures should be coupled with understanding of the physical processes and how these processes are connected.

- Increase in steepness of slope and amount of sand content are found to deteriorate RF performance in two flow regime groups. This demonstrates catchment characteristics can cause variability in performance of the model and should be considered in both predictor selection and evaluation of the model.

Considering the current and future vulnerabilities of the Pacific Northwest to flooding caused by extreme precipitation and significant snowmelt events [Ralph et al., 2014], skillful streamflow forecasts can have important implications. Due to its practical applications, RF and RF-based algorithms continue to gain popularity in hydrological studies [Tyralis et al., 2019]. Given the promising results from our study, RF can be used as part of an ensemble of models to achieve better generalization ability and accuracy not only in streamflow forecast but also in other water-related applications in this region.

APPENDIX

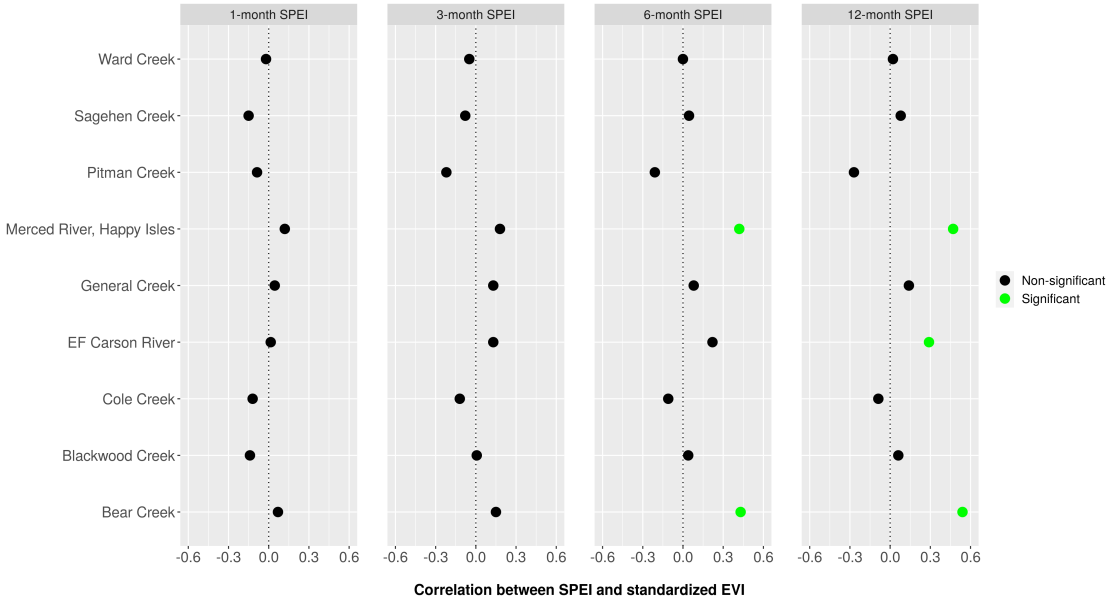
APPENDIX

APPENDIX A

A.1 Vegetation response to drought at different time scales

We observe in Fig. A.1 that the vegetations in the 9 watersheds are relatively resilient to drought at short time scales. We found no significant correlation between SPEI and EVI was found at 1-month time scale. At 3-month, we observe a weak correlation ($r = 0.26$) between EVI and SPEI at Bear Creek. At 6- and 12-month time scales, we observe significant vegetation response to SPEI in 3 watersheds: Merced River, Bear Creek, and EF Carson River. The observed pattern is likely due to the fact that longer SPEI time scales (6- and 12-month) account for the water deficit accumulated from the winter season where the majority of precipitation occurs. Furthermore, these three watersheds, which locate in the central and southern parts of Sierra Nevada, also have higher coverage of shrubs and grasslands (Fig. 1.4) compared to the other densely forested watersheds. More specifically, [Dong et al., 2019] reported a geographical difference between vegetation responses in Northern versus Southern California where the sensitivity of the vegetation to drought are larger in the southern part. This is also consistent with previous research where grassland ecosystems are more sensitive to drought than coniferous forests [Zha et al., 2010, Vicente-Serrano et al., 2010].

Figure A.1 Correlation between EVI and 1-, 3-, 6-, 12-month SPEI values during the growing season (May-Sep) for the period 2001-2018. EVI time series were standardized, according to the average and the standard deviation of the values for each month. Relationship is considered significant at $\alpha = 0.05$).



A.2 Correlation between streamflow elasticity and catchment characteristics

Climate elasticity of streamflow, a non-parametric indicator commonly used to quantify the sensitivity of streamflow to changes in climate, is defined as the proportional change in streamflow, Q, to the proportional change in a climatic variable such as precipitation P [Sankarasubramanian et al., 2001] and can be expressed as:

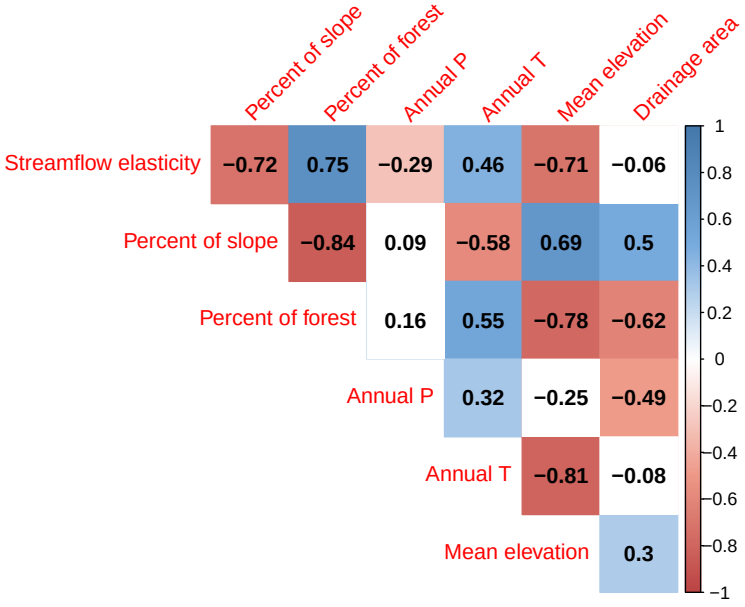
$$\epsilon = median(\frac{Q_t - \bar{Q}}{P_t - \bar{P}} \frac{\bar{P}}{\bar{Q}}) \tag{A.1}$$

where Q_t is the annual runoff, P_t is the annual precipitation, $\bar{Q}$ is the long-term average annual streamflow, $\bar{P}$ is the long-term average annual precipitation. $\bar{Q}$ and $\bar{P}$ were calculated using 30 years of data for 1987-2018 period.

To explore the effects of catchment characteristics on the streamflow elasticity, we computed Pearson's correlation between ϵ and the respective catchment variable. Results are

shown in Fig. A.2. Cells with a white background indicate the relationship is not significant at ($\alpha = 0.05$).

Figure A.2 Correlation between streamflow elasticity (ϵ) and selected catchment characteristics. Blank tiles indicate the relationship is not significant at $\alpha = 0.05$.



BIBLIOGRAPHY

BIBLIOGRAPHY

Jan F Adamowski. Development of a short-term river flood forecasting method for snowmelt driven floods based on wavelet and cross-wavelet analysis. *Journal of Hydrology*, 353(3-4): 247–266, 2008.

KJ Allen, John Ogden, BM Buckley, ER Cook, and PJ Baker. The potential to reconstruct broadscale climate indices associated with southeast australian droughts from athrotaxis species, tasmania. *Climate Dynamics*, 37(9-10):1799–1821, 2011.

Douglas G Altman and J Martin Bland. Statistics notes variables and parameters. *Bmj*, 318 (7199):1667, 1999.

David Aubert, Cecile Loumagne, and Ludovic Oudin. Sequential assimilation of soil moisture and streamflow data in a conceptual rainfall–runoff model. *Journal of Hydrology*, 280(1-4): 145–161, 2003.

Francesco Avanzi, Joseph Rungee, Tessa Maurer, Roger Bales, Qin Ma, Steven Glaser, and Martha Conklin. Climate elasticity of evapotranspiration shifts the water balance of mediterranean climates during multi-year droughts. *Hydrology and Earth System Sciences*, 24(9):4317–4337, 2020.

Roger C Bales, Michael L Goulden, Carolyn T Hunsaker, Martha H Conklin, Peter C Hartsough, Anthony T O'Geen, Jan W Hopmans, and Mohammad Safeeq. Mechanisms controlling the impact of multi-year drought on mountain hydrology. *Scientific Reports*, 8(1): 690, 2018.

Tim P Barnett, Jennifer C Adam, and Dennis P Lettenmaier. Potential impacts of a warming climate on water availability in snow-dominated regions. *Nature*, 438(7066):303–309, 2005.

Emilie Beaulieu, Yann Lucas, Daniel Viville, François Chabaux, Philippe Ackerer, Yves Goddéris, and Marie-Claire Pierret. Hydrological and vegetation response to climate change in a forested mountainous catchment. *Modeling Earth Systems and Environment*, 2(4):1–15, 2016.

Santiago Beguería, M Sergio, and Lazyload Yes. Calculation of the standardised precipitation-evapotranspiration index. 2013.

Santiago Beguería, Sergio M Vicente-Serrano, Fergus Reig, and Borja Latorre. Standardized precipitation evapotranspiration index (spei) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. *International journal of climatology*, 34(10):3001–3023, 2014.

Simon Bernard, Laurent Heutte, and Sébastien Adam. Influence of hyperparameters on random forest accuracy. In *International Workshop on Multiple Classifier Systems*, pages 171–180. Springer, 2009.

Douglas P Boyle, Hoshin V Gupta, and Soroosh Sorooshian. Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. *Water Resources Research*, 36(12):3663–3674, 2000.

Kate A Brauman, Gretchen C Daily, T Ka'eo Duarte, and Harold A Mooney. The nature and value of ecosystem services: an overview highlighting hydrologic services. *Annu. Rev. Environ. Resour.*, 32:67–98, 2007.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

Mikhail Ivanovich Budyko. *Climate and life*. Academic Press, Inc., 1974.

M Luz Calle and Víctor Urrea. Letter to the editor: stability of random forest importance measures. *Briefings in bioinformatics*, 12(1):86–89, 2010.

Alejandra M Carmona, Murugesu Sivapalan, Mary A Yaeger, and Germán Poveda. Regional patterns of interannual variability of catchment water balances across the continental us: A budyko framework. *Water Resources Research*, 50(12):9177–9193, 2014.

Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

Daniel R Cayan, Kelly T Redmond, and Laurence G Riddle. Enso and hydrologic extremes in the western united states. *Journal of Climate*, 12(9):2881–2893, 1999.

Xi Chen and Hemant Ishwaran. Random forests for genomic data analysis. *Genomics*, 99 (6):323–329, 2012.

Lei Cheng, Zongxue Xu, Dingbao Wang, and Ximing Cai. Assessing interannual variability of evapotranspiration at the catchment scale using satellite-based evapotranspiration data sets. *Water Resources Research*, 47(9), 2011.

Eunsang Cho and Jennifer M Jacobs. Extreme value snow water equivalent and snowmelt for infrastructure design over the contiguous united states. *Water Resources Research*, page e2020WR028126, 2020.

John W Coulston, Christine E Blinn, Valerie A Thomas, and Randolph H Wynne. Approximating prediction uncertainty for random forest regression models. *Photogrammetric*

Engineering & Remote Sensing, 82(3):189–197, 2016.

Aiguo Dai. Increasing drought under global warming in observations and models. *Nature climate change*, 3(1):52–58, 2013.

Christopher Daly, Michael Halbleib, Joseph I Smith, Wayne P Gibson, Matthew K Doggett, George H Taylor, Jan Curtis, and Phillip P Pasteris. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous united states. *International Journal of Climatology: a Journal of the Royal Meteorological Society*, 28 (15):2031–2064, 2008.

Anthony W D'Amato, John B Bradford, Shawn Fraver, and Brian J Palik. Effects of thinning on drought vulnerability and climate response in north temperate forest ecosystems. *Ecological Applications*, 23(8):1735–1742, 2013.

Christian W Dawson, Robert J Abrahart, Asaad Y Shamseldin, and Robert L Wilby. Flood estimation at ungauged sites using artificial neural networks. *Journal of hydrology*, 319 (1-4):391–409, 2006.

J Demarty, Fi Chevallier, AD Friend, Ni Viovy, Shilong Piao, and P Ciais. Assimilation of global modis leaf area index retrievals within a terrestrial biosphere model. *Geophysical research letters*, 34(15), 2007.

Michael D Dettinger, Daniel R Cayan, Mary K Meyer, and Anne E Jeton. Simulated hydrologic responses to climate variations and change in the merced, carson, and american river basins, sierra nevada, california, 1900–2099. *Climatic Change*, 62(1-3):283–317, 2004.

Mauro Di Luzio, Gregory L Johnson, Christopher Daly, Jon K Eischeid, and Jeffrey G Arnold. Constructing retrospective gridded daily precipitation and temperature datasets for the conterminous united states. *Journal of Applied Meteorology and Climatology*, 47 (2):475–497, 2008.

Yonas B Dibike and Dimitri P Solomatine. River flow forecasting using artificial neural networks. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, 26(1):1–7, 2001.

Kamel Didan. Mod13q1 modis/terra vegetation indices 16-day l3 global 250m sin grid v006. *NASA EOSDIS Land Processes DAAC*, 10, 2015.

Noah S Diffenbaugh, Daniel L Swain, and Danielle Touma. Anthropogenic warming has increased drought risk in california. *Proceedings of the National Academy of Sciences*, 112 (13):3931–3936, 2015.

S Lawrence Dingman. *Physical hydrology*. Waveland press, 2015.

Chunyu Dong, Glen M MacDonald, Katherine Willis, Thomas W Gillespie, Gregory S Okin, and A Park Williams. Vegetation responses to 2012–2016 drought in northern and southern california. *Geophysical Research Letters*, 46(7):3810–3821, 2019.

Randall Donohue, Michael Roderick, Tim R McVicar, et al. On the importance of including vegetation dynamics in budyko's hydrological model. 2007.

Marketa M Elsner, Lan Cuo, Nathalie Voisin, Jeffrey S Deems, Alan F Hamlet, Julie A Vano, Kristian EB Mickelson, Se-Yeun Lee, and Dennis P Lettenmaier. Implications of 21st century climate change for the hydrology of washington state. *Climatic Change*, 102 (1-2):225–260, 2010.

James A Falcone. Gages-ii: Geospatial attributes of gages for evaluating streamflow. Technical report, US Geological Survey, 2011.

Wei Feng, Hongwei Lu, Tianci Yao, and Qing Yu. Drought characteristics and its elevation dependence in the qinghai–tibet plateau during the last half-century. *Scientific Reports*, 10(1):1–11, 2020.

Douglas A Frank and Richard S Inouye. Temporal variation in actual evapotranspiration of terrestrial ecosystems: patterns and ecological implications. *Journal of Biogeography*, pages 401–411, 1994.

Christopher C Funk, Andrew Hoell, and Daithi Stone. Examining the contribution of the observed global warming trend to the california droughts of 2012/13 and 2013/14. *Bulletin of the American Meteorological Society*, 95(9):S11–S15, 2014.

Michael L Goulden and Roger C Bales. Mountain runoff vulnerability to increased evapotranspiration with vegetation expansion. *Proceedings of the National Academy of Sciences*, 111(39):14071–14075, 2014.

ML Goulden, RG Anderson, RC Bales, AE Kelly, M Meadows, and GC Winston. Evapotranspiration along an elevation gradient in california's sierra nevada. *Journal of Geophysical Research: Biogeosciences*, 117(G3), 2012.

Chris B Graham, Holly R Barnard, Kathleen L Kavanagh, and James P McNamara. Catchment scale controls the temporal connection of transpiration and diel fluctuations in streamflow. *Hydrological Processes*, 27(18):2541–2556, 2013.

Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, 27(3):659–678, 2017.

Hoshin V Gupta, Harald Kling, Koray K Yilmaz, and Guillermo F Martinez. Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of hydrology*, 377(1-2):80–91, 2009.

Hoshin Vijai Gupta, Soroosh Sorooshian, and Patrice Ogou Yapo. Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. *Journal of Hydrologic Engineering*, 4(2):135–143, 1999.

Minxue He, Mitchel Russo, and Michael Anderson. Hydroclimatic characteristics of the 2012–2015 california drought from an operational perspective. *Climate*, 5(1):5, 2017.

Barbara FF Huang and Paul C Boutros. The parameter sensitivity of random forests. *BMC bioinformatics*, 17(1):331, 2016.

Alfredo Huete, Kamel Didan, Tomoaki Miura, E Patricia Rodriguez, Xiang Gao, and Laerte G Ferreira. Overview of the radiometric and biophysical performance of the modis vegetation indices. *Remote sensing of environment*, 83(1-2):195–213, 2002.

Carolyn T Hunsaker, Thomas W Whitaker, and Roger C Bales. Snowmelt runoff and water yield along elevation and temperature gradients in california's southern sierra nevada 1. *JAWRA Journal of the American Water Resources Association*, 48(4):667–678, 2012.

Seok Hwan Hwang, Dae Heon Ham, and Joong Hoon Kim. A new measure for assessing the efficiency of hydrological data-driven forecasting models. *Hydrological sciences journal*, 57 (7):1257–1274, 2012.

Hemant Ishwaran and Min Lu. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in medicine*, 38 (4):558–582, 2019.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An introduction to statistical learning, volume 103 xiv of, 2013.

James A Johnstone. A quasi-biennial signal in western us hydroclimate and its global teleconnections. *Climate dynamics*, 36(3-4):663–680, 2011.

Daniel J Karran, Efrat Morin, and Jan Adamowski. Multi-step streamflow forecasting using data-driven non-linear methods in contrasting climate regimes. *Journal of Hydroinformatics*, 16(3):671–689, 2013.

Wouter JM Knoben, Jim E Freer, and Ross A Woods. Inherent benchmark or not? comparing nash–sutcliffe and kling–gupta efficiency scores. *Hydrology and Earth System Sciences*, 23(10):4323–4331, 2019.

John F Knowles, Noah P Molotch, Ernesto Trujillo, and Marcy E Litvak. Snowmelt-driven trade-offs between early and late season productivity negatively impact forest carbon uptake during drought. *Geophysical Research Letters*, 45(7):3087–3096, 2018.

N Knowles, M Dettinger, and D Cayan. Trends in snowfall versus rainfall for the western

united states, 1949-2001. prepared for california energy commission public interest energy research program. *Trends in Snowfall Versus Rainfall for the Western United States, 1949-2001. Prepared for California Energy Commission Public Interest Energy Research Program*, 2007.

Noah Knowles, Michael D Dettinger, and Daniel R Cayan. Trends in snowfall versus rainfall in the western united states. *Journal of Climate*, 19(18):4545–4559, 2006.

Randal D Koster and Max J Suarez. A simple framework for examining the interannual variability of land surface moisture fluxes. *Journal of Climate*, 12(7):1911–1917, 1999.

Sebastian A Krogh, Patrick D Broxton, Patricia N Manley, and Adrian A Harpold. Using process based snow modeling and lidar to predict the effects of forest thinning on the northern sierra nevada snowpack. *Frontiers in Forests and Global Change*, 3:21, 2020.

Max Kuhn et al. Building predictive models in r using the caret package. *Journal of statistical software*, 28(5):1–26, 2008.

L Labudová, M Labuda, and J Takáč. Comparison of spi and spei applicability for drought impact assessment on crop production in the danubian lowland and the east slovakian lowland. *Theoretical and Applied Climatology*, 128(1-2):491–506, 2017.

David R Legates and Gregory J McCabe Jr. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water resources research*, 35(1):233–241, 1999.

Dongyue Li, Melissa L Wrzesien, Michael Durand, Jennifer Adam, and Dennis P Lettenmaier. 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, 2017.

Weiguang Li, Meiting Hou, Huilin Chen, and Xiaomin Chen. Study on drought trend in south china based on standardized precipitation evapotranspiration index. *Journal of natural disasters*, 21(4):84–90, 2012.

Xue Li, Jian Sha, and Zhong-Liang Wang. Comparison of daily streamflow forecasts using extreme learning machines and the random forest method. *Hydrological Sciences Journal*, 64(15):1857–1866, 2019.

Xu Lian, Shilong Piao, Laurent ZX Li, Yue Li, Chris Huntingford, Philippe Ciais, Alessandro Cescatti, Ivan A Janssens, Josep Peñuelas, Wolfgang Buermann, et al. Summer soil drying exacerbated by earlier spring greening of northern vegetation. *Science advances*, 6(1): eaax0255, 2020.

Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

Scott R Loarie, Philip B Duffy, Healy Hamilton, Gregory P Asner, Christopher B Field, and David D Ackerly. The velocity of climate change. *Nature*, 462(7276):1052–1055, 2009.

Steven P Loheide, Richard S Deitchman, David J Cooper, Evan C Wolf, Christopher T Hammersmark, and Jessica D Lundquist. A framework for understanding the hydroecology of impacted wet meadows in the sierra nevada and cascade ranges, california, usa. *Hydrogeology Journal*, 17(1):229–246, 2009.

Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees. In *Advances in neural information processing systems*, pages 431–439, 2013.

Jessica D Lundquist, Michael D Dettinger, Iris T Stewart, and Daniel R Cayan. Variability and trends in spring runoff in the western united states. *Climate warming in western North America: evidence and environmental effects. University of Utah Press, Salt Lake City, Utah, USA*, pages 63–76, 2009.

Lifeng Luo, Deanna Apps, Samuel Arcand, Huating Xu, Ming Pan, and Martin Hoerling. Contribution of temperature and precipitation anomalies to the california drought during 2012–2015. *Geophysical Research Letters*, 44(7):3184–3192, 2017.

David A Lytle and N LeRoy Poff. Adaptation to natural flow regimes. *Trends in ecology & evolution*, 19(2):94–100, 2004.

Michael E Mann and Peter H Gleick. Climate change and california drought in the 21st century. *Proceedings of the National Academy of Sciences*, 112(13):3858–3859, 2015.

Nathan Mantua, Ingrid Tohver, and Alan Hamlet. Impacts of climate change on key aspects of freshwater salmon habitat in washington state. 2009.

Antonio Manzano, Miguel A Clemente, Ana Morata, M Yolanda Luna, Santiago Beguería, Sergio M Vicente-Serrano, and M Luisa Martín. Analysis of the atmospheric circulation pattern effects over spei drought index in spain. *Atmospheric Research*, 230:104630, 2019.

Clifford Mass. *The weather of the Pacific Northwest*. University of Washington Press, 2015.

Tessa Maurer, Francesco Avanzi, Steven D Glaser, and Roger C Bales. Drivers of drought-induced shifts in the water balance through a budyko approach. *Hydrology and Earth System Sciences*, 26(3):589–607, 2022.

Josué Medellín-Azuara, Julien J Harou, Marcelo A Olivares, Kaveh Madani, Jay R Lund, Richard E Howitt, Stacy K Tanaka, Marion W Jenkins, and Tingju Zhu. Adaptability and adaptations of california's water supply system to dry climate warming. *Climatic Change*, 87(1):75–90, 2008.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881, 2016.

Ashok K Mishra and Vijay P Singh. Drought modeling–a review. *Journal of Hydrology*, 403 (1-2):157–175, 2011.

Marion P Mittermaier. The potential impact of using persistence as a reference forecast on perceived forecast skill. *Weather and forecasting*, 23(5):1022–1031, 2008.

Amir Mosavi, Pinar Ozturk, and Kwok-wing Chau. Flood prediction using machine learning models: Literature review. *Water*, 10(11):1536, 2018.

Philip W Mote, Sihan Li, Dennis P Lettenmaier, Mu Xiao, and Ruth Engel. Dramatic declines in snowpack in the western us. *Npj Climate and Atmospheric Science*, 1(1):1–6, 2018.

Qiaozhen Mu, Maosheng Zhao, and Steven W Running. Improvements to a modis global terrestrial evapotranspiration algorithm. *Remote sensing of environment*, 115(8):1781–1800, 2011.

Qiaozhen Mu, Maosheng Zhao, and Steven W Running. Modis global terrestrial evapotranspiration (et) product (nasa mod16a2/a3). *Algorithm Theoretical Basis Document, Collection*, 5, 2013.

Kristin K Nicodemus. Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. *Briefings in bioinformatics*, 12(4):369–373, 2011.

Malcolm North. Managing sierra nevada forests. *Gen. Tech. Rep. PSW-GTR-237. Albany, CA: US Department of Agriculture, Forest Service, Pacific Southwest Research Station. 184 p*, 237, 2012.

Renee Obringer and Roshanak Nateghi. Predicting urban reservoir levels using statistical learning techniques. *Scientific reports*, 8(1):5164, 2018.

Thais Mayumi Oshiro, Pedro Santoro Perez, and José Augusto Baranauskas. How many trees in a random forest? In *International workshop on machine learning and data mining in pattern recognition*, pages 154–168. Springer, 2012.

Thomas C Pagano, David C Garen, Tom R Perkins, and Phillip A Pasteris. Daily updating of operational statistical seasonal water supply forecasts for the western us 1. *JAWRA Journal of the American Water Resources Association*, 45(3):767–778, 2009.

Mahesh Pal. Random forest classifier for remote sensing classification. *International Journal*

of Remote Sensing, 26(1):217–222, 2005.

Ming Pan, Justin Sheffield, Eric F Wood, Kenneth E Mitchell, Paul R Houser, John C Schaake, Alan Robock, Dag Lohmann, Brian Cosgrove, Qingyun Duan, et al. Snow process modeling in the north american land data assimilation system (nldas): 2. evaluation of model simulated snow water equivalent. *Journal of Geophysical Research: Atmospheres*, 108(D22), 2003.

Georgia A Papacharalampous and Hristos Tyralis. Evaluation of random forests and prophet for daily streamflow forecasting. *Advances in Geosciences*, 45:201–208, 2018.

Jeffrey T Payne, Andrew W Wood, Alan F Hamlet, Richard N Palmer, and Dennis P Lettenmaier. Mitigating the effects of climate change on the water resources of the columbia river basin. *Climatic change*, 62(1-3):233–256, 2004.

David Peterson, Richard Smith, Iris Stewart, Noah Knowles, Chris Soulard, and Stephen Hager. Snowmelt discharge characteristics sierra nevada, california. *US Geological Survey Scientific Investigations Report*, pages 1–13, 2005.

K Podolak, D Edelson, S Kruse, B Aylward, M Zimring, and N Wobbrock. Estimating the water supply benefits from forest restoration in the northern sierra nevada. *An unpublished report of the nature conservancy prepared with ecosystem economics. San Francisco, CA*, 2015.

NJ Potter, C Petheram, and L Zhang. Sensitivity of streamflow to rainfall and temperature in south-eastern australia during the millennium drought. In *19th International Congress on Modelling and Simulation, Perth, Dec*, pages 3636–3642, 2011.

Philipp Probst, Marvin N Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1301, 2019.

FM Ralph, M Dettinger, A White, D Reynolds, D Cayan, T Schneider, R Cifelli, K Redmond, M Anderson, F Gherke, et al. A vision for future observations for western us extreme precipitation and flooding. *Journal of Contemporary Water Research & Education*, 153 (1):16–32, 2014.

Kabir Rasouli, William W Hsieh, and Alex J Cannon. Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology*, 414: 284–293, 2012.

Satish Kumar Regonda, Balaji Rajagopalan, Martyn Clark, and John Pitlick. Seasonal cycle shifts in hydroclimatology over the western united states. *Journal of climate*, 18(2): 372–384, 2005.

M Renner, R Seppelt, and C Bernhofer. Evaluation of water-energy balance frameworks to predict the sensitivity of streamflow to climate change. *Hydrology and Earth System Sciences*, 16(5):1419–1433, 2012.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

Ignacio Rodriguez-Iturbe. Ecohydrology: A hydrologic perspective of climate-soil-vegetation dynamies. *Water Resources Research*, 36(1):3–9, 2000.

Edward B Royce and Michael G Barbour. Mediterranean climate effects. i. conifer water use across a sierra nevada ecotone. *American Journal of Botany*, 88(5):911–918, 2001.

Mohammad Safeeq, Guillaume S Mauger, Gordon E Grant, Ivan Arismendi, Alan F Hamlet, and Se-Yeun Lee. Comparing large-scale hydrological model predictions with observed streamflow in the pacific northwest: effects of climate and groundwater. *Journal of Hydrometeorology*, 15(6):2501–2521, 2014.

Margarita Saft, Andrew W Western, Lu Zhang, Murray C Peel, and Nick J Potter. The influence of multiyear drought on the annual rainfall-runoff relationship: An a ustralian perspective. *Water Resources Research*, 51(4):2444–2463, 2015.

Eric P Salathé Jr, Alan F Hamlet, Clifford F Mass, Se-Yeun Lee, Matt Stumbaugh, and Richard Steed. Estimates of twenty-first-century flood risk in the pacific northwest based on regional climate model simulations. *Journal of Hydrometeorology*, 15(5):1881–1899, 2014.

A Sankarasubramanian, Richard M Vogel, and James F Limbrunner. Climate elasticity of streamflow in the united states. *Water Resources Research*, 37(6):1771–1781, 2001.

Heidi Seibold, Christoph Bernau, Anne-Laure Boulesteix, and Riccardo De Bin. On the choice and influence of the number of boosting steps for high-dimensional linear cox-models. *Computational Statistics*, 33(3):1195–1215, 2018.

Julie E Shortridge, Seth D Guikema, and Benjamin F Zaitchik. Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences*, 20(7):2611–2628, 2016.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*, 2017.

Bernard A Silverman. An evaluation of eleven operational cloud seeding programs in the watersheds of the sierra nevada mountains. *Atmospheric Research*, 97(4):526–539, 2010.

Jan Sitterson, Chris Knightes, Rajbir Parmar, Kurt Wolfe, Brian Avant, and Muluken Muche. An overview of rainfall-runoff model types. 2018.

Anurag Srivastava, Joan Q Wu, William J Elliot, Erin S Brooks, and Dennis C Flanagan. Modeling streamflow in a snow-dominated forest watershed using the water erosion prediction project (wepp) model. *Transactions of the ASABE. 60 (4): 1171-1187.*, 60(4): 1171–1187, 2017.

Jens T Stevens, Brandon M Collins, Jonathan W Long, Malcolm P North, Susan J Prichard, Leland W Tarnay, and Angela M White. Evaluating potential trade-offs among fuel treatment strategies in mixed-conifer forests of the sierra nevada. *Ecosphere*, 7(9):e01445, 2016.

Iris T Stewart, Daniel R Cayan, and Michael D Dettinger. Changes toward earlier streamflow timing across western north america. *Journal of climate*, 18(8):1136–1155, 2005.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

Daniel L Swain. A tale of two california droughts: Lessons amidst record warmth and dryness in a region of complex physical and human geography. *Geophysical Research Letters*, 42 (22):9999–10, 2015.

Charles Warren Thornthwaite. An approach toward a rational classification of climate. *Geographical review*, 38(1):55–94, 1948.

Ingrid M Tohver, Alan F Hamlet, and Se-Yeun Lee. Impacts of 21st-century climate change on hydrologic extremes in the pacific northwest region of north america. *JAWRA Journal of the American Water Resources Association*, 50(6):1461–1476, 2014.

Bryan A Tolson and Christine A Shoemaker. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, 43(1), 2007.

Hakan Tongal and Martijn J Booij. Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. *Journal of hydrology*, 564:266–282, 2018.

Hristos Tyralis, Georgia Papacharalampous, and Andreas Langousis. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water*, 11(5):910, 2019.

U.S. Geological Survey. U.s. geological survey, 2019, national hydrography dataset (ver. usgs national hydrography dataset best resolution (nhd) for hydrologic unit (hu) 4 - 2001), 2020.

Ype Van der Velde, Nikki Vercauteren, Fernando Jaramillo, Stefan C Dekker, Georgia Destouni, and Steve W Lyon. Exploring hydroclimatic change disparity via the budyko framework. *Hydrological Processes*, 28(13):4110–4118, 2014.

Jan N Van Rijn and Frank Hutter. Hyperparameter importance across datasets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2367–2376, 2018.

Julie A Vano, Bart Nijssen, and Dennis P Lettenmaier. Seasonal hydrologic responses to climate change in the p acific n orthwest. *Water Resources Research*, 51(4):1959–1976, 2015.

Anoop Valiya Veettil, Goutam Konapala, Ashok K Mishra, and Hong-Yi Li. Sensitivity of drought resilience-vulnerability-exposure to hydrologic ratios in contiguous united states. *Journal of hydrology*, 564:294–306, 2018.

Naga M Velpuri, Gabriel B Senay, Ramesh K Singh, Stefanie Bohms, and James P Verdin. A comprehensive evaluation of two modis evapotranspiration products over the conterminous united states: Using point and gridded fluxnet and water balance et. *Remote Sensing of Environment*, 139:35–49, 2013.

Sergio M Vicente-Serrano, Santiago Beguería, and Juan I López-Moreno. A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *Journal of climate*, 23(7):1696–1718, 2010.

S Vicuna and JA Dracup. The evolution of climate change impact studies on hydrology and water resources in california. *Climatic Change*, 82(3-4):327–350, 2007.

Sebastian Vicuña, Rebecca Leonardson, MW Hanemann, LL Dale, and John A Dracup. Climate change impacts on high elevation hydropower generation in california's sierra nevada: a case study in the upper american river. *Climatic Change*, 87(1):123–137, 2008.

James M Vose, Chelcy Ford Miniat, Charles H Luce, Heidi Asbjornsen, Peter V Caldwell, John L Campbell, Gordon E Grant, Daniel J Isaak, Steven P Loheide II, and Ge Sun. Ecohydrological implications of drought for forests in the united states. *Forest Ecology and Management*, 380:335–345, 2016.

Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651, 2014.

Zhaoli Wang, Chengguang Lai, Xiaohong Chen, Bing Yang, Shiwei Zhao, and Xiaoyan Bai. Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, 527: 1130–1141, 2015.

RH Waring, NC Coops, W Fan, and JM Nightingale. Modis enhanced vegetation index predicts tree species richness across forested ecoregions in the contiguous usa. *Remote Sensing of Environment*, 103(2):218–226, 2006.

Seth J Wenger, Charles H Luce, Alan F Hamlet, Daniel J Isaak, and Helen M Neville. Macroscale hydrologic modeling of ecologically relevant flow metrics. *Water Resources Research*, 46(9), 2010.

Frank Wilcoxon, SK Katti, and Roberta A Wilcox. Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1:171–259, 1970.

David M Wolock and Gregory J McCabe. Explaining spatial variability in mean annual runoff in the conterminous united states. *Climate Research*, 11(2):149–159, 1999.

Chen Yang, You-Kuan Zhang, and Xiuyu Liang. Analysis of temporal variation and scaling of hydrological variables based on a numerical model of the sagehen creek watershed. *Stochastic environmental research and risk assessment*, 32(2):357–368, 2018.

Dawen Yang, Fubao Sun, Zhiyu Liu, Zhentao Cong, Guangheng Ni, and Zhidong Lei. Analyzing spatial and temporal variability of annual water-energy balance in nonhumid regions of china using the budyko hypothesis. *Water Resources Research*, 43(4), 2007.

Meijian Yang, Denghua Yan, Yingdong Yu, and Zhiyong Yang. Spei-based spatiotemporal analysis of drought in haihe river basin from 1961 to 2010. *Advances in Meteorology*, 2016, 2016.

Ning Yao, Yi Li, Tianjie Lei, and Lingling Peng. Drought evolution, severity and trends in mainland china over 1961–2013. *Science of the total environment*, 616:73–89, 2018.

Tianshan Zha, Alan G Barr, Garth van der Kamp, T Andy Black, J Harry McCaughey, and Lawrence B Flanagan. Interannual variation of evapotranspiration from forest and grassland ecosystems in western canada in relation to drought. *Agricultural and Forest Meteorology*, 150(11):1476–1484, 2010.

Xiaohui Zheng, Qiguang Wang, Lihua Zhou, Qing Sun, and Qi Li. Predictive contributions of snowmelt and rainfall to streamflow variations in the western united states. *Advances in Meteorology*, 2018, 2018.