Enhancing Water Supply Forecasts with SHAP and What-If Analyses

Christoph Molnar (kurisu)

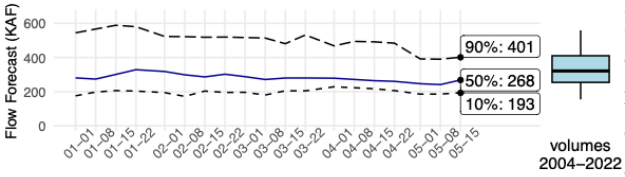
This report describes communication and explainability outputs for the water supply forecasting competition. It introduces Forecast Plots for visualizing water supply forecasts at the 10%, 50%, and 90% quantiles, along with Context Plots showing watershed conditions (SWE, PPT, Antecedent Flow) throughout the year compared to historical conditions (2004-2022). Explanations of forecasts and uncertainties are provided through What-If Plots and SHAP Waterfall Plots, offering insights into the impact of the features on forecasts and uncertainties.

1 Technical Approach

1.1 Forecast and uncertainty quantification

1.1.1 The Forecast Plot

The first figure in the forecast reports, the "Forecast Plot" (see Figure 1), shows the currently predicted seasonal flow at the 50% quantile, along with uncertainty bounds at the 10% and 90% quantiles. It also illustrates how the forecast has developed over the year.



Forecast Plot: The lines represent the forecasted quantiles and how they changed up until the current issue date. The boxplot to the right of the current predictions shows the historical volumes, allowing the current forecasts to be set in a historical context.

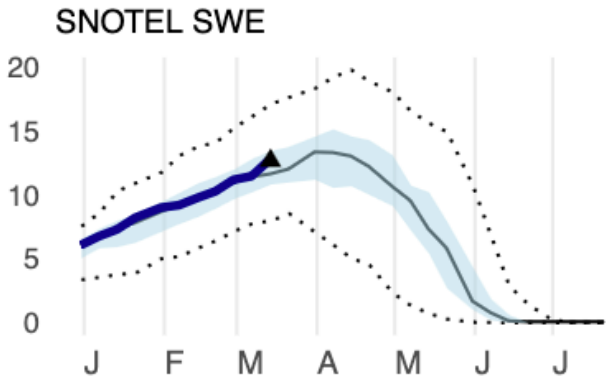
Figure 1

Across plots in the forecast report, the colors are used consistently: light blue for historical conditions and dark blue for the current year.

1.1.2 The Context Plot

The Context Plots (see Figure 2) provide supporting information to make sense of the forecasts. Each of the four Context Plot displays how conditions in the watershed have developed over the calendar year. The following Context Plots are included in the forecast reports:

- SWE estimates
  - Based on SNOTEL, averaged over stations within 40 miles of the site
  - Based on SWANN, spatially averaged using Hydrologic Unit Codes (HUC)
- Accumulated precipitation (SWANN)
- Antecedent flow (NRCS/RFCs): The site's flow from previous month



The **Context Plot** shows watershed conditions over the year and compares them to historical conditions (2004-2022).

- Red line: The SWE/flow/PPT for the current year
- Black dotted lines: Historical min and max
- Black straight line: Historical median
- Blue ribbon: Historical 25%-75% range
- Triangle: feature date

Figure 2: Context Plot

In contrast to the Forecast Plot, each Context Plot shows the entire issue date range from January to July. The Context Plot informs about the relevant variables in the watershed, but also more directly about the features the model uses: `antecedent_flow`: The last months flow; `snotel_swe_conditional`: SWE estimate up until feature date; `swann_swe_conditional`: SWE estimate up until feature date; `swann_ppt_conditional`: Accumulated precipitation up until feature date; `swann_ppt_unaccounted`: Accumulated precipitation since feature date. The feature date[1] is marked with a dot in the plots.

The Context Plots bridge the forecast and the forecast explanation: The plots provide supporting information to understand why a certain predictions was made. This includes setting the predictions in context with historical conditions.

1.2 Explainability Metrics and Communication

The forecast report contains the following explanations:

[1]For before-season issue dates, the feature date is the day before the issue date. For in-season dates, the feature date is the day before the antecedent flow is updated, for example, for 05-15, the feature date is 04-30. Further explanations for this can be found in the final report under section **Fixing the "End-of-Month Melt Bias" with a conditional date**

- Forecast Explanation:
  - Explaining why the current forecast deviates from historical forecasts
  - Explaining why the current forecast deviates from last week's
- Uncertainty Explanation: Explaining why the current uncertainty deviates from historical uncertainty.

All explanations are visualized with two methods/figures: The **What-If Plots** show how changes in the features affect forecast and uncertainty; The **SHAP Waterfall Plots** attribute the difference between current forecast (or uncertainty) and a reference forecast (or uncertainty) to the features. The reference here is past years (2004-2022) or last week.

1.2.1 The What-If Plots

The What-If Plots (Figure 3) display how changes in individual features change the 50%-forecast (and the 90%-10% range, respectively). It's a simple yet effective way of investigating how the model would react to slightly different circumstances and to contrast the current prediction with "what-if"-scenarios.

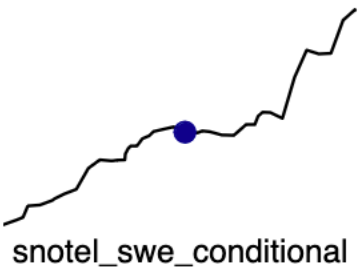


Figure 3

The **What-If Plot** shows how manipulating a feature, here `antecedent_flow`, changes the forecast while keeping the other features fixed.

- The x-axis shows the feature range
- The y-axis the forecasted values
- The dark blue dot is the current prediction
- The black line is the prediction function when changing the current feature

How to compute the What-If plot for a feature j for current data point $x_{current}$:

1. Create grid for feature j from min to max (e.g. 30 values)
2. For each grid value i:
   - Replace the j-th feature in $x_{current}$ with the grid value i to create data point $x_{what-if}$
   - Get the 50% forecast for this new data point: $f(x_{what-if})$.
3. Plot a line from the grid and the forecasts

This procedure can also be used to compute the What-If Plot for explaining the uncertainty range instead of the forecast: At step 3, instead of computing the 50% forecast, get the 90% and 10% forecasts and compute the difference: $f_{90}(x_{what-if}) - f_{10}(x_{what-if})$. And at step 4, plot this quantity instead on the y-axis. For each feature, two What-If Plots are included, one to explain the 50%-forecast, and one to explain the uncertainty range.[2]

[2]Exception: For before-season issue dates, `swann_ppt_unaccounted` is always zero and therefore not included in before-season issue date reports.

The What-If Plots give a general impression of how a feature affects the prediction. However, it's also a limited analysis, as it only changes one feature at a time and doesn't give a sense of importance of each feature for the prediction. Design-wise the plots are kept minimalistic, because they are only intended to show tendencies.

1.2.2 The SHAP Waterfall Plots

To get a sense of importance and influence of each feature on the forecast, I used Shapley Additive Explanations (Lundberg and Lee 2017), or SHAP for short. This technique originates from Game Theory (Shapley et al. 1953) where it is used for team games to fairly assign payouts to the team members. In the machine learning context, SHAP is used to fairly attribute the forecast to the individual features, compared to a reference forecast. With a modification, SHAP can also be used to explain the uncertainty range. I visualized the SHAP values with a waterfall plot (Figure 4).

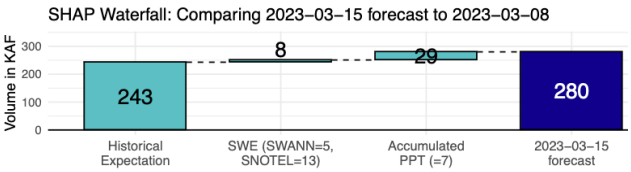


Figure 4

The **Waterfall Plot** shows how each feature value contributed to the forecast compared to the reference forecast (historical or from last week).

- The left bar shows the reference forecast.
- The right bar the current forecast.
- The steps between are SHAP values.
- For in-season dates, a dotted rectangle shows the already observed seasonal flow.

When used for explaining uncertainty, the y-axis shows the uncertainty range (Q90 - Q10).

The Waterfall puts the reference forecast (or uncertainty range) in direct comparison with the current forecast (or uncertainty range) and the SHAP values explain the differences between the two.

SHAP values work by computing contributions of each feature towards the difference between current forecast and a reference forecast. Given the 5-page restriction, I can't explain SHAP theory and estimation in detail here. In short, SHAP computes values of coalitions of features by creating hybrid data points from current data and reference data. The sum of SHAP values for the features yields the difference between current and reference forecasts. For more information about SHAP, see Molnar (2022). By cleverly setting reference data and defining the function to be explained, we can leverage SHAP for different interpretation targets.

Setting the historical data (years 2004-2022, same issue day, same site) as reference, SHAP values explain why the current forecast deviates from the historical forecasts. Interpretation example:

The 50%-forecast for Pueblo Reservoir, issue date March 15th, deviates from historical forecasts by -8 KAF, to which Antecedent Flow=19 contributed -2.

Setting the last week's data point as reference, SHAP values explain why the current forecast deviates from last week. Interpretation example:

The 50%-forecast for Pueblo Reservoir, issue date March 15th, deviates from the last week's forecast by +37 KAF, to which Accumulated PPT=7 contributed +29.

SHAP is model-agnostic, meaning it works for any model. By defining the forecast range $f_{range}(x) = f_{90}(x) - f_{10}(x)$ as our model (instead of $f_{50}(x)$), we can leverage SHAP to explain why the current uncertainty range deviates from the historical uncertainty range.

The 90%-10% forecast range for Pueblo Reservoir, issue date March 15th, deviates from the historical range by -27, to which SWE (SWANN=5, SNOTEL=13) contributed -14.

A few more notes on interpretation

- **Previous flow**: For dates after May 1st, the total forecasts consist of measured previous in-season flow plus the forecast of the remaining flow. Since this previous flow can also differ between the current year and the historical data, one of the bars in the waterfall plot represents the difference between current and expected (April) flow.
- **Antecedent flow as feature**: Antecedent flow is also a feature in the model (and waterfall plot). For e.g. 2023-05-15, the April flow and `antecedent_flow` coincide. But the SHAP value for `antecedent_flow` is about the features' influence on forecasted remaining flow.
- **Non-causal attribution**: SHAP values don't represent physical/causal attributions, since features like `antecedent_flow` aren't causal.
- **Feature Correlation**: When feature are strongly correlated, SHAP may create implausible data, which poses a challenge to interpretability (Aas, Jullum, and Løland 2021). Since SNOTEL and SWANN are very strongly correlated and measure the same concept, I computed a shared SHAP value for them, by treating them as one feature. However, other features, such as accumulated PPT and SWE, are also moderately to strongly correlated and would benefit from deeper investigation in collaboration with hydrologists.
- **Not counterfactual**: SHAP may not be interpreted as "local counterfactual analysis", meaning we can't infer what would happen if we increased a feature with a positive SHAP value, as demonstrated by Bilodeau et al. (2024). That's why I included What-If Plots.
- My current modeling solution uses issue date (in form of day in year) and site as features to differentiate time and place. For the communication outputs, these two are simply fixed.

Generalizability of the explainability solution

My final prize model is an ensemble of 10 xgboost models – not so interpretable. That left as option only so-called "post-hoc interpretation methods" like SHAP and What-If analysis, which "probe" the model to describe their behavior. Post-hoc methods are typically model-agnostic, meaning they can be used with any model and architecture. This also means they are also easily generalized to other sites and issue dates. It's a simple matter of subsetting the reference data to a different site and issue date. The flexibility of the reference data also means that one can, for the shap plots, be more creative when creating explanations: We could also create explanations that compare the current prediction to the last 40 years. Or to compare the 90% forecast with the 90% forecast from a specific year.

References

Aas, Kjersti, Martin Jullum, and Anders Løland. 2021. "Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values." *Artificial Intelligence* 298 (September): 103502. https://doi.org/10.1016/j.artint.2021.103502.

Bilodeau, Blair, Natasha Jaques, Pang Wei Koh, and Been Kim. 2024. "Impossibility Theorems for Feature Attribution." *Proceedings of the National Academy of Sciences* 121 (2): e2304406120. https://doi.org/10.1073/pnas.2304406120.

Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–77. NIPS'17. Red Hook, NY, USA: Curran Associates Inc.

Molnar, Christoph. 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed. https://christophm.github.io/interpretable-ml-book.

Shapley, Lloyd S et al. 1953. "A Value for n-Person Games."