Optimizing spatial distribution of watershed-scale hydrologic models using Gaussian Mixture Models⋆

Tessa Maurer[a,*], Francesco Avanzi[a,b], Carlos A. Oroza[c], Steven D. Glaser[a], Martha Conklin[d] and Roger C. Bales[a,d]

[a]Department of Civil and Environmental Engineering, University of California, Berkeley, 94720, Berkeley, California, USA
[b]CIMA Research Foundation, via Armando Magliotto 2, 17100, Savona, Italy
[c]Department of Civil and Environmental Engineering, University of Utah, 84112, Salt Lake City, UT, USA
[d]Sierra Nevada Research Institute and School of Engineering, University of California, Merced, California, USA

ARTICLE INFO

Keywords:
physically based hydrologic models
spatial distribution
Gaussian Mixture Models
statistical learning

ABSTRACT

Common methods for spatial distribution, such as hydrologic response units, are subjective, time-consuming, and fail to capture the full range of basin attributes. Recent advances in statistical-learning techniques allow for new approaches to this problem. We propose the use of Gaussian Mixture Models (GMMs) for spatial distribution of hydrologic models. GMMs objectively select the set of modeling locations that best represent the distribution of watershed features relevant to the hydrologic cycle. We demonstrate this method in two hydrologically distinct headwater catchments of the Sierra Nevada and show that it meets or exceeds the performance of traditionally distributed models for multiple metrics across the water balance at a fraction of the time cost. Finally, we use univariate GMMs to identify the most-important drivers of hydrologic processes in a basin. The GMM method allows for more robust, objective, and repeatable models, which are critical for advancing hydrologic research and operational decision making.

Highlights

- Gaussian Mixture Models can be used for spatially distributing hydrologic models
- GMM is objective, efficient, and rooted in physical basin characteristics
- GMM meets the performance of traditional models at a fraction of the cost
- Univariate GMMs can be used to identify drivers of hydrologic processes in a basin

tmaurer13@berkeley.edu (T. Maurer)
ORCID(s): 0000-0003-3547-9624 (T. Maurer)

1. Introduction

Spatial heterogeneity of hydrologic processes within a watershed is fundamentally impacted by basin topography. Topographic variations affect vegetation characteristics directly via climatic controls and indirectly via impacts on soil profile and water and nutrient availability (Fan et al., 2020; Tian et al., 2020; Zhang et al., 2011; Qiu et al., 2001). All three – topography, soil, and vegetation – combine to impact hydrologic processes including evapotranspiration, infiltration, runoff, and interflow (see, e.g., Ghestem et al., 2011; Wilcke et al., 2011; Obojes et al., 2015; Young et al., 1997). Both vegetation and topographic variations such as slope and aspect impact snow accumulation and ablation patterns through controls on short- and longwave radiation, wind, and interception (Lundquist et al., 2013; Maxwell et al., 2019; Varhola et al., 2010). In montane regions, heterogeneity of the landscape can have profound implications for all portions of the water balance: orographic effects can create dramatic differences in precipitation rates on either side of mountain ranges as well as influencing the phase (rain versus snow) of that precipitation (Roe, 2005). Landscape variability in these regions is of particular interest due to the role these river basins play in the "waterscape" connecting natural headwaters with human needs (Karpouzoglou and Vij, 2017). These "water towers of the world," supply water to over half of the human population (Mountain Partnership, 2014; Immerzeel et al., 2020; Viviroli et al., 2007a). Understanding the variations of hydrologic processes that contribute to the timing and quantity of streamflow from these basins is a fundamental goal for both scientific researchers and, increasingly, operational forecasters in the water management sector. These questions have become all the more pressing in regions where climate change is inducing shifts in the water balance not previously seen.

In order to meet these needs, and spurred by increases in computational resources, the use of physically based, spatially distributed hydrologic models is becoming more common. Physical models indicate a bottom-up approach in which mass and energy balances are resolved; with a spatially distributed set-up, these simulations are performed at multiple points across a river basin and then aggregated. Models may be partially (or semi-) distributed, where some model components (e.g. input data) are varied across the landscape but others (e.g. parameters) are held constant, or fully distributed (all components are spatially variable). In spatially lumped models, on the other hand, all hydrologic processes occurring within a basin are simulated at a single point and output is given as a single time series for the basin (usually streamflow at the basin outlet).

Much research has focused on calibration approaches for distributed models in an effort to address concerns of overparameterization, non-identifiability of parameters (equifinality) and scale consistency (Beven et al., 1988; Beven, 1989; Wood et al., 1988; Blöschl and Sivapalan, 1995; Andréassian et al., 2012; Bai et al., 2009; Pianosi et al., 2015). Comparisons between lumped, semi-distributed, and fully distributed calibration techniques (e.g. Khakbaz et al., 2012; Reed et al., 2004; Lobligeois et al., 2014; Carpenter and Georgakakos, 2006) have attempted to characterize the relationship between spatial distribution and model performance, with mixed results. While more modeling points can better capture variations in topography, vegetation, and climate, they also introduce a greater number of free parameters, which can contribute to overparameterization issues. A high number of free parameters mean these models can often be calibrated to a baseline level of performance for average conditions regardless of the quality of the input data, but the resulting parameters may have not be reflective of actual physical conditions. As a result, model performance will decline for previously unobserved conditions. The extent to which distributed or lumped models are most appropriate may depend on the landscape and application of the model (Lobligeois et al., 2014). Research on scale consistency focuses on reconciling parameter values across scales in an effort to avoid sudden changes in results when spatial resolution is changed (Sivapalan and Kalma, 1995). These approaches may be top-down (calibrating a lumped model and disaggregating parameter values; e.g. Tran et al., 2018) or bottom-up ("regionalization"; e.g. Blöschl and Sivapalan, 1995; Arsenault and Brissette, 2014; Hundecha et al., 2016; Samaniego et al., 2010).

This work on reconciling parameter values across scales largely focuses on the calibration step of model set-up, but less attention has been given to the prior step of selecting which and how many specific locations within the basin to include in the model. This critical first step of spatially distributing a model impacts all subsequent set-up, including input data distribution across the basin and parameter definition and calibration. Methods proposed in the literature for selecting modeling locations or otherwise partitioning the basin include Representative Elementary Areas (REAs), an intermediate scale at which neither small- nor large-scale processes dominate (Wood et al., 1988); Representative Elementary Watersheds (REWs), units derived based on the streamflow network and over which equations of mass and energy fluxes are integrated (Reggiani et al., 2000; Reggiani and Rientjes, 2005); and landform classes based on the UPNESS index (Summerell et al., 2005; Roberts et al., 1997). These methods showed promise in capturing spatial variability, but were limited by detailed data or catchment-monitoring requirements, inability to simulate multiple hydrologic processes rather than runoff alone, and/or assumptions in the derivation process. More recently, pixel-based distribution approaches have risen in popularity to be compatible with gridded remote-sensing products. Though convenient, this approach is disconnected from the physical characteristics of a basin: pixels may straddle discontinuities in topography or land use, introducing uncertainty into simulations and/or runoff routing. In addition, pixel-based approaches typically result in hundreds or even thousands of simulation points for a moderately sized basin (see, e.g., Tran et al., 2018), since model resolution is frequently dictated by input data resolution. Not only does the high number of modeling locations raise equifinality concerns, these models often have higher simulation times and increase computational requirements. This can be particularly problematic for time- or resource-constrained applications such as real-time flood forecasting. In montane regions, elevation bands are sometimes used as a simple alternative to capture spatial variability (e.g., Bongio et al., 2016; Valéry et al., 2014), but are also often arbitrarily defined and may not align with topographic features.

Alongside pixel-based methods, the most widely used approach for spatially representing a basin is Hydrologic Response Units (HRUs) (Leavesley et al., 1983; Flügel, 1995, 1997), defined as areas of a basin that can be considered homogeneous in all respects influencing the water balance (e.g. topography, land cover and vegetation density, and soil type). Conceptually simple, HRUs are favored by some modelers as having a stronger connection to physical basin characteristics than pixel-based models. HRUs are the default distribution method in several major hydrologic models, including the Precipitation-Runoff Modeling System (PRMS; Markstrom et al., 2015), the Soil and Water Assessment Tool (SWAT; see, e.g., Kalcic et al., 2015; Teshager et al., 2016; Qi et al., 2017), Precipitation-Runoff Evapotranspiration Hydrotope Model (PREVAH; Viviroli et al., 2007b), the Sacramento Soil Accounting Model (SAC-SMA; National Oceanic and Atmospheric Administration, 2002) and the Regional Hydro-Ecological Simulation System (RHESSys; Tague and Band, 2004). In addition, HRUs are used by many large water-management agencies that rely on physical hydrologic models, including California's Department of Water Resources (DWR) and Pacific Gas & Electric (PG&E) energy company.

Despite their popularity, HRU-based distribution presents both theoretical and practical problems. There is inherent tension between having more, smaller HRUs that are more likely to conform to the assumption of homogeneity and the need to reduce unnecessary model complexity. In addition, though HRUs are meant to represent a distributed sub-area of a basin, hydrologic processes are simulated at a particular point, usually the geometric centroid of the HRU. This necessarily limits the points of the basin that can be simulated with HRUs; for example, the geometric centroid will always be lower than a peak or ridge, meaning that the model is likely to miss the highest elevations. HRUs are frequently delineated using a GIS-based approach, starting with a digital elevation model and using topography, including drainage divides, slope, and aspect, to partition the study area (see, e.g., Flügel, 1995, 1997; Koczot et al., 2005). Though tools such as the ArcMap-based tool GIS Weasel (Viger and Leavesley, 2007) have been built to assist with this process, this method of HRU delineation involves significant subjective decision-making, such as selecting minimum HRU size and stream-segment resolution. All of this can translate to multiple days of hands-on work. Other methods have been proposed for HRU delineation, including Khan et al. (2013) and Khan et al. (2016), who overlaid soil and stream-network data on a set of identified landform classes, and Fiddes and Gruber (2012), who used a subgrid sampling method to include the effects of topography in a lumped model. While promising, these approaches both rely on assumptions that are not generalizable across catchments and/or all aspects of the water balance. Ultimately, HRU delineation (and, by extension, selection of modeling locations) involves subjective decisions and significant time investment.

Given these issues, there is need for a simple, rapid, and objective approach to selecting modeling locations for spatially distributed hydrologic models. Recent advances in statistical-learning algorithms have made possible alternative approaches to this problem. Such algorithms, broadly speaking, are used to identify and characterize patterns in data, particularly those that are not obvious or that would be too labor intensive to test individually (Shen, 2018). Use of statistical learning is increasing in the field of hydrology (see, e.g., Oroza et al., 2018; Avanzi et al., 2019; Schmidt et al., 2020; Kim et al., 2020), but understanding which algorithms and for what applications it is most appropriate is an ongoing area of research (Shen, 2018; Kim et al., 2020). For example, how statistical models compare to and interact with traditional physically based hydrologic models has not been comprehensively tested (Oyebode and Stretch, 2019). Physical models are usually mechanistic and process-based by design, since it is frequently important for hydrologists to understand the causes and relationships underlying an observed phenomena. Statistical learning, on the other hand, typically identifies correlations and associations between variables without suggesting causality. The implications of this discrepancy and whether they matter for using these two model types together is an open research question (Oyebode and Stretch, 2019; Schmidt et al., 2020).

Here, we use statistical learning to return to the question of selecting modeling locations for a physically based model, and we propose a method that is grounded in physical properties and does not obscure process understanding. Given the relatively recent introduction of statistical learning to hydrology, there are many possible approaches that have yet to be tested, but as a first step, we focus on mixture models, a type of algorithm that has emerged as a way of optimally identifying a set of underlying components that best describes a population. We propose the application of Gaussian Mixture Models (GMMs) as an objective, efficient, and physically based spatial-distribution technique that addresses both the theoretical and practical shortcomings of existing methods for selecting modeling locations in a basin. Using basin characteristics that influence the water balance, mixture models identify a set of modeling locations that optimally characterize the the water balance throughout the basin. Gaussian Mixture Models have been successfully used to capture spatial patterns in other hydrologic contexts such as snow-water equivalent (SWE) distribution at a single site (Oroza et al., 2016), but have not yet been tested in conjunction with a physically based hydrologic model. This is also the first time it has been tested at landscape scale across the diverse topography of montane river basins.

We demonstrate the GMM-based distribution method in two contrasting headwater catchments of the Sierra Nevada using the Precipitation-Runoff Modeling System (PRMS), a physically based rainfall-runoff model commonly used in water management. Owing to their widespread use by researchers and forecasters, we use a GIS HRU-based PRMS model as a baseline to compare performance of the GMM-based models. In the research reported here, we address the following:

1. What is the measurable impact of a GMM-based spatial-distribution method versus an HRU-based method on predictive accuracy?
2. Are these spatial-distribution methods robust to unobserved, extreme hydrologic events? Which hydrologic process(es) drive improvements or declines in modeled performance?
3. What attributes are the most important drivers of predictive accuracy in montane catchments?

2. Methods and data

2.1. Study Area

We focus on Almanor and the East Branch, two headwater catchments of the North Fork of the Feather River, the northernmost basin of the California Sierra Nevada (see Figure 1). The Feather River is important for water resources and energy production. Pacific Gas & Electric (PG&E), California's largest utility company, operates a series of hydropower plants on the North Fork totaling 740 MW of installed capacity, about 19% of company's overall hydropower portfolio. The basin also drains to Lake Oroville, the primary storage reservoir for the State Water Project operated by the California DWR and serving drinking water and agricultural water needs in the central and southern parts of the state.

As a lower-elevation Sierra Nevada basin (peak elevation 2950 m), the Feather is susceptible to climate change effects as more precipitation falls as rain rather than snow. The Feather River can therefore be thought of as an early example of how other basins in the Sierra Nevada may change with rising temperatures (Freeman, 2011). The main stem of the North Fork of the Feather originates in the Almanor catchment to the northwest and is regulated at the outlet of Lake Almanor. Almanor drains an area of approximately 1150 km^2 and contains Mount Lassen, the highest and wettest point in the Feather River at about 2900 m elevation and 3000 mm of annual precipitation (Koczot et al., 2005). Geologically, Almanor is part of the Cascade Mountain range rather than the Sierra Nevada, making it distinct from the rest of the basin. The subsurface is largely comprised of more-permeable volcanic rocks, and baseflow makes up a higher percentage of flow than in other subbasins (Freeman, 2008).

The East Branch is a tributary of the North Fork and drains an area of approximately 2650 km^2. It meets the North Fork south of Lake Almanor. The East Branch is rain-shadowed due to the eastern ridge of the Upper North Fork Canyon on its western edge and is thus considerably drier than Almanor, with an average annual precipitation of about 300 mm. The subbasin has a largely granitic subsurface and low baseflow (Freeman, 2008). It is also mostly unregulated.

2.2. Gaussian Mixture Models for spatial distribution

Gaussian Mixture Models are a statistical-learning algorithm used to identify a subset of discrete points that best represent a feature space.

Here, "feature" is a measurable characteristic that describes a phenomenon being observed (Bishop, 2006). For example, in describing runoff from a basin, a feature may be the basin's elevational distribution. A "feature space" is the single (if there is only one feature) or multidimensional (if there is more than one) range collectively defined by the feature data. For example, in Figure 2, there are two features, elevation and slope, creating a two-dimensional feature space. All data points fall (shown as dots) somewhere in the feature space. Features must be continuous numeric variables for use in a standard GMM, but otherwise may be defined at the discretion of the modeler.

The GMM algorithm selects the optimal points across the feature space by assuming that the feature space can be represented by superimposing a finite number (M) of "latent components," which are normally distributed. Figure 2a shows three latent components, each of which can be uniquely described by a mean (expected value; μ) and covariance (Σ). Each is also assigned a mixing parameter (π) based on the prior probability of observing that component (essentially, a weighting factor). Thus, the parameters that are defined in fitting a GMM to a particular dataset are the means, covariances, and mixing parameters for each latent component.

We take the expected values of the latent components as the set of points that optimally describes the feature space. In concrete terms, and relating the example schematic in Figure 2a, the three μ values (shown as red X's) are the points that best represent the distribution of elevation and slope in this hypothetical basin. However, a point that exists in feature space (say, for example, an elevation of 2300 m and a slope of 85) may not exist physically in the basin. Thus, once the means have been identified, we use a Nearest Neighbors approach to find the physical location that is closest to the means of feature space (Figure 2b). These locations define the spatial distribution of the GMM-based PRMS models (henceforth, "modeling locations") and are analogous to the HRU centroids that define the spatial distribution of traditional models (Figure 2c). Maps of the actual selected modeling locations for each subbasin are available in the Supplementary Information (Figures S1 and S2).

Formally, the ability of a GMM latent component to represent the feature space is modeled as a multivariate normal distribution, $\mathcal{N}$, with expected value μ and covariance Σ applied to a D-dimensional vector of empirical data $\mathbf{x}$ (equation 1).

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\} \tag{1}$$

The collective ability of the M components to reproduce the feature space is calculated by superimposing each $\mathcal{N}_m$, weighted with its mixing parameter, π_M. The number of features (i.e., the length of $\mathbf{x}$) determines the dimension of each $\mathcal{N}_m$ distribution. The expected values, covariance, and mixing parameters that best represent the data are identified by maximizing the likelihood function of the superimposed multivariate normal distributions, given by equation 2. In other words, the GMM maximizes the following objective function:

$$\ln \mathrm{p}(\mathbf{x}_n|\pi,\mu,\Sigma) = \sum_{n=1}^{N}\ln\left\{\sum_{m=1}^{M}\pi_m\mathcal{N}(\mathbf{x}_n|\mu_m,\Sigma_m)\right\} \tag{2}$$

subject to:

$$\sum_{m=1}^{m}\pi_m = 1 \tag{3}$$

In this study, the number of dimensions was five, and the features were basin elevation, slope, aspect, vegetation coverage, and soil hydraulic conductivity (k_{sat}). Together these features capture the major drivers of the water balance endogenous to the basin (i.e., not driven by climate or weather inputs), including spatial distribution of the snowpack, an important if not dominant component of the hydrologic cycle in the Feather River; evapotranspiration; and infiltration characteristics.

Elevation, slope, and aspect were defined using the USGS USGS National Elevation Dataset (NED; EROS Data Center, 1999), and vegetation data were obtained from the 2013 U.S. Forest Service LANDFIRE dataset (LANDFIRE, 2013a,b), respectively. (See Supplementary Information Section S1.2 for full details on the LANDFIRE dataset). Topographic and vegetation data rasters were both at 30-m resolution. These rasters were masked to the extent of the subbasins in the study and filtered to remove pixels with undefined values (for example, flat areas with undefined aspect) and passed through a 1-in-2 resampling algorithm to make processing computationally feasible (pandas.DataFrame.sample; The pandas Development Team, 2019). Subsampling was performed with a uniform distribution, without replacement, and with an initial seed for reproducibility.

Soil hydraulic conductivity, k_{sat}, was derived from STAGO2 data (Soil Survey Staff Natural Resources Conservation Service, 2019), which are available as shapefiles indicating the extent of different "geologic groups"; i.e., clusters of one or more soil types, each of which is associated with a set of unique soil properties. Properties were first depth integrated, then spatially averaged using the percent of each soil type in a geologic group. This averaged property (i.e., k_{sat}) was assumed to be spatially homogeneous across the geologic group. The distribution of the averaged k_{sat} was used as the GMM input feature. While this is relatively simple as a descriptor of soil type, it demonstrates how soil properties may be incorporated into GMM modeling. Future work, particularly in basins with large contributions of groundwater to streamflow, could include more detailed assessment with multiple soil characteristics as GMM inputs.

Using the five rasters as inputs, the GMM algorithm was run using Scikit-learn's mixture.GaussianMixture class (Pedregosa et al., 2011b). Covariance parameters were trained using the "spherical" setting, meaning that a single covariance value was calculated for each component. This was selected to in order to simplify the analysis and to decrease the runtime of the algorithm, but future work should explore the implications of other covariance options. Since many of the landscape features that control movement of water co-vary, but sometimes to differing degrees across the landscape (Beven et al., 1988), it is possible that applying the different covariance parameters for each component would improve model performance.

The mixture.GaussianMixture class uses the Expectation Maximization algorithm, an iterative gradient descent method, to optimize of Equation (2) and identify the most likely mixing parameters, covariance, and means to explain the data (McLachlan and Peel, 2004; Pedregosa et al., 2011a). The optimization terminates when a maximization step no longer increases the log-likelihood. As noted, the optimal expected values (means) of the latent components in feature space were translated to physical modeling locations using a nearest neighbors algorithm. Features were scaled with equal weight to prevent features with higher magnitude values from dominating the nearest-neighbor search. In addition to these five-dimensional multivariate GMMs, we ran five additional GMMs in each basin, each driven by only one of the features (univariate GMMs). This was done to assess the usefulness of each individual GMM input feature, and we show that this can inform feature selection and their relevance to different hydrologic processes.

GMMs fall into a general category of statistical learning models called clustering algorithms, which aim to find areas of relative homogeneity of data in feature space. While any number of clustering algorithms could also be effective for spatial distribution of modeling locations, we used GMMs for this analysis because every latent component in the GMM is affected by every data point. This is not necessarily true of other clustering algorithms such as k-means, where means are calculated based only on the points assigned to a given cluster. The former thus had a better potential to represent all points across the feature space.

2.3. Precipitation-Runoff Modeling System

The Precipitation-Runoff Modeling System (PRMS) is a distributed-parameter hydrologic model developed by the U.S. Geological Survey (Markstrom et al., 2015). The model runs on the daily time step, taking as inputs daily precipitation and minimum and maximum temperature, which are distributed to each HRU either ahead of time by the modeler or through an interpolation scheme within the model. PRMS simulates mass and energy balances beginning with calculation of solar radiation and precipitation phase partitioning and ending with computation of total streamflow. Intermediate processes include snow accumulation and ablation, canopy interception and evapotranspiration, infiltration, surface runoff, interflow, and groundwater recharge. A description of the major processes included in PRMS and their calculations can be found in the Supplementary Information Section S1.5. PRMS is executed in a linear fashion at each time step in the simulation, with each hydrologic process represented by a module of code. For some processes, users may specify a desired calculation method by selecting from multiple possible modules.

The spatial distribution in PRMS is achieved by partitioning the modeling area into HRUs, represented by a specific geographical point in the basin (by default, the geographic centroid; Koczot et al., 2005; Markstrom et al., 2015) to which input data are distributed and at which the water balance is simulated. The water balance is simulated separately at each HRU and scaled according to the surface area of the HRU. Outflow is aggregated across the basin based on the selected streamflow routing method.

PRMS is currently used throughout the California Sierra Nevada for streamflow modeling by PG&E (Richards, 2018). It is also being actively developed for new river basins by DWR (see, e.g., Burley and Fabbiani-Leon, 2018). Its widespread use for water-resources planning as well as its commonalities with other distributed-parameter models makes it ideal for this study. The model is publicly available at https://www.usgs.gov/software/precipitation-runoff-modeling-system-prms. The latest release is version 5, but at the time of this research, version 4.0.3 was the most updated available. Major changes between the two versions do not affect the modules used in this study.

2.3.1. *PRMS models used in this study*

Required input data for PRMS are daily temperature range and precipitation amount, which are spatially distributed based on a user-selected method. Here, both temperature and precipitation were pre-distributed to each HRU or modeling location before executing PRMS. Precipitation was distributed using an algorithm called DRAPER, in which spatially distributed long-term average monthly precipitation surfaces from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) dataset are adjusted using daily ground-based values (Daly et al., 2008). Each day, the PRISM surface is multiplied by the ground-based measurements as a percent of long-term daily average. This effectively "tilts" the PRISM surface to reflect the daily observations.

For more details on the DRAPER algorithm, see Donovan and Koczot (2019). For a full discussion of the implementation of DRAPER in this study, see Supplementary Information Section S1.4.

Daily minimum and maximum temperature were also distributed externally to PRMS. This procedure was selected based on an analysis of temperature variability across the Feather River basin that showed both temporal (sub-monthly) and spatial (intra-subbasin) variability in lapse rates (Avanzi et al., 2020). Internal PRMS distribution methods did not permit this variability to be accounted for. Instead, temperature was distributed as a two-step process in which we regressed elevation against minimum and maximum temperature for several training stations in each subbasin to establish daily basin-wide lapse rates. Residuals between temperature predicted using these lapse rates and observed temperature at a set of evaluation stations were distributed using multilinear regression with elevation and the temperature at a designated seed station. The final values of maximum and minimum temperature were obtained by subtracting the residual from the first-guess temperature obtained using the lapse rates computed in the first step. For more details on this process, see Supplementary Information (Section S1.4).

In PRMS, some process calculations are predetermined, while for others, the user may select from a variety of options. In this study, solar radiation was calculated using a degree-day approach, which uses daily maximum air temperature to obtain actual daily solar radiation (*ddsolrad_hru* module). Evapotranspiration (ET) was calculated using the Jensen-Haise formulation (*potet_jh*). PRMS also requires values for several dozen parameters that may be spatially and temporally global or may be set on a per-month or per-HRU basis. Values based on topography and canopy cover were computed based on the USGS NED (EROS Data Center, 1999) and the US Forest Service LANDFIRE dataset (LANDFIRE, 2013a,b), respectively. Remaining non-calibration parameters were computed based on available data, set to default values, or retained from the original USGS version of PRMS on the Feather River. Details of this process for both the GIS and GMM versions can be found in the Supplementary Information (Section S1).

In this study, we use a traditionally designed PRMS model as a benchmark for model performance. This model is henceforth referred to as the "GIS" model, in reference to the geographic information system framework typically used to delineate HRUs. The GIS model was based on the Feather River PRMS model (version 2) designed by the USGS in the early 2000s, in which HRUs were delineated using standard methods (see Koczot et al., 2005, for details). We updated the model to PRMS version 4.0.3; as part of this process, we made minor manual updates in the HRU boundaries to better reflect drainage divides. In addition, PRMS version 4 allows for greater functionality in terms of streamflow routing, which was introduced in lieu of straight summation of HRU outflows at each time step as was done in version 2. Details of the new model design, including how non-calibration parameter values were selected, are given in the Supplementary Information (Section S1).

Other than the process for selecting modeling locations, the PRMS set-up for the GMM models was largely the same as for the GIS model. Topographic and vegetation parameters were based on the values at the selected modeling locations. Unlike HRUs, which have each represent a different surface area, each modeling location was assigned an equal portion of the headwater catchment based on the GMM feature rasters. This decision was made in order to test the baseline effectiveness of the GMM, but future work could test the effectiveness of scaling based on the surface areas that are represented by relatively homogeneous areas of feature space. The GMM models did not employ a streamflow-routing method since modeling locations are not coupled with drainage areas. Instead, runoff from each modeling location was summed for each time step. For details on the model design, see Supplementary Information (Section S1).

In order to explore the first two research questions (assessing the accuracy of the distribution methods and robustness to unobserved events), we trained four models for each subbasin with varying numbers of target modeling locations: one (lumped case, for comparison), 50, 100, and 200. This process allowed us to identify how the number of modeling locations changes the performance of the model. In order to address the third research question (identifying important drivers of predictive accuracy), we trained univariate-GMM models (i.e., each using only one of the five variables from the multivariate versions). This allowed identification of the most important driving variables for each subbasin. For this step, models were run for the best-performing number of modeling locations from the multivariate GMMs.

2.4. Calibration and Evaluation Strategy

We used a multi-step, multi-objective method to calibrate the models in order to avoid the overfitting that is common when calibrating on streamflow alone (Hay et al., 2006; Gupta et al., 1998). In order to isolate questions of basin partitioning from issues of distributed-parameter calibration, all calibration targets were basinwide; in other words, we did not calibrate to internal basin gauges or other internal targets. Based on the availability of data (~20-year period of record) and the dominant hydrologic processes in the Feather River, we chose to calibrate on

SWE, ET, and basin outflow. Five calibration steps were used, each with a specific target variable, objective function, and set of calibration parameters. The order of the steps reflects the modeling order of hydrologic processes in PRMS. The objective functions and target variables for each step are as follows: daily RMSE of SWE; cumulative annual bias of ET; daily weighted sum of Kling-Gupta Efficiency (KGE, Kling et al., 2012; Gupta et al., 2009) and Log Nash-Sutcliffe Efficiency (LogNSE Nash and Sutcliffe, 1970) of full-natural flow (FNF); monthly weighted sum of KGE and LogNSE of FNF; and cumulative annual bias of FNF. For any objective functions using the KGE metric, all rows with missing observations were removed before the calculation. Daily FNF values were weighted between two metrics in order to capture the performance for both high and low flows. The objective functions for each step are listed in Table 1.

Daily SWE maps published by Margulis et al. (2016) were masked to the Almanor and East Branch subbasins and then aggregated to catchment-wide values. Annual distributed evapotranspiration data were calculated following Roche et al. (2020) on a 30-m basis and were aggregated to catchment-wide values. Finally, daily runoff values for the outlet of each subbasin were computed from FNF values provided by PG&E. FNF is a reconstructed time series of runoff that would have occured in the absence of diversions or other human activity. Uncertainties in sensor readings can result in negative FNF values, so the values for the period of record were smoothed using a five-day moving-average window. Any FNF values that were still negative after the smoothing were masked to NaN.

PRMS, like many large-scale hydrologic models, has hundreds of parameters available for calibration, some of which may be individually calibrated for different months of the year or on a per-HRU-basis. We selected calibration parameters on the basis of previous sensitivity analyses of PRMS (Markstrom et al., 2016) and the availability of informative target data on which to calibrate them (Avanzi et al., 2020). While some parameters are calibrated separately for different months, all are spatially lumped. The parameters calibrated at each step are presented in Table 1. Phase partitioning parameters *tmax_allsnow*, *tmax_allrain*, and *adjmix_rain* (which determine the percentages of precipitation that falls as rain and snow) were calibrated to basin-wide SWE. In addition, we calibrated *freeh2o_cap*, the free-water holding capacity of the snowpack. Subsurface parameters that are related to how much water is directed to the capillary soil layer, were calibrated to ET. The capillary layer is the only soil layer from which ET can occur, so these parameters govern the largest source of ET in the basin (transpiration by plants). Though this neglects parameters related to evaporation from intercepted storage or sublimation from snow, these contributors to ET are much lower than transpiration by plants from soil storage. Finally, other subsurface parameters governing overland flow, interflow, and groundwater flow were calibrated to streamflow at various time steps (daily, monthly, and annual), reflecting the timescales over which we expect these processes to occur. For details on the use and physical meaning of these parameters, see Markstrom et al. (2015).

Based on availability of input and calibration data, the total calibration period included water years 1998-2016. Instead of a traditional split-sample approach using continuous calibration and validation periods, a stochastic, multi-split process was employed to avoid biases that might arise from arbitrarily selecting a calibration period. Eleven of the nineteen water years in the period of record were randomly selected for calibration, a process which was repeated to give five 70%/30% calibration/validation splits (see Table S2 in the Supplementary Information for specific years used for each split).

All calibrations were performed using the Shuffled Complex Evolution (SCE) algorithm, a well-established method that was specifically developed for large hydrologic models with many degrees of freedom (Duan et al., 1993, 1994). This algorithm was designed to handle arbitrary objective functions, differences in parameter sensitivities, and the presence of non-optimal local minima in the feasible space. In brief, SCE works as follows: randomly selected sample points are partitioned into complexes, which each evolve separately, allowing the parameter space to be explored more efficiently. Periodically, complexes are combined, shuffled, and re-partitioned into new complexes. This step allows for sharing of the information gained separately by each complex as it evolved. The algorithm stops when one of several possible specified convergence criteria is met (see lines marked with asterisks in Table S1 in the Supplementary Information). This process is performed sequentially for each of the calibration steps as listed in Table 1. One iteration through all calibration steps is a “calibration round”. Users may set multiple calibration rounds; here, we used five. Thus, calibration is both sequential and iterative; while not guaranteed to reach a Pareto front or other global optimum, this accounts for trade-offs between the objective functions and prevents overcalibration to any single objective function. The various metaparameters of the SCE algorithm, including number of complexes, number of calibration rounds, and convergence criteria, may be individually set for each step and were selected based on a combination of suggested default values (Hay and Umemoto, 2006) and trial-and-error. Table S1 in the Supplementary Information lists the metaparameters used for each step and a short explanation of how they were chosen. More details of the SCE algorithm can be found in (Duan et al., 1992, 1993, 1994).

Each GIS or GMM model was calibrated five times according to each of the splits and each calibration was assessed separately across its validation period. The metrics used in model-performance assessment were daily absolute bias of SWE, cumulative-annual absolute bias of ET, and daily KGE, LogNSE, and root-mean-squared error (RMSE) of streamflow. RMSE gives more weight to accurately reproducing peaks in the time series, while LogNSE gives more weight to baseflow periods. It should be noted that SWE bias was calculated across all pixels, including those without snow, to be consistent with the PRMS model's calculation of basin-wide metrics. The performance of the models relative to observed values may therefore appear artificially good, but values are appropriate for comparison across models. Metrics and scores values are shown in figures in the main text as absolute values. Signed scores are reported in the Supplementary Information Tables S3 and S4.

To allow for comparison across metrics with different units, all twenty-five calibrations (five for each of the five models) were ranked for each metric. Then, the average rank across all five metrics for each calibration was calculated and the median average rank was the "score" of that model. Unless otherwise noted, all performance values reported are for the calibration with the median average rank.

Other performance metrics focusing on conditions that are of particular interest to forecasters and modelers, such as peak SWE and flood periods, are also presented, but were not used in scoring model performance. These include the Relative Error of High Flows (REHF; Silvestro et al., 2018) to assess reproduction of peak flows, peak SWE error, date of peak SWE, and baseflow error. Peak flows were identified as the top 5% of flows across the period of record and days of peak flow were the days these occurred. Only the peaks occurring within a calibration's validation period were used to calculate REHF. All other metrics were calculated separately for each validation year and averaged.

3. Results

This section presents results on GIS and GMM model performance. Section 3.1 gives results under average conditions (i.e., metrics computed across validation periods), which were used to rank overall performance of the models. It also discusses computational resources required to run the GMM algorithm. Next, we present further analysis aiming to verify GMM performance during periods of extreme conditions (Section 3.2). Finally, we present the results of the univariate GMM models, demonstrating how these can inform feature selection for different parts of the water balance (Section 3.3).

3.1. GMM model set-up and performance

In both subbasins, the GMM model performance is comparable to and, in some cases, better than the GIS models with respect to all metrics calculated for the validation periods (Figure 3). The median average rank (Figure 3a) shows that the GIS model is best-performing (lowest ranked) in Almanor and the 200-location GMM model is best in the East Branch. The best-performing GMM model in Almanor is the 100-location one.

Within a subbasin, there is little variability in performance across models, including both GIS and GMM models. SWE and ET bias in both catchments, for example, vary by less than 20 mm. Runoff RMSE is variable between the two catchments, reflecting the difference is average flow, but models for a given catchment show similar performance. However, no single model consistently performs best for all components of the water balance. Performance rankings are more consistent in the East Branch than Almanor, particularly with respect to streamflow metrics. Here, the 200-location GMM generally performs best with the exception of SWE. In both subbasins, lumped GMM model, which represents a baseline from which to assess the improvement provided due to any type of spatial distribution, ranks somewhere in the middle with respect to SWE and ET performance, but consistently performs poorly with respect to streamflow.

In the East Branch, overall ranking consistently improves with higher numbers of GMM modeling locations, but in Almanor, this is only true up to 100 modeling locations. Thus, the best-performing GMM uses half the number of modeling locations in Almanor than the East Branch (Figure 3a), possibly reflecting the difference in catchment area (the area of the East Branch is approximately double that of Almanor). The optimal number of modeling locations may also be influenced by the input-data raster resolutions. Further discussion of the implications of this result can be found in Section 4.1.

The computational times required to run the GMM algorithm were between 80 and 1500 seconds on a single core of a high-performance computing cluster (3.9 GHz; Figure 4). This time includes all steps from raster sub-sampling, scaling, fitting the GMM model, and the nearest-neighbor search for the physical modeling locations. Times varied by subbasin (East Branch models took longer to run than Almanor due to the larger raster size) and number of components.

3.2. Extreme and peak periods

As with average-condition metrics, the GMM-based approach yields comparable, if not better, performance than the GIS method during extreme periods (Figure 3b).

In particular, the GMM models that perform best in each subbasin under average conditions (100-location in Almanor and 200-location in the East Branch; Figure 3a) also match or exceed GIS performance during extreme periods. We note that the lumped model performs worse in both subasins with regard to extreme periods than the average conditions, which is further discussed in Section 4.1.

Peak SWE marks the transition from accumulation to ablation season in the Sierra; both timing and magnitude of peak SWE are important seasonal benchmarks in snow-dominated basins. On average, peak SWE was better simulated in Almanor, but day of peak SWE was better simulated in East Branch. Peak SWE tended to be overestimated in both basins (see Table S4 in the Supplementary Information). Day of peak SWE was estimated later than observed in the East Branch; there was no consistent pattern in Almanor.

July-September flow was used to capture baseflow performance, and REHF was used to assess peak flows. Like the SWE metrics, performance was better for both in Almanor. For baseflow in particular, observed flows are lower on average in the East Branch, making the performance in Almanor even more comparatively strong. Peak flows were generally challenging to capture in both subbasins, with minimum REHF of 0.3 (100-location GMM in Almanor). Though there was little consistency in model rankings between baseflow and peak flows, the best-performing GMM models were again able to meet or exceed GIS performance.

3.3. Univariate GMMs

Univariate GMM performance was generally worse than multivariate models (Figures 3a and 5), though metrics fell within the same order of magnitude, reflective of a baseline level of performance that can usually be achieved for average conditions by calibrating models with a high number of free parameters. In the East Branch, the elevation-driven GMM performed best (as based on median average rank) followed by (in order) the models driven by slope, aspect, saturated hydraulic conductivity, and vegetation, with median rank of slope-driven model only slightly larger than elevation-based model. In Almanor, models in order of performance from best to worst were based on slope, aspect, saturated hydraulic conductivity, elevation, and vegetation. Slope performed particularly well with respect to runoff RMSE and KGE, while aspect performed better with respect to SWE bias and runoff LogNSE. Possible reasons for the differences between the two subbasins are discussed in Section 4.4.

4. Discussion

The GMM method has a number of significant logistical and theoretical advantages compared to GIS models, but its usefulness is still contingent on its ability to replicate the performance standards of traditional models. In the following sections, we first discuss both the performance of the GMM models and under what conditions they are able to meet or exceed the performance of traditional models (Section 4.1). Next, we describe the logistical advantages of GMM and the implications for users of hydrologic models (Section 4.2). In the final two sections of the discussion, we look more specifically at the spatially distributed performance of the GMM models (Section 4.3) and the performance and use of univariate GMM models (Section 4.4).

4.1. Model performance

The GMM method provides a sound, objective basis for spatially distributing hydrologic models. GMM-based models match the performance of the GIS models under both average and extreme conditions (Figure 3). They accurately simulate water-balance components over time and provide more-accurate spatial distribution of streamflow generation than do GIS models. We focus this discussion on the overall performance of the GMM models as indicated by the model score, since, in most cases, users need and expect models to perform well across all components of the water balance in order to ensure that physical processes are being accurately simulated. However, there may be cases in which modelers would have more specific needs, for which metrics related to a certain component like SWE or ET would be more informative.

The best-performing GMM model based on model score exceeds the performance of the GIS model in the East Branch, but not in Almanor (Figure 3a). Thus, the variables selected for GMM prediction may be more relevant for water-balance partitioning in the East Branch than Almanor due to the particular hydrology of the catchments. Almanor is more subsurface-dominated than the East Branch, and saturated hydraulic conductivity, the only predictor used here related subsurface conditions, is a relatively limited characterization of soil and groundwater flow. Thus, the variables used in this study were likely a more-complete characterization of hydrologic processes in the East Branch. Another factor may be the relative importance of the GMM variables: the most-informative variables in Almanor, based on the performance of the univariate models, were aspect and slope (Figure 5). The GIS and GMM distributions of these factors were similar and largely consistent with the raster values (see Supplementary Information Figure S4). The most-informative variable in the East Branch was elevation. Here, the GMM models span a greater range than the GIS distributions, which may have contributed to improved performance. In addition, the overall lower elevation of the East Branch and the fact that its highest elevations are rain-shadowed means that a greater proportion of the precipitation in East Branch falls in the rain/snow transition than in Almanor (40% versus 33%, where the rain/snow transition is defined between 1300 and 2200 m; Cui et al., 2020). This means that uncertainties in modeling precipitation phase (well reported in the literature; e.g. Harpold et al., 2017; Jennings and Molotch, 2019; Feiccabrino et al., 2015) will affect the East Branch more than Almanor. Thus, the East Branch may be more sensitive to tuning the elevational distribution of modeling locations than Almanor; in other words, there may be more potential for the GMM approach to improve results.

In addition to capturing temporally averaged metrics, the GMM models demonstrated the ability to accurately reproduce periods of extreme or peak conditions. Good performance during extreme flow periods is particularly important for applications like flood forecasting, but increasingly necessary for all streamflow modeling as climate change increases year-to-year variability and induces more severe weather events. Day of peak SWE, for example, has traditionally been estimated as April 1st (Montoya et al., 2014). This estimate has always been uncertain due to seasonal weather characteristics and elevation effects, but is becoming increasingly inexact due to climate-change-induced shifts in precipitation (Margulis et al., 2016). Thus, it is valuable for forecasters to be able to model the date of peak SWE rather than relying on the April 1st estimate. The ability to reproduce baseflows – and, correspondingly, low-flow periods – is also of greater concern as length and severity of dry periods in arid regions are projected to increase (Williams et al., 2020; Woodhouse et al., 2010; Cayan et al., 2010). In each of these cases, the best-performing GMM model is able to meet or exceed the performance of the GIS model. Furthermore, there was consistency between GMM models that performed well under average conditions and those that performed well under peak conditions, meaning that forecasters would not need to rely on a separate model for extreme periods. Since models calibrated to average periods do not always work well under extreme conditions (see, e.g., Vaze et al., 2010), the consistent performance of the GMM-based models is a significant advantage.

Finally, the relative performance of the GMM models varied between Almanor and the East Branch. In both subbasins, the lumped GMM model had the worst overall performance, particularly with regards to streamflow, but the best-performing GMM models differed between the two catchments. The uniformly poor performance of the lumped models reflects the added value of spatially distributed models in diverse topography, where a single location is not sufficient to capture the variations across the landscape that impact the hydrologic process. Since lumped model performance is especially poor for streamflow metrics, the variations the lumped models fail to capture may be related in particular to uncertainties in the simulation of subsurface processes (e.g. model structure and lack of data). Moreover, the lumped models perform even worse with respect to other models during extreme periods, which may be reflective of the inability of these models to capture variations across the landscape that are especially important for extreme periods (for example, rapid snowmelt from high elevations that contributes to flooding). For forecasters in particular, accurate modeling of streamflow and the ability to capture both flood and drought conditions is imperative to optimize dam operations and to protect infrastructure and communities downstream. Our results are consistent with other studies (e.g., Lobligeois et al., 2014) that have shown that spatially distributing hydrologic models can yield significant improvements over lumped models in basins with heterogeneities climatic inputs. This an argument in favor of performing the additional steps required to run the GMM algorithm and so obtain a spatially distributed model for montane regions.

Unlike the lumped models, the spatially distributed models do not show this uniform drop in performance, but there is evidence that adding more modeling locations may only be helpful to a point. In Almanor, the 100-location GMM model outperforms the 200-location GMM, which could be explained by equifinality problems in the 200-location mdoel outweighing the information gain from the additional modeling locations. This pattern is not seen in the larger East Branch, where the 200-location GMM is the best-performing model. As noted in Section 3.1, the optimal number of modeling locations (at least among the models tested in this study) represent approximately the same surface area in each subbasin, since the East Branch is approximately double the area of Almanor. Thus, it is possible that for this particular model and resolution of input data, these catchments are essentially reaching a saturation point for modeling locations, though more tests would need to be run on the East Branch in particular to determine if performance drops with more locations. The impact of number of modeling locations on model performance, while not drastic, is enough to support further attention being given to the distribution step (i.e., selection of modeling locations) of model set-up.

4.2. Modeling set-up

A key advantage of the GMM method is its efficiency and repeatability, especially when compared to traditional methods of HRU delineation. GMM requires only rasters of input variables, thus combining the data-processing advantages of pixel-based models while still being based in physical basin characteristics. Once the rasters are prepared, running the GMM algorithm from start to finish, including subsampling and saving the outputs, required less than half an hour on a high-performance-computing core. As long as a seed is set in the random number generator for the GMM optimization, the process is also repeatable. While the GMM method does not address questions of scalability of parameters or automatically identify optimal resolutions, this efficiency can be leveraged to test multiple spatial resolutions and allow modelers to understand how the resolution influences their results. The ease of setup also allows modelers to test different combinations of input variables to their model and understand the drivers of hydrologic processes in their basin. This can help inform what variables to use as inputs to the GMM. The GMM algorithm requires no specific software and can be implemented through open-source products, as was done for this study. There are no theoretical limits on modeling locations using GMM and all locations necessarily represent equal areas to comply with the multivariate selection process. This removes questions of relative HRU size and decisions about the maximum range of HRU areas. Finally, since the GMM method is separate from calibration, it can be applied for any number of calibration designs, including different algorithms, single- or multi-objective functions; and semi- or fully distributed parameters.

Traditional HRU delineation, on the other hand, is necessarily subjective: delineation usually begins by identifying areas with similar topography using a DEM, but there are few norms or guidelines to selecting the number of HRUs to use (and, by extension, their average size) other than the resolution of the input data and the computational power available to run the model. Once initial HRUs are delineated, smaller HRUs are generally merged into neighboring larger ones so sizes fall within a similar range; which HRUs to merge and where is entirely subjective. Some common software tools including GIS Weasel automatically and randomly merge smaller HRUs, but do not contain the ability to set the seed of their random number generators, making this process impossible to replicate (Viger and Leavesley, 2007). Though the time required for GIS-based HRU delineation is not consistent, since it depends on the size and topography of each basins, our experience in this study and conversations with modelers and forecasters suggest the process is on the order of days to weeks. Since the spatial-distribution process is so labor intensive, the assumptions made during this process cannot be easily tested by creating alternative versions of the model.

GIS-based HRU delineation also presents theoretical problems, including the fact that hydrologic processes are simulated at the geometric center of a supposedly homogeneous HRU. This means that extreme elevations will never be represented by the model, potentially missing areas that are significant contributors to runoff production. Another issue is that if an HRU is not convex, its geometric center is not guaranteed to fall within the HRU or even within the river basin itself.

Given these issues, more attention should be paid to novel methods of spatial distribution in hydrologic models. Though related to issues of scalability and overparameterization, this has not received the same attention in the literature. Future work should focus on further exploring how GMMs interacts with physically based models; for example, one area that was outside the scope of this study but would be highly useful for modelers is a method for model selection (i.e. for determining *a priori* the optimal number of modeling locations) to couple with GMM. Tools like information criterion may provide some insights, but since these are usually calculated with respect to the statistical algorithm itself, they may not capture the effects of calibration and equifinality in the physically based models (see Section 4.1). In addition, as noted in Section 2.2, the GMM method is only one of several clustering algorithms that may be appropriate for this, so another next step is to explore the use of other related algorithms.

4.3. Spatially distributed performance

Optimized GMM modeling locations also lead to more realistic spatial representation of the basin in PRMS, which has implications for model interpretation and distributed performance. For example, the extent of the elevations represented in the GIS model is less than half the true range (the GIS model covers 1391 to 2155 m in Almanor and 1103 to 2001 m in the East Branch, while the range of the DEMs is 1365 to 2950 m in Almanor and 700 to 2550 m in the East Branch). Due to this limited elevational range, all hydrologic processes in the GIS model can occur only up to 2200 m in Almanor and between 1100 and 2100 m in the East Branch. Compared to the GIS models, the best-performing GMM models cover a 42% greater range in Almanor (1406 to 2492 m) and 38% greater range in the East Branch (1016 to 2263 m). Though elevation gradients are only one of many types of spatial heterogeneity, they are particularly relevant due to the strong orographic influence on precipitation in our study area (Roe, 2005; Roe and Baker, 2006). Their greater range means the GMM models are better positioned to capture processes with strong elevational dependence, including SWE distribution, vegetation, and timing of runoff generation. This finding applies to these study sites in particular, but based on the theoretical limits on HRU elevations as discussed in Section 4.2, we expect the GMM method to give broader elevational representation than the GIS method in any other montane catchment.

The implications of using GMM versus GIS for spatial distribution are clear when we examine elevational trends in model performance (Figures 6 and 7). In Figure 6, average daily bias shows how well the models match overall volume at different elevations, while the Pearson correlation coefficient shows how well temporal patterns are simulated. Bias here may be driven by two factors: 1) errors in data or modeling assumptions or 2) biased elevational distribution of area in the model. We see that the ET correlation in Almanor shows a clear "U-shaped" pattern across all models, while the ET bias starts moderately negative, decreases after 1500 m, and rises again until about 2100 m, where it becomes consistently positive. Since these general trends are common across all models, it is likely that these errors are related to problems with input data and/or to model-structural error. The better performance of all models at the lowest elevations may be related to vegetation patterns. Grass and bare lands are more common at low elevations, while forests dominate the middle elevations. Since only forests intercept snow, model structural errors in the snow interception and sublimation calculations would lead to errors in calculating ET. Another possibility is that input climate data, which also impact evapotranspiration calculations, are more accurate at lower elevations, where most data collections stations are located. In particular, the steady increase in R-value and reduction in bias magnitude between 1500 and 2100 meters suggests biases that are correlated with elevation. Input temperature data, which was calculated using seed stations and lapse rates, may show such a bias and would impact SWE representation in the model. SWE, in turn, interacts with ET by influencing rates of sublimation during winter and the timing and amount of water available for transpiration in the spring and summer growing season. At the highest elevations (above 2300 m), the GMM models reveal a significant drop-off in correlation values (Figure 6a). Notably, this drop in correlation performance occurs at about the same elevation above which the ET bias of the Almanor models becomes consistently positive (Figure 6b). We hypothesize that these patterns are related to ET modeling above and below the tree line, which since the highest portions of Almanor, including Mount Lassen above about 2400 m, are largely free of vegetation. Elevations below the tree line are transpiration dominated, while those above the tree line are evaporation dominated; thus, ET calculations in PRMS appears to underestimate the transpiration component and overestimate the evaporation component. In the evaporation-dominated higher elevations, structural issues may include estimating sublimation from the snowpack or evaporation from soil storage. Below the tree line, bias may be related to underestimation of the depth of the root zone or other problems with subsurface modeling. Importantly, this pattern at high elevations is not captured by the GIS model, which does not capture any location above the tree line, nor is it seen in the East Branch, where elevations do not exceed 2300 m (see Supplementary Information Figure S3).

Simulation of the distribution of runoff production across elevations also benefits from the broader spatial range in the GMM models (Figure 7). Due to misrepresentation of area per elevation band, all models tend to overproduce runoff at mid-to-low elevations (1200 to 1600 m in the East Branch) as compared to observed precipitation minus evapotranspiration. (P-ET is a first-order estimate of runoff production, which is not directly observable by elevation band.) Most models show an overestimation at mid-elevations that compensates for underestimations at higher elevations (particularly elevations not represented at all where, by default, runoff production is zero), in order to match overall runoff volume. This error is greatest in the GIS model, which represents the narrowest range of elevations of any model. The 200-component GMM model, which performed best with respect to the spatially lumped metrics, also shows the overall best match with observed P-ET and, as such, would be the best candidate to represent the spatial distribution of runoff production in PRMS.

This misrepresentation of contributing area may lead simulated runoff to interact with other water-balance components in non-physical ways. Over- or under-generation of runoff may lead to errors in partitioning infiltration versus runoff, potentially impacting ET simulation since the majority of ET in vegetated areas is transpiration from soil storage and generally receives priority allocation of runoff over streamflow (Bales et al., 2018). Moreover, misrepresentation of contributing area may lead to particularly poor representation during extreme periods like drought. It has been shown that lower elevations of some Northern Sierra basins may become water limited during droughts, even as the basin as a whole is energy limited. Thus, failure to simulate these lower elevations may lead to the models overestimating runoff during droughts; on the other hand, failure to capture the higher elevations may mean the models miss an important drought mitigation factor (Bales et al., 2018).

4.4. Univariate models

Univariate models performance metrics fell within the same order of magnitude as multivariate model metrics for average conditions (again, reflective of equifinality challenges in hydrologic models), but their performance is still measurably worse. This drop in performance is unsurprising since the multivariate models capture more the factors that influence the water balance in montane catchments (Figure 5). Thus, univariate models would be largely inappropriate for forecasting or process simulation, but we propose that they can provide insights into the most important drivers for different hydrologic processes in the basins. For example, the top-performing univariate model in the East Branch is elevation-based. Since the East Branch sits largely in the rain/snow transition zone, elevation is a critical factor for determining runoff timing by way of precipitation phase. However, both the aspect and slope models performed better than the elevation model with regard to overall SWE bias, suggesting that the rain-shadowed nature of the East Branch and strong directional precipitation patterns are important factors in influencing accumulation and ablation. In addition, slope and aspect may influence the timing and shape of the SWE ablation curve, since they influence the amount of incident solar radiation a site will receive (Maxwell et al., 2019).

In Almanor, the top-performing GMM models were based on slope and aspect, followed by those based on saturated hydraulic conductivity, elevation, and vegetation. Almanor is a higher-elevation basin than East Branch with more area above the rain/snow transition zone. Thus, elevation may be less informative since precipitation phase is more consistent than in the East Branch. In these high-elevation regions, snowmelt starts later in the season and is radiation-dominated (Bales et al., 2006), so slope and aspect are greater controls on the timing of snow accumulation and melt and, by extension, runoff. In addition, baseflow fed by groundwater is a larger component of streamflow in Almanor than the East Branch, so soil characteristics (i.e., saturated hydraulic conductivity) may be more relevant for determining streamflow.

The relatively good performance of the aspect-driven models in both catchments may be due to rain-shadowing effects: direction of slope matters not only for snow ablation due to solar radiation but also for snow accumulation, since the wettest parts of both basins are the western-facing, non-rain-shadowed portions along the main stem of the North Fork. The saturated-hydraulic-conductivity-based model performed reasonably well in both subbasins, but, as expected, topographic features were still overall most relevant for runoff generation. The vegetation-density model gave poor results across both basins, indicating that subsurface conditions may be stronger drivers of ET variation in the Feather River. Since the majority of the land cover in both subbasins is forest, vegetation density may be less informative due to relatively little variation across the landscape.

The univariate models and their differing performance in each subbasin demonstrate the physical basis of GMM-based models, showing how some factors exert more control on the hydrologic process than others depending on the subbasin. We further show how these results can lead to greater processes understanding in headwater catchments. We suggest that univariate GMMs could be used in practice to assess the most-relevant input features before running a multivariate GMM to distribute a new model. This is relevant for both modelers and scientist seeking to improve forecasting performance, prioritize data collection, and better understand the hydrologic cycle. At this stage, selecting this initial set of input features to test is left to the expert knowledge of the modelers. A more rigorous assessment is outside the scope of this study, but options such as the use of an information content criterion should be the subject of future work.

5. Conclusion

We introduce a new method for spatial distribution of hydrologic models using the Gaussian Mixture Models (GMM) algorithm and demonstrate its use in two geologically distinct headwater catchments of the Sierra Nevada. Unlike traditional GIS-based methods, the GMM method is objective, repeatable, and computationally fast (on the order of minutes). The method identifies the set of modeling locations that best represent the basin as a whole, leveraging an efficient statistical-learning tool while being grounded in physical basin properties. Analysis shows that GMM-based models are able to match or exceed the performance of traditional, GIS-based models with respect to both average and extreme conditions for both streamflow and other water-balance components. Furthermore, we show that the modeling locations selected using GMM better represent the geometry of the basin and thus more accurately reproduce the spatial distribution of processes such as runoff production. Thus, GMM-based models are closer to being "right for the right reasons" (Kirchner, 2006). Finally, we show how the method can be adapted to test multiple feature combinations and identify the relative importance of a basin's hydrologic drivers. An elevation-based GMM model performed best in the study basin that sits primarily on the rain/snow transition zone, while slope- and aspect-based models performed best for the higher-elevation catchment. Further research should investigate how different input components, climates, and topographies influence GMM performance. The best model performance in the two headwater catchments where this method was tested was achieved with different numbers of modeling locations (100 in Almanor and 200 in the East Branch). However, these numbers are likely basin-specific, so future work should consider methods for identifying the optimal number of modeling locations.

The improved spatial representation of GMM-based hydrologic models create a more-robust decision-making and process-understanding tool for water-supply agencies, utility companies, and flood-control operators, especially in topographically heterogeneous basins. This can help mitigate risk and reduce costs to downstream users, residents, and infrastructure. In addition to enhancing models directly, the efficiency of the GMM method can facilitate improvements by encouraging more-regular model upgrades. This allows agencies to stay abreast of changes to their basins, such as land use or vegetation coverage, and advances model structure and data collection technology. The resource- and time-intensive nature of updating hydrologic models mean that many agencies do so infrequently, often with more than a decade between upgrades. The rapid and repeatable nature of the GMM method would reduce time and labor associated with model updates. Overall, the GMM method provides the basis for objective, efficient, process-based model set-up with the same capabilities as traditional semi-distributed models. Leveraging advances in statistical learning, it is a powerful and promising new tool for hydrologic modeling.

Acknowledgements

This work was partially supported by the California Energy Commission under contract EPC-14-067 (T.M. and F.A.) and the National Science Foundation Graduate Research Fellowship under grant DGE 1106400 (T.M.). The authors would like to thank Kevin Richards, Joseph Rungee, and Qin Ma for their assistance with data collection and cleaning.

References

V. Andréassian, N. Le Moine, C. Perrin, M.-H. Ramos, L. Oudin, T. Mathevet, J. Lerat, and L. Berthet. All that glitters is not gold: the case of calibrating hydrological models. *Hydrological Processes*, 26(14):2206–2210, 2012. doi: 10.1002/hyp.9264.

R. Arsenault and F. P. Brissette. Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. *Water Resources Research*, 50(7):6135–6153, 2014. doi: 10.1002/2013WR014898.

F. Avanzi, R. C. Johnson, C. A. Oroza, H. Hirashima, T. Maurer, and S. Yamaguchi. Insights Into Preferential Flow Snowpack Runoff Using Random Forest. *Water Resources Research*, 2019. ISSN 19447973. doi: 10.1029/2019WR024828.

F. Avanzi, T. Maurer, S. D. Glaser, R. C. Bales, and M. H. Conklin. Information content of spatially distributed ground-based measurements for hydrologic-parameter calibration in mixed rain-snow mountain headwaters. *Journal of Hydrology*, 582:124478, 2020. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2019.124478.

Y. Bai, T. Wagener, and P. Reed. A top-down framework for watershed model evaluation and selection under uncertainty. *Environmental Modelling & Software*, 24(8):901–916, 2009. ISSN 1364-8152. doi: 10.1016/j.envsoft.2008.12.012.

R. C. Bales, N. P. Molotch, T. H. Painter, M. D. Dettinger, R. Rice, and J. Dozier. Mountain hydrology of the western United States. *Water Resources Research*, 42(8):1–13, 2006. ISSN 00431397. doi: 10.1029/2005WR004387.

R. C. Bales, M. L. Goulden, C. T. Hunsaker, M. H. Conklin, P. C. Hartsough, A. T. O'Geen, J. W. Hopmans, and M. Safeeq. Mechanisms controlling the impact of multi-year drought on mountain hydrology. *Scientific Reports*, 8(1):1–8, 2018. ISSN 20452322. doi: 10.1038/s41598-017-19007-0.

K. Beven. Changing ideas in hydrology - the case of physically-based models. *Journal of Hydrology*, 105:157–172, 1989.

K. Beven, E. Wood, and M. Sivapalan. On hydrological heterogeneity - catchment morphology and catchment response. *Journal of Hydrology*, 100:353–375, 1988.

C. Bishop. *Pattern recognition and machine learning*. Springer, Berlin, 2006. ISBN 0-387-31073-8.

G. Blöschl and M. Sivapalan. Scale issues in hydrological modelling: a review. *Hydrological Processes*, 9(3-4):251–290, 1995. ISSN 10991085. doi: 10.1002/hyp.3360090305.

M. Bongio, F. Avanzi, and C. De Michele. Hydroelectric power generation in an Alpine basin: future water-energy scenarios in a run-of-the-river plant. *Advances in Water Resources*, 94:318–331, 2016. ISSN 0309-1708. doi: 10.1016/j.advwatres.2016.05.017.

N. Burley and A. Fabbiani-Leon. PRMS San Joaquin Model Update. In *California Snow Surveys 64th Annual Meeting of Cooperators*, Kings Beach, CA, 2018. California Cooperative Snow Survey Program. URL `http://cdec.water.ca.gov/snow/meeting/2018/Wednesday/17_PRMS_SanJoaquin_Model_Update.pdf`.

T. M. Carpenter and K. P. Georgakakos. Intercomparison of lumped versus distributed hydrologic model ensemble simulations on operational forecast scales. *Journal of Hydrology*, 329(1):174–185, 2006. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2006.02.013.

D. R. Cayan, T. Das, D. W. Pierce, T. P. Barnett, M. Tyree, and A. Gershunov. Future dryness in the southwest US and the hydrology of the early 21st century drought. *Proceedings of the National Academy of Sciences*, 107(50):21271–21276, 2010. ISSN 0027-8424. doi: 10.1073/pnas.0912391107.

G. Cui, R. Bales, R. Rice, M. Anderson, F. Avanzi, P. Hartsough, and M. Conklin. Detecting rain-snow transition elevations in mountain basins using wireless-sensor networks. *Journal of Hydrometeorology*, 21(9):2061–2081, 2020. doi: 10.1175/JHM-D-20-0028.1.

C. Daly, M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, March, 2008. doi: 10.1002/joc.1688.

J. M. Donovan and K. M. Koczot. User's Manual for the Draper Climate-Distribution Software Suite with Data-Evaluation Tools. Technical report, U.S. Geological Survey Techniques and Methods 7-C22, 2019. URL `https://doi.org/10.3133/tm7C22`.

Q. Duan, S. Sorooshian, and V. Gupta. Effective and Efficient Global Optimization. *Water Resources Research*, 28(4):1015–1031, 1992. doi: 10.1029/91WR02985.

Q. Duan, S. Sorooshian, and V. K. Gupta. Optimal use of the SCE-UA global optimization method for calibrating watershed models. *Journal of Hydrology*, 158(3-4):265–284, 1994. ISSN 00221694. doi: 10.1016/0022-1694(94)90057-4.

Q. Y. Duan, V. K. Gupta, and S. Sorooshian. Shuffled complex evolution approach for effective and efficient global minimization. *Journal of Optimization Theory and Applications*, 76(3):501–521, mar 1993. ISSN 1573-2878. doi: 10.1007/BF00939380.

EROS Data Center. National Elevatoin Dataset, 1999. URL `https://www.usgs.gov/core-science-systems/national-geospatial-program/national-map`.

J. Fan, Y. Xu, H. Ge, and W. Yang. Vegetation growth variation in relation to topography in Horqin Sandy Land. *Ecological Indicators*, 113:106215, 2020. ISSN 1470-160X. doi: 10.1016/j.ecolind.2020.106215.

J. Feiccabrino, W. Graff, A. Lundberg, N. Sundström, and D. Gustafsson. Meteorological Knowledge Useful for the Improvement of Snow Rain Separation in Surface Based Models. *Hydrology*, 2(4):266–288, 2015. doi: 10.3390/hydrology2040266.

J. Fiddes and S. Gruber. TopoSUB: A tool for efficient large area numerical modelling in complex topography at sub-grid scales. *Geoscientific Model Development*, 5(5):1245–1257, 2012. ISSN 1991959X. doi: 10.5194/gmd-5-1245-2012.

W.-a. Flügel. Delineating hydrological response units by geographical information system analyses for regional hydrological modelling using PRMS/MMS in the drainage basin of the river Bröl, Germany. *Hydrological Process*, 9(September 1994):423–436, 1995. doi: 10.1002/hyp.3360090313.

W.-a. Flügel. Combining GIS with regional hydrological modelling using hydrological response units (HRUs): An application from Germany. *Mathematics and Computers in Simulation*, 43:297–304, 1997. doi: 10.1016/S0378-4754(97)00013-X.

G. J. Freeman. Runoff Impacts of Climate Change on Northern California's Watersheds as Influenced by Geology and Elevation. In *Proceedings of 76th Annual Western Snow Conference*, pages 23–34, Hood River, OR, 2008. Western Snow Conference. URL `sites/westernsnowconference.org/PDFs/2008Freeman.pdf`.

G. J. Freeman. Climate change and the changing water balance for California's North Fork Feather River. In *Proceedings of the 79th Annual Western Snow Conference*, pages 71–82, Stateline, NV, 2011. Western Snow Conference. URL `sites/westernsnowconference.org/PDFs/2011Freeman.pdf`.

M. Ghestem, R. C. Sidle, and A. Stokes. The Influence of Plant Root Systems on Subsurface Flow: Implications for Slope Stability. *BioScience*, 61(11):869–879, nov 2011. ISSN 0006-3568. doi: 10.1525/bio.2011.61.11.6.

H. V. Gupta, S. Sorooshian, and P. O. Yapo. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, 34(4):751–763, 1998. ISSN 00431397. doi: 10.1029/97WR03495.

H. V. Gupta, H. Kling, K. K. Yilmaz, and G. 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. ISSN 00221694. doi: 10.1016/j.jhydrol.2009.08.003.

A. A. Harpold, M. L. Kaplan, P. Zion Klos, T. Link, J. P. McNamara, S. Rajagopal, R. Schumer, and C. M. Steele. Rain or snow: Hydrologic processes, observations, prediction, and research needs. *Hydrology and Earth System Sciences*, 21(1):1–22, 2017. ISSN 16077938. doi: 10.5194/hess-21-1-2017.

L. E. Hay and M. Umemoto. Multiple-objective stepwise calibration using Luca. *U. S. Geological Survey Open-File Report*, page 25, 2006. URL https://pubs.usgs.gov/of/2006/1323/.

L. E. Hay, G. H. Leavesley, M. P. Clark, S. L. Markstrom, R. J. Viger, and M. Umemoto. Step wise, multiple objective calibration of a hydrologic model for a snowmelt dominated basin. *Journal of the American Water Resources Association*, 42(4):877–890, 2006. ISSN 1093474X. doi: 10.1111/j.1752-1688.2006.tb04501.x.

Y. Hundecha, B. Arheimer, C. Donnelly, and I. Pechlivanidis. A regional parameter estimation scheme for a pan-European multi-basin model. *Journal of Hydrology: Regional Studies*, 6:90–111, 2016. ISSN 2214-5818. doi: 10.1016/j.ejrh.2016.04.002.

W. W. Immerzeel, A. F. Lutz, M. Andrade, A. Bahl, H. Biemans, T. Bolch, S. Hyde, S. Brumby, B. J. Davies, A. C. Elmore, A. Emmer, M. Feng, A. Fernández, U. Haritashya, J. S. Kargel, M. Koppes, P. D. A. Kraaijenbrink, A. V. Kulkarni, P. A. Mayewski, S. Nepal, P. Pacheco, T. H. Painter, F. Pellicciotti, H. Rajaram, S. Rupper, A. Sinisalo, A. B. Shrestha, D. Viviroli, Y. Wada, C. Xiao, T. Yao, and J. E. M. Baillie. Importance and vulnerability of the world's water towers. *Nature*, 577(7790):364–369, 2020. ISSN 1476-4687. doi: 10.1038/s41586-019-1822-y.

K. Jennings and N. Molotch. The sensitivity of modeled snow accumulation and melt to precipitation phase methods across a climatic gradient. *Hydrology and Earth System Sciences*, 23:3765–3786, 2019. doi: 10.5194/hess-23-3765-2019.

M. M. Kalcic, I. Chaubey, and J. Frankenberger. Defining Soil and Water Assessment Tool (SWAT) hydrologic response units (HRUs) by field boundaries. *International Journal of Agricultural and Biological Engineering*, 8(3):1–12, 2015. ISSN 19346352. doi: 10.3965/j.ijabe.20150803.951.

T. Karpouzoglou and S. Vij. Waterscape: a perspective for understanding the contested geography of water. *Wiley Interdisciplinary Reviews: Water*, 4(3):e1210, 2017. doi: 10.1002/wat2.1210.

B. Khakbaz, B. Imam, K. Hsu, and S. Sorooshian. From lumped to distributed via semi-distributed: Calibration strategies for semi-distributed hydrologic models. *Journal of Hydrology*, 418-419:61–77, 2012. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2009.02.021.

U. Khan, N. K. Tuteja, and A. Sharma. Delineating hydrologic response units in large upland catchments and its evaluation using soil moisture simulations. *Environmental Modelling & Software*, 46:142–154, 2013. ISSN 1364-8152. doi: 10.1016/j.envsoft.2013.03.005.

U. Khan, N. K. Tuteja, A. Sharma, S. Lucas, B. Murphy, and B. Jenkins. Applicability of Hydrologic Response Units in low topographic relief catchments and evaluation using high resolution aerial photograph analysis. *Environmental Modelling and Software*, 81:56–71, 2016. ISSN 13648152. doi: 10.1016/j.envsoft.2016.03.010.

T. Kim, J. Y. Shin, H. Kim, and J. H. Heo. Ensemble-Based Neural Network Modeling for Hydrologic Forecasts: Addressing Uncertainty in the Model Structure and Input Variable Selection. *Water Resources Research*, 2020. ISSN 19447973. doi: 10.1029/2019WR026262.

J. W. Kirchner. Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, 42(3):1–5, 2006. ISSN 00431397. doi: 10.1029/2005WR004362.

H. Kling, M. Fuchs, and M. Paulin. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, 424-425:264–277, 2012. ISSN 00221694. doi: 10.1016/j.jhydrol.2012.01.011.

K. M. Koczot, A. E. Jeton, B. J. Mcgurk, and M. D. Dettinger. Precipitation-Runoff Processes in the Feather River Basin , Northeastern California , with Prospects for Streamflow Predictability, Water Years 1971 – 97 U . S . Department of the Interior. Technical report, U.S. Geological Survey Scientific Investigations Report, Reston, VA, 2005. URL https://pubs.usgs.gov/sir/2004/5202/sir2004-5202.pdf.

LANDFIRE. Existing Vegetation Type, 2013a. URL https://landfire.gov/evt.php.

LANDFIRE. Existing Vegetation Cover, 2013b. URL https://landfire.gov/evc.php.

G. H. Leavesley, R. Lichty, B. Troutman, and L. Saindon. Precipitation-Runoff Modeling System: User's manual. Technical report, U.S. Geological Survey, Denver, CO, 1983. URL https://pubs.usgs.gov/wri/1983/4238/report.pdf.

F. Lobligeois, V. Andréassian, C. Perrin, P. Tabary, and C. Loumagne. When does higher spatial resolution rainfall information improve streamflow simulation? An evaluation using 3620 flood events. *Hydrology and Earth System Sciences*, 18(2):575–594, 2014. ISSN 10275606. doi: 10.5194/hess-18-575-2014.

J. D. Lundquist, S. E. Dickerson-Lange, J. A. Lutz, and N. C. Cristea. Lower forest density enhances snow retention in regions with warmer winters: A global framework developed from plot-scale observations and modeling. *Water Resources Research*, 49(10):6356–6370, 2013. ISSN 00431397. doi: 10.1002/wrcr.20504.

S. A. Margulis, G. Cortés, M. Girotto, and M. Durand. A Landsat-Era Sierra Nevada Snow Reanalysis (1985–2015). *Journal of Hydrometeorology*, 17(4):1203–1221, 2016. ISSN 1525-755X. doi: 10.1175/JHM-D-15-0177.1.

S. L. Markstrom, R. S. Regan, L. E. Hay, R. J. Viger, R. M. Webb, R. A. Payn, and J. H. LaFontaine. PRMS-IV, the precipitation-runoff modeling system, version 4. In *U.S. Geological Survey Techniques and Methods*, chapter B7, page 158. U.S. Geological Survey, 2015. doi: 10.3133/tm6B7.

S. L. Markstrom, L. E. Hay, and M. P. Clark. Towards simplification of hydrologic modeling : identification of dominant processes. *Hydrology and Earth System Sciences*, 20:4655–4671, 2016. doi: 10.5194/hess-20-4655-2016.

J. D. Maxwell, A. Call, and S. B. St. Clair. Wildfire and topography impacts on snow accumulation and retention in montane forests. *Forest Ecology and Management*, 432:256–263, 2019. ISSN 0378-1127. doi: https://doi.org/10.1016/j.foreco.2018.09.021.

G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, New York, 2004. ISBN 0471006262.

E. L. Montoya, J. Dozier, and W. Meiring. Biases of April 1 snow water equivalent records in the Sierra Nevada and their associations with large-scale climate indices. *Geophysical Research Letters*, 41(16):5912–5918, 2014. ISSN 19448007. doi: 10.1002/2014GL060588.

Mountain Partnership. Mountains as the water towers of the world: a call for action on the sustainable development goals. Technical report, Food and Agriculture Organization of the United Nations, 2014. URL http://www.fao.org/fileadmin/templates/mountain_partnership/doc/POLICY_BRIEFS/SDGs_and_mountains_water_EN.pdf.

J. E. Nash and J. V. Sutcliffe. River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3):282–290, 1970. ISSN 0022-1694. doi: 10.1016/0022-1694(70)90255-6.

National Oceanic and Atmospheric Administration. Sacramento-

Soil Moisture Accounting Model, 2002. URL https://www.nws.noaa.gov/oh/hrl/nwsrfs/users_manual/part2/_pdf/23sacsma.pdf?referrer=Baker.

N. Obojes, M. Bahn, E. Tasser, J. Walde, N. Inauen, E. Hiltbrunner, P. Saccone, J. Lochet, J. C. Clément, S. Lavorel, U. Tappeiner, and C. Körner. Vegetation effects on the water balance of mountain grasslands depend on climatic conditions. *Ecohydrology*, 8(4):552–569, 2015. ISSN 19360592. doi: 10.1002/eco.1524.

C. A. Oroza, Z. Zheng, S. D. Glaser, D. Tuia, and R. C. Bales. Optimizing embedded sensor network design for catchment-scale snow-depth estimation using LiDAR and machine learning. *Water Resources Research*, 52(10):8174–8189, oct 2016. ISSN 00431397. doi: 10.1002/2016WR018896.

C. A. Oroza, R. C. Bales, E. M. Stacy, Z. Zheng, and S. D. Glaser. Long-Term Variability of Soil Moisture in the Southern Sierra: Measurement and Prediction. *Vadose Zone Journal*, 2018. ISSN 1539-1663. doi: 10.2136/vzj2017.10.0178.

O. Oyebode and D. Stretch. Neural network modeling of hydrological systems: A review of implementation techniques, 2019. ISSN 19397445.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning*, 12:2825–2830, 2011a. URL http://jmlr.org/papers/v12/pedregosa11a.htm.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011b.

F. Pianosi, F. Sarrazin, and T. Wagener. A Matlab toolbox for Global Sensitivity Analysis. *Environmental Modelling & Software*, 70:80–85, 2015. ISSN 1364-8152. doi: 10.1016/j.envsoft.2015.04.009.

J. Qi, S. Li, Q. Yang, Z. Xing, and F. R. Meng. SWAT Setup with Long-Term Detailed Landuse and Management Records and Modification for a Micro-Watershed Influenced by Freeze-Thaw Cycles. *Water Resources Management*, 31(12):3953–3974, 2017. ISSN 15731650. doi: 10.1007/s11269-017-1718-2.

Y. Qiu, B. Fu, J. Wang, and L. Chen. Soil moisture variation in relation to topography and land use in a hillslope catchment of the Loess Plateau, China. *Journal of Hydrology*, 240(3):243–263, 2001. ISSN 0022-1694. doi: 10.1016/S0022-1694(00)00362-0.

S. Reed, V. Koren, M. Smith, Z. Zhang, F. Moreda, D.-J. Seo, and A. DMIP Participants. Overall distributed model intercomparison project results. *Journal of Hydrology*, 298(1):27–60, 2004. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2004.03.031.

P. Reggiani and T. H. Rientjes. Flux parameterization in the representative elementary watershed approach: Application to a natural basin. *Water Resources Research*, 41(4):1–18, 2005. ISSN 00431397. doi: 10.1029/2004WR003693.

P. Reggiani, M. Sivapalan, and S. M. Hassanizadeh. Conservation equations governing hillslope responses: Exploring the physical basis of water balance. *Water Resources Research*, 36(7):1845–1863, 2000. ISSN 00431397. doi: 10.1029/2000WR900066.

K. Richards. Improving snow & streamflow modeling on the Feather River using intelligent information systems. In *California Snow Surveys 64th Annual Meeting of Cooperators*, Kings Beach, CA, 2018. California Cooperative Snow Survey Program. URL http://cdec.water.ca.gov/snow/meeting/2018/Wednesday/18_CCSS_Richards_20181114.pdf.

B. D. W. Roberts, T. I. Dowling, and J. Walker. FLAG : A Fuzzy Landscape Analysis GIS Method for Dryland Salinity Assessment. Technical Report 8, CSIRO Water and Land, Canberra, Australia, 1997. URL http://www.clw.csiro.au/publications/technical97/tr8-97.pdf.

J. W. Roche, Q. Ma, J. Rungee, and R. C. Bales. Evapotranspiration mapping for forest management in California's Sierra Nevada. *Frontiers for Global Change*, 2020. doi: 10.3389/ffgc.2020.00069.

G. H. Roe. Orographic Precipitation. *Annual Review of Earth and Planetary Sciences*, 33(1):645–671, 2005. doi: 10.1146/annurev.earth.33.092203.122541.

G. H. Roe and M. B. Baker. Microphysical and Geometrical Controls on the Pattern of Orographic Precipitation. *Journal of the Atmospheric Sciences*, 63(3):861–880, 2006. doi: 10.1175/JAS3619.1.

L. Samaniego, R. Kumar, and S. Attinger. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, 46(5), 2010. doi: 10.1029/2008WR007327.

L. Schmidt, F. Heße, S. Attinger, and R. Kumar. Challenges in Applying Machine Learning Models for Hydrological Inference: A Case Study for Flooding Events Across Germany. *Water Resources Research*, 2020. ISSN 19447973. doi: 10.1029/2019WR025924.

C. Shen. A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists, 2018. ISSN 19447973.

F. Silvestro, A. Parodi, L. Campo, and L. Ferraris. Analysis of the streamflow extremes and long-term water balance in the Liguria region of Italy using a cloud-permitting grid spacing reanalysis dataset. *Hydrology and Earth System Sciences*, 22(10):5403–5426, 2018. ISSN 16077938. doi: 10.5194/hess-22-5403-2018.

M. Sivapalan and J. D. Kalma. Scale problems in hydrology: Contributions of the Robertson workshop. *Hydrological Processes*, 9(3-4):243–250, 1995. ISSN 10991085. doi: 10.1002/hyp.3360090304.

Soil Survey Staff Natural Resources Conservation Service. Web Soil Survey, 2019. URL http://websoilsurvey.nrcs.usda.gov/.

G. K. Summerell, J. Vaze, N. K. Tuteja, R. B. Grayson, G. Beale, and T. I. Dowling. Delineating the major landforms of catchments using an objective hydrological terrain analysis method. *Water Resources Research*, 41(12):1–12, 2005. ISSN 00431397. doi: 10.1029/2005WR004013.

C. Tague and L. Band. RHESSys: Regional Hydro-Ecologic Simulation System—An object-oriented approach to spatially distributed modeling of carbon, water, and nutrient cycling. *Earth Interactions*, 8(19):1–42, 2004. doi: 10.1175/1087-3562(2004)8<1:RRHSSO>2.0.CO;2.

A. D. Teshager, P. W. Gassman, S. Secchi, J. T. Schoof, and G. Misgna. Modeling Agricultural Watersheds with the Soil and Water Assessment Tool (SWAT): Calibration and Validation with a Novel Procedure for Spatially Explicit HRUs. *Environmental Management*, 57(4):894–911, 2016. ISSN 14321009. doi: 10.1007/s00267-015-0636-4.

The pandas Development Team. pandas-dev/pandas: Pandas, aug 2019. URL https://doi.org/10.5281/zenodo.3509134.

Q. Tian, D. Wang, D. Li, L. Huang, M. Wang, C. Liao, and F. Liu. Variation of soil carbon accumulation across a topographic gradient in a humid subtropical mountain forest. *Biogeochemistry*, 2020. ISSN 1573-515X. doi: 10.1007/s10533-020-00679-2.

Q. Q. Tran, J. De Niel, and P. Willems. Spatially Distributed Conceptual Hydrological Model Building: A Generic Top-Down Approach Starting From Lumped Models. *Water Resources Research*, 54(10):8064–8085, 2018. ISSN 19447973. doi: 10.1029/2018WR023566.

A. Valéry, V. Andréassian, and C. Perrin. 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 1 – Comparison of six snow accounting routines on 380 catchments. *Journal of Hydrology*, 517:1166–1175, 2014. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2014.04.059.

A. Varhola, N. C. Coops, M. Weiler, and R. D. Moore. Forest canopy effects on snow accumulation and ablation: An integrative review of empirical results. *Journal of Hydrology*, 392:219–233, 2010. doi: 10.1016/j.jhydrol.2010.08.009.

J. Vaze, D. A. Post, F. H. S. Chiew, J.-M. Perraud, N. R. Viney, and J. Teng. Climate non-stationarity – Validity of calibrated rainfall–runoff models for use in climate change studies. *Journal of Hydrology*, 394(3):447–457, 2010. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2010.09.018.

R. Viger and G. H. Leavesley. The GIS Weasel User's Manual. In *Techniques and Methods Book 6*, chapter B4, page 201. U.S. Geological Survey, 2007. URL `https://pubs.usgs.gov/tm/2007/06B04/`.

D. Viviroli, H. H. Dürr, B. Messerli, M. Meybeck, and R. Weingartner. Mountains of the world, water towers for humanity: Typology, mapping, and global significance. *Water Resources Research*, 43(7):1–13, 2007a. ISSN 00431397. doi: 10.1029/2006WR005653.

D. Viviroli, J. Gurtz, and M. Zappa. The Hydrological Modelling System PREVAH. Part I - Overview and Selected Applications. Technical report, Institute of Geography, University of Berne, Berne, Switzerland, 2007b. URL `https://dviviroli.github.io/documents/Viviroli_et_al_2007_GB_P40_I.pdf`.

W. Wilcke, J. Boy, R. Goller, K. Fleischbein, C. Valarezo, and W. Zech. Effect of topography on soil fertility and water flow in an Ecuadorian lower montane forest. In F. N. Scatena, L. A. Bruijnzeel, and L. S. Hamilton, editors, *Tropical Montane Cloud Forests: Science for Conservation and Management*, International Hydrology Series, pages 402–409. Cambridge University Press, Cambridge, 2011. ISBN 9780521760355. doi: 10.1017/CBO9780511778384.045.

A. P. Williams, E. R. Cook, J. E. Smerdon, B. I. Cook, J. T. Abatzoglou, K. Bolles, S. H. Baek, A. M. Badger, and B. Livneh. Large contribution from anthropogenic warming to an emerging North American megadrought. *Science*, 368(6488):314–318, 2020. ISSN 0036-8075. doi: 10.1126/science.aaz9600.

E. F. Wood, M. Sivapalan, K. Beven, and L. Band. Effects of spatial variability and scale with implications to hydrologic modeling. *Journal of Hydrology*, 102(1-4):29–47, 1988. ISSN 00221694. doi: 10.1016/0022-1694(88)90090-X.

C. A. Woodhouse, D. M. Meko, G. M. MacDonald, D. W. Stahle, and E. R. Cook. A 1,200-year perspective of 21st century drought in southwestern North America. *Proceedings of the National Academy of Sciences*, 107(50):21283–21288, 2010. ISSN 0027-8424. doi: 10.1073/pnas.0911197107. URL `https://www.pnas.org/content/107/50/21283`.

K. L. Young, M. K. Woo, and S. A. Edlund. Influence of local topography, soils, and vegetation on microclimate and hydrology at a High Arctic site, Ellesmere Island, Canada. *Arctic and Alpine Research*, 29(3):270–284, 1997. ISSN 00040851. doi: 10.2307/1552141.

S. Zhang, X. Zhang, T. Huffman, X. Liu, and J. Yang. Influence of topography and land management on soil nutrients variability in Northeast China. *Nutrient Cycling in Agroecosystems*, 89(3):427–438, 2011. ISSN 1573-0867. doi: 10.1007/s10705-010-9406-0.

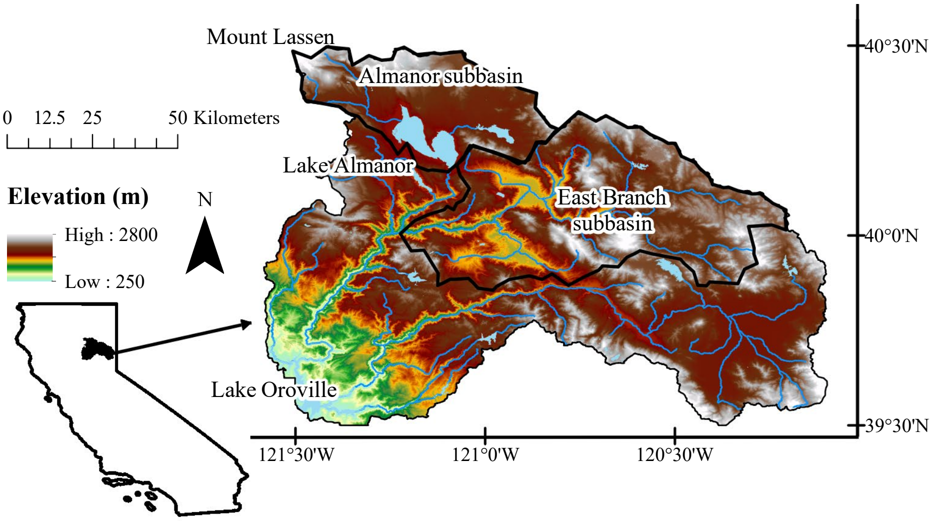
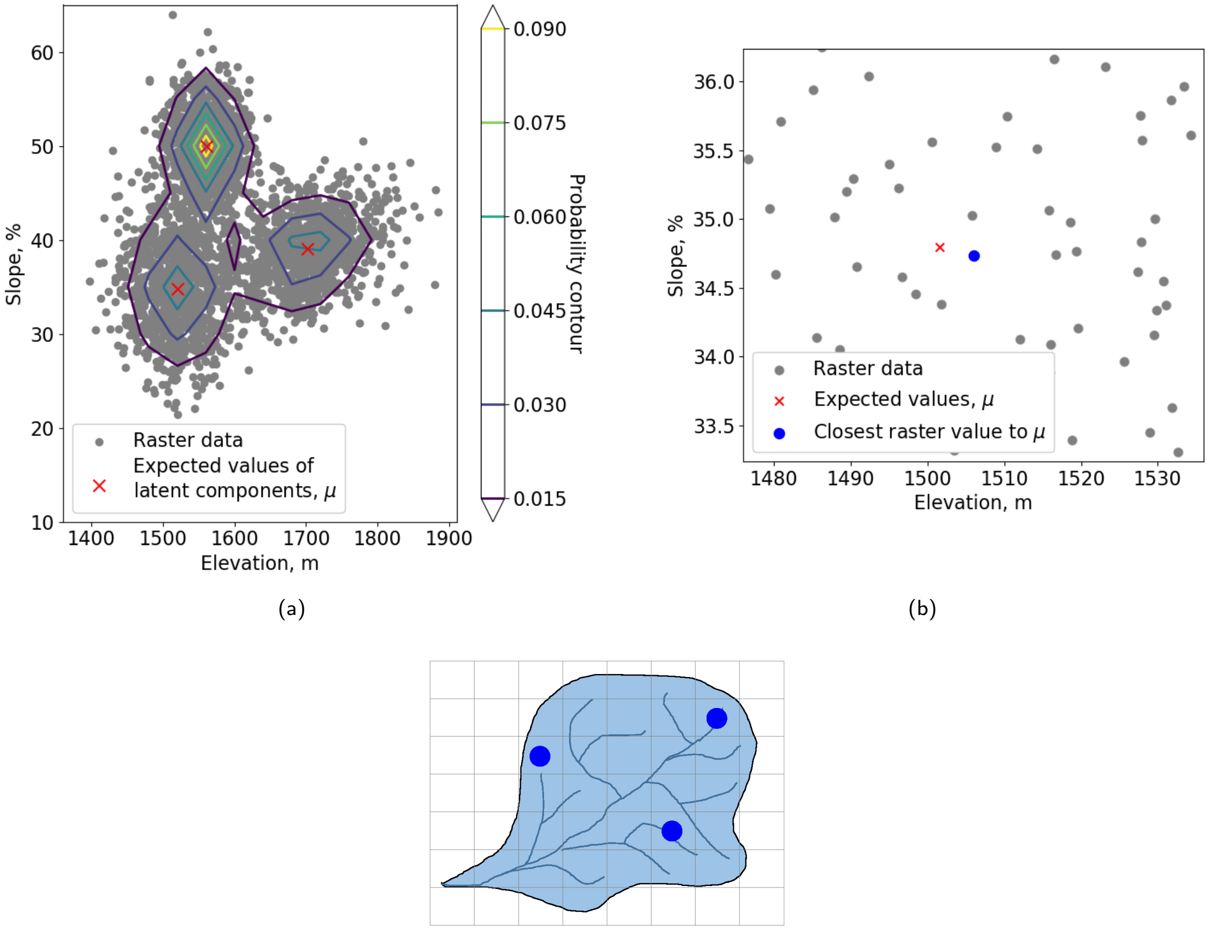


Figure 1: Map of California, indicating extent of Feather River and the headwater catchments, Almanor and East Branch, that were the subject of this study.

Table 1
Calibration details for PRMS models

Step	Calibration variable	Parameters calibrated	Objective function
1	SWE	tmax_allsnow tmax_allrain (Jan-Dec) adjmix_rain (Oct-May) freeh2o_cap	Daily RMSE
2	ET	pref_flow_den soil_moist_max soil_rechr_max	Annual cumulative absolute bias
3	FNF (daily)	carea_max smidx_coef smidx_exp K_coef (per stream segement)*	0.75 × daily KGE + 0.25 × daily LogNSE
4	FNF (monthly)	fastcoef_lin fastcoef_sq sat_threshold slowcoef_lin slowcoef_sq soil2gw_max ssr2gw_exp ssr2gw_rate	0.75 × monthly mean KGE + 0.25 × monthly mean LogNSE
5	FNF (annual)	gwflow_coef gwsink_coef gwstor_min	Annual cumulative absolute bias

Asterisk indicates parameters that were only calibrated in the GIS PRMS model, as they do not appear in the GMM-based models. See Markstrom et al. (2016) for details on PRMS parameters.

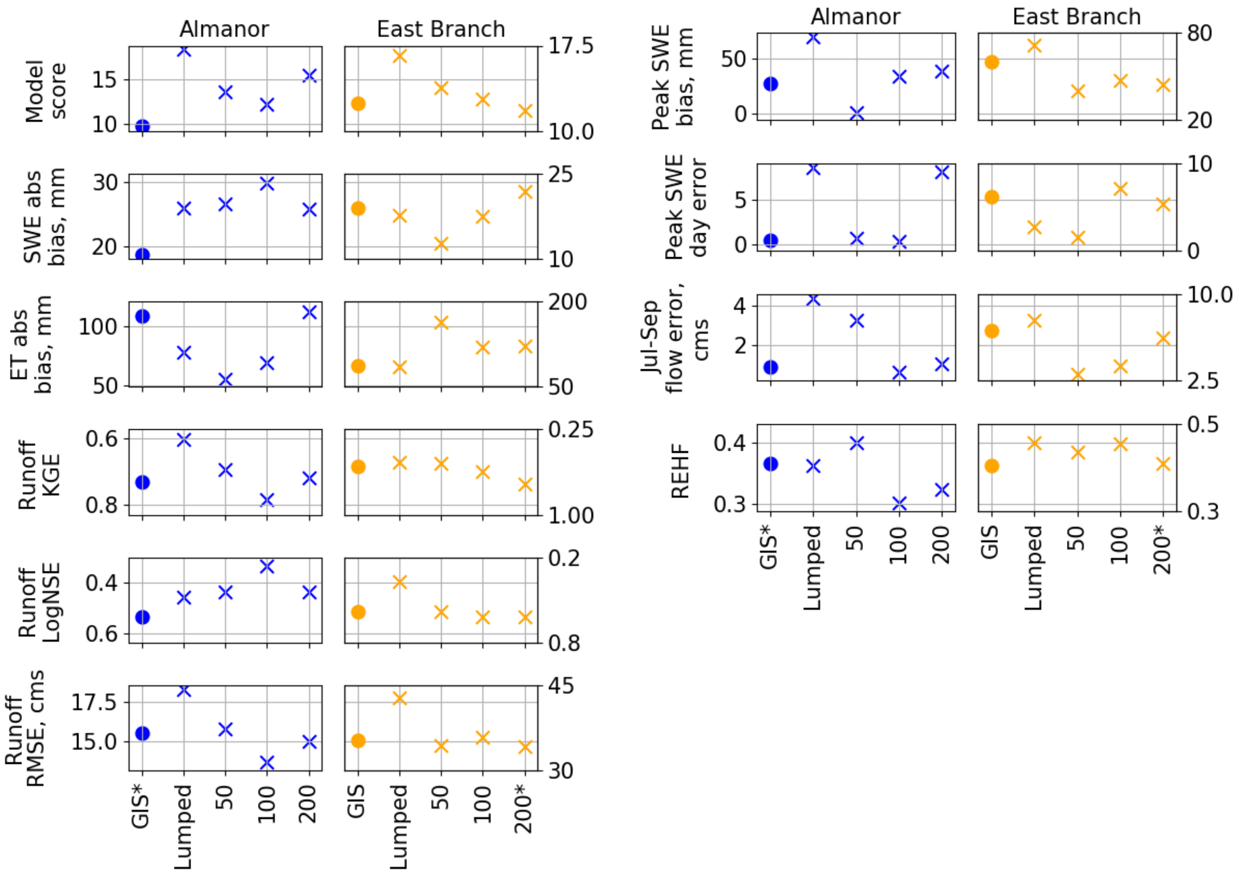


(a)

(b)

(c)

Figure 2: Conceptual schematic illustrating the process of selecting modeling locations using GMM. (a) Example of two-dimensional GMM with three latent components. Lines show equal probability contours. Red X's indicate the expected values of each component. (b) Close-up of Figure 2a showing the Nearest-Neighbor selection of the datapoint closest to the expected value of a latent component. The blue dot represents the raster pixel that is selected as the modeling location. (c) Hypothetical river basin indicating pixels selected as modeling locations



(a) All validation periods

(b) Extreme and peak periods

Figure 3: Model performance metrics Almanor and the East Branch (note differing axis extents for the subbasins). Axes are oriented such that the best performance will appear lowest on the plot. × denotes GMM models. The best-performing model (lowest average median rank across calibrations) in each subbasin is marked with an asterisk (*).

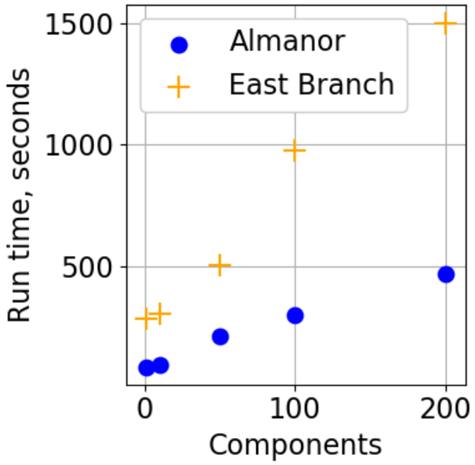


Figure 4: GMM algorithm runtimes by subbasin and number of components

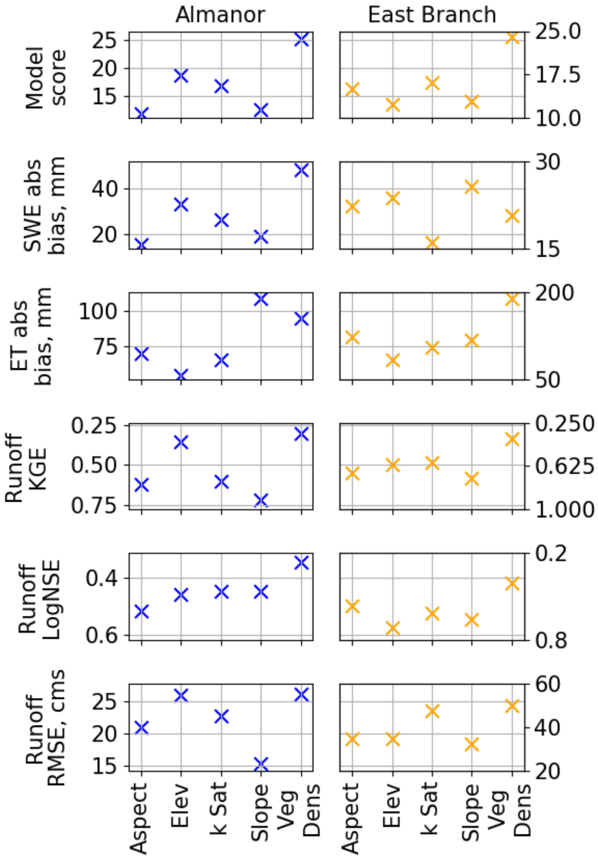
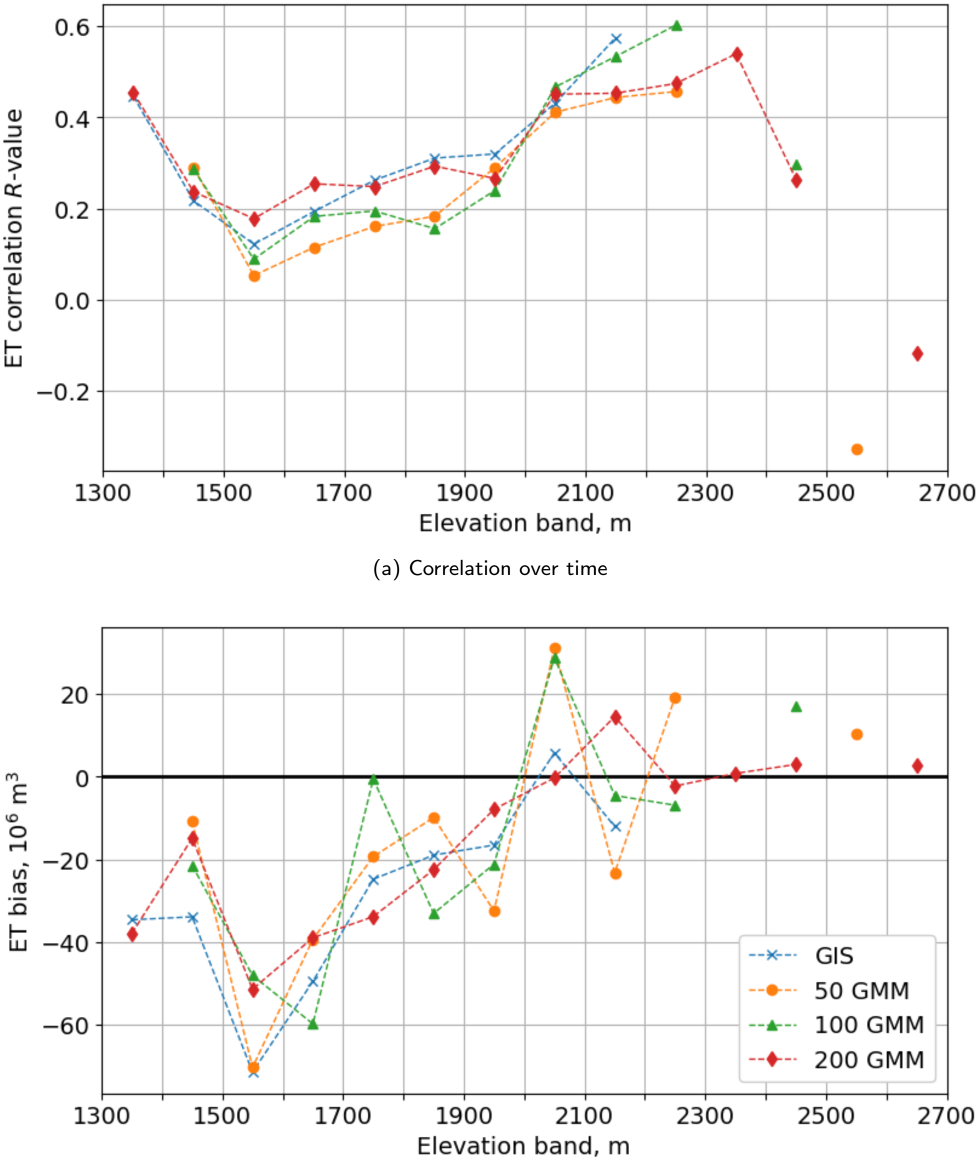


Figure 5: Performance metrics in Almanor and the East Branch for univariate models (note differing axis extents for the subbasins). Axes are oriented such that the best performance will appear lowest on the plot. × denotes GMM models.



(a) Correlation over time

(b) Volumetric bias

Figure 6: ET performance for each elevation band in Almanor, averaged across calibrations. Tick marks indicate upper and lower bounds of 100-m elevation bands. The lumped GMM model is not shown.

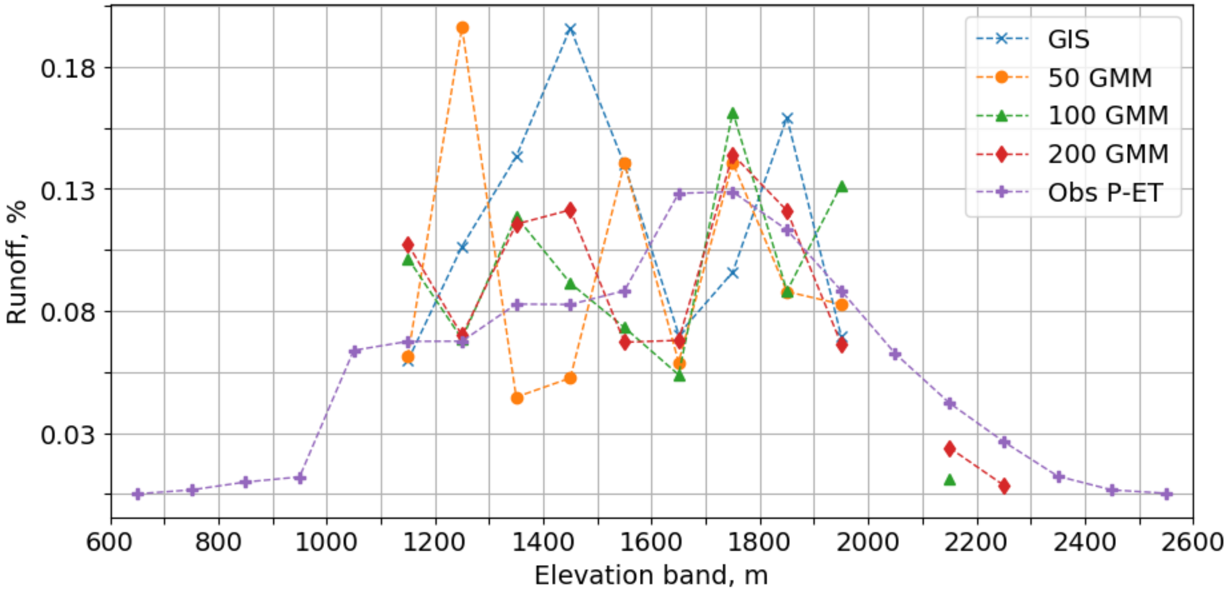


Figure 7: Percent runoff by elevation band in the East Branch. Tick marks indicate upper and lower bounds of 100-m elevation bands. The lumped GMM model is not shown.