Water Supply Forecast Rodeo - Model Explainability Report

Author: Rasyid Ridha (rasyidstat)
Model ID: `ens_all`
Model date: 2024-03-17

Summary

We use SHAP (SHapley Additive exPlanations) to calculate the percentage of feature contribution and relative contribution as explainability metrics for a given location and issue date. We present forecast predictions starting from the initial forecast update in January, along with historical median, average, minimum, and maximum values. In the explainability summary, we describe explainability metrics with the latest two consecutive issue dates as a comparison, including values for each feature, typically expressed as percent of normal.

Forecast and uncertainty communication

We present the water supply forecast and its uncertainty bounds, starting from the initial forecast update in January, using a line plot. The visualization provides an overview of the forecast's evolution and indirectly describes how the predictors, especially snowpack, change throughout the season. In addition, we provide additional context data of historical water supply volume statistical summary such as median, average, minimum, and maximum value to show how the forecast compares to historical value.

The median and average values describe the normal condition of the water supply. These values calculations use the last 30 years of historical data from 1991-2020, commonly used by the official and as required by the World Meteorological Organization (WMO).[1,2] In the narrative detail, we also include percent of normal number beside the original forecast output. We use the median value to calculate percent of normal because it is not affected by extreme values and is also commonly used by NRCS.[2] However, we include both information in the plot as a comparison to give an idea of the skewness and impact of extreme values, usually coming from wet years.

The minimum and maximum values describe the possibility of extreme forecast results and provide a possible water supply range based on historical data. Forecast results near or even outside of this range should be given attention. Unlike median and average values, minimum and maximum values calculations use all years used for model training. Internally, this provides a view of model capability in predicting extreme values outside the historical minimum and maximum range (e.g., Merced River in 2023).

Finally, we include the latest known historical volume within the seasonal month target, typically from April to July. In the narrative detail, we emphasize that the original forecast generated within the seasonal month target excludes the latest known historical value. We also include the percentage between the latest known historical volume and forecast to give an idea of the remaining water supply forecast for the rest of the season, which is coming from snowpacks and potentially from future precipitation.

Explainability metrics and communication

SHAP method

We use SHAP (SHapley Additive exPlanations) as the base explainability metric to understand why our model generates specific forecast output for a given location and issue date. SHAP is based on Shapley values from cooperative game theory, which attributes the payout (in this case, the forecast prediction) fairly among the features based on their contribution. The additive feature attribution method used by SHAP ensures that sum of all feature contributions equals to the model's prediction, providing a linear and interpretable approximation for any machine learning model.[3]

SHAP calculation for ensemble

There are two approaches to calculating SHAP values for our ensemble of LightGBM models, KernelSHAP and TreeSHAP. At first, we tried to use KernelSHAP because we can use the forecast output of the ensemble directly as the approximation to obtain SHAP values. However, it is very slow to calculate. Hence, we use TreeSHAP applied to all single LightGBM model variants which is faster to calculate.

SHAP values generated from all single LightGBM model variants are then aggregated, the same way as we did in the ensemble, benefiting from the additive feature attribution method used by SHAP. It means that we can simply average SHAP values for the deterministic forecast (quantile 0.5). However, we need to adjust the calculations for probabilistic forecasts (quantiles 0.1 and 0.9). With nine models, the 10th and 90th percentile ensemble for the probabilistic forecast is equivalent to a weighted ensemble of two models, where the models with the 1st and 2nd highest/lowest forecast are used. Specifically, SHAP values are calculated as 0.2 * 1st lowest + 0.8 * 2nd lowest for 0.1 quantile forecast, and 0.2 * 1st highest + 0.8 * 2nd highest for 0.9 quantile forecast. As a result, depending on which models are used for the weighted ensemble, not all features will have SHAP contribution values.

Derived explainability metrics

On top of SHAP values, we introduce two derived explainability metrics, such as percentage of feature contribution and relative feature contribution. The percentage of feature contribution is calculated by taking SHAP value for a feature divided by sum of absolute SHAP values for all features. In the explainability summary, we use percentage of feature contribution rather than raw SHAP value due to simple, flexible, and comparative nature of percentage. In addition, we can also compute relative feature contribution (named "Rel" in the explainability summary), which is calculated based on percentage of feature contribution relative to the average of all sites for a particular issue date. Relative feature contribution is only calculated for latest issue date based on feature contribution from deterministic forecast model.

The percentage of feature contribution should not be interpreted as the amount of water supply value originating from the features or predictors. It can be interpreted as the percentage of attribution from a specific feature amongst all features used in the model. For example, if the percentage of feature contribution from snowpacks is 50%, it does not mean that half of the seasonal water supply volume originated from snowpacks. Instead, the forecast output has a 50% contribution from snowpack features among all features used in the model. We should highlight that the percentage is calculated within the context of all features used in the model.

Relative feature contribution is used to understand whether any specific minor or major feature contributes to forecast output relative to the average of all sites. A relative contribution higher than 1 means that the feature has more impact compared to the average of all sites, while a relative contribution lower than 1 means that the feature has less impact compared to the average of all sites.

Explainability metrics in forecast summaries

We summarize the explainability metrics in a table for each feature along with the output from the previous issue date as a comparison. We aggregate the explainability metrics into feature groups such as base, snowpack, precipitation, drought, upcoming weather (SEAS51 forecast), and others. This will give a better understanding of how a specific feature group, for example, snowpack, contributes to the forecast output regardless of the data source used.

The base feature group includes features derived from issue date, site, supplemental, and static features (metadata). These features can be interpreted as base or average predictions if no predictors are included in the model. A higher percentage of feature contribution from base features indicates that a basin has less forecastability and is less affected by snowpack since snow cover area is lower (Figure 5, Figure 6). Internally, it can highlight room for improvement in the model building. For example, although snow cover area is high, Libby Reservoir Inflow forecasts are still mainly contributed from base features. As we recommend in the Final Model Report, developing a sub-model might improve the prediction because averaging for larger basin area size will lose more information. Besides, only partial predictors are also available for this location which might explain the higher contribution from base features.

Percent of feature contribution for an individual feature can be positive or negative depending on its value and relationship with water supply output (Figure 7, Figure 8). A positive value indicates that the feature is contributing positively to the model's prediction, pushing the prediction towards a higher value. Conversely, a negative value indicates that the feature is contributing negatively to the model's prediction, pushing the prediction towards a lower value. For example, in wet years, a higher amount of snowpack will contribute positively to the model's prediction. Also, depending on when the forecast is made, snowpack can contribute negatively to the model's prediction because it's already melted in later issue dates. In general, the relationship and direction of feature contribution align well with physical intuition (Figure 7, Figure 8).

Unlike individual features, we present the contribution of feature groups in absolute form, which sums up to 100%. At the feature group level, we are more interested in the amount of contribution rather than the overall contribution direction. In addition, direction detail is already represented and explained in individual features. It's possible that snowpack from different data sources disagrees in the direction of contribution, especially in later issue dates during the melting season, and the total contribution will still represented in absolute form.

In addition to explainability metrics, we include feature values presented in percent of normal for additional context. However, for certain features, using percent of normal may not be appropriate due to negative values and the scale. These features are:

- PDSI → raw value, classified as one of 10 levels of wet/dry conditions[4,5]
- Snow cover area → raw value and its average
- Temperature → deviation with its average (in degrees Celcius)
- Leaf area index → raw value (in m^2 m^{-2})

We use the range between the lowest and highest value for features that have multiple sources, such as the SEAS51 forecast (mean, 10th, and 90th percentile from 26 ensemble members), UA-SWANN (grid average, HUC-6 and HUC-8 based), leaf area index (low, high vegetation), and soil water volume (layer 1-4).

We present the forecast summary this way to provide a brief, compact, and easier-to-read summary. For a longer report, we can add more details of the feature explanations. For example, we can include details of SNOTEL/CDEC stations used to measure the snowpack, basin geovisualization of snowpack from gridded data (ERA5-Land and UA-SWANN), and other details of feature input.

Generalization and scalability

The SHAP method provides flexibility, as the values can be aggregated in different ways to provide an intuitive understanding of global feature contribution and feature contribution differences across model quantiles, years, sites, issue dates, etc.

For example, we can compare global feature contribution across model quantiles (Figure 1, Figure 2) where interestingly, snow cover area is the predictor that contributes higher to the upper uncertainty bound (90th percentile). The intuition is that snowpack measurement, e.g., coming from SNOTEL/CDEC might be the same, however, snow cover area might vary, which can lead to different possibilities of translated water supply in the upper uncertainty bound. This also applies to predictors from SEAS51 forecast, which mostly influence uncertainty bounds.

We can also use SHAP to understand the evolution of feature contribution based on different issue dates (Figure 3, Figure 4). Snowpack contribution is lower at the beginning of the issue date because snowpack which translates into water supply is not fully formed. It gradually gets bigger until the end of the accumulation period, and then gets smaller again after the accumulation period. From both figures, we also see that snowpack feature contribution in wet years (2023) is higher than in all years (2004-2023).

The examples above show that SHAP method can be applied to a wide range of use cases beyond understanding the local explanation of the forecast prediction. Combined with the global model approach, the local explanation can be easily extended to different locations, issue dates, and years, providing a consistent and intuitive explanation of the relationship between predictors and water supply output.

Appendix

Figure 1: Global feature contribution (2023) excluding base features

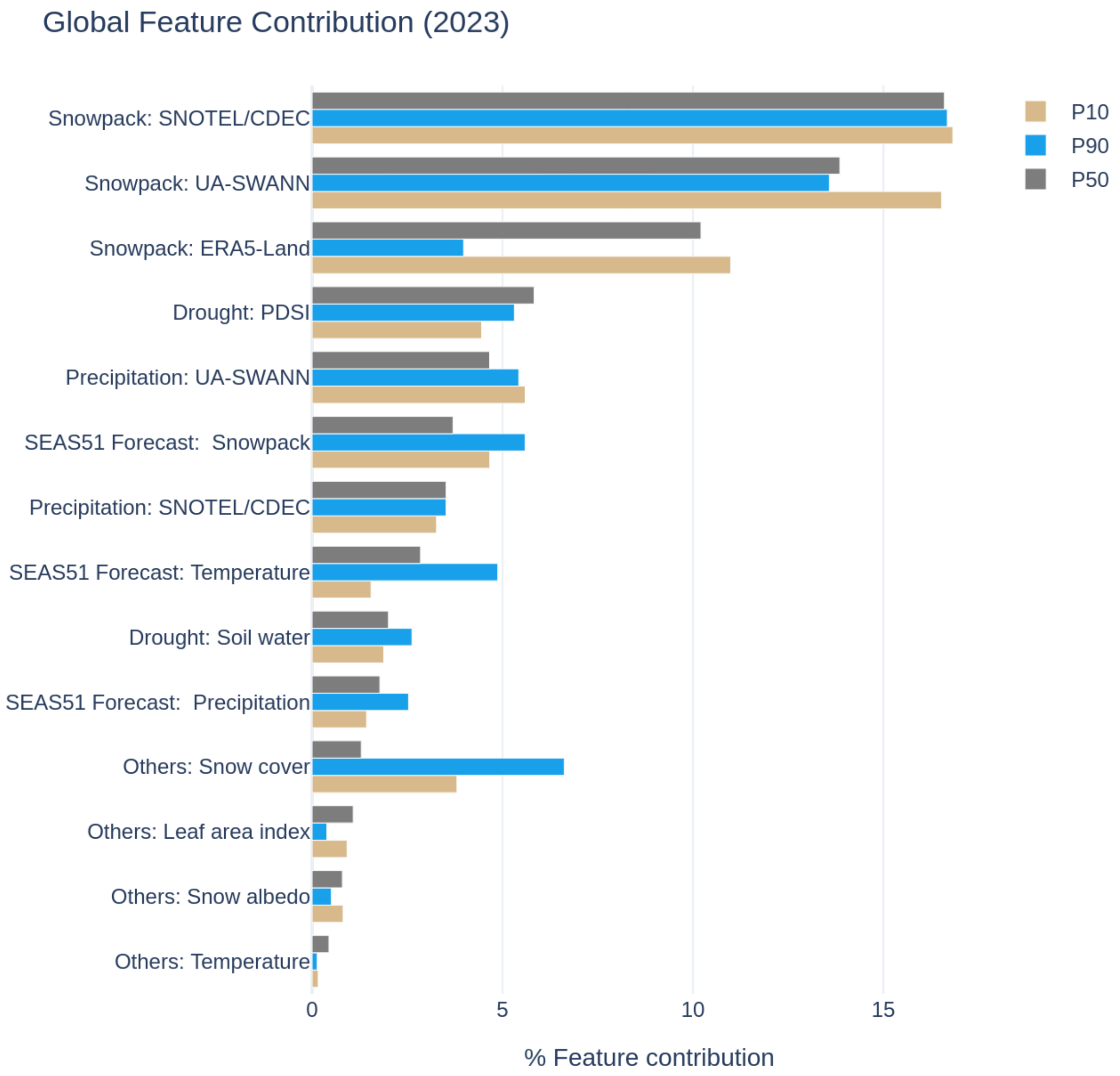


Figure 2: Global feature contribution (2004-2023) excluding base features

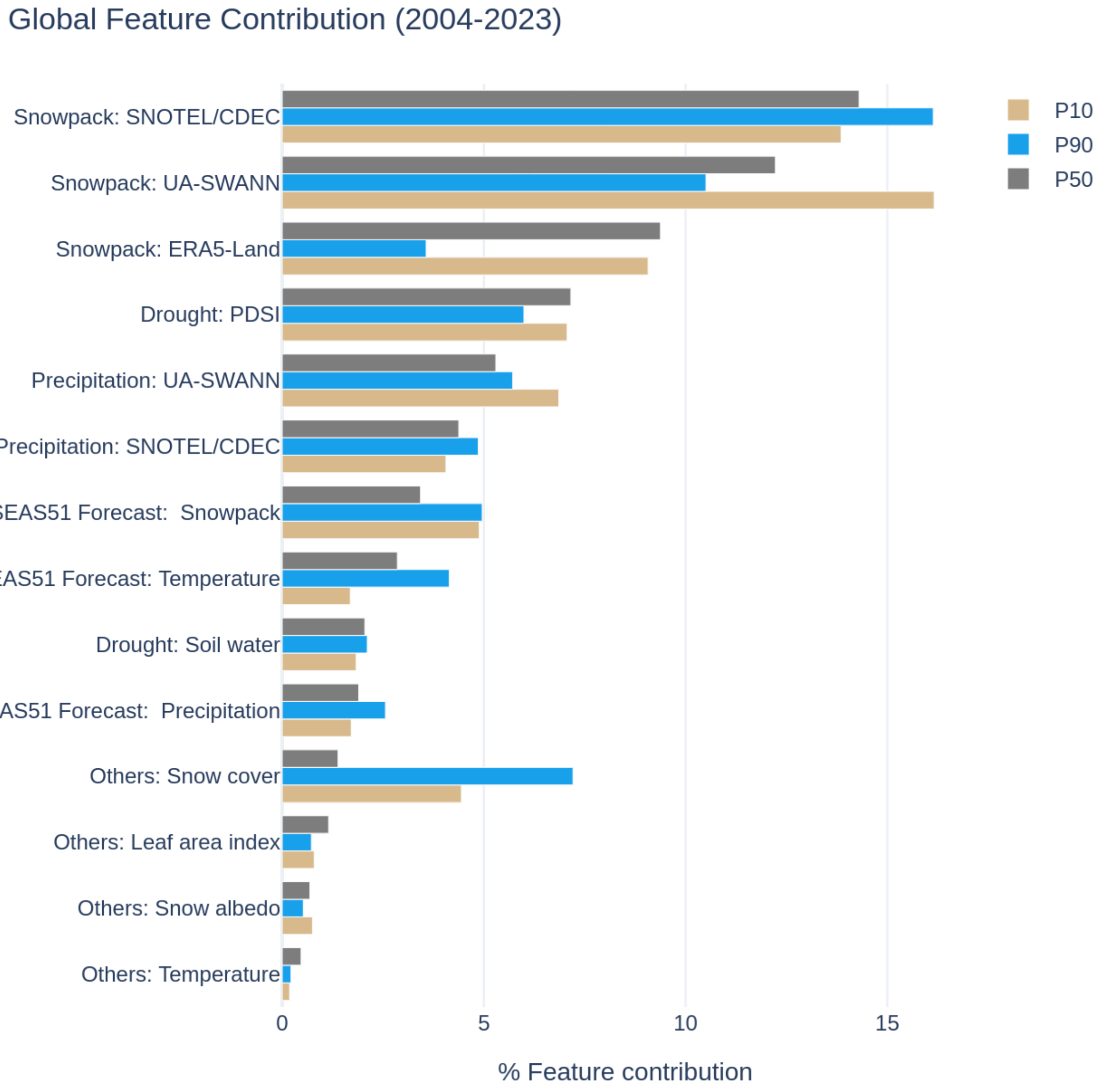


Figure 3: Feature contribution across different months (2023)

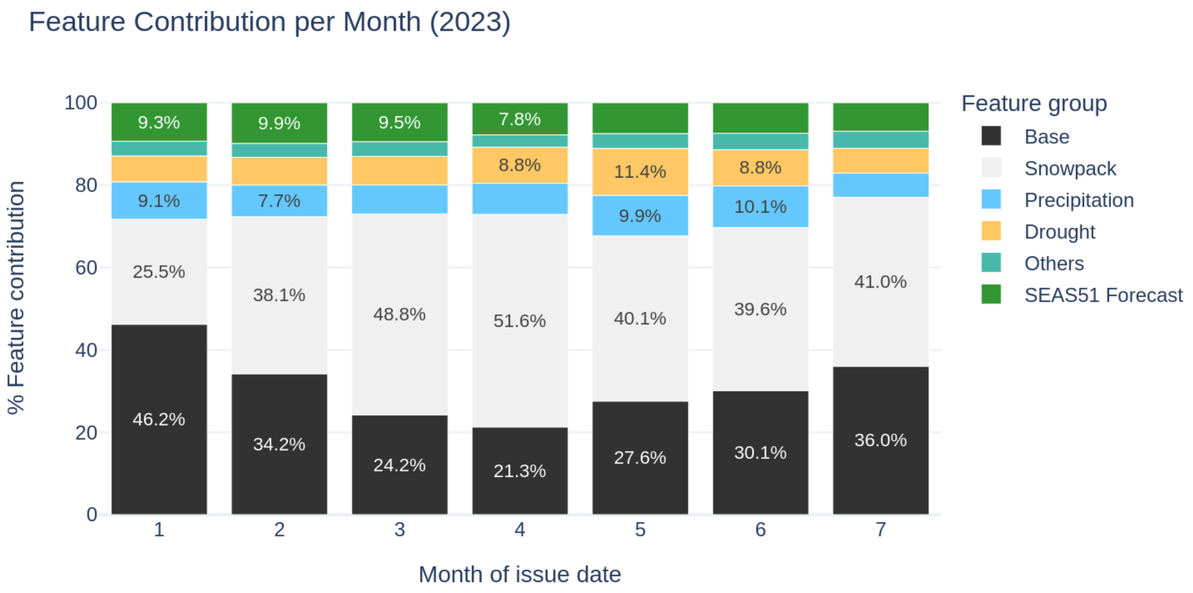


Figure 4: Feature contribution across different months (2004-2023)

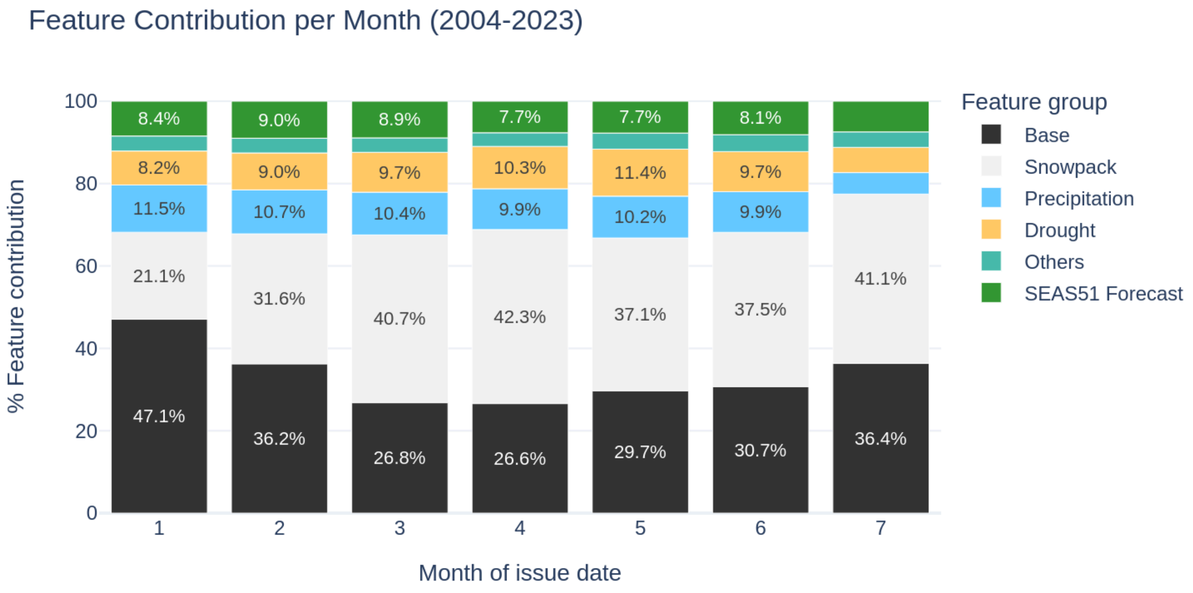


Figure 5: Scatter plot of base feature contribution and average % snow cover area

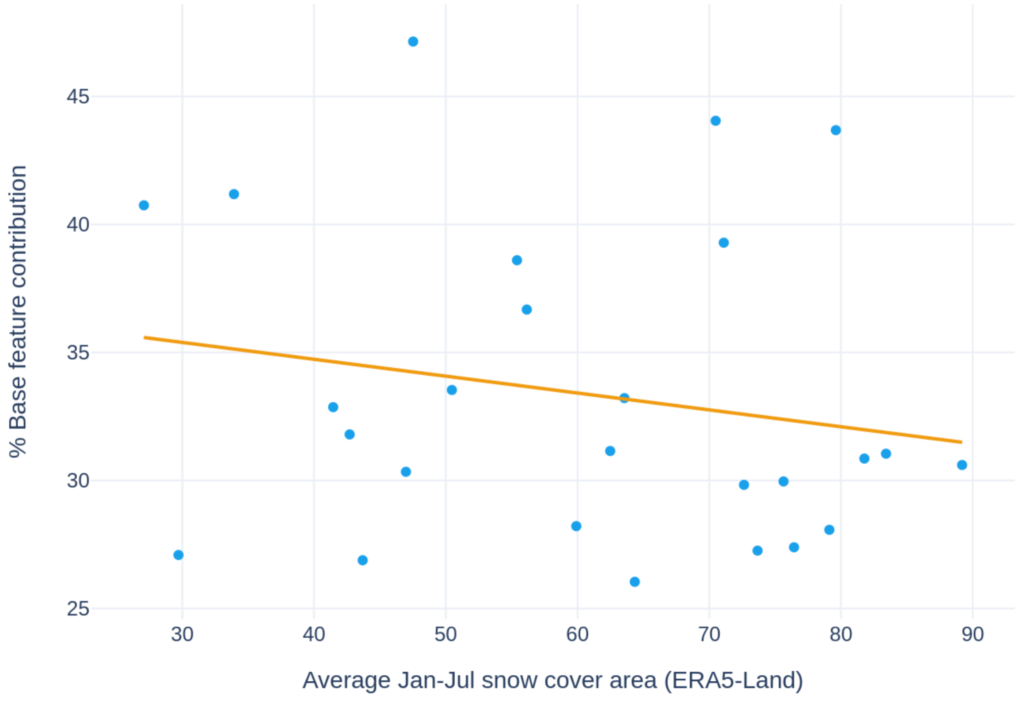


Figure 6: Scatter plot of snowpack feature contribution and average % snow cover area

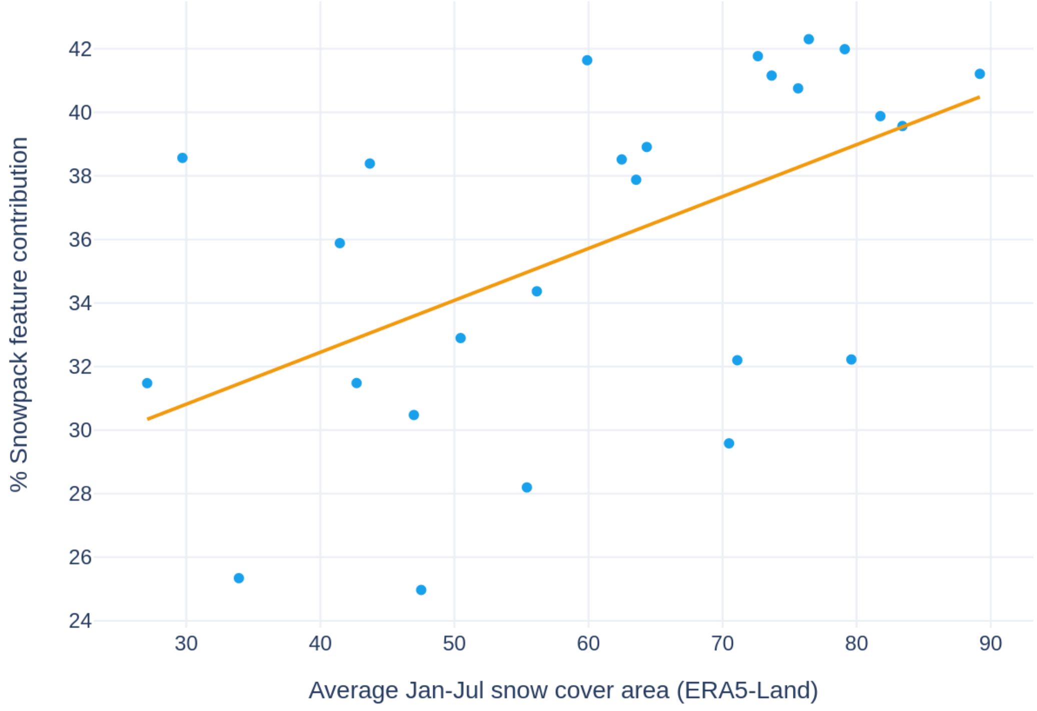


Figure 7: SHAP dependence plot of quantile 0.5 model (2023) - swe_lag1 (SNOTEL/CDEC SWE), sd (ERA5-Land SWE), prec_cml_lag1 (SNOTEL/CDEC cumulative precipitation), pdsi_lag5 (gridMET PDSI), t2m_mean (SEAS51 one month temperature forecast), t2m (ERA5-Land temperature)

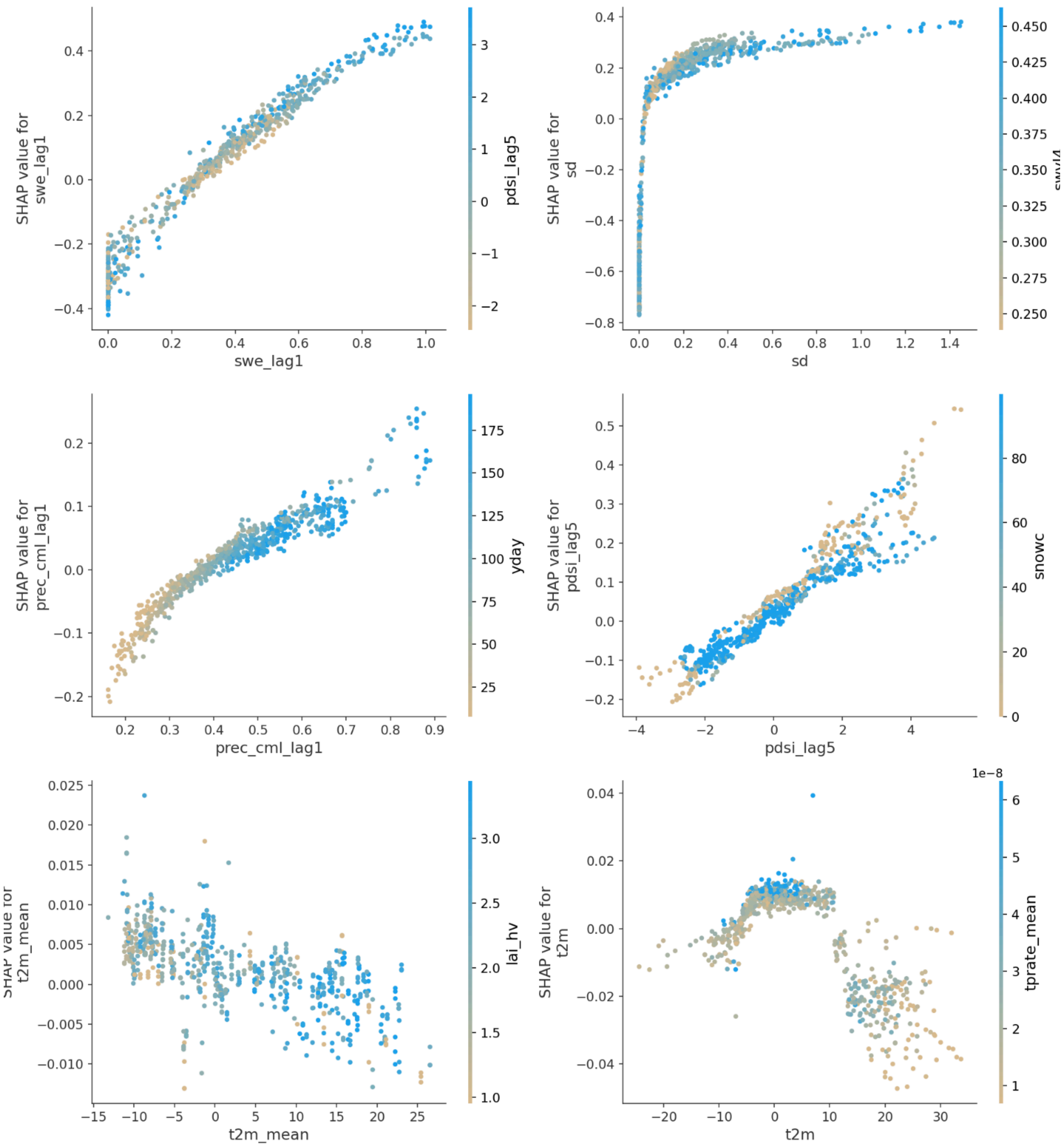
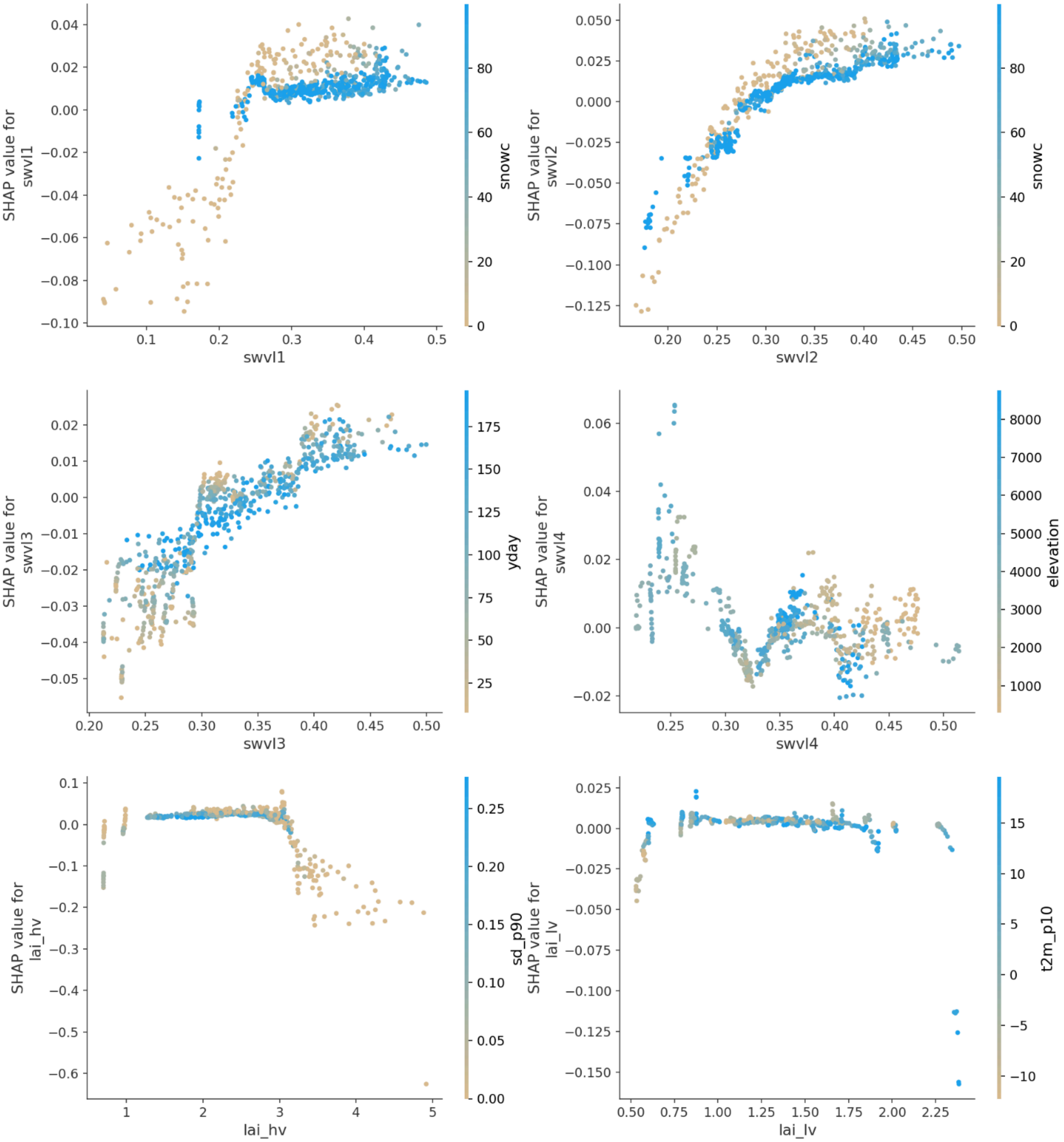


Figure 8: SHAP dependence plot of quantile 0.5 model (2023) - swvl1~swvl4 (ERA5-Land volumetric soil water layer 1-4; 0-7 cm, 7-28 cm, 28-100 cm and 100-289 cm), lai_hv (ERA5-Land leaf area index high vegetation), lai_lv (ERA5-Land leaf area index low vegetation)



References

1. https://www.weather.gov/tbw/newnormals
2. https://www.nrcs.usda.gov/wps/portal/wcc/home/snowClimateMonitoring/30YearNormals/30YearNormalsFaqs
3. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. arXiv. https://doi.org/10.48550/arxiv.1705.07874
4. https://en.wikipedia.org/wiki/Palmer_drought_index#cite_note-10
5. https://www.drought.gov/data-maps-tools/us-gridded-palmer-drought-severity-index-pdsi-gridmet