Upstream Tech

Reach-based Daily Actual and Natural Flow Predictions

Prepared for California Wildlife Conservation Board
Managed by: The Nature Conservancy, California
February 2025

Laura Read, Mostafa Elkurdy, Alden Keefe Sampson

Executive Summary

Upstream Tech utilizes a machine learning hydrologic modeling methodology which has been shown in literature and in previous phases of development with TNC CA to perform well in a variety of hydrologic regimes and in basins with human alteration. Upstream Tech has developed two models for predicting in ungauged basins - one that is suitable for natural/unaltered basins – the *Unimpaired Flows model*; and the second that can make predictions in basins with human-induced alteration – the *Actual Flows model*. The goal of this phase of work was to thoroughly assess the performance of these two models and draw conclusions on the conditions and locations for applying them to address a wide range of water management needs in California.

The main tasks and deliverables for this project are:

- Scale the Unimpaired and Actual Flows models from lumped, basin scale in discrete basins (previously Phase 3 covered all USGS gauges) to distributed scale across the full state of California.
- Implement methodology to downscale predictions from sub-basin scale to reach scale, covering all NHD+ river reaches.
- Produce 20-year daily reanalysis streamflow records at all river reaches in California for Unimpaired Flows and Actual Flows models.
- Evaluate the two models at a set of validation sites and analyze the results via meaningful categories agreed upon by the external Technical Advisory Committee.

This work evaluated the performance over the 2000-2021 period of the daily Unimpaired Flows model at 53 unseen test sites and the Actual Flows model at 100 unseen test (validation) sites. High-level findings include:

- The dataset provides daily predictions which meet or exceed existing dataset accuracy from previous phases of work in a wide range of conditions including snowmelt and perennial rain regimes in a range of drainage areas and in unimpaired and altered systems.
- The Unimpaired Flows model performs consistently best in snowmelt and perennial rain basins across the full hydrograph (all flow) and even better in baseflows.
- Drainage area is an important indicator in performance, where larger basins (above 500km2) show consistently better performance across metrics than in smaller basins, where there are a mixture of high and low performing sites. This is true for both the Unimpaired and Actual Flows model.
- Spatially across the state, both models perform best in the northern and central (inland and coast) regions with snow cover, mixed snow/rain and perennial rain. In the southern part of the state in basins with intermittent and flashy rain, the model had challenges in predicting the flow magnitudes.

- The Actual Flows model learned:
    - the altered behavior from canals and upstream reservoir density (management); metrics were on par and in some cases even slightly better in basins with these characteristics, however directly below dams at sites without a gauge in between did not show high performance;
    - the model did better in undeveloped (low developed) land cover compared with developed lands.
- Mean NSE skill scores for both models at test sites were above zero in both baseflow and all flows, which could be interpreted as: generally the predictions are more accurate than if the observed mean flow was used as a predictor, *and in all of these test sites, the observed mean is unknown.*

Introduction

In response to a call for proposals by the Wildlife Conservation Board of California for advancements in streamflow management, The Nature Conservancy California (TNC CA) and Upstream Tech partnered to build a state of the art set of models and datasets that use the latest hydrologic advancements in AI as their foundation. The type of hydrologic model employed by Upstream Tech utilizes an AI driven modeling methodology which has been shown in literature and in previous phases of development with TNC CA to perform well in a variety of hydrologic regimes and in basins with human alteration.

Objectives

The goal of this work was to produce two spatially and temporally consistent and complete datasets which describe streamflow at all stream and river reaches in California under two scenarios: streamflow in the absence of human impacts and actual streamflow.

The deliverables of this work are:

- Produce daily, gap-filled 20-year historic records of streamflow in altered basins using the Actual Flows model **and** in natural systems using the Unimpaired Flows model at every river reach in California, aligned to the NHD+ medium resolution Flowline dataset.
- Provide an evaluation of the two models at pre-selected validation sites to understand the range of performance expected, and how performance varies by certain hydrologic and basin characteristics.
- Create and engage an external Technical Advisory Committee to offer feedback, direction, and input on the utility of these models and the potential use cases for their respective groups. Deliver the data and metrics in a format that can facilitate disaggregation into functional flow components for comparison with other models.

The main expected impacts and uses of this data are:

- Help inform stream restoration priority basins by quantifying alteration
- Identify areas of the state that need more gauging or customized modeling techniques to represent streamflow behavior
- Utilize AI hydrology to create a statewide dataset of historical predictions from a single model that can be utilized by many stakeholders for analysis, unifying previous efforts.

Model Methodology

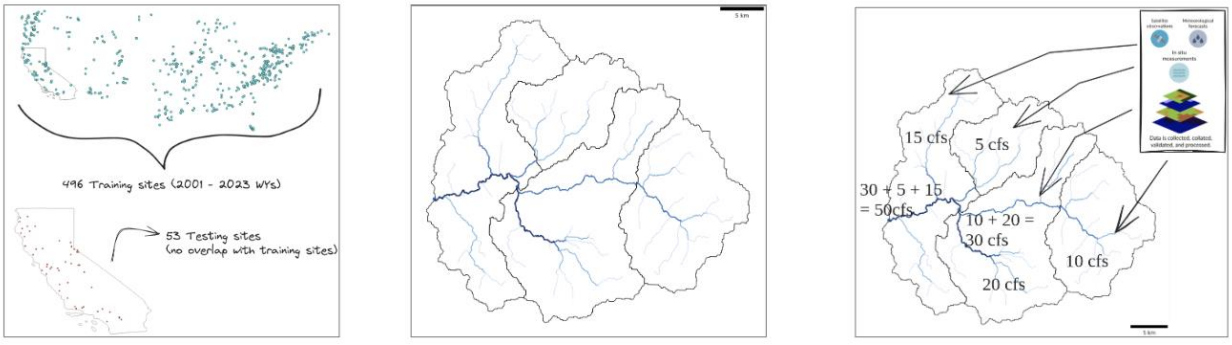
This section describes the development of the models, including an overview of the neural network model, the training and test site locations and details on their selection, and the process

for producing predictions at each stream reach. Model evaluation methods are in the next section.

High level model architecture overview

The methodology used to develop both the Unimpaired and Actual Flows models follows the principles of a 'theory-guided' machine learning approach to hydrologic modeling, described fully in Kratzert et al., 2019. Broadly, this approach takes advantage of the benefits of using machine learning algorithms to learn hydrologic relationships from data, while also ensuring that hydrologic processes are represented. The Unimpaired Flows and Actual Flows models employ a type of neural network model called a Long Short-Term Memory (LSTM) that is particularly suited for time series data and learning relationships between inputs. Kratzert et al., 2019 demonstrates how this approach improves on traditional process-based hydrologic models, especially in making predictions in ungauged locations. The Unimpaired and Actual Flows models share the same architecture, but differ in their inputs and how they handle upstream streamflow observations.

Using this framework, the Unimpaired and Actual Flows model development occurred in three main steps: 1) base model development, 2) routing model development, and 3) downscale to reach predictions. The next sections describe the details of each step. Figure 1 illustrates a high level workflow for each step of this process.



Step 1: Train and evaluate base model

Step 2: Tune routing models using available observed data

Step 3: Downscale the tuned models to reach scale.

Figure 1. Workflow diagram for the three steps in model development

Step 1: Base model development

In neural network or AI hydrology, a "base" model is a term that refers to a generalized model developed by collecting large amounts of data that can teach the network how input variables affect a target output variable (streamflow in this case). The goal of the base model is to learn the

relationships between inputs and streamflow under a diverse set of hydrologic conditions so that when a prediction is made in a completely new and ungauged setting, the model uses its knowledge of hydrology in addition to the unique basin information to make accurate predictions. The inputs to the base model are summarized here, and a full list of the input variables is provided in the Appendix.

Model inputs used in both base models

- **Weather.** The ERA5-Land reanalysis product supplied meteorological variables. This dataset provides a long daily record (1982-present) of the best estimate of "observed" weather.
- **Surface observations.** The models received daily MODIS satellite observations of the daytime and night-time land surface temperatures, snow cover (NDSI) and vegetation vigor (NDSI) to learn how runoff and infiltration are impacted by these important dynamics.
- **Historic streamflow observations.** For the training sites, USGS daily observations were used to guide the model in learning how the inputs translated into streamflow.
- **Basin characteristics.** Static variables that describe the climatology, topography, soils, and land cover, to guide learning of hydrologic processes. These are largely derived from the ERA5-Land dataset and topography.

Additional Actual Flows model inputs

- **Enhanced basin characteristics.** The Actual Flows model was trained to learn about anthropogenic activity and alteration using variables from the EPA StreamCat dataset. See the Appendix for the list, which includes data on dams, canals, etc.
- **Dynamic basin-wide land cover.** Annual MODIS Landcover[1] (500m resolution) to ensure changes in alteration were captured.

Training and Test Sites

In order for the model to learn the diverse set of hydrologic conditions in California, the Unimpaired and Actual Flows models were trained on as many sites (watersheds) as are available, pulling from across the United States. The training sites for the Unimpaired Flows model were a subset of the "Reference" gauges from the USGS Gages II dataset (USGS, 2011). The Reference gauges are the "least-disturbed watersheds" according to a set of hundreds of characteristics that take into account human alteration, and local knowledge about the watersheds themselves. The Actual Flows sites were selected from all USGS gauge in the United States. The criteria for including a site in either model's training sets were: 1) a USGS gauge with a flow record of at least one year in 2000 to 2022, 2) a drainage area < 25,000 km^2. This filtering resulted in the sites summarized in the table below and Figures 2 and 3.

[1] Source: https://lpdaac.usgs.gov/products/mcd12q1v006/

Model	Training sites (nationwide)	Test sites (all in California)	Source
Unimpaired Flows	496 *[44 in California]*	53	USGS GagesII sites that met criteria listed above
Actual Flows	1,779 *[238 in California]*	100	USGS sites that met criteria listed above

Figure 2 illustrates a map of the training and test sites for the Unimpaired Flows model. The test sites were selected through a random tessellated sampling method of available sites in California, at the suggestion of the Technical Advisory Committee.

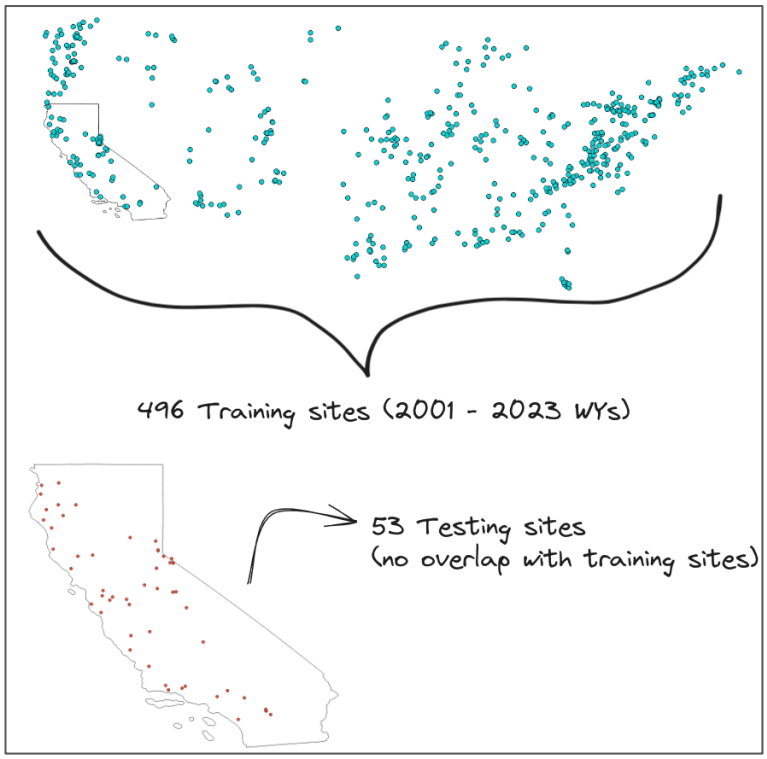


Figure 2. Unimpaired Flows training and test site distribution

Figure 3 illustrates the locations of the Actual Flows training and test sites. The test sites were selected through a random tessellated sampling method of available sites in California, at the suggestion of the Technical Advisory Committee.

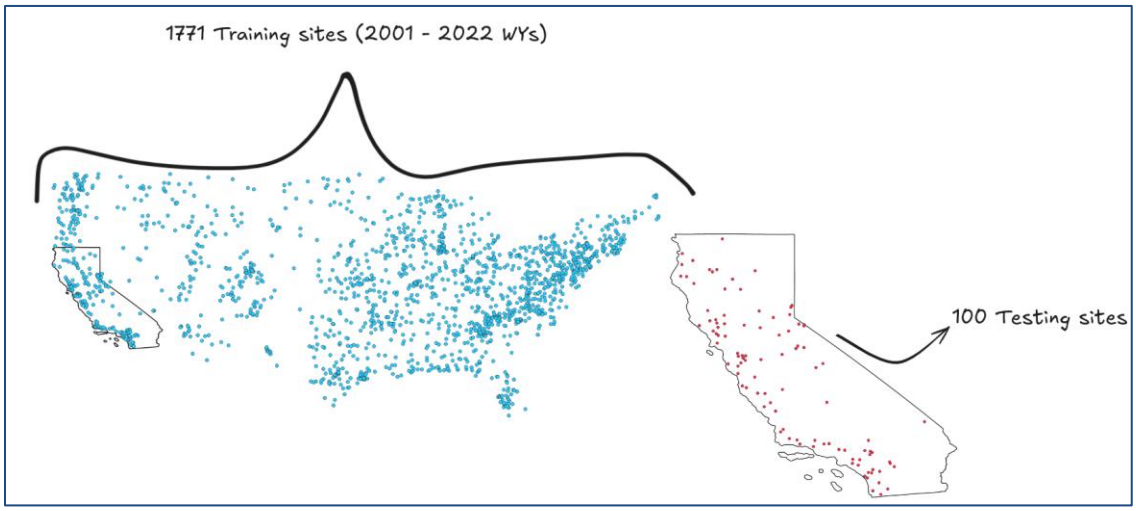


Figure 3. Actual Flows training and test site distribution

Step 2: Routing model development

We built the "base" model neural network to learn general hydrology, using meteorological and satellite data to predict streamflow. To predict at a particular downstream point within a small basin, we average these inputs over the entire upstream drainage. For larger basins, this averaging causes the loss of important information that is contained in the high spatial resolution of the source data. To maintain this information as much as possible, we split basins into smaller sub-basin units, make stream flow predictions on each unit, and route the flows down through the river network. The first step in the overall routing was to delineate all basins in the state. To complete this, we built a custom module that delineates drainage basins upstream of all river outlets within the state of California. The delineation produces both subbasins for all watersheds in the state as well as a river network for all reaches that contribute to and within those subbasins. For each of these river networks, we split basins where there are active USGS gauges with recent data available, at confluences where rivers (of any size) converge, and with a size limit ensuring that all sub-basins were below 500 km^2. These sub-basins are shown in Figure 4:

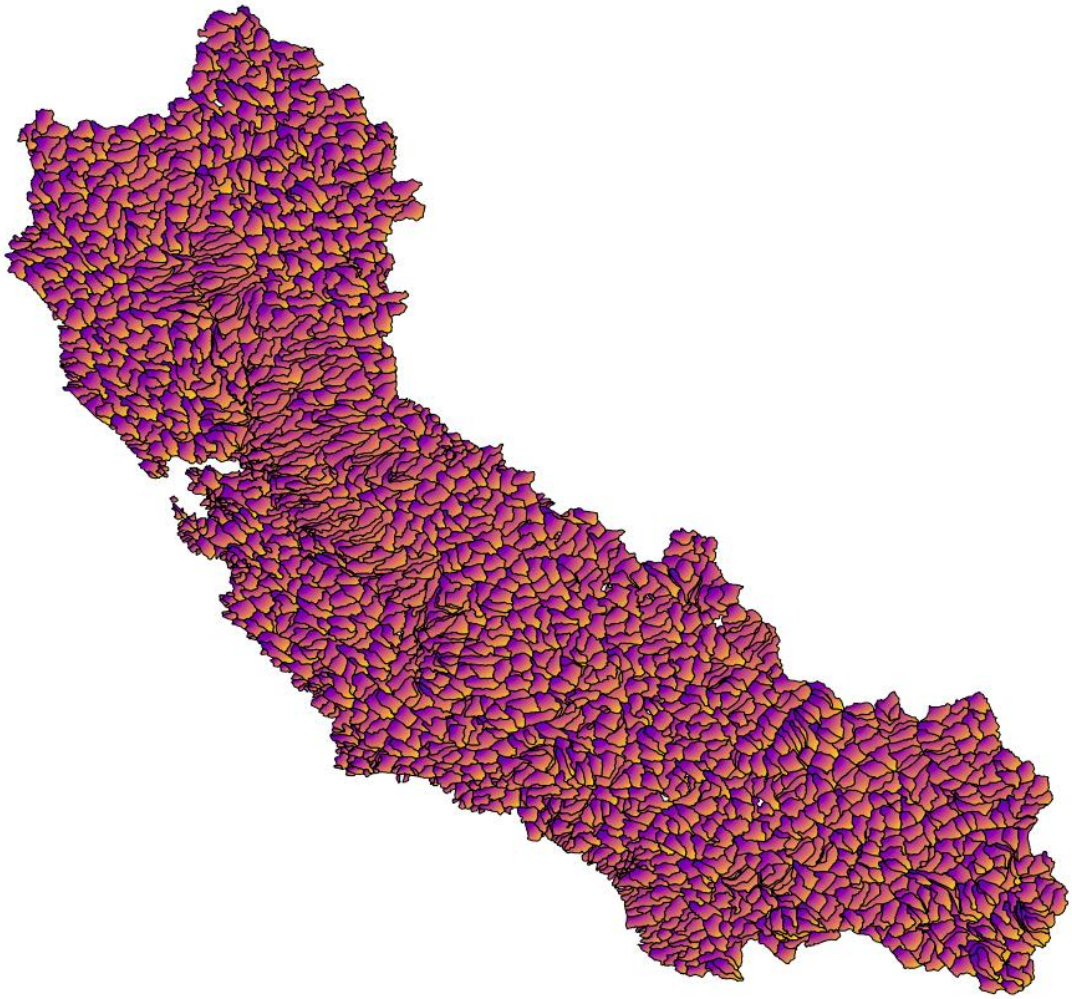


Figure 4. Map of delineated sub-basins

In addition to the delineation of all sub-basins, we produced a network graph representing the basin connectivity, which the model uses as a map to route sub-basin flows. Once the subbasins were delineated and the river network was established, the next steps in producing predictions and routing flows were as follows:

1) Process inputs for all sub-basins: For each of the subbasins produced across the state, we ingested the base model inputs averaged over each subbasin.
2) Produce predictions across all sub-basins: Using these inputs, we produced base model predictions for every sub-basin.
3) Route sub-basin flows through network: A simple representation of the routing process is represented in Figure 5 below.

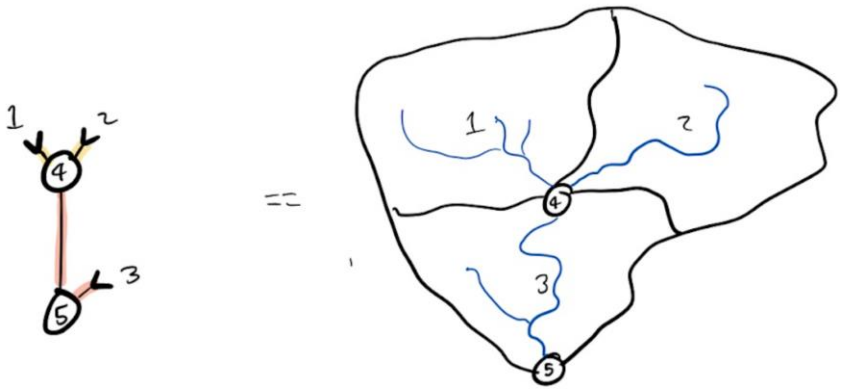


Figure 5. Simple representation of how a river basin split into sub-basins (on the right) is represented as nodes and edges in the graph structure used by the model (left). Sub-basin outflows (1-3) and confluences (4-5) are nodes.

Using the base model runoff contribution produced in step #2 for each of the sub-basins 1, 2, and 3, we can route flows between sub-basins (step 3). To route flows, we first produce an 'adjacency layer matrix', which represents the connectivity across the basin nodes and includes all sub-basins (1, 2, and 3 in Figure 5 above) as well as confluences (4 and 5 in Figure 5 above). To get the flows at the confluence nodes (outlets of subbasins) we sum flows from upstream areas. For compute efficiency, rather than summing all the upstream subbasins, if there is a nearby confluence that already includes the sum of some of the upstream subbasins, we use the flow from that confluence instead. Thus, confluences with the fewest number of upstream nodes get routed first until we reach the bottom of the longest path across the river networks (4 needs to be processed before 5). The flow routing is applied across all nodes with the same number of upstream nodes in parallel by applying a matrix multiplication between the adjacency layer matrix and the sub-basin flows. This allows us to parallelize the calculations to apply the flow routing as efficiently as possible.

Finally, at confluences where we have USGS gauges with observations, we can replace the model predictions with these observations to produce more accurate downstream predictions. For the Actual Flows model, we take advantage of this functionality for the final predictions, while for the Unimpaired Flows model we do not since we would be routing observations that may include unnatural modifications.

Step 3: Downscale to reach level predictions

At the end of Step 2, we have routed flow predictions for every sub-basin and confluence node. Using these predictions, information about the subbasins and their connectivity, and drainage area metadata from NHD v2 flowlines, we apply a downscaling algorithm to produce NHD reach-scale predictions. The general logic is depicted in Figure 6.

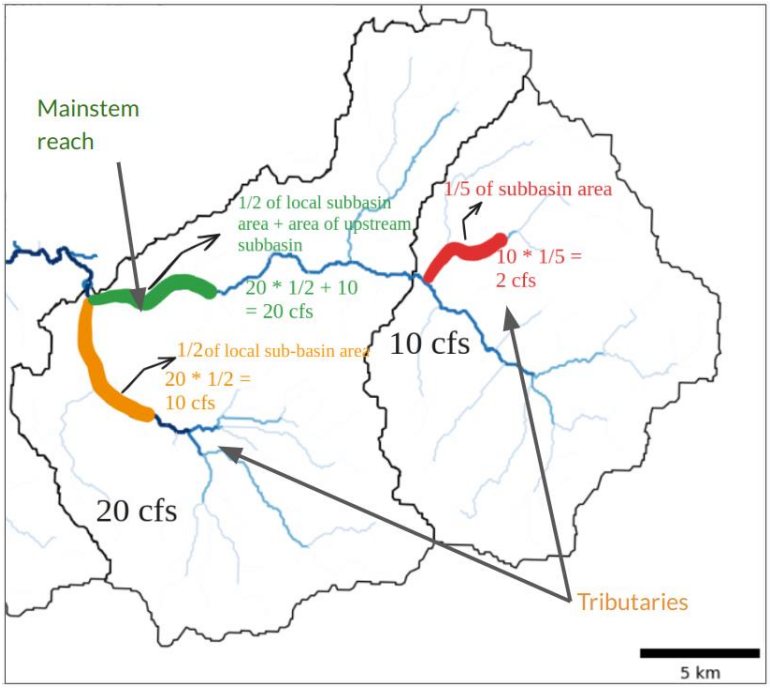


Figure 6. Depiction of downscaling sub-basin flows to reach scale

The algorithm iterates over all NHD flowlines within each sub-basin and calculates the predicted flow along every stream segment by first determining if the sub-basin containing the flowline has any upstream nodes or not. If it does not (e.g. red flowline in the figure above), we apply a conventional drainage area scaling method:

Qflowline = Q_subbasin * DA_flowline / DA_subbasin

Where DA_subbasin is the area of the subbasin, DA_flowline is the area flowing to the reach using the flowline metadata in the NHD v2 dataset, Q_subbasin is the local flow predicted for that subbasin. We use the same equation for any tributary reach, (e.g. orange flowline in Figure 6), which is defined as a flowline that starts and ends in the same subbasin.

If the sub-basin containing the flowline does have upstream nodes in the river network, meaning it is a reach that starts in one sub-basin and ends in another (e.g. the green flowline in Figure 6, a mainstem river) then we use the following equation to calculate the flow for this reach:

Q_flowline = Q_flow from upstream nodes + Q_subbasin * (DA_flowline - DA_upstream subbasins) / DA_subbasin,

Where Q_flow from upstream nodes is the flow from the confluence most directly upstream of the subbasin containing the flowline, and DA_upstream subbasins is the sum of the area of all subbasins upstream.

Some NHD flowlines do not have a drainage area provided as part of the dataset metadata. These flowlines were skipped. Additionally, for about 1% of the flowlines which intersected our delineation, the metadata for drainage area in the NHD and our delineation's calculation of drainage area did not agree. Results for these flowlines were omitted from the final dataset, leaving predictions for a total of 154,754 flowlines in total. Qualitatively these were parallel flowlines indicating artificial routing.

Figures 7 and 8 show the distribution of the drainage areas of all California sub-basins relative to the drainage areas in each model. In Figure 7 the Unimpaired flows train (yellow) and test (green) sets show that the train/test sets for the model encompassed very small basins as well as larger ones due to gauge availability. A similar pattern is true for Actual Flows (Figure 8).

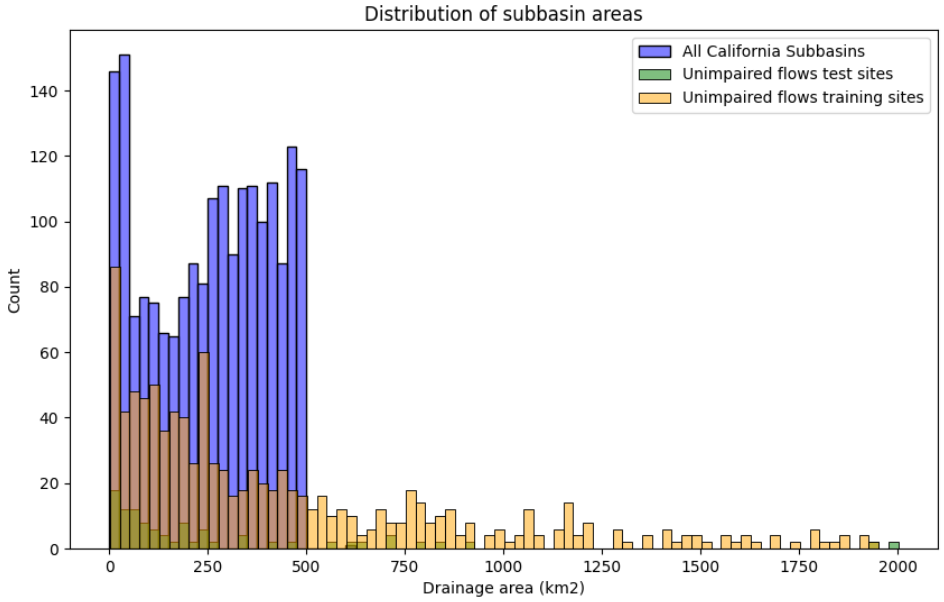


Figure 7. Distribution of drainage areas from all California sub-basins denoted into categories of Unimpaired Flows model train and test sites.

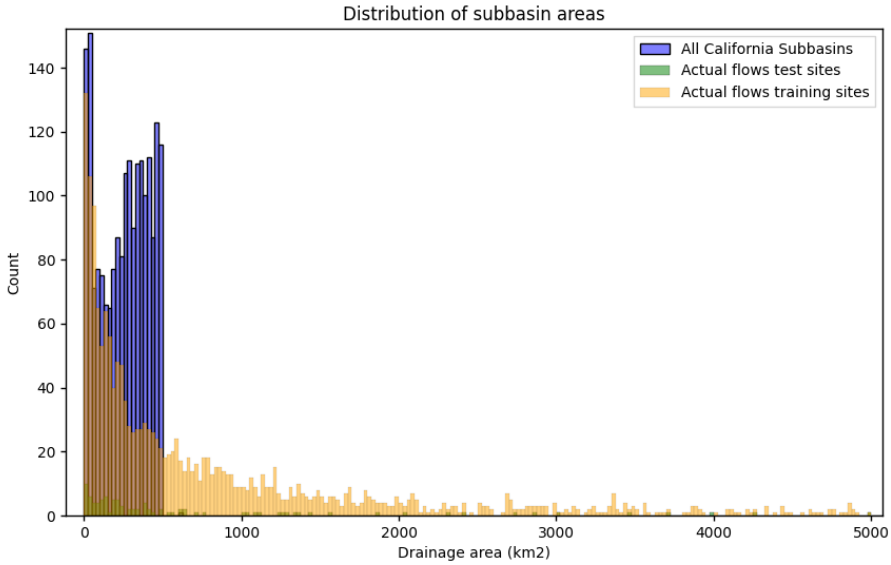


Figure 8. Distribution of drainage areas from all California sub-basins, denoted into categories of Actual Flows model train and test sites.

Evaluation Methods

The goal of model evaluation is to quantify model performance at locations where observations are available and to use these "test" sites to summarize how performance varies by certain factors, e.g. hydrologic conditions, alteration, etc. This summary is intended to serve as a set of recommendations for users of how well the model works in ungauged basins. The project team and advisory committee jointly selected a set of metrics and categorizations by which to dissect the performance.

Metrics

The following metrics were selected as representative of understanding model performance and in line with common hydrological model evaluation:

- Kling-Gupta Efficiency (KGE)
- Nash Sutcliffe Efficiency (NSE)
- Bias, normalized
- Correlation (Pearson's r)
- Root mean square error, normalized
- Confidence interval metric: hit rate, defined as the percent of observed values that fall within the interquartile range

In the plots which show a single value of the model's performance, we default to assessing the mean of the distribution, since this is the model's best estimate. The median is a useful bar for understanding the value of which there's a 50% chance of being above or below.

For the hit rate metric, we assess whether the width of the model's confidence intervals (CIs) are accurately sized. Ideally, CIs are narrow enough to be useful for decision making and wide enough to indicate the model's uncertainty. Ideally, the model's CIs capture the observed flows the correct fraction of the time, e.g. as shown in Figure 9 for the 50% confidence interval. Explicitly, 50% of the time the observations fall inside the model's prediction range, and 50% of the time the observations are outside of the 50% CI.

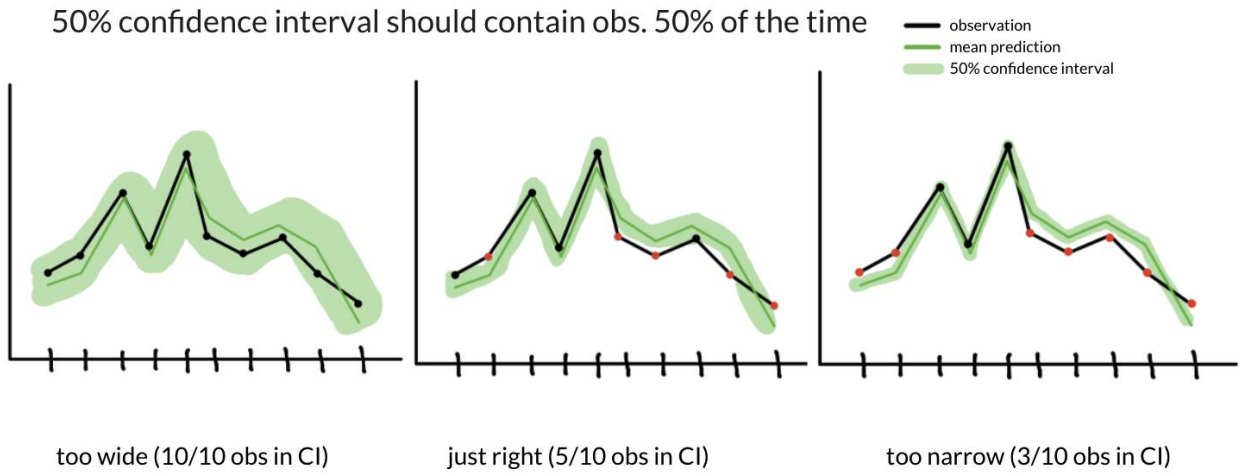


Figure 9. Depiction of hit rate assessment of confidence intervals, for the 50% case

Prior to calculating performance metrics, the flows from each site were separated into 'base flow' and 'all flows' according to a method published by Lyne & Hollick (1979) and suggested by the TAC. To help inform the types of use cases for the data, the project team and advisory committee identified key watershed characteristics that could affect model performance. The characteristics were selected from readily available fields in the StreamCat dataset (Hill et al., 2018). Once the separation method was applied, performance metrics were evaluated using the following watershed characteristics:

Watershed Characteristics

- Drainage area
- Hydrologic regimes (snow, intermittent rain, perennial rain, mixed snow and rain) based on Lane et al. 2018 with two changes: the nine classes were condensed to three by type of precipitation (snow, rain, or mixed), and the NHD classification for intermittent vs perennial was added to subdivide the rain category into intermittent rain and perennial rain.
- Level of alteration (Actual Flows only)
    - Canal presence/absence

- Upstream storage (none, low, high)
- Developed /undeveloped land

Table 1. Reach count in California by class

Hydrologic regime	Reach count in California
Intermittent Rain	24,124 (35.2%)
Mixed snow and rain	19,546 (28.5%)
Perennial Rain	15,828 (23.1%)
Snowmelt	9,006 (13.1%)

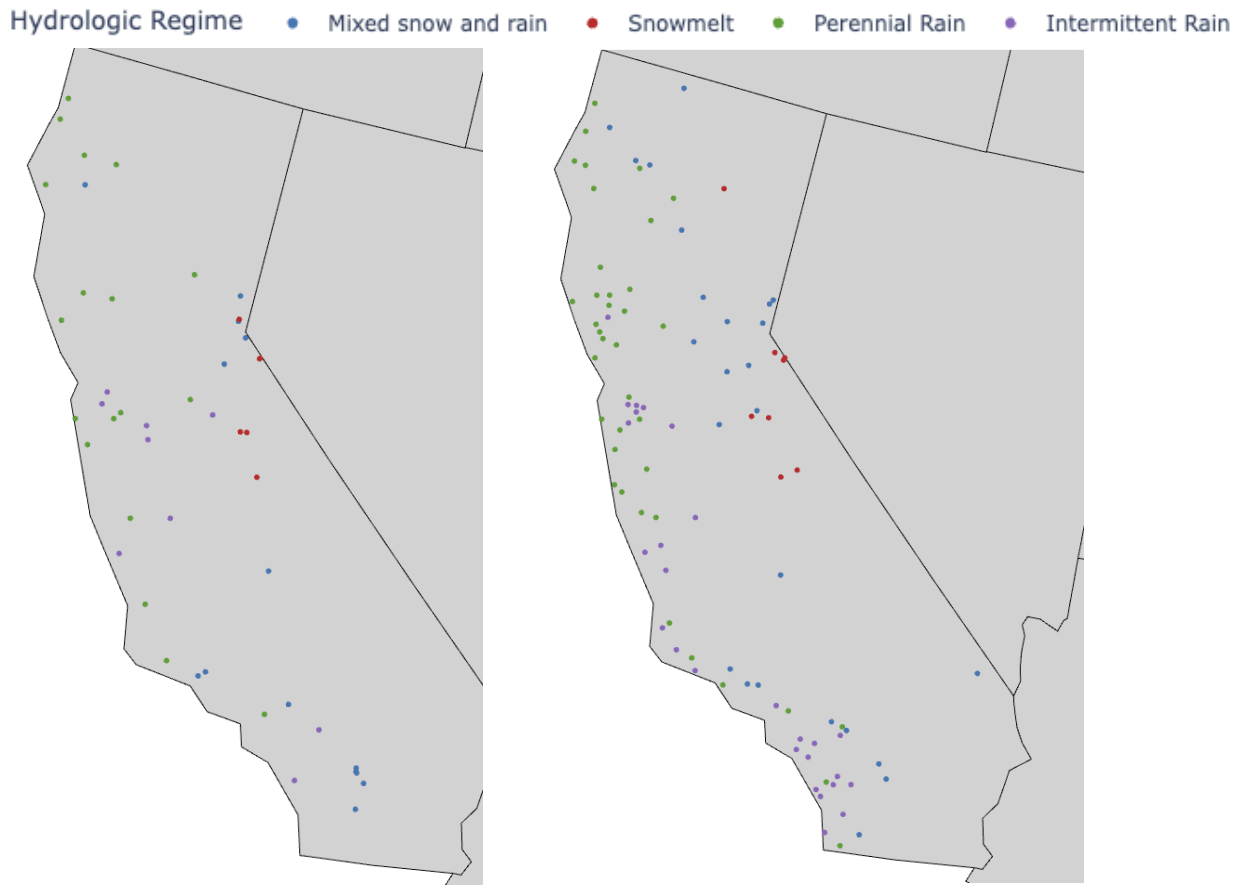


Figure 10. Maps of test sites from unimpaired (left) and actual flows (right) models showing the geographic locations of each sites classified by hydrologic regime (mixed rain and snow = blue, snowmelt = red, intermittent rain = green, perennial rain = purple).

The alteration variables were derived from the StreamCAT dataset and the categories were divided into groups defined as:

Canal presence

Using the variable that defines presence or absence of canals (*CanalDensWs*):

- False = 0
- True if > 0

Upstream storage

Using the variable that represents the volume of reservoir storage per total drainage area into the location (*DamNrmStor*):

- None = 0 m3
- Low = 0 -100,000 m3
- High > 100,000 m3

Developed/Undeveloped land

- Developed is defined as subbasins where the sum of all urban land use types (open space, high, medium, and low intensity) is greater than 5%.
- Undeveloped is defined as subbasins where the sum of all urban land use types (open space, high, medium, and low intensity) is less than or equal to 5%.

Results

This section summarizes the performance at test sites for the Unimpaired Flows and the Actual Flows models. The section begins with a summary of performance and then presents detailed results for each model according to the evaluation plan.

Unimpaired Flows Model Results

Metrics Performance Summary

Hydrologic Regime	Flow type	KGE (median, mean)	Normalized Bias (median, mean)	Correlation (median, mean)	Normalized RMSE (median, mean)
Snow (n=8)	Base	0.30, -0.08	-0.29, -0.14	0.89, 0.86	0.94, 0.88
	All	0.20, -0.19	-0.31, -0.16	0.89, 0.87	1.03, 1.06
Perennial rain (n=21)	Base	0.48, 0.31	-0.25, -0.27	0.87, 0.79	0.91, 1.13
	All	0.24, 0.17	-0.37, -0.38	0.81, 0.74	2.17, 2.63
Intermittent Rain (n=10)	Base	0.17, 0.07	-0.34, -0.29	0.76, 0.63	1.23, 1.47
	All	-0.15, -0.05	-0.55, -0.49	0.62, 0.56	3.49, 4.15
Mixed snow and rain (n=15)	Base	-0.24, -0.08	-0.39, -0.37	0.52, 0.51	1.21, 1.41
	All	-0.30, -0.14	-0.57, -0.48	0.54, 0.54	2.06, 3.19

Key takeaways

Highest confidence in utilizing the model data:

- Base flows
- Snowmelt driven basins
- Mixed snow and rain driven basins in the Northern part of the state

Good performance, especially considering alternatives:

- Large basins
- Intermittent rain driven basins (base flows)
- Perennial rain driven basins (high + base flows)

Lower confidence, representing conditions where the performance and/or observations did not show a clear signal:

- Highly intermittent rain driven/desert dominated basins (high flows)
- Mixed snow and rain driven basins in the Southern part of the state (high + base flows)

- Very dry and small basins, e.g. those with a normalized q75 (the value below which 25% of the data points in a dataset fall) lower than 0.0005 (see explanation in Results section), which are primarily located in dry, flashy watersheds in the Southern region where validation sites were limited.

To further understand the characteristics and conditions where the model's performance was lower, we dug further into the relationships between metrics, drainage area, and hydrologic regime (especially flashiness).

One final overall takeaway is that the model tended to be low biased across regimes. The reason is that is the model's baseflow predictions were low, and thus when downscaled to the reach level and compared with the USGS observations, the model's magnitude was low biased.

By hydrologic regime

Overall performance by hydrologic regime across Unimpaired Flows test sites indicates overall higher skill scores in the snowmelt and perennial rain categories, though there is wider variability at perennial rain sites. The LSTM utilized by the Unimpaired and Actual Flows models is particularly well suited for snowmelt basins because snowmelt has a memory, which this type of model tracks well, based on its ability to store information from historical precipitation over a long warmup period. The model has more variable and lower performance in the intermittent rain and mixed snow/rain category. The intermittent rain and mixed snow/rain regimes contain sites with very flashy behavior and sometimes < 1 cfs baseflow, which can be difficult for the model to capture and are also highly penalized by most of the performance metrics. Figure 11 shows the distribution of NSE across hydrologic regime with baseflow separated into its own distribution (orange).

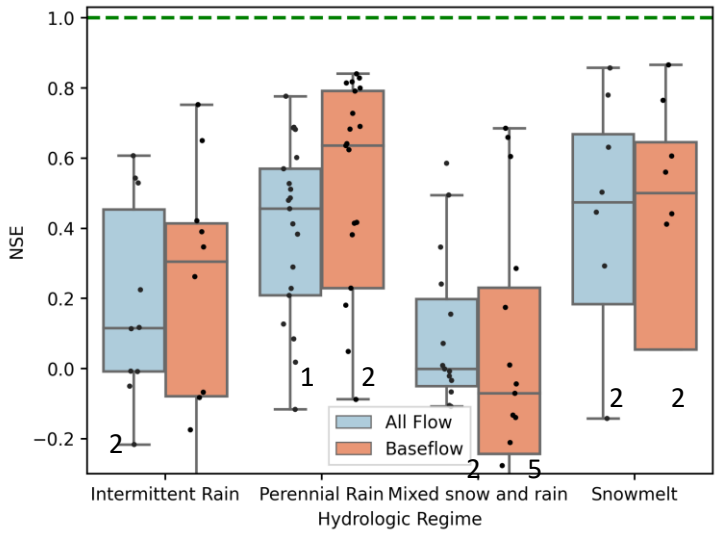


Figure 11. Unimpaired Flows model NSE scores across test sites (n = 53) split by hydrologic regime. The number of sites off the plot boundaries are shown as annotations in the plots.

Figures 12-14 illustrate the normalized bias, RMSE, and correlation for unimpaired sites. The model's bias was consistent across baseflow and all flows, slightly low biased in general with snowmelt being the least biased, suggesting that the model consistently slightly underpredicts flows. The model had lower RMSE in the baseflow predictions across all regimes with the most variability in the intermittent rain category, which indicates that some of the magnitude during flashy rain peaks were missed. Correlation shown in Figure 14 corroborates the model performing well in snowmelt regions, and more varied performance, though overall still high values in intermittent rain (e.g. the peaks were captured but the magnitudes were off).

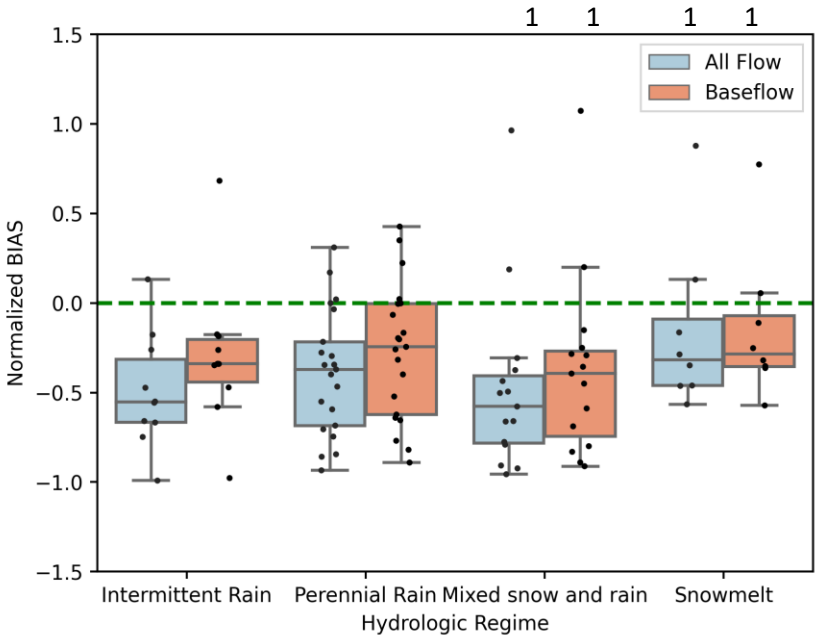


Figure 12. Unimpaired Flows model normalized bias across test sites (n = 53) split by hydrologic regime. The number of sites off the plot boundaries are shown as annotations in the plots.

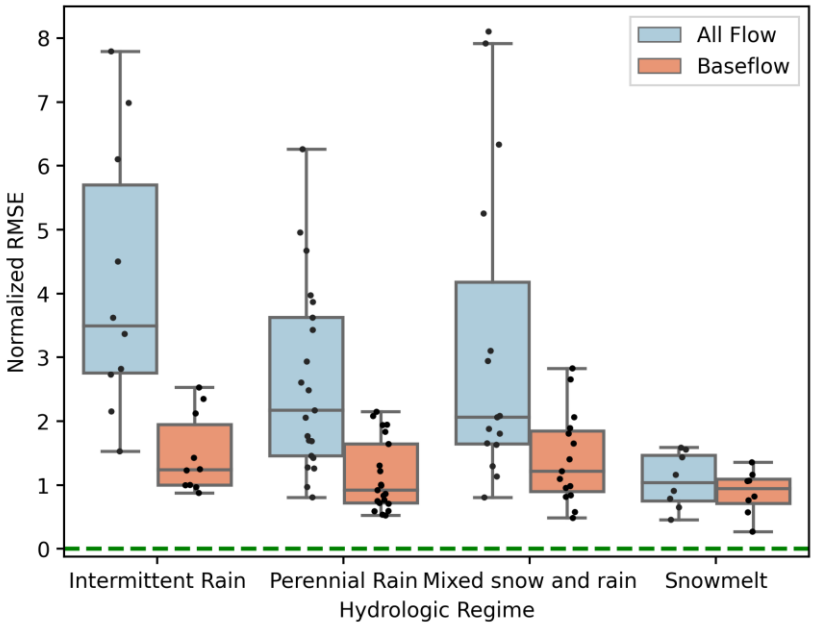


Figure 13. Unimpaired Flows model normalized RMSE across test sites (n = 53) split by hydrologic regime.

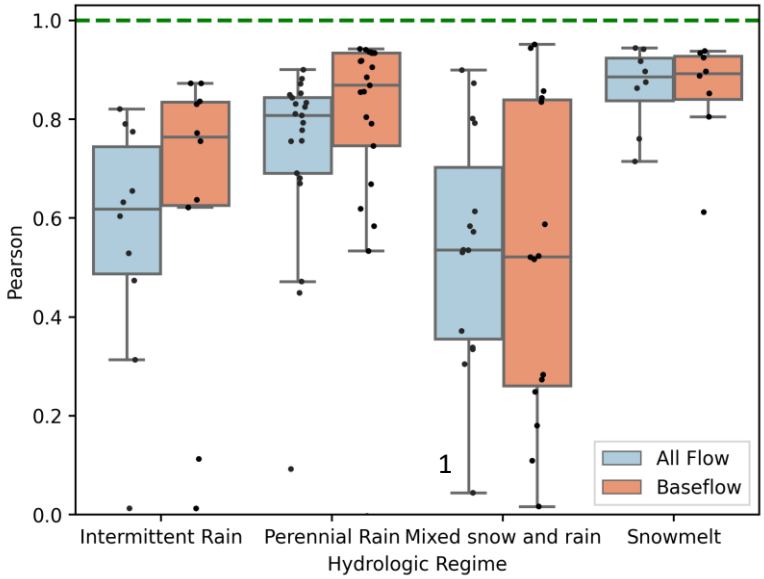


Figure 14. Unimpaired Flows model correlation coefficient (Pearson's r) across test sites (n = 53) split by hydrologic regime. The number of sites off the plot boundaries are shown as annotations in the plots.

By drainage area

The following plots (Figures 15-18) show scatterplots of the models' mean performance for each metric (NSE, bias, RMSE, correlation) across all drainage areas, with the dots colored by hydrologic regime. From this view, we note that the intermittent rain and mixed snow/rain sites generally had smaller drainage area distributions than the other regimes. Basins larger than 500

km2 generally had higher NSEs, e.g. > 0.4 in baseflow and > 0.1 for all flows. Basins smaller than 500 km2 had mixed performance where the model performed well and others where it had NSEs less than zero (primarily in the mixed snow/rain regime). The low performing sites were those with near zero flows for much of the year and then very flashy rain events, occurring in the Southern part of the state.

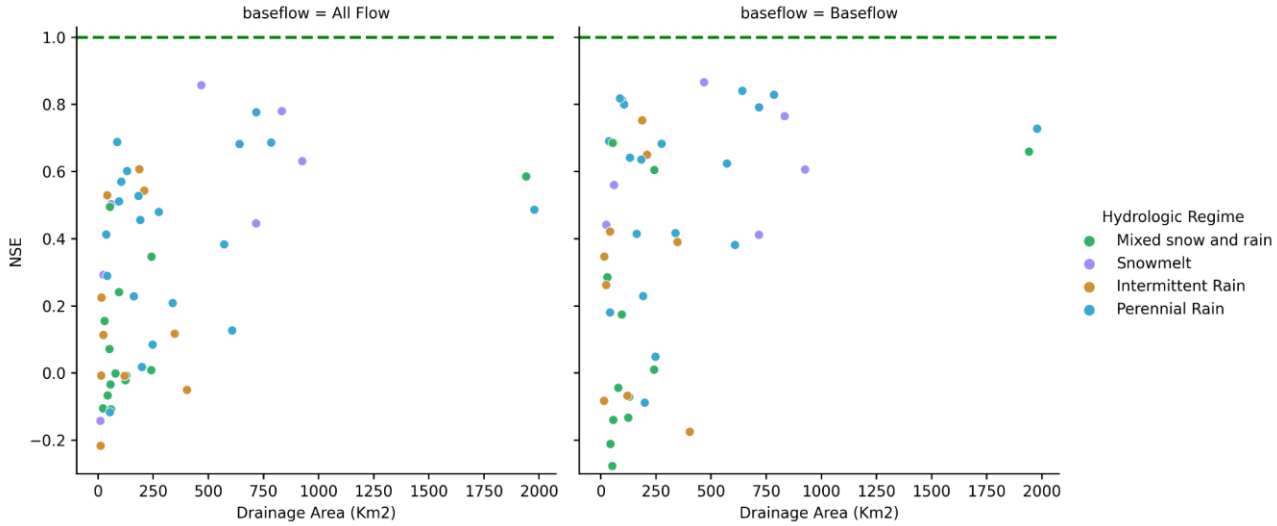


Figure 15. Unimpaired Flows model NSE scores across test sites (n = 53) versus drainage area, split by hydrologic regime, and All Flows (left) and baseflow only (right).

Figures 16, 17 and 18 show normalized bias, RMSE and correlation coefficient by drainage area calculated on the mean flow, separated by baseflow (left) and all flow (right). In term of bias, there is more variability in smaller basins, while larger basins have more consistent performance, generally slightly low biased between 0 and −0.5. This is true for all hydrologic regimes. In terms of RMSE, the variability is higher in the all flows than the baseflow. Otherwise the pattern is similar to bias, where larger sites have consistently lower RMSE and smaller than 500 km2 basins have more variability. All sites with higher RMSEs are in small basins. The correlation coefficient shows a similar pattern, where basins larger than 500 km2 have consistently higher correlation coefficients.

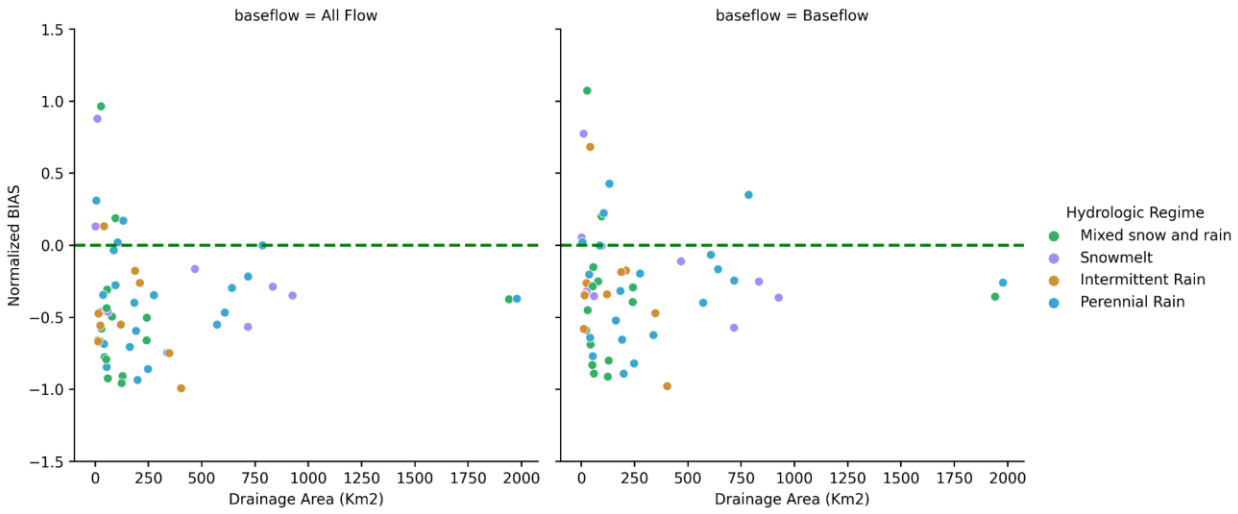


Figure 16. Unimpaired Flows model normalized bias across test sites (n = 53) versus drainage area, split by hydrologic regime, and All Flows (left) and baseflow only (right).

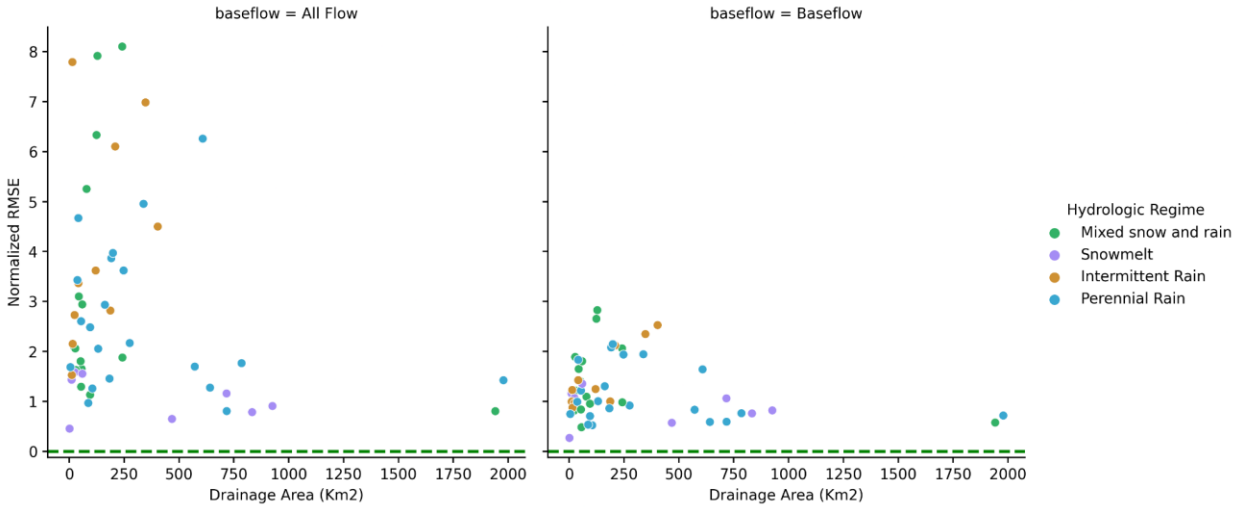


Figure 17. Unimpaired Flows model normalized RMSE across test sites (n = 53) versus drainage area, split by hydrologic regime, and All Flows (left) and baseflow only (right).

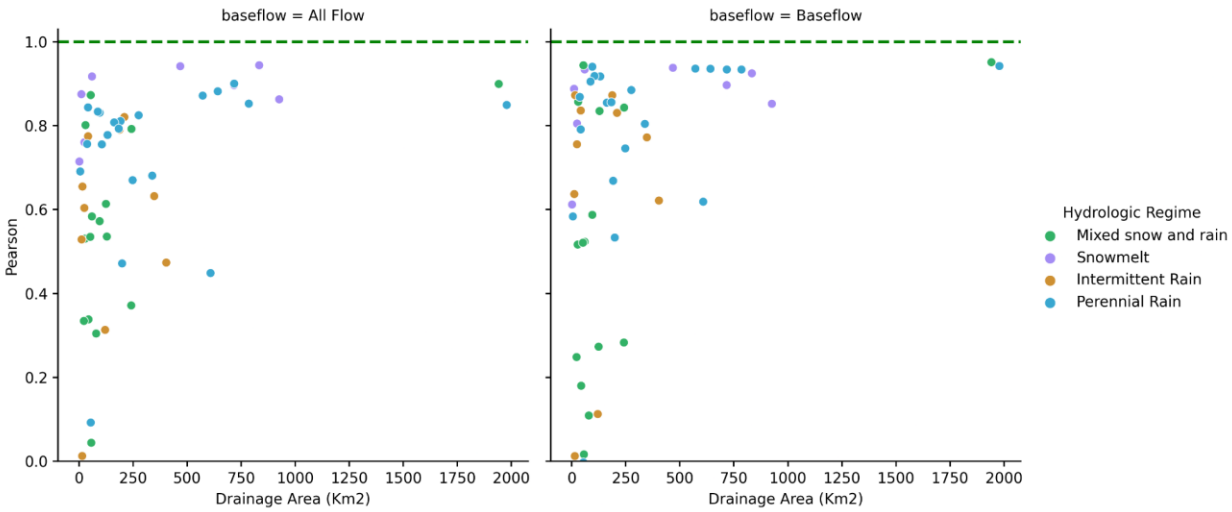


Figure 18. Unimpaired Flows model correlation coefficient (Pearson) across test sites (n = 53) versus drainage area, split by hydrologic regime, and All Flows (left) and baseflow only (right).

Figure 19 shows the 50% confidence interval metric: hit rate, by drainage area. Across all drainage areas and hydrologic regimes, the bounds are narrow. The bounds are closest to the expected 50% in the perennial rain category.

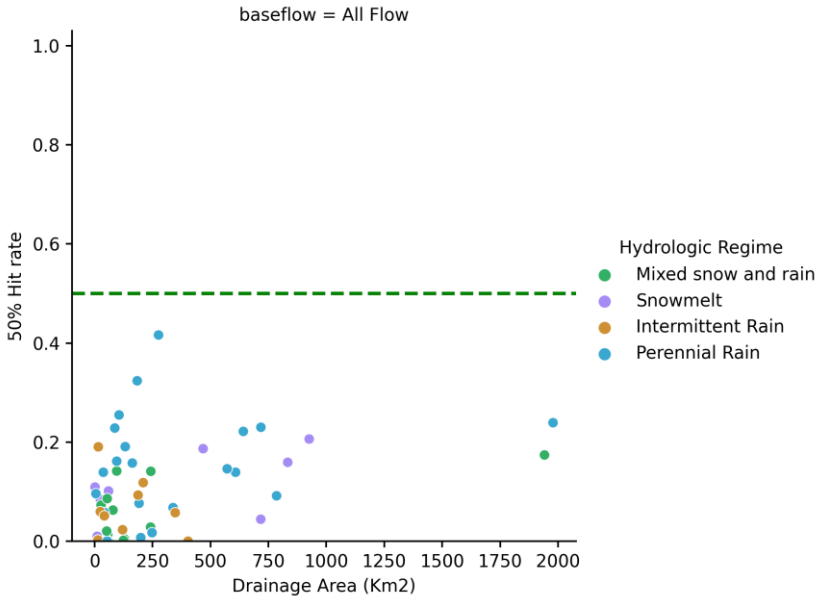


Figure 19. Unimpaired Flows model hit rate across test sites (n = 53) versus drainage area, split by hydrologic regime.

Spatial visualizations

Figures 20-23 illustrate performance metrics by spatial location in the state for the n = 53 test sites. We note that the southern part of the state with many of the sites in the mixed snow/rain regime are difficult for the model due to the flashy behavior, whereas the central and northern regions show high KGEs and correlations, and lower RMSEs and bias.

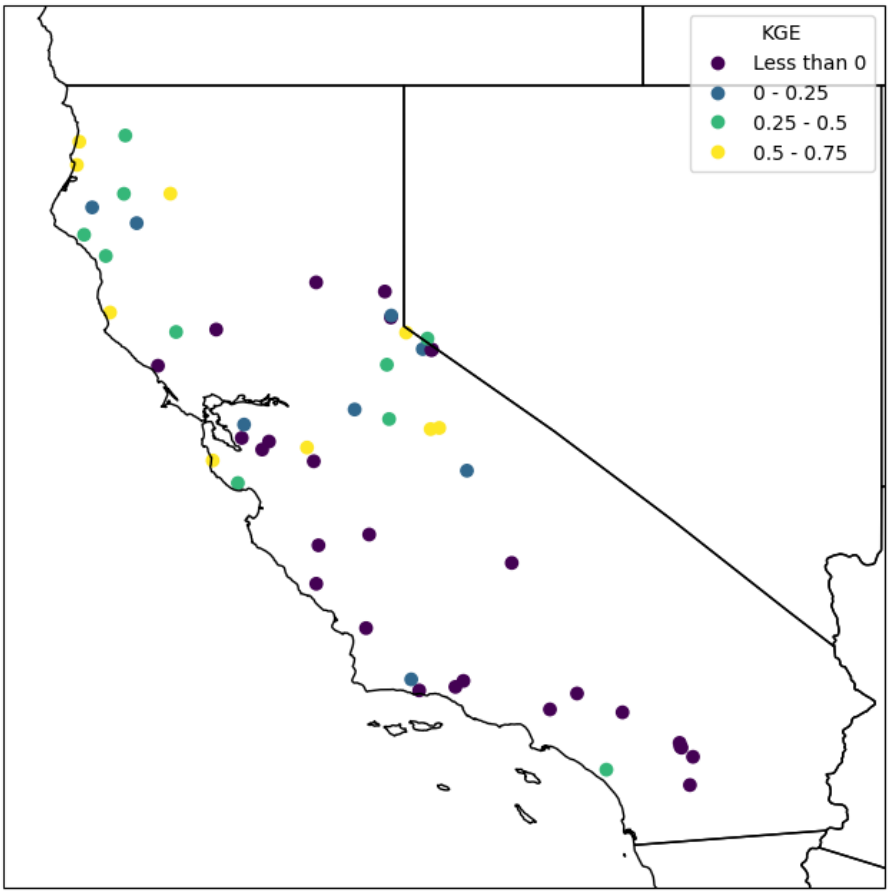


Figure 20. Unimpaired Flows model KGE across test sites (n = 53).

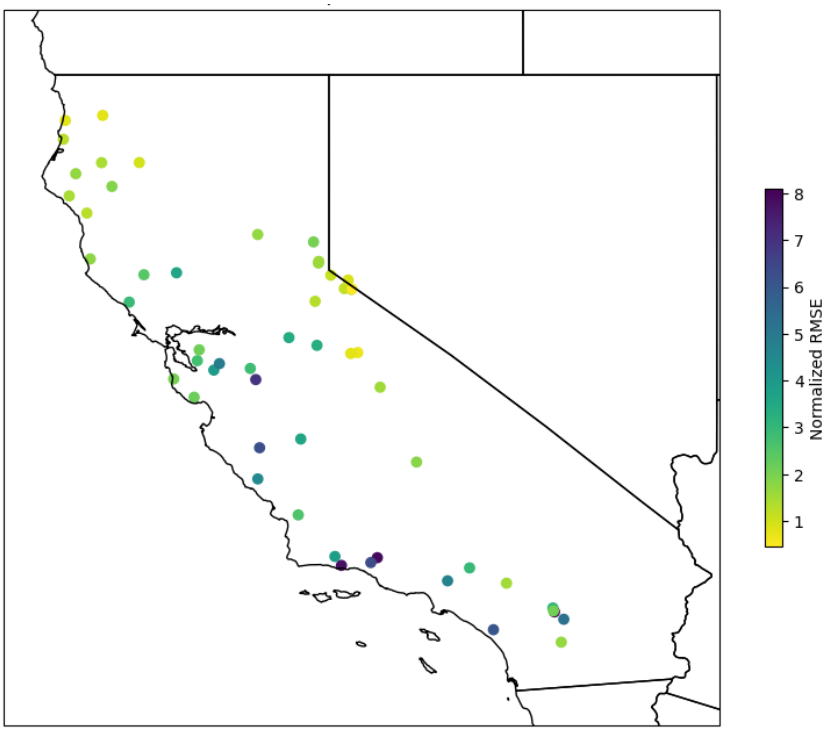


Figure 21. Unimpaired Flows model normalized RMSE across test sites (n = 53).

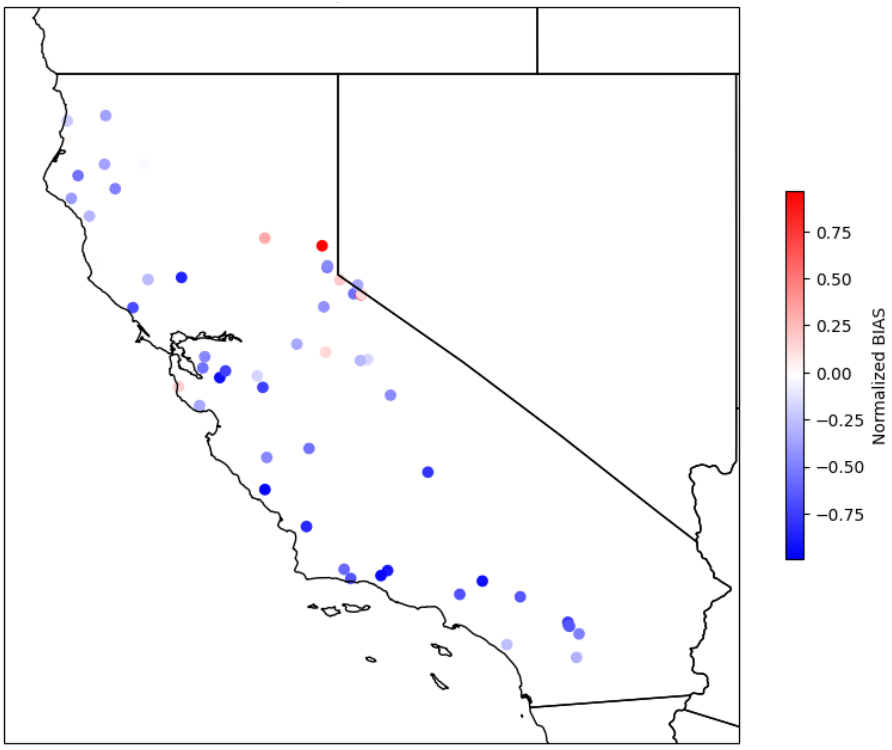


Figure 22. Unimpaired Flows model normalized bias across test sites (n = 53).

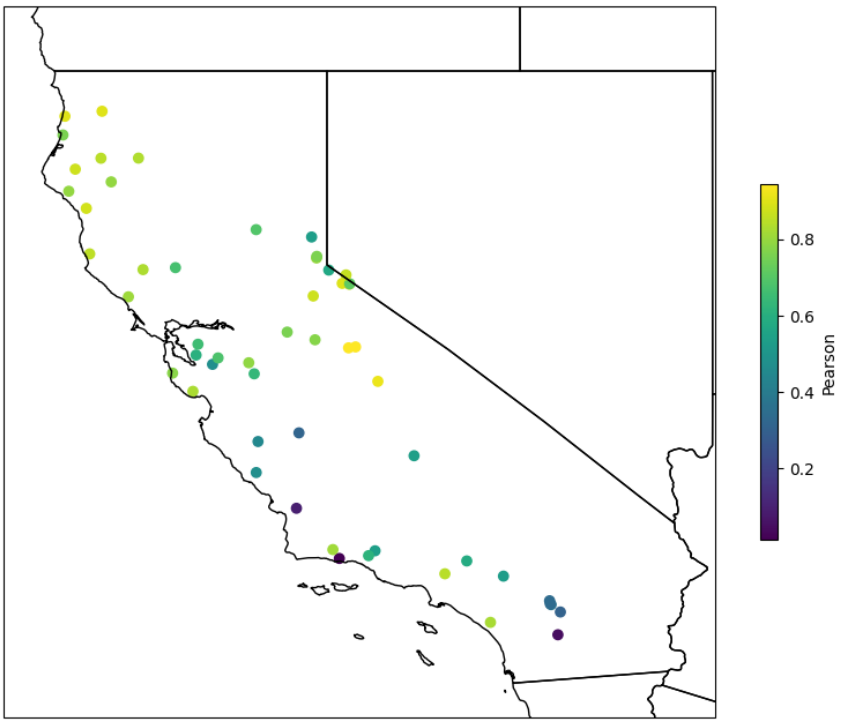


Figure 23. Unimpaired Flows model correlation coefficient across test sites (n = 53).

Additional guidance

In an attempt to further guide users of the data on which areas/regions and characteristics require caution when using this data, we explored several indices and thresholds to summarize performance outside of the metrics above. Since there are relatively clear areas where the model's performance is variable and/or relatively unexplored, we overlaid the validation site normalized RMSE with a variety of flow indices, thresholds, etc. The most meaningful cross-section that we identified is categorizing the sub-basin scale's normalized q75 flow, or in other words the q75 (75th percentile prediction) divided by the sub-basin area. This value represents a way to parse the wetness and dryness of sub-basin flow, also taking into account the size of the sub-basin. The reason q75 was selected versus other quantiles (or the median/mean) is that even the median in many basins is very close to zero in intermittent streams, and this caused the alignment between performance and normalized flow to not align with a useful threshold. Figure 24 shows a categorized map showing low to high normalized q75.

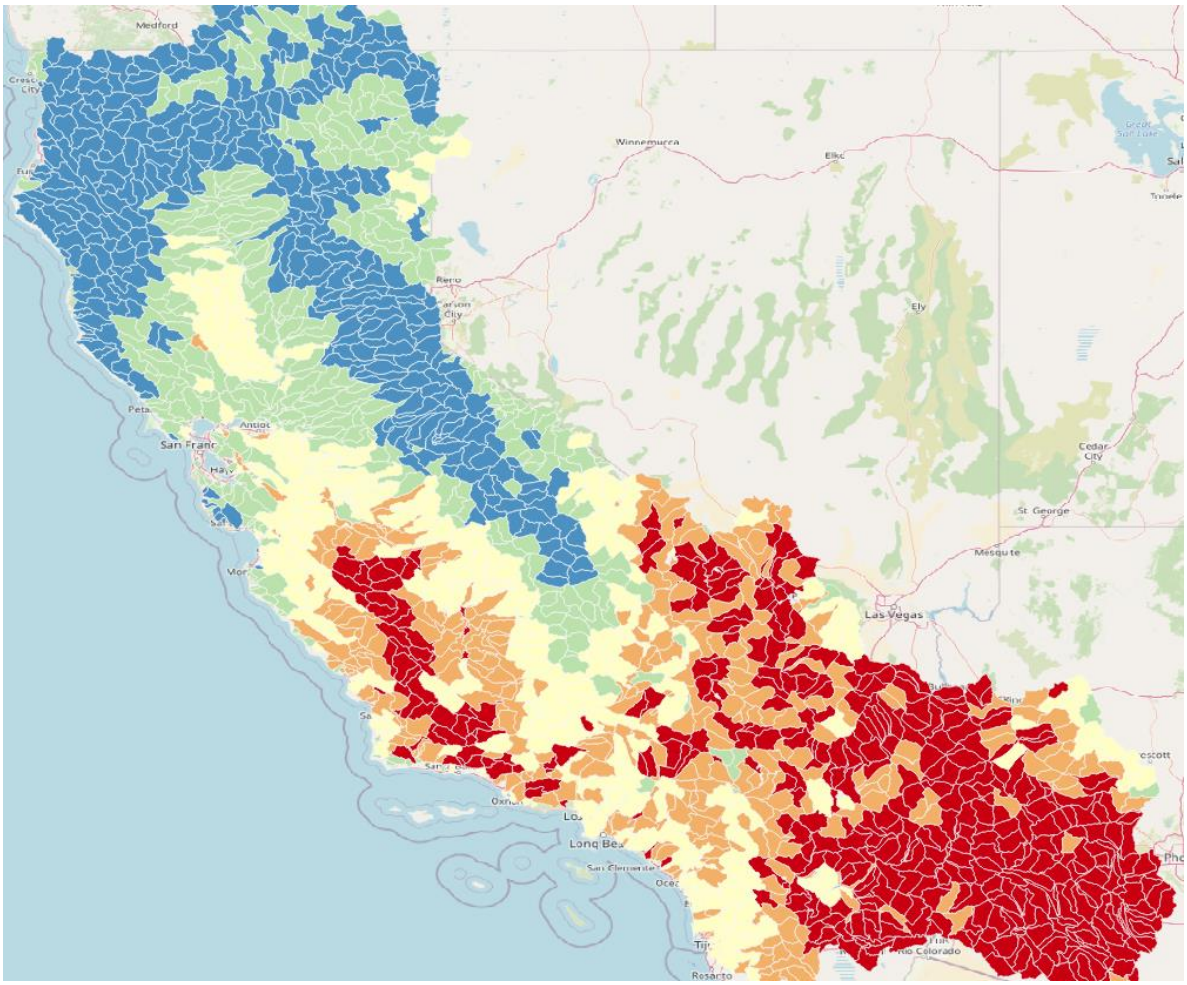


Figure 24. Sub-basins showing normalized q75 flow. Red basins have a normalized flow below 0.0005, and are in intermittent rain areas with few validation (test) sites. Upstream Tech's guidance is to use caution in utilizing predictions directly in these regions.

Figure 25 shows a closer image of the underlying areas below red sub-basins in the South, along with the validation site locations as white circles. Note the lack of validation sites in the Southeastern most part of the state, which is largely due to the lack of reliable gauges, which are less common on intermittent streams.

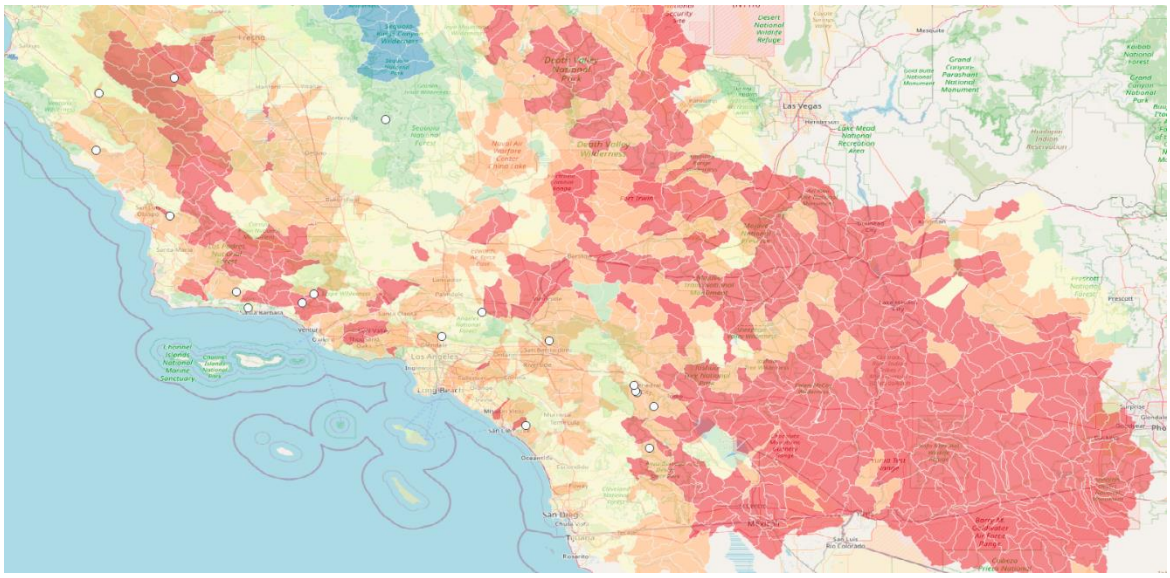


Figure 25. Closer view of the Southeastern part of the state without validation (test) sites and with very low normalized q75 flows. The guidance is to use caution in directly utilizing the predictions in these sub-basins and reaches within.

Actual Flows Model Results

Metrics Performance Summary

Hydrologic Regime	Flow type	KGE (median, mean)	Normalized Bias (median, mean)	Correlation (median, mean)	Normalized RMSE (median, mean)
Snow (n=8)	Base	0.43, -0.35	-0.12, -0.25	0.88, 0.78	1.03, 0.93
	All	0.38, -0.80	-0.24, -0.26	0.84, 0.77	1.23, 1.30
Perennial rain (n=38)	Base	0.39, 0.19	-0.34, -0.18	0.82, 0.75	1.26, 1.49
	All	0.09 ,0.10	-0.47, -0.38	0.76, 0.69	2.64, 1.49
Intermittent Rain (n=27)	Base	0.17, -0.06	-0.10, 0.03	0.63, 0.61	1.28, 1.69
	All	-0.08, 0.05	-0.44, -0.37	0.63, 0.61	3.84, 4.33
Mixed snow	Base	0.08, 0.11	-0.30, -0.24	0.72, 0.60	1.33, 1.38

and rain (n=27)	All	-0.07, 0.04	-0.35, -0.35	0.67, 0.56	1.66, 2.50

Note that these metrics include sites with upstream gauges that were utilized by the Actual Flows model. The list of comids can be provided separately to know which reaches had this benefit and in a later section, a sensitivity analysis shows the impact of the gauge/routing.

Key Takeaways

Highest confidence in utilizing the model data:

- Base flows
- Snowmelt driven basins
- Below gauges

Good performance, especially considering alternatives:

- Large basins
- Intermittent rain driven basins (base flows)
- Mixed snow and rain driven basins (all flows + base flows)
- Perennial rain driven basins (all flows + base flows)
- Ungauged locations (all flows + base flows)

Lower confidence, representing conditions where the performance and/or observations did not show a clear signal:

- Highly intermittent rain driven/desert dominated basins (all flows)
- Directly below regulating dams, before reaching another USGS gauge
- Southern coast, highly urbanized sites

Figure 26 shows the NSE skill scores for all test sites (n=100) by hydrologic regime, separating baseflow from all flows. High level takeaways are: NSE medians are above zero, indicating the model has more predictive power than the observed mean (*if the observed mean was known*) and the distribution of performance is slightly higher for baseflows than all flows. As with the Unimpaired Flows model, the model has the strongest performance in the snowmelt regime.

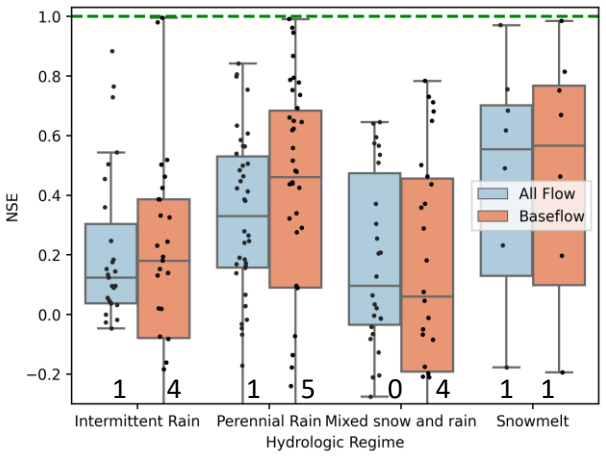


Figure 26. Actual Flows model NSE scores across test sites (n = 100) split by hydrologic regime. The number of sites off the plot boundaries are shown as annotations in the plots.

Figures 27- 29 show normalized bias, normalized RMSE and correlation across all Actual Flows test sites split by regime and baseflow / all flows. We observe slightly higher performance in the baseflow in all metrics, especially higher in RMSE, and with snowmelt as a consistently strong regime for the model.

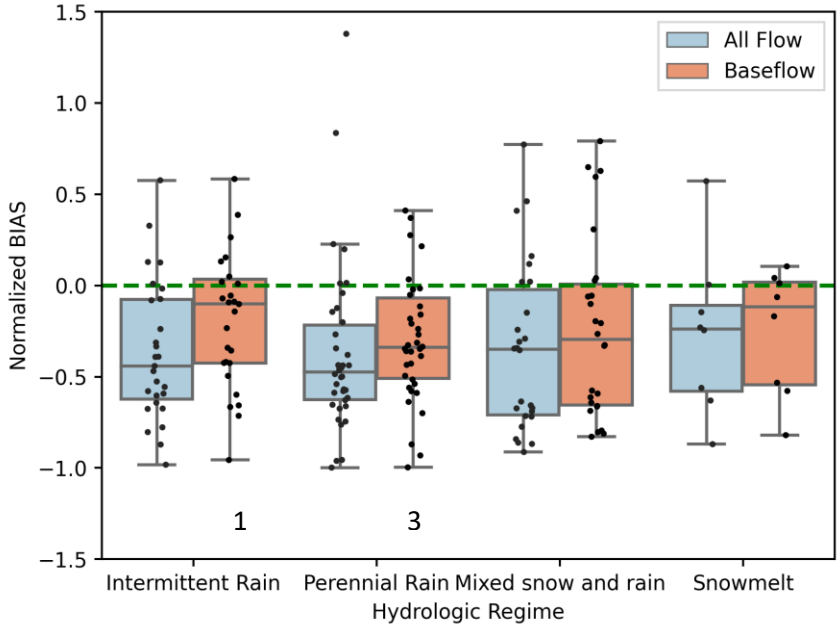


Figure 27. Actual Flows model normalized bias across test sites (n = 100) split by hydrologic regime.

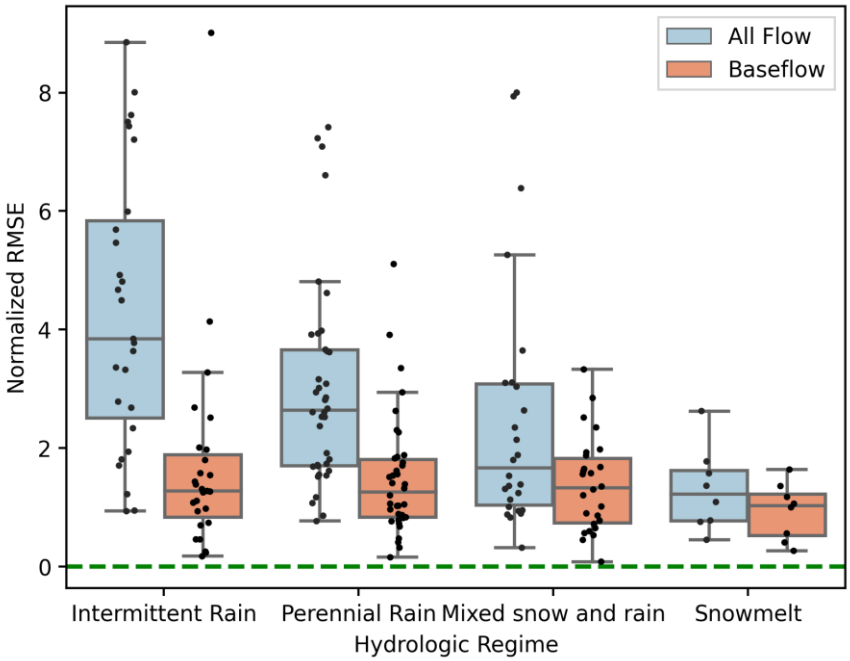


Figure 28. Actual Flows model normalized RMSE across test sites (n = 100) split by hydrologic regime.

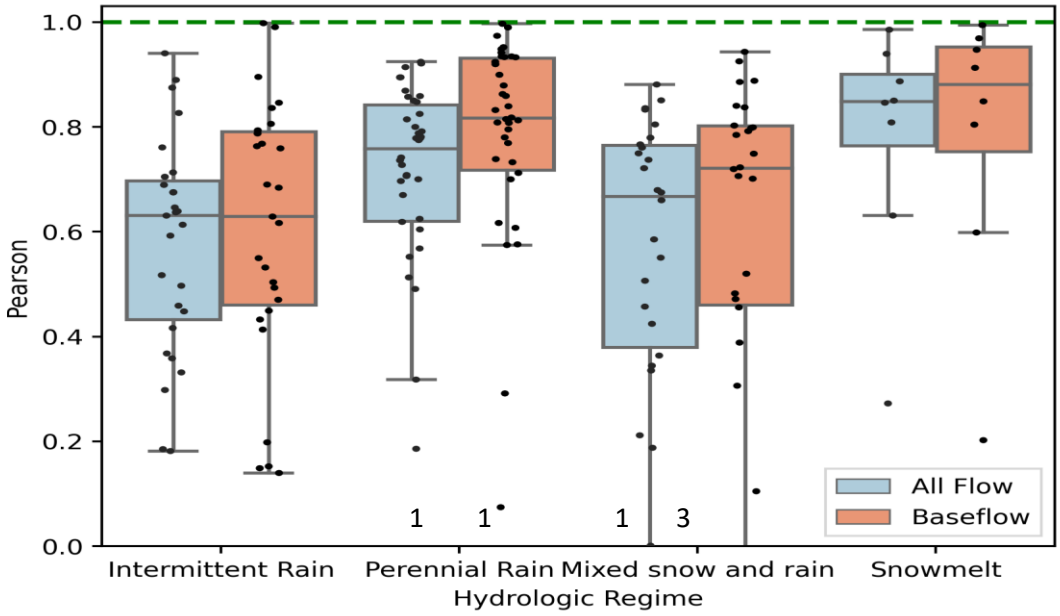


Figure 29. Actual Flows model correlation across test sites (n = 100) split by hydrologic regime.

By drainage area

The following plots (Figures 30-33) show scatterplots of performance for each metric (NSE, bias, RMSE, correlation) across all drainage areas, with the dots colored by hydrologic regime.

Generally, as with Unimpaired Flows, the model has more consistent lower (better) RMSE in larger basins. NSE and bias do not show a clear performance pattern by drainage area, with mixed performance in both base and all flows. There are small sites with higher NSEs and lower NSEs. The bias plots indicate that generally the model underpredicts across the majority of sites. All sites with low correlation values are small sites, likely with a handful of flashy rain events that the model did not capture.

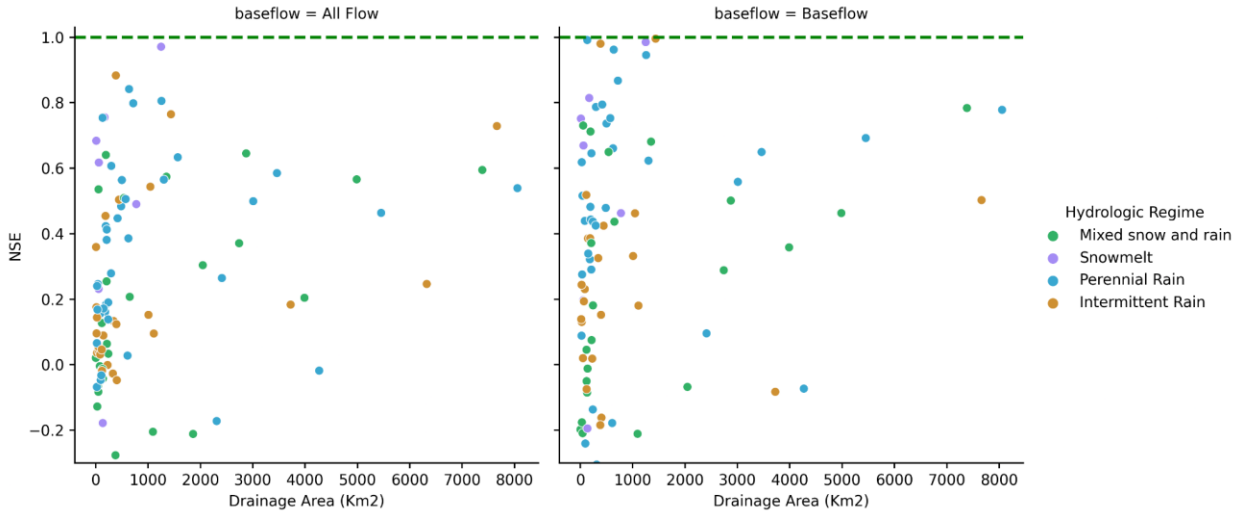


Figure 30. Actual Flows model NSE across test sites (n = 100) vs drainage area, split by hydrologic regime.

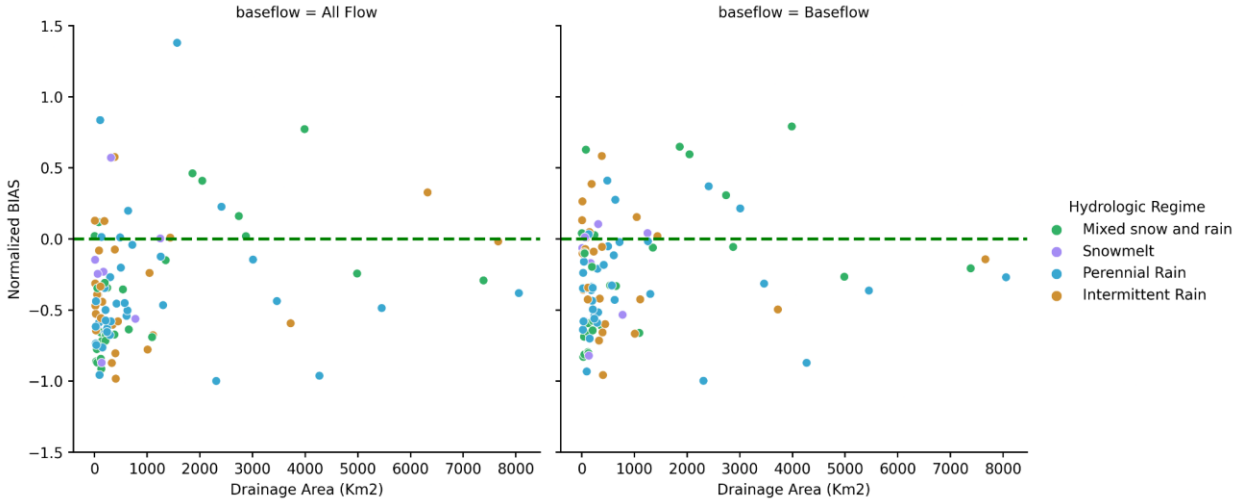


Figure 31. Actual Flows model bias across test sites (n = 100) vs drainage area, split by hydrologic regime.

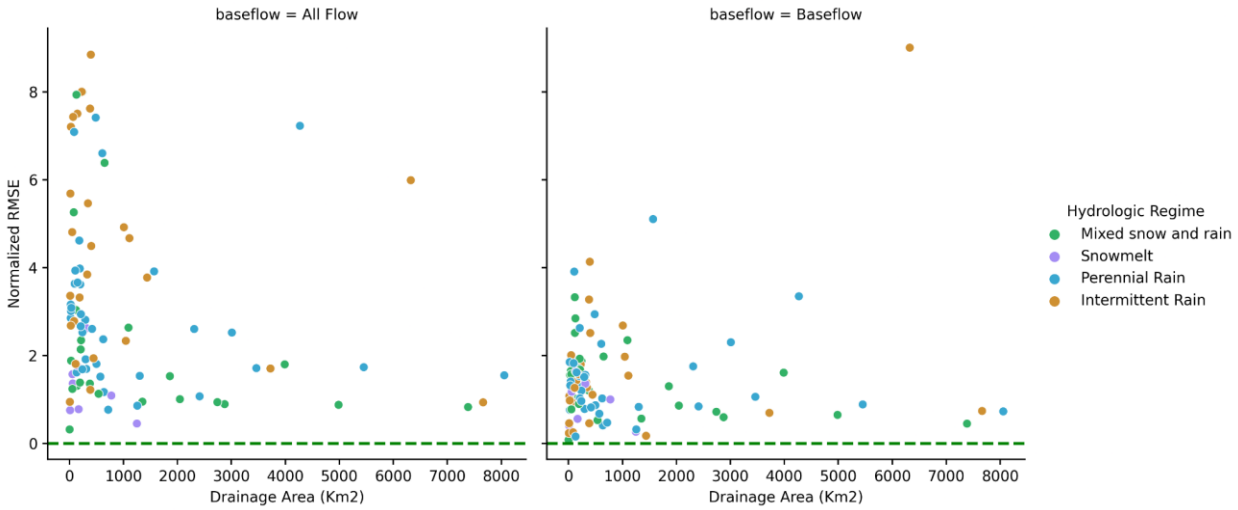


Figure 32. Actual Flows model RMSE across test sites (n = 100) versus drainage area, split by hydrologic regime.

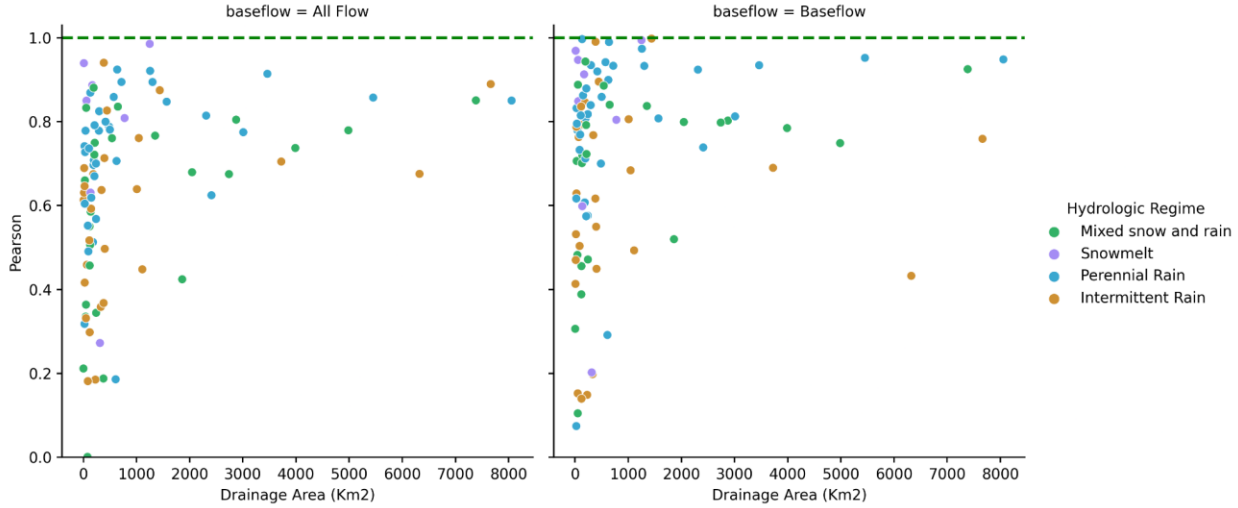


Figure 33. Actual Flows model correlation across test sites (n = 100) vs drainage area, by hydrologic regime.

In examining confidence interval accuracy and drainage area, Figure 34 shows no particular pattern by regime or drainage area, with the 50% confidence interval too narrow across the board.

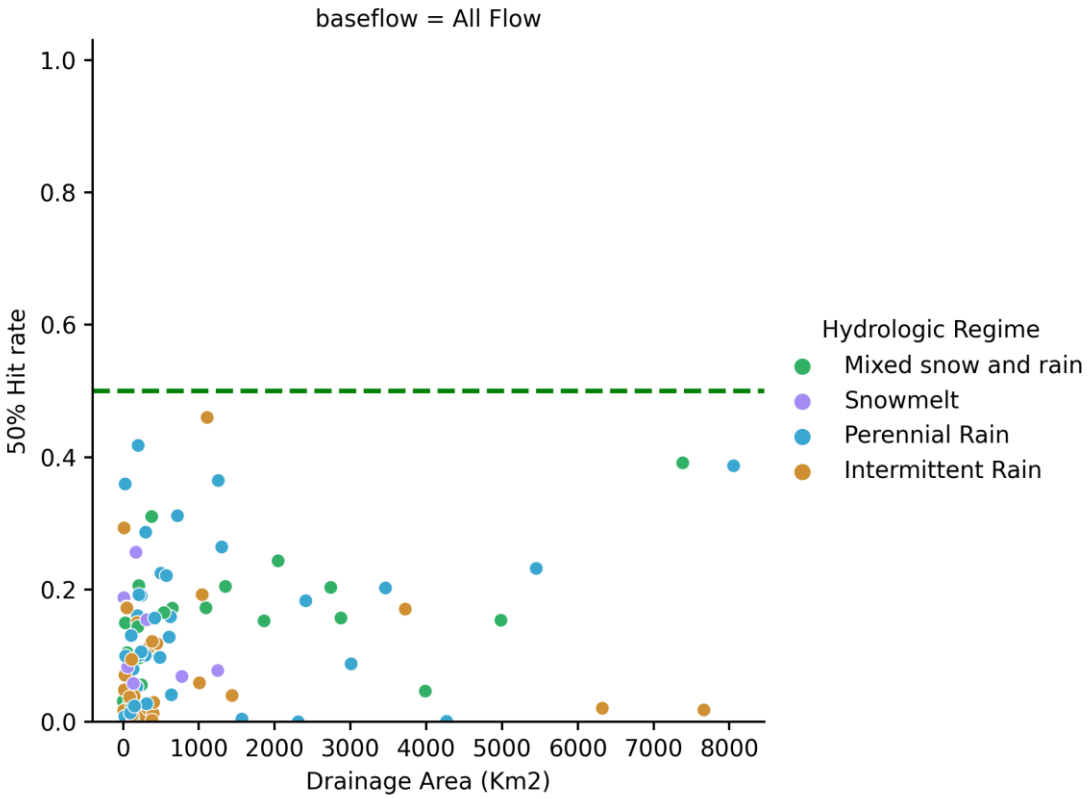


Figure 34. Actual Flows model confidence interval hit rate across test sites (n = 100) versus drainage area, split by hydrologic regime.

By alteration category

As noted in the Evaluation Metrics section, we split the Actual Flows metrics into particular alteration categories (upstream storage, canal density and developed land). Figures 35-37 show performance metrics by **upstream storage**, where we note the following:

- The distribution of performance does not degrade in either baseflow or all flows because of upstream storage; in fact the bias distribution improves with upstream storage.
- The RMSE in the baseflow is consistent across the categories of upstream storage (none, low, high), whereas with all flows, the RMSE is highest in *none* and lowest in *high*. This indicates the model did learn flow magnitude changes in basins with upstream dams.

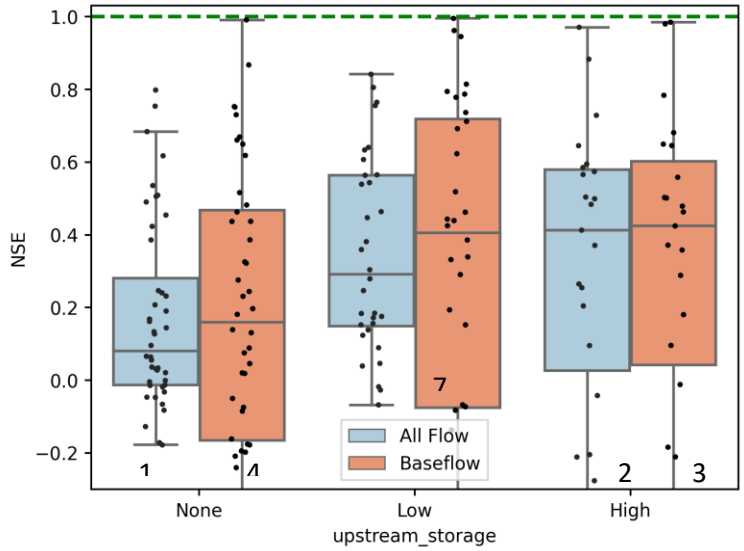


Figure 35. Actual Flows model NSE across test sites (n = 100) split into upstream storage categories (none, low, high), and by baseflow and all flows.

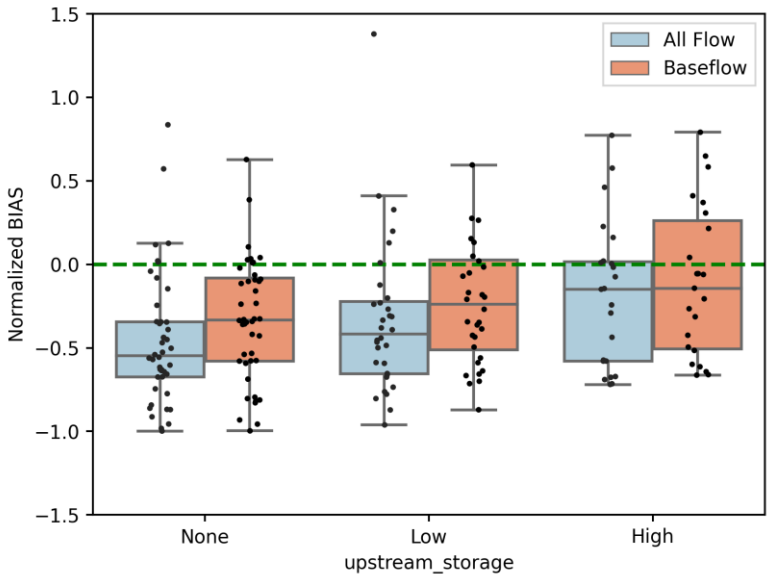


Figure 36. Actual Flows model bias across test sites (n = 100) split into upstream storage categories (none, low, high), and by baseflow and all flows.

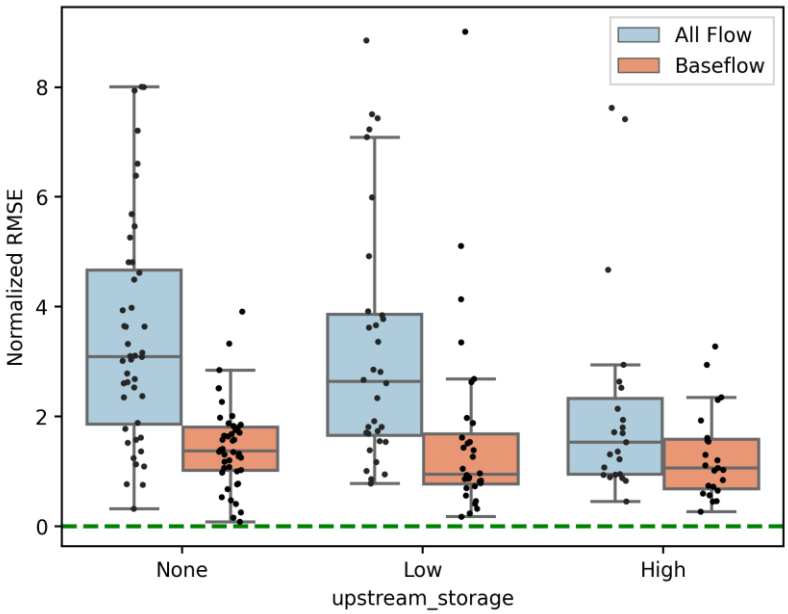


Figure 37. Actual Flows model normalized RMSE across test sites (n = 100) split into upstream storage categories (none, low, high), and by baseflow and all flows.

Figures 38-39 show **canal presence** at Actual Flows test sites, and we note:

- The model learned to pick up on the impacts of canals, with performance higher in test basins with canals present.
- The model continues to show strong performance in baseflows, even at sites with canals.

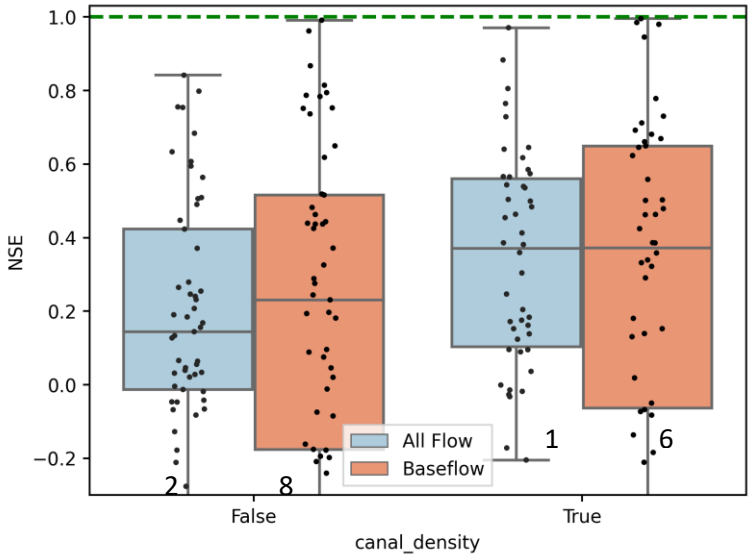


Figure 38. Actual Flows model NSE across test sites (n = 100) split into canal presence (true, false), and by baseflow and all flows.

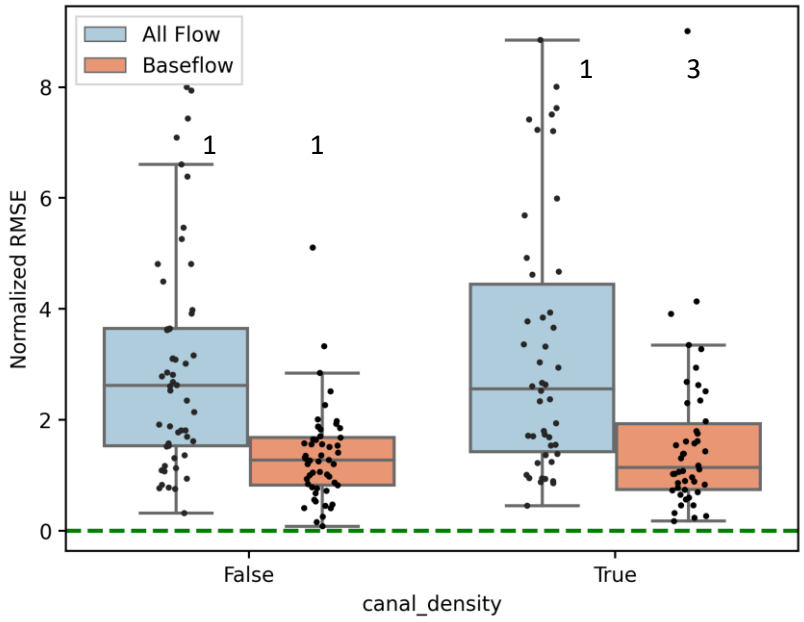


Figure 39. Actual Flows model bias across test sites (n = 100) split into canal presence (true, false), and by baseflow and all flows.

Developed land (above or below 5%) is shown in Figures 40-42 at Actual Flows test sites. We note:

- All flows performance is better in less developed basins (NSE is higher, RMSE and bias are lower).
- Baseflow only performance is more nuanced, but overall still shows slightly better performance in less developed sites: the bias distributions between developed and less developed are relatively similar; NSE has a higher median value in less developed sites with wide variability in the distribution; the median RMSEs are quite similar, slightly higher in the more developed sites.

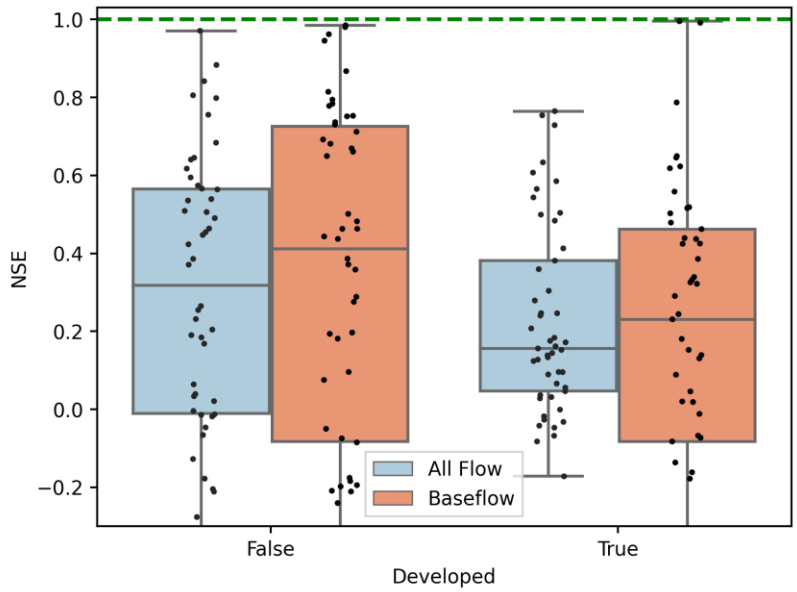


Figure 40. Actual Flows model NSE across test sites (n = 100) split into developed land (true = > 5%, false = < 5%), and by baseflow and all flows.

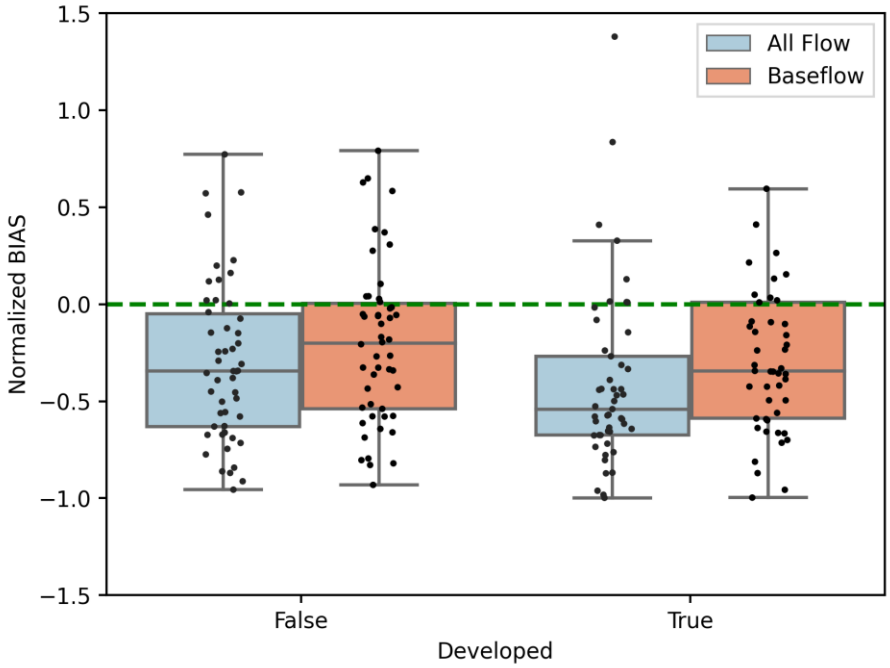


Figure 41. Actual Flows model bias across test sites (n = 100) split into developed land (true = > 5%, false = < 5%), and by baseflow and all flows.

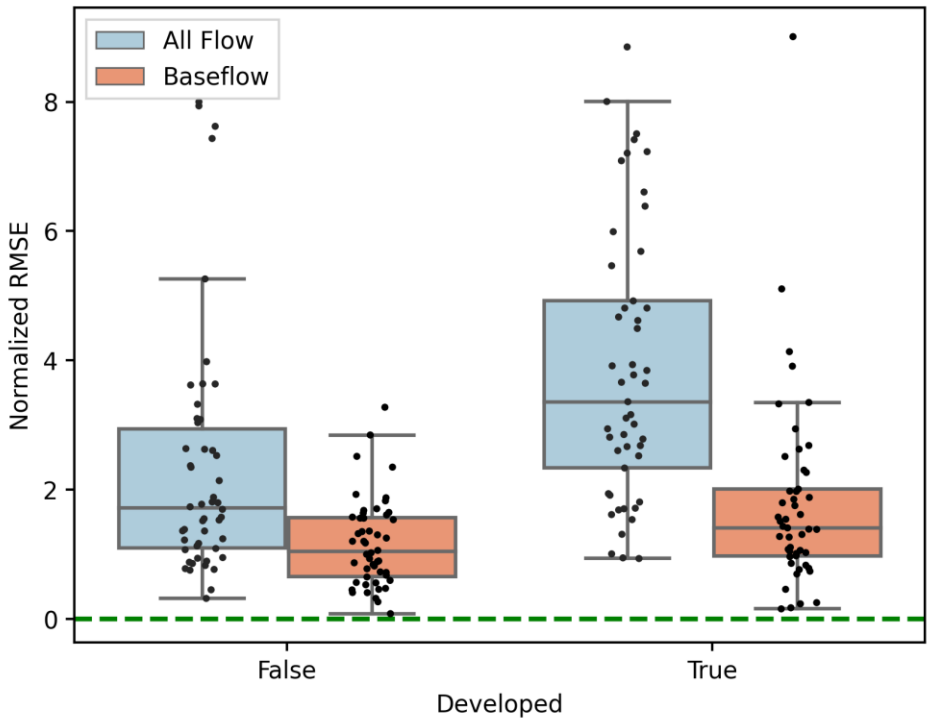


Figure 42. Actual Flows model RMSE across test sites (n = 100) split into developed land (true = > 5%, false = < 5%), and by baseflow and all flows.

Spatial visualizations

Figures 43-46 illustrate each metric spatially across the state, for n=100 sites. The southern part of the state, which has flashy intermittent rain and desert type hydrology shows lower KGE scores and a mix of performance according to RMSE, bias and correlation. The central and northern part (inland and coast) has generally higher metrics and we have high confidence in basins in this region, even in highly impaired basins. There are a few sites with lower KGEs, but in general this region is a higher performing area in the state.

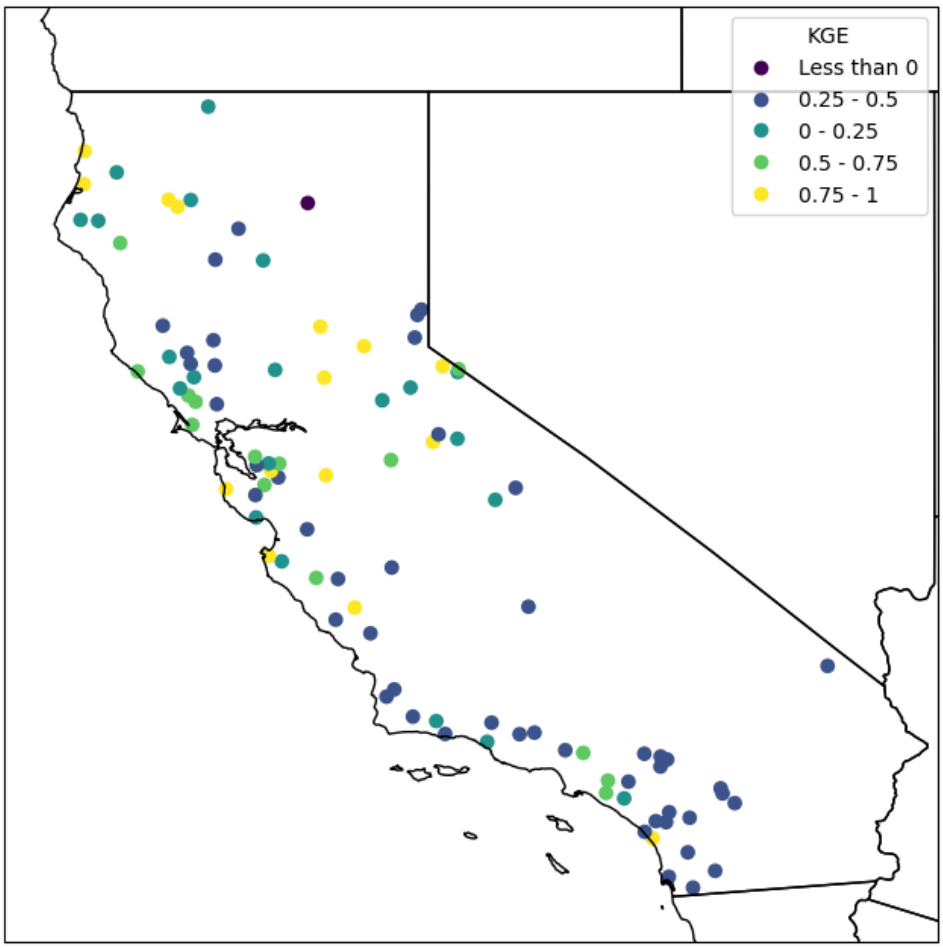


Figure 43. Actual Flows model KGE across test sites (n = 100).

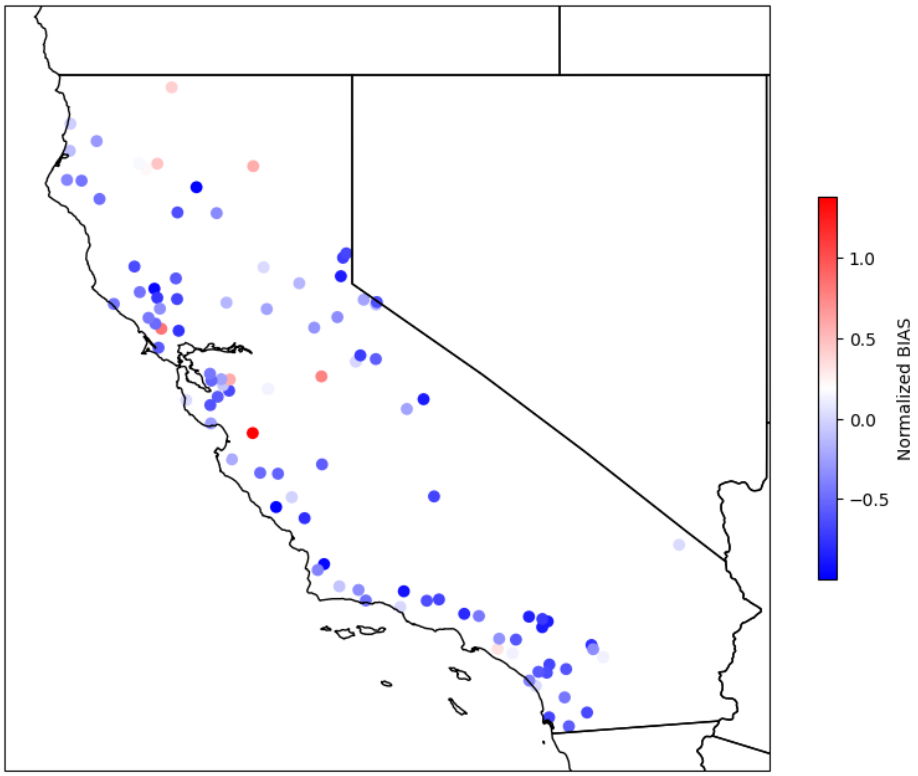


Figure 44. Actual Flows model normalized bias across test sites (n = 100).

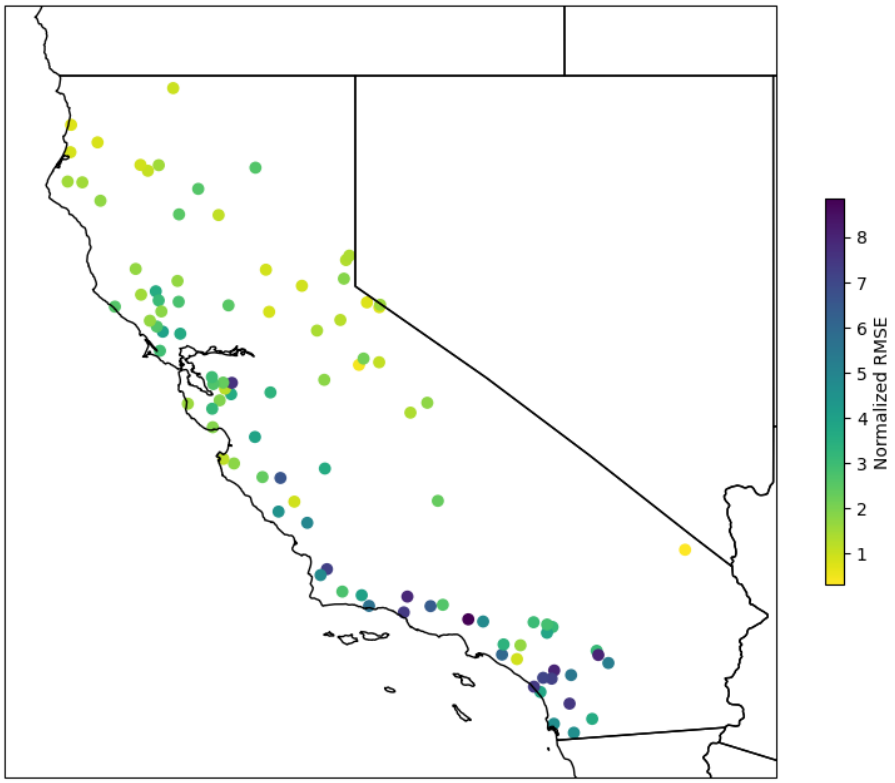


Figure 45. Actual Flows model normalized RMSE across test sites (n = 100).

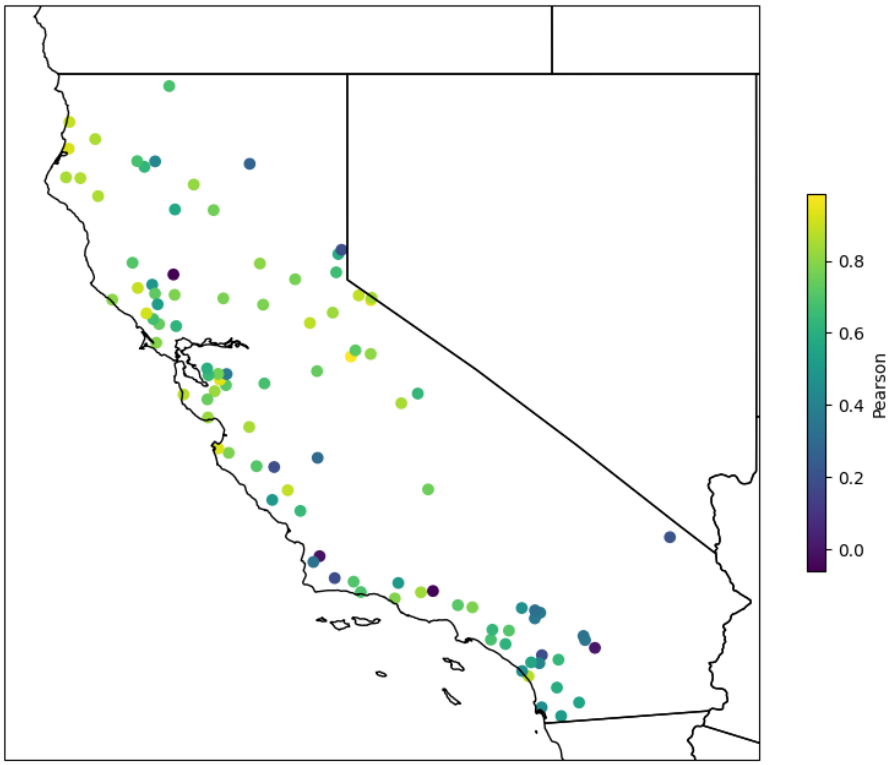


Figure 46. Actual Flows model correlation coefficient across test sites (n = 100).

Sensitivity Analysis for Routed/Non-routed sub-basins

The model identified six USGS gauges that were active and available with data over the historical period within the Actual Flows test sites. The test site basins with upstream gauges are shown on the map in Figure 47, which are all within the San Francisco bay area.

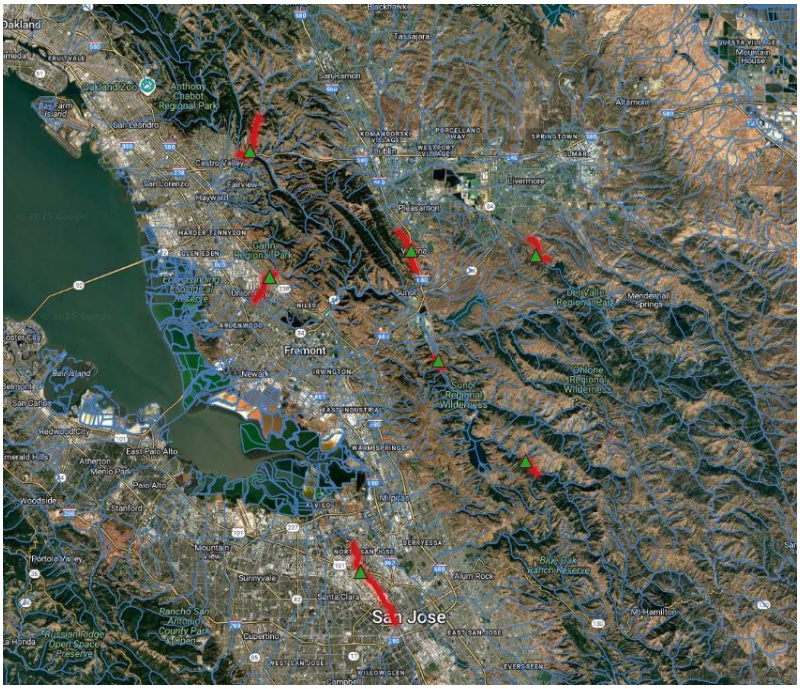


Figure 47. Map of USGS gauges (green triangles) in test set Actual Flows basins, which benefitted from routing of flow during the historical time period.

We ran the model twice, once with routing these gauged flows and again without the gauged flow routing to understand the difference in performance between these two scenarios. As expected, the test sites with upstream basins generally had better performance metrics: this was true at five of the six sites. The table below summarizes the average percent difference (improvement if positive, degradation if negative) for the six sites.

Flows	KGE	NSE	Pearson	RMSE
Baseflow	24.8	49.2	10.5	492.6
All Flows	64.0	49.5	-7.8	23.5

Generally, routing gauged flow leads to improvement in both flow categories, though since the gauges and basins were in the same hydrologic regime and geographic area the broader messages about impact are somewhat limited. With additional gauges embedded, we might expect model performance to increase, and this would be a goal in any future revision of this work.

Conclusions & Future Work

Overall, this modeling effort produced *two billion* time series data points, and is the first of its kind in using a machine learning hydrologic model to create a state-of-the-art daily historical streamflow prediction dataset in every river of California. The next steps of this work are to

publish the flows dataset on TNC's web-based tools and for TNC to calculate functional flows for ecological e-flow uses and to work with stakeholders to apply the data to their use case.

Producing outputs at this scale required a phased approach to build a system capable of processing inputs, making base model predictions and routing them all within a single model architecture. One of the challenges in this work was aligning the flows to NHD v2 flowline dataset, which have some reliability issues with its delineation, and some reaches did not specify drainage attributes. These made it difficult to map as expected to the topology and led to removal of about 1% of flowlines. Another challenge in some of the larger sub-basins and parts of the state with many braided small streams is that the size of the sub-basins relative to the size of the flowlines led to a smoothing of the downscaling to the reach scale. Smaller sub-basins in dense reach areas would likely improve the downscaling transition. One final challenge is that *tuning* a fully routed flow model is extremely computationally intensive, and would likely improve the base model predictions. This is an area of future work.

In terms of future work to improve the model, as with any research and science work there is more to be explored with experimentation. A few specific follow on ideas if there are additional phases of work include:

- Improvement of confidence interval width in base flow predictions. CIs were too narrow. These can be improved by adding noise in the training process to introduce more uncertainty and adjusting predicted bounds where observations are available based on known uncertainty.
- Increase in the training/test sites and inputs/variables in intermittent and flashy basins to improve the model learning these patterns.
- Add a second weather reanalysis source in addition to ERA5-Land to rely less on a single source of precipitation and add uncertainty to the predictions based on natural weather phenomena.

References

Geological, U.S., 2011, GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow: U.S. Geological Survey data release, https://doi.org/10.5066/P96CPHOT.

Hill, Ryan A., Marc H. Weber, Scott G. Leibowitz, Anthony R. Olsen, and Darren J. Thornbrugh, 2016. The Stream-Catchment (StreamCat) Dataset: A Database of Watershed Metrics for the Conterminous United States. Journal of the American Water Resources Association (JAWRA) 52:120-128. DOI: 10.1111/1752-1688.12372.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. Water Resources Research, 55(12), 11344-11354.

Lane, B.A., S. Sandoval-Solis, E.D. Stein, S.M. Yarnell, G.B. Pasternack, and H.E. Dahlke, 2018, Beyond metrics? The role of hydrologic baseline archetypes in environmental water management. Environmental Management. doi: 10.1007/s00267-018-1077-7

Appendix

Full list of model inputs

Unimpaired Flows inputs

```
# ERA5 Land weather
```

- land-temperature
- land-dewpoint
- humidity
- precipitation
- solar radiation
- wind
- wind-u
- wind-v
- pressure
- evaporation
- runoff
- sensible_heat_flux
- latent_heat_flux
- net_solar_radiation
- snowmelt
- snowfall
- snow water equivalent
- snow_temperature
- snow_cover
- snow_density

- soil_water_content
- soil_temperature

Satellite Land Surface Observations (daily)
- vegetation vigor (NDVI)
- snow extent
- daytime land surface temperature
- nighttime land surface temperature

Characteristics of the basin - static, single values
- average daytime land surface temperature
- average nighttime land surface temperature
- average vegetation presence and vigor
- average snow extent
- average precipitation
- fraction of precipitation which falls in large storms
- area of the drainage basin
- fraction of the precipitation which falls as snow
- average elevation in meters from sea level
- average elevation slope of the basin

Soil related variables - a set of 20 - summarized here
- permeability
- clay % and type
- sand % and type
- mean seasonal water depth of soils in catchment
- AgK factors
- Kf factors
- Average depth to bedrock
- Compressive strength

Lithology variables - a set of 25 - that include the percents from different lithology classes (coast, glacial lakes, alluvial, etc.)

Ungauged Actual Flows inputs

All of the unimpaired flows model inputs + what is listed below:

The main source for these additional variables is the EPA's StreamCat dataset, from which we include ~ 400 variables. We input them into our model using a technique that groups/clusters them. See the definitions and complete list here: https://www.epa.gov/national-aquatic-resource-surveys/streamcat-metrics-and-definitions

Land use information - groups of variables summarized below
Includes many land use types & percents, e.g. % of watershed - woody wetland (about 100 variables here)
- Road density data, e.g. density of roads within watershed
- Agriculture/irrigation human activity, e.g. fertilizer applied per area
- Population density
- Density of mines, and coal mines specifically

Hydro properties
- Total dam storage volume

- Normal dam storage volume
- Base flow index
- Canal density
- Percent open water
- Topographic wetness index

Climatological average weather variables
- 30-year mean precipitation
- 30-year max temp
- 30-year mean temp
- 30-year min temp