Water Resources Research

Supporting Information for

Forecast Informed Reservoir Operations Using Ensemble Streamflow Predictions for a Multi-Purpose Reservoir in Northern California

Chris Delaney[1], Robert Hartman[2], John Mendoza[1], Michael Dettinger[3], Luca Delle Monache[3], Jay Jasperse[1], F. Martin Ralph[3], Cary Talbot[4], James Brown[5], David Reynolds[6], Simone Evett[7]

[1]Sonoma Water, Santa Rosa, California, USA, [2]Robert K. Hartman Consulting Services, Roseville, California, [3]Scripps Institution of Oceanography, Center for Western Weather and Water Extremes, University of California San Diego, San Diego, California, [4]U.S. Army Engineer Research and Development Center, Vicksburg, Mississippi, [5]Hydrologic Solutions Limited, Winchester, United Kingdom, [6]University of Colorado at Monterey, Monterey, California, [7]Independent Researcher, Pasadena, California

Contents of this file

1 Lake Mendocino Operations Model Inputs and Constraints

The Lake Mendocino Operations (LMO) model defines data inputs for all primary sources and sinks of the system water balance to simulate current conditions of Lake Mendocino and the Upper Russian River. These data inputs include ensemble streamflow hindcasts, observed and forecasted unimpaired reach gains from natural runoff, water imports from the Eel River through the Potter Valley Project, and reach water loss due to consumption and other sinks. The LMO model also incorporates primary system constraints that govern actual operations of the reservoir.

1.1 HEFS Hindcast

One of the primary inputs to the LMO model for the simulation of Ensemble Forecast Operations (EFO) is the ensemble streamflow hindcasts from the National Weather Service (NWS) Hydrologic Ensemble Forecast System (HEFS). HEFS leverages the models and states used to generate single-value forecasts and forces them with an ensemble of equally likely precipitation and temperature scenarios modulated by the current weather forecast, a statistically calibrated relationship between the precipitation and temperature forecast (model or forecaster driven) and historical observed precipitation and temperature for the basin of interest. The fundamental science behind this approach can be found in Demargne et.al, 2012. An overview of the HEFS process and workflow is provided in Figure 1.

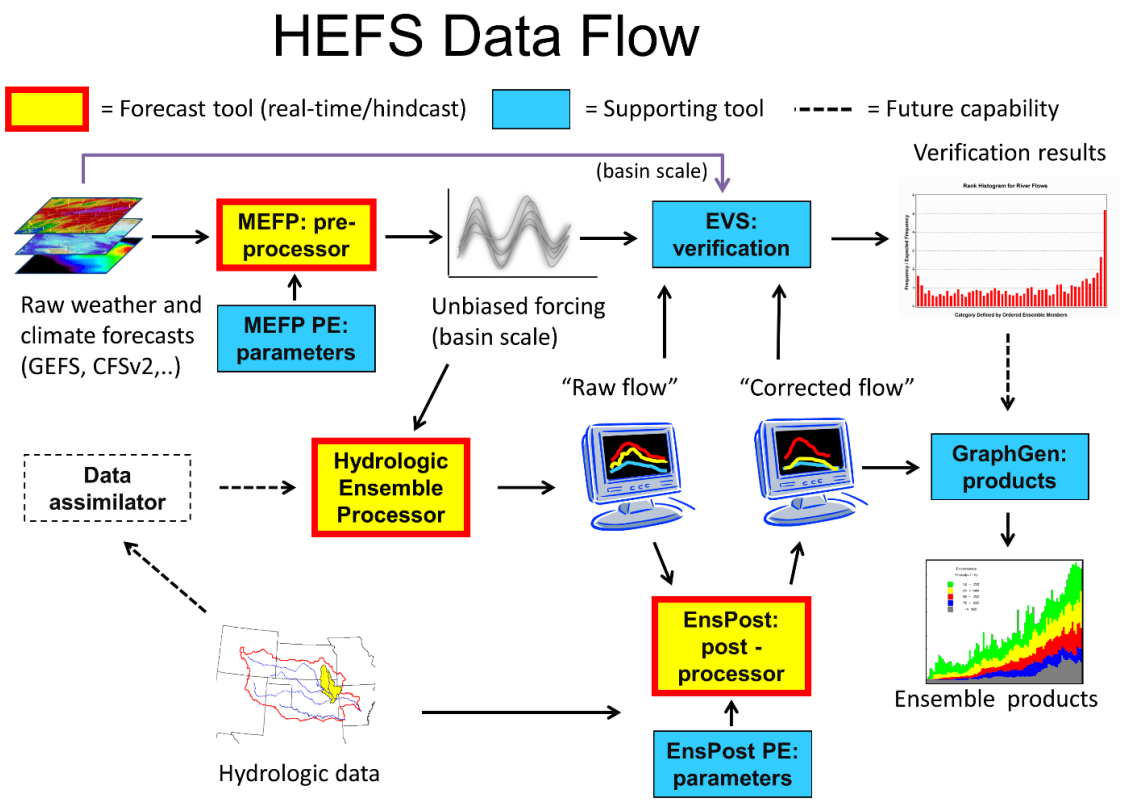

Figure 1. Overview of HEFS workflow.

The NWS HEFS has been operational for the Russian River basin since 2012, which only provides a limited period to evaluate forecast based alternatives. To provide a more comprehensive period for analysis, retrospective ensemble forecasts (i.e., hindcasts) of Lake Mendocino inflow and the downstream watersheds were generated by the California Nevada River Forecast Center (CNRFC) using the HEFS model over a 26-year period. To develop these hindcasts, the CNRFC hydrology models were forced with the Meteorological Ensemble Forecast Processor (precipitation and temperature) from the National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS Version 10) 2012 reforecast data set (Hamill et al, 2013). The full hindcast process creates an hourly, 61-member flow forecast from 1985 to 2010, with a forecast horizon of 15 days.

The hydrology and atmospheric models used in the hindcast process are as consistent as possible with what is used operationally (Hamill et al., 2013); however, the hindcasting procedure is automated, so it does not fully capture the forecast process used by the

CNRFC in real-time operations. For example, in real-time operations, a hydrologist may make minor adjustments to model states and/or suspect observations to ensure proper simulation of observed discharge up through the current time when the forecast is issued. These adjustments are often iterative until the forecaster is satisfied that the modeled location is functioning properly and that the single-value forecasts, given the forecast meteorology, are credible and appropriate. Interestingly, the forecaster only sees the single-value forecast several days into the future and not the full ensemble distribution generated by the follow-on ensemble process. Tuning of model states and observations given the ensemble forecast distribution is not currently practiced at the CNRFC. In this respect, the hindcast dataset is consistent with real-time operations. The hindcasts are not, however, an exact representation of operational methods, but are a consistent, realistic, and likely skill-conservative sample of forecasts for testing alternative reoperation strategies.

1.2 Unimpaired Flows

The CNRFC developed historical unimpaired flow hydrology of the Russian River from 1985 to 2010. Unimpaired flows are estimates of reach gains from actual precipitation runoff, which are unaffected by man-made influences such as water demands or reservoir operations. These data were originally developed to assess forecast quality and reliability; however, they have been applied in this study to simulate operational alternatives of Lake Mendocino from 1985 to 2010. For example, to simulate EFO, at each daily time step the LMO model uses the hindcast (described above) to forecast future conditions and calculate a flood release, but once the release has been calculated, simulated storage and flows are calculated using the historical unimpaired flows.

Unimpaired inflow into Lake Mendocino were estimated by the CNRFC using water balance methods with observed inflows from the US Army Corps of Engineers Sacramento District (USACE-SPK) and observed diversions into the East Fork Russian River from the Potter Valley Project to estimate inflow from natural runoff. Historical local unimpaired flows for the remaining model junctions (West Fork, Hopland, Cloverdale and Healdsburg) were simulated by running the CNRFC hydrology model for the Russian

River continuously with historical observed weather conditions (temperature and precipitation).

1.3 Potter Valley Project Diversions

In the fall of 2006, operations of the Potter Valley Project changed substantially as a result of a 2004 Federal Energy Regulatory Commission license amendment for the Potter Valley Project (FERC, 2004). Consequently, historical diversions from 1985 to 2006 would not be representative of current operations of the Potter Valley Project. Modeled Potter Valley Project diversions were developed to simulate current, post-2006 regulation. The estimated Potter Valley Project diversions from the Eel River Model are provided as inputs into the LMO model at the Potter Valley Project model junction.

1.4 Reach Losses

The LMO model accounts for system losses at four model junctions: Lake Mendocino, Hopland, Cloverdale, and Healdsburg. Model junction losses applied in the model consist of water balance derived losses and metered municipal and industrial (M&I) diversions. Water balance losses, which account for all water sinks including riparian evapotranspiration, evaporation, surface water/groundwater interactions, and consumptive use from agricultural diversions and M&I diversions, were calculated for the dry season months from May to October. During the remaining months, November through April, it is assumed that water loss from most sinks declines significantly. Therefore, for these months, it is assumed that M&I loss estimates are accurate estimates of the total reach depletions.

The unimpaired flow for most reaches typically diminishes to low or no flow conditions by early to mid-May in dry years and June in wetter years. During the wet season unimpaired flows are large, and the magnitude of these flows is significantly greater than the magnitude of losses. Because of these differences, gage error and unimpaired flow estimation error can be significant compared with the magnitude of losses and obscure a reliable calculation of water balance derived losses. Additionally, in the springtime of wetter years, agricultural water use is very low so that corresponding stream losses are also low, making water balance loss calculations unreliable. For these reasons, water

balance losses are only calculated for the months of May through October for incorporation into the total loss time series.

Water balance losses were calculated for water years 2000 through 2010 using daily observed diversions from the Potter Valley Project, releases from Lake Mendocino, flow data from USGS gages, and modeled unimpaired flow from the CNRFC. Daily water balance losses were calculated at each model junction using the following equation:

$$Loss_{WaterBalance} = Q_{Upstream} - Q_{Downstream} + Q_{Unimpaired} \quad (1)$$

Where $Q_{Upstream}$ and $Q_{Downstream}$ are the observed flows at the USGS gages corresponding to the model junctions, and $Q_{Unimpaired}$ is the CNRFC modeled unimpaired flow (as previously described) in the reach between the upstream and downstream model junctions.

Losses due to M&I diversions were estimated from metered diversion data collected from 11 public water systems in the Upper Russian River basin for the years 2009 through 2013. Total monthly M&I diversions were calculated for each reach using pumping data for the points of diversions located in that reach. Monthly average metered diversions were estimated for the months of April through November.

Annual loss hydrographs were calculated for the Lake Mendocino, Hopland, Cloverdale and Healdsburg model junctions. The average monthly losses are shown for the Hopland model junction in Figure 2. The annual monthly losses were applied in the model by equally distributing the monthly average loss to each day of the month to develop daily loss estimates.

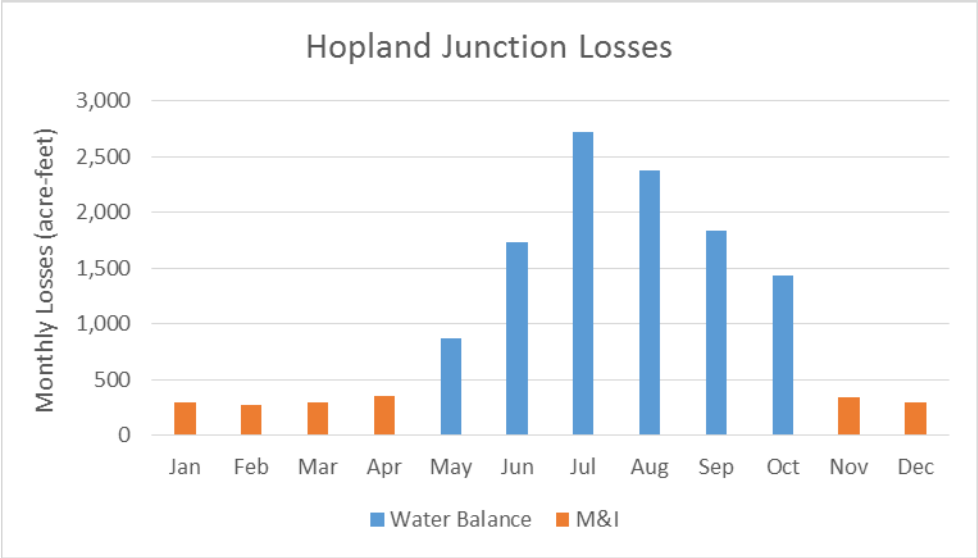

Figure 2. Monthly loss pattern for Hopland model junction.

Losses due to reservoir surface evaporation were accounted for in the model using an annually repeating pattern of monthly evaporation rates. The monthly evaporation rates were calculated based on monthly mean pan evaporation estimates and the monthly evaporation coefficients provided in the Lake Mendocino Water Control Manual (WCM) (USACE, 2003). Daily reservoir surface water evaporation is simulated in the LMO model by taking the product of simulated water surface area of Lake Mendocino by the monthly evaporation rate.

1.5 Lake Mendocino Release Constraints

To simulate releases from Lake Mendocino the LMO model incorporates the primary system constraints that govern actual operations of the reservoir. These constraints include physical capacity of the outlet structures, and release rules defined for flood control and water supply operations.

The controlled outlet of Lake Mendocino consists of a single conduit approximately 720 feet long and 11 feet in diameter. There are three pairs of hydraulically operated release gates, with each pair consisting of a service gate and an emergency gate. Maximum release capacity of the controlled outlet is approximately 7,500 cubic feet per second (cfs) when the water surface elevation is within the Emergency Release Pool

(above elevation 773 feet mean sea level). The controlled outlet rating curve defined in the WCM (USACE, 2003) is shown in Figure 3.

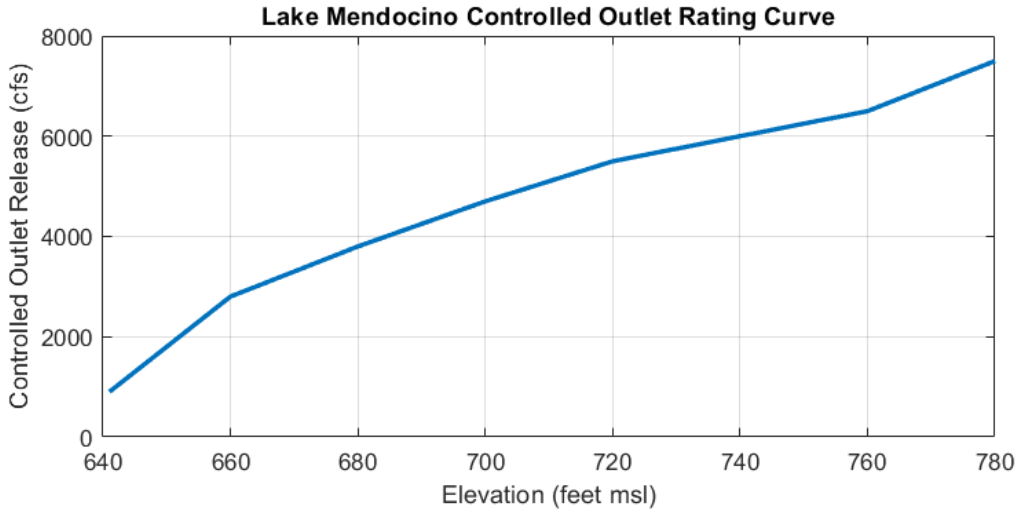

Figure 3. Lake Mendocino controlled outlet rating curve.

The flood release schedules defined in the WCM (USACE, 2003) set the maximum release rates from the reservoir's controlled outlet structure for different reservoir storage levels. The flood release schedules, as shown in Table 1, were defined in the LMO model to constrain releases during simulated flood operations.

Table 1. Lake Mendocino maximum flood release schedule.

Flood Release Schedule	**Elevation (feet)**	**Storage (ac-ft)**	**Max Release (cfs)**
Schedules 1 and 2	737.5 to 755	68,400 to 98,700	4,000
Schedule 3	755 to 771	98700 to 128,100	6,400

The emergency spillway crest for Lake Mendocino is at storage level 116,500 acre-feet (ac-ft). The emergency spillway does not have a release control structure, so emergency spillway releases are uncontrolled. When simulated storage levels are above the crest of the emergency spillway, the LMO model simulates spillway releases according to the spillway rating curve defined in the WCM (see Figure 4). The maximum rated capacity of the spillway (top of dam elevation water surface elevation) is 47,300 cfs.

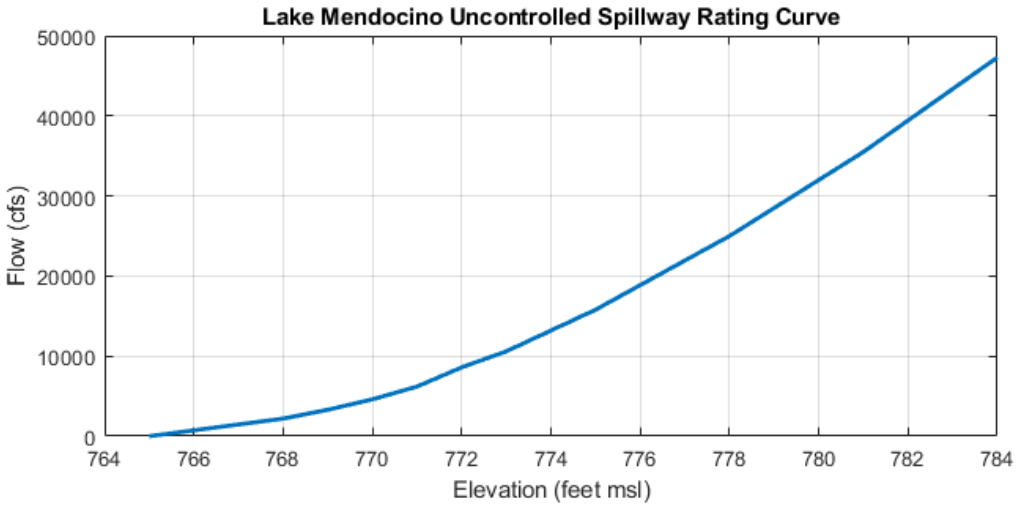

Figure 4. Lake Mendocino uncontrolled spillway rating curve.

When simulated storage in Lake Mendocino is within the conservation pool for the Existing Operations (EO) alternative or forecasted risk is below the risk tolerance curve for EFO, the LMO model simulates releases according to the constraints defined for water supply operations. To simulate water supply operations conducted by Sonoma Water, the LMO model simulates releases consistent with current operations. These constraints include requirements defined in Sonoma Water's water rights permits (Sonoma Water, July 2016) and a 2008 Biological Opinion (BiOp) issued by the National Oceanic and Atmospheric Administration's (NOAA's) National Marine Fisheries Service (NMFS, 2008). The model simulates releases during water supply operations by calculating the required release to supply all downstream reach losses, including a buffer release, to ensure minimum instream flows are met at all downstream model junctions (Hopland, Cloverdale and Healdsburg).

The primary model assumptions for simulating water supply release decisions for Lake Mendocino include the hydrologic index and the Upper Russian River minimum instream flow requirements. The hydrologic index is a metric that is defined in Sonoma Water's water rights permits (Sonoma Water, 2016), which sets the water supply condition and minimum instream flow schedules for the Russian River. The water supply condition (*Normal*, *Dry* or *Critical*) is determined through an evaluation of water year cumulative inflow into Lake Pillsbury (located on the Eel River). The water supply condition is used to set the minimum instream flow schedules as shown in Table 2. The Upper Russian

River minimum instream flow requirements vary from 185 cfs for the wettest (*Normal*) Water Supply Conditions to 25 cfs for the driest (*Critical*) water supply conditions.

Table 2. Upper Russian River minimum instream flow schedules.

Water Supply	Month												
Condition	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct 1-15	Oct 16-31	Nov	Dec
Normal	150	150	150	185	125	125	125	125	125	125	150	150	150
Normal-Dry Spring 1										75	75	75	75
Normal-Dry Spring 2						75	75	75	75	75	75	75	75
Dry	75	75	75	75	75	75	75	75	75	75	75	75	75
Critical	25	25	25	25	25	25	25	25	25	25	25	25	25

The water supply conditions used in the LMO model were calculated from historical Lake Pillsbury inflow from 1985 to 2010, which are shown in Figure 5. Under *Normal* water supply conditions from June through December, Lake Pillsbury and Lake Mendocino are evaluated for low storage levels to potentially trigger reduced flow schedules (*Normal-Dry Spring 1* or *Normal-Dry Spring 2*) as shown in Table 2.

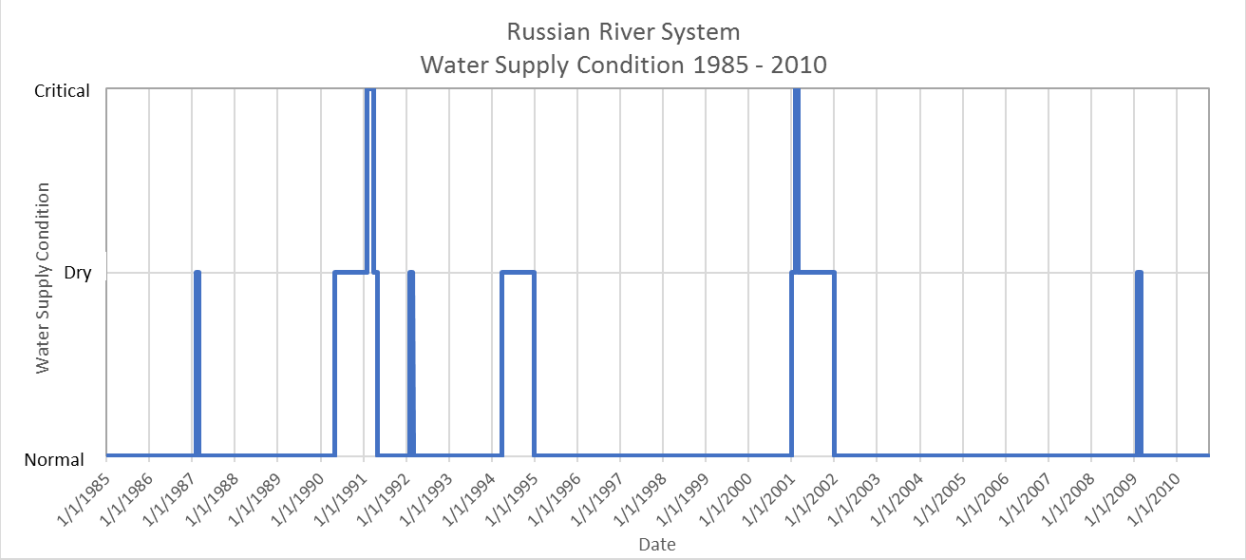

Figure 5. Russian River System Water Supply Conditions from 1985 to 2010 calculated from observed Lake Pillsbury inflows.

Operationally, additional buffer releases are made from Lake Mendocino to account for the dynamic variability of flows downstream of the reservoir and to help prevent flows from dropping below the downstream minimum instream flow requirements. These buffer releases are accounted for in the model through constant values which are added to simulated releases. The buffer releases applied in the model vary between 9 and 20 cfs

depending on the month and Water Supply Condition and were developed through an analysis of observed releases and river flows from 2000 to 2013.

Rapid changes in release rate from Lake Mendocino can result in rapid changes in stage downstream, which can potentially strand juvenile salmonids on gravel bar surfaces or in off-channel habitats. Juvenile salmonids that are stranded in off-channel habitat or in cobble substrates are subject to increased mortality (NMFS, 2008). Increasing and decreasing rate of change release criteria (ramping rates) are defined in the WCM and a 2016 letter to the USACE from NMFS in an effort to limit impacts to fish downstream. The LMO model incorporates these constraints in the simulation of releases from Lake Mendocino, which can be seen in Table 3.

Table 3. Lake Mendocino release rate of change guidance.

Period	Release	Hourly		Daily
		IROC	DROC	DROC
	(cfs)	(cfs/hour)	(cfs/hour)	(cfs/day)
March 15 to May 15	>0 & ≤250	1,000	25	50
May 16 to March 14	>0 & ≤250	1,000	25	-
All Year	>250 & ≤1,000	1,000	100	-
All Year	>1,000 & ≤2,500	2,000	100	-
All Year	>2,500	2,000	250	-

IROC - Increasing Rate of Change
DROC - Decreasing Rate of Change

2 Assessment of the HEFS Hindcast Reliability

The NWS has completed verification studies of the HEFS meteorological forcings of temperature and precipitation (Brown et al., 2014a) as well as streamflow forecasts (Brown et al., 2014b), but the Russian River basin was not included as one of the verification sites. Additionally, the CNRFC did not complete a validation study associated with the Russian River hindcast dataset. However, to support the findings in this study, evaluations of the Lake Mendocino inflow hindcast have been completed to assess the reliability for the purposes of flood control management of Lake Mendocino.

One assessment of reliability was to evaluate how well the central tendency and spread of the hindcasted Lake Mendocino inflow ensemble match the central tendency

and spread of observed Lake Mendocino inflows. Figure 6 shows the distributions of 3-day inflow volumes for the 5 to 7-day and 9 to 11-day lead time hindcasts conditioned to the forecast-ensemble median and 95% quantiles. Results from this figure indicate (panels a and c) good skill of the ensemble median forecast, and (panels b and d) the HEFS ensemble spread provides a good representation of the forecast uncertainty (probability distribution function).

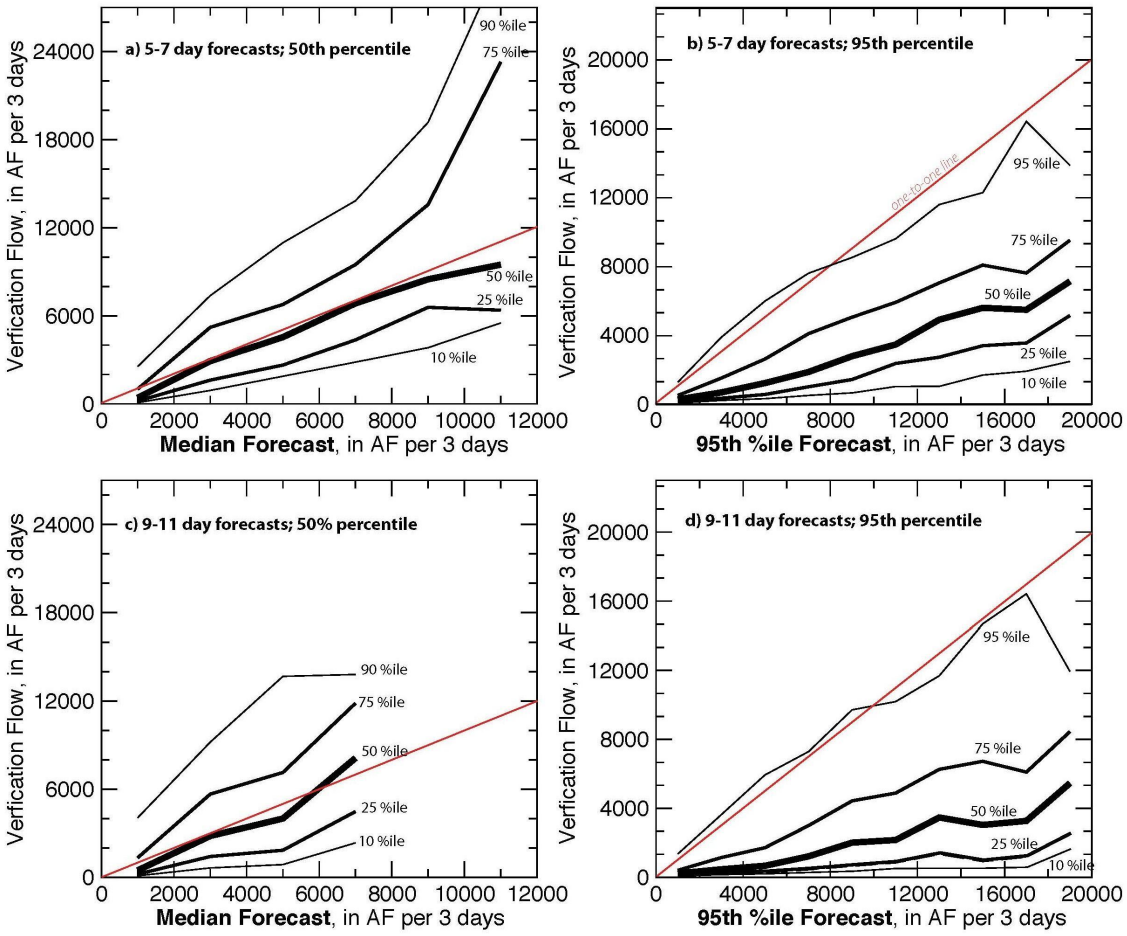

Figure 6 Observed and forecasted November through April 3-day Lake Mendocino inflow volume mean for 5 to 7-day (panels a and b) and 9 to 11-day (panels c and d) lead times for median and 95 percentile forecasts conditioned on the ensemble median.

Figure 7 provides a summary of the same kinds of results for multiple quantiles (25, 50, 75, 90 and 95 percentile) for the 5 to 7-day and 9 to 11-day lead times. It displays the ratio of observed to forecast quantiles (i.e., the 95% observed values [y-axis] divided by the 95% ensemble values [x-axis] from panel b Figure 6, as a function of the forecast magnitude). The grey bars indicate the band where the ratios are within 10% (darker grey)

and 20% (lighter grey) of perfect. Results of this figure show that much of the range of flows for the 5 to 7-day and 9 to 11-day lead times fall within 20% of perfect. However, the lower range of flow forecasts (<2,500 ac-ft in 3-days) and quantiles (<75%) are biased to under forecast, and the lower to middle range of forecasts (0 to 10,000 ac-ft in 3-days) for the upper quantile of observation (90 and 95%) show a tendency to over forecast. Most importantly for the flood-risk management issues that most often dictate the EFO release decisions, the middle to upper range of forecasts (>10,000 ac-ft in 3-days) mostly fall within the light grey shaded region, indicating forecasts that lie roughly within 20% of perfect, but show a general bias toward under forecasting.

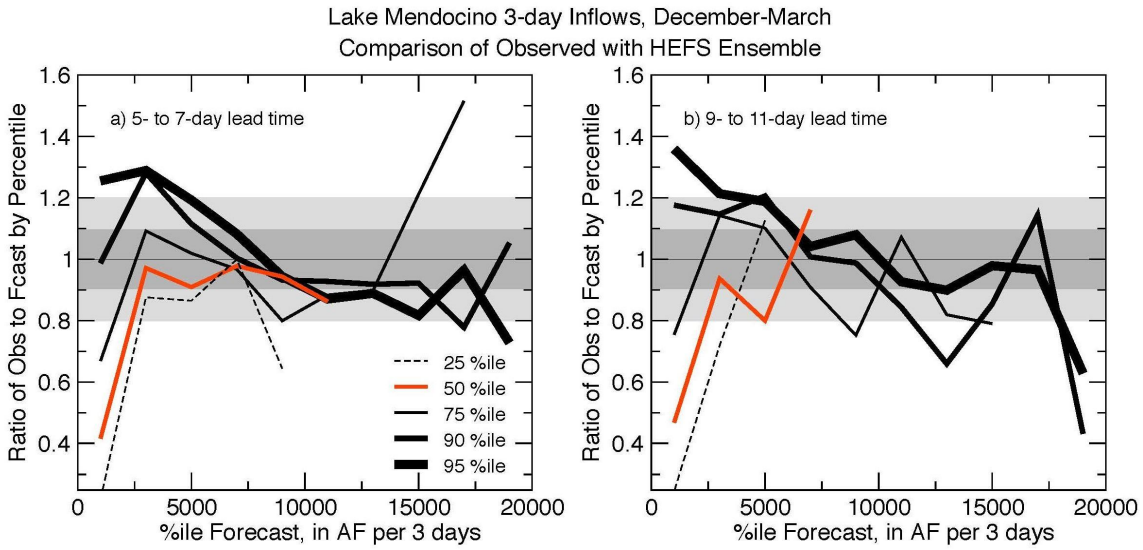

Figure 7 Ratio of observed to forecasted 3-day Lake Mendocino inflow volumes for 25, 50, 75 and 95 percentiles for 5.5 (panel a) and 9.5-day (panel b) lead times.

Rank histograms are a commonly used tool to evaluate the statistical reliability (Type I conditional bias) of ensemble forecast systems (e.g., Hamill and Colucci 1997, 1998; Delle Monache et al., 2013; Elmore, 2005), and according to Hamill (2000), they are useful for determining the reliability of ensemble forecasts and for diagnosing errors in its mean and spread. As an additional assessment of hindcast reliability, cumulative rank frequency plots (provided in Figure 8), which are similar to rank histograms, were developed using the ensemble hindcast and observed Lake Mendocino 3-day inflow volumes from 1985 to 2010 for the months of December through March, which are the months that EFO most heavily relies on forecasts to inform release decisions.

To calculate the rank frequency, each daily observed inflow volume is ranked by the order in which it falls within the range of the ensemble spread. Since there are 61 ensemble members (degrees of freedom) in the hindcast, there are 62 possible ranks with ranks 1 and 62 accounting for observations that fall either less than (rank 1) or greater than (rank 62) the ensemble range. The calculated frequency of each rank is displayed in the plots shown as the monotonic accumulation ordered by increasing rank. The frequency plots were generated for subsets of the hindcast for 5 different forecast lead times (1 to 3, 4 to 6, 7 to 9, 10 to 12, and 13 to 15 days) shown as rows in Figure 8. Bellier et al. (2017) showed that stratification of forecast-observation pairs can be used to learn how forecasts behave under specific conditions. Stratification was applied here to sample different quantile ranges (shown as columns in the figure), which are conditioned by the ensemble median. The first column of frequency plots shows the entire range of hindcasts (0 to 100%), while columns 2 through 4 stratify the hindcasts into ranges of quantiles (50 to 75, 75 to 90 and 90 to 100%). The upper and lower volumes that were used to condition each stratified subset are provided in the title of each subplot.

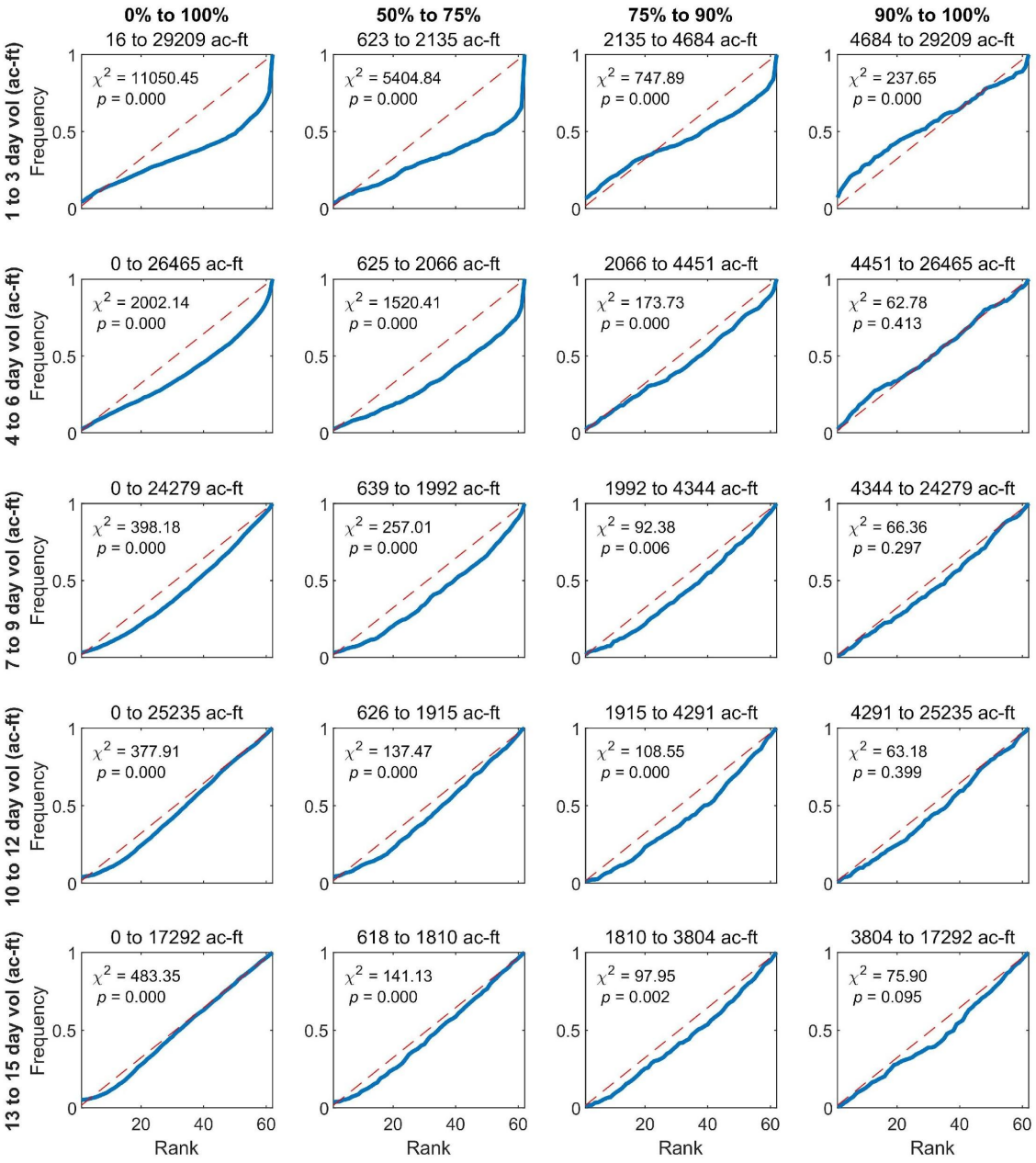

Figure 8 Cumulative rank frequency diagrams of December through March Lake Mendocino 3-day inflow volume for 5 lead times (as rows) and 4 stratified subsets (as columns) conditioned on the ensemble median; upper and lower bound volumes (in ac-ft) used to stratify each subset are included in the titles of each subplot.

An ideal forecast would produce ensemble members that are equally probable, which is presented in the frequency plots as the red dashed lines, which show a straight line with a slope equal to the expected equal probability (1 divided by the 62 possible ranks). These results show a strong tendency for the observation to fall in the upper tail of the ensemble distribution (higher ranks) of the hindcast for 1 through 6 days of lead time, as

evidenced by the hindcast cumulative frequency falling well below the idealized cumulative frequency for most of the distribution and then increasing significantly for the top ranks, especially for the 0 to 100%, 50 to 75% and the 75 to 90% stratified subsets. Forecast reliability improves for the 7 to 15-day lead times with decreasing bias; however, there is still an asymmetric pattern, which implies that the central tendency of the forecasts is systematically too low (i.e., an under forecasting issue). Forecast reliability improves for all lead times for the 90 to 100% subset. The 1 to 3-day lead time of this subset shows possible issues of under dispersion, with higher frequencies occurring in the upper and lower tails of the ensemble distribution. The 4 to 6-day lead time of the 90 to 100% subset shows the best reliability with the cumulative frequency of the hindcast closely matching the idealized cumulative frequency.

To test the assumption of equally probable ensemble members, a chi-square goodness of fit test was completed to evaluate whether the observed frequencies of the hindcast are statistically different from the expected frequencies of an idealized forecast with equally probable ensemble members. The sample subsets for each of the cumulative rank frequency plots (rows of lead time and columns of quantile ranges) were stratified to meet the conservatively defined required sample size of the chi-square test to provide a minimum expected count of at least 5 for each bin. The minimum subset sample size is 312 for the 90 to 100% ranges, which exceeds the required amount of 310 for a 61-member ensemble. The results of the chi-square test are included as insets in each of the cumulative frequency plots with the associated p-value, which is derived from a theoretical chi-square distribution with 61 degrees of freedom.

Results of the chi-square test (a lower result is better and 0 is a perfect score) show that forecast reliability consistently improves with increasing stratified quantile ranges of the ensemble median, showing the best reliability for the highest sample range of 90 to 100%. Assuming a significance level of 0.05, most of the subsets show p-values below the significance level and therefore do not meet the criteria of equally probable ensemble members. However, the 90 to 100% subset for forecast lead times from 4 to 15-days show higher p-values than the significance level and therefore meet criteria of reliability for equal probability.

In summary, these evaluations have demonstrated a systematic bias of the hindcast to under forecast conditions for the months December through March, which are the primary months that EFO simulates flood control release decisions. Bias decreases with increasing days of lead time and for the higher range of forecasts when conditioned on the ensemble median. Based on results of a chi-square test, the hindcast shows the greatest reliability for the 90 to100% stratified subset, especially for lead times of 4 to 15 days which met the assumption of reliability assuming a significance level of 0.05. It should be noted that the stratification volume ranges for the 90 to 100% subsets includes the range of hindcasts (>10,000 ac-ft in 3-days) that would inform many of the flood control pre-release decisions of EFO.

According to Elmore (2005), the chi-square test is insensitive to the nature of the departure from the null distribution, and cannot distinguish between a noisy departure and an ordered systematic departure. However, given the large systematic bias present in the hindcast, a visual inspection of the distribution was adequate to qualify the nature of the departure. Elmore (2005) and Joliffe and Primo (2007) have identified other tests such as the Cramer-von Mises family of statistics, which are better suited for evaluating rank histograms. Future research could use these methods to evaluate the hindcasts of Lake Mendocino and other forecast locations in the Russian River watershed and evaluate the possibility of statistical post processing of the HEFS forecasts to reduce bias.

3 Testing of the Lake Mendocino Operations Model

A historical scenario was developed from 2000 to 2010, the period for which historical operational data was readily accessible for the purpose of testing the LMO model. This scenario simulated historical hydrology and operational constraints for comparison to observed storage levels and system flows to evaluate the accuracy of the primary model assumptions such as the unimpaired flows, estimated reach losses, and reservoir release constraints. Several hydrologic and reservoir operation assumptions were modified for this scenario because they were not stationary during the testing period or consistent with current operations.

The modeled Potter Valley Project diversions used in the LMO model for the Forecast Informed Reservoir Operations (FIRO) alternatives analysis are designed to simulate current operations, post-2006 (as previously discussed). This dataset was not used for the test scenario, because it would not accurately simulate Potter Valley Project operations from 2000 to 2006. For this reason, observed Potter Valley Project releases were used for the test scenario from 2000 to 2010.

The guide curve for Lake Mendocino was not consistent for all of the years of the testing period. Prior to 2007, the USACE operated using a different guide curve with a maximum conservation storage of 86,400 ac-ft, and actual storage levels were observed to reach up to 90,000 ac-ft. To improve simulation of historical flood control operations, the test scenario uses a guide curve which varies over the simulation period with a maximum conservation storage of 90,000 ac-ft from 2000 to 2006 and 111,000 ac-ft (current guide curve maximum storage level) from 2007 to 2010.

Historical water supply operations of the Russian River System from 2000 to 2010 varied due to changes in regulatory compliance such as temporary emergency actions taken for conservation of water supply and/or changes in minimum flow requirements to comply with the BiOp. To capture the operational variability during the testing period, actual historical minimum instream flow requirements were used for the test scenario in place of the simulated minimum instream flow requirements used for the analysis of alternatives.

Results of simulated Lake Mendocino storage for the test scenario were compared to observed storage from 2000 to 2010, as shown in Figure 9. Simulated storage levels closely follow observed storage levels; however, water years 2000 to 2003, 2008 and 2010 show lower peak simulated storage than observed storage. The higher observed storage for these years is the result of reservoir operators allowing storage levels to encroach into the Lake Mendocino flood control pool that was not accounted for in the model.

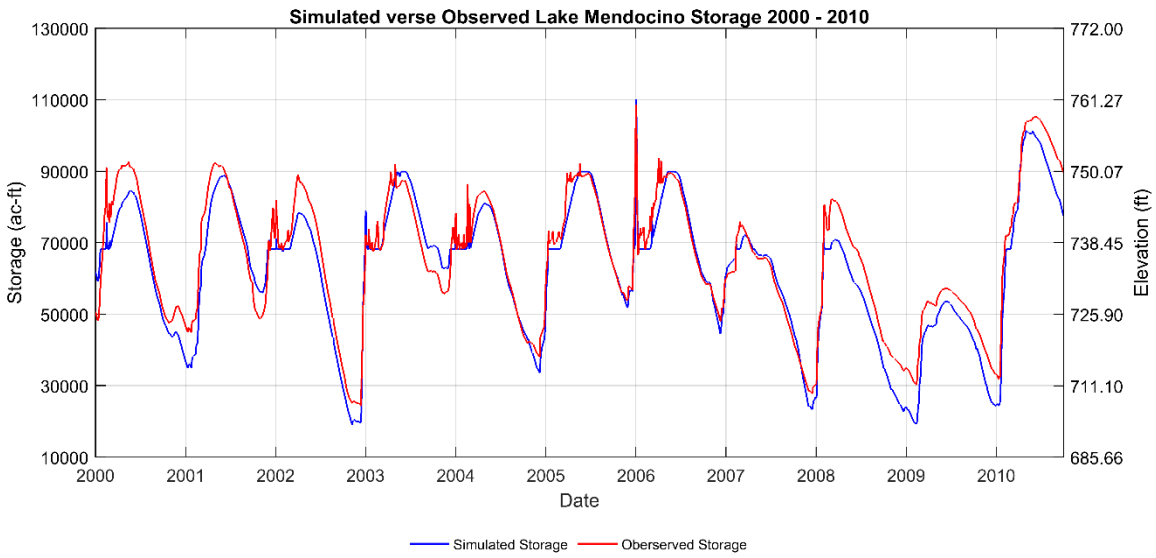

Figure 9. Simulated and observed Lake Mendocino storage (ac-ft) levels for 2000-2010.

Results of the daily simulated flows at the Hopland model junction from the test scenario were compared to observed flows at the U.S. Geologic Survey (USGS) Russian River near Hopland Gage (USGS 111462500) from 2000 to 2010. A scatter plot of simulated flows versus observed Hopland Gage flows is provided in Figure 10. These results show a least-squares linear regression fit with a slope of approximately 0.9 and an R^2 of 0.89. Some of the unexplained variation in the simulated flows is the result of the timing of simulated releases and flow travel times not precisely matching observed conditions.

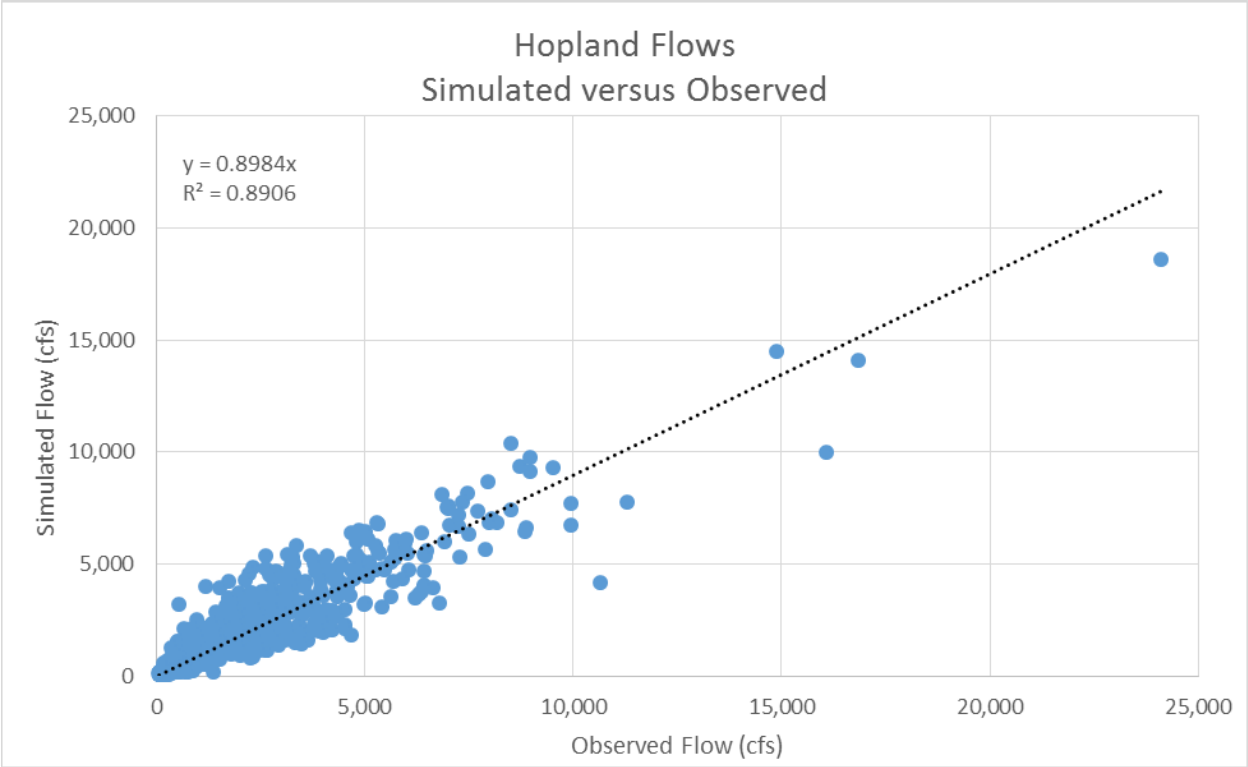

Figure 10. Observed versus simulated Hopland gage flow from 2000 to 2010.

A hydrograph of Hopland junction flow for the 2006 New Year's Day event is shown in Figure 11 which is the largest in the test scenario period, and the fourth largest observed flow since the gage was installed in 1940. Observed flows reach a peak flow of 24,100 cfs on New Year's Day. Simulated flows fall below observed conditions for this event reaching a peak flow of 18,600 cfs, but closely follow observed flows for the days leading up to and following the peak flow event. The difference in peak flow is largely attributable to an under-simulation of Hopland unimpaired flows provided by the CNRFC.

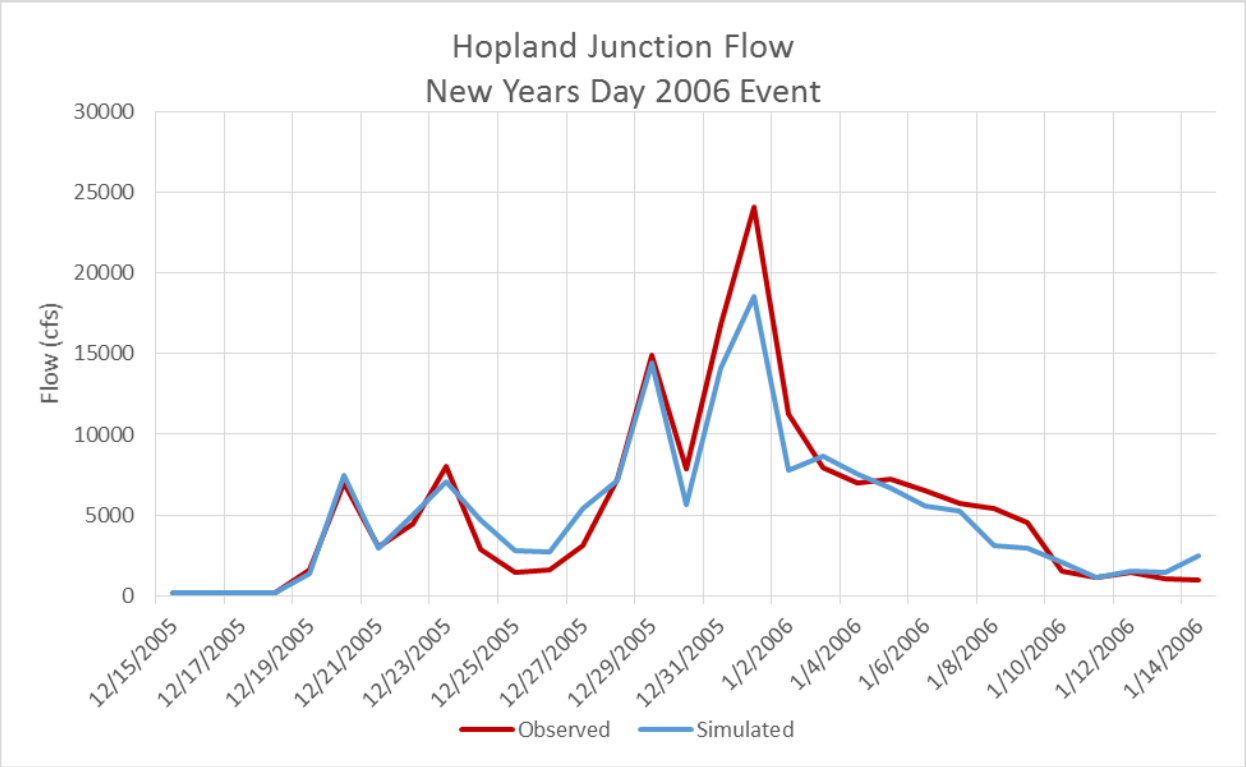

Figure 11. Simulated and Observed flows New Year's Day 2006 flood event for the Hopland model junction for the verification scenario.

Results of the simulated flows at the Healdsburg model junction were also compared to observed flows at the USGS Russian River near Healdsburg Gage (USGS 1146400) from 2000 to 2010. A scatter plot of simulated flows versus observed flows is provided in Figure 12. These results show a least-squares linear regression with a slope of approximately 1.02 and an R^2 of 0.95.

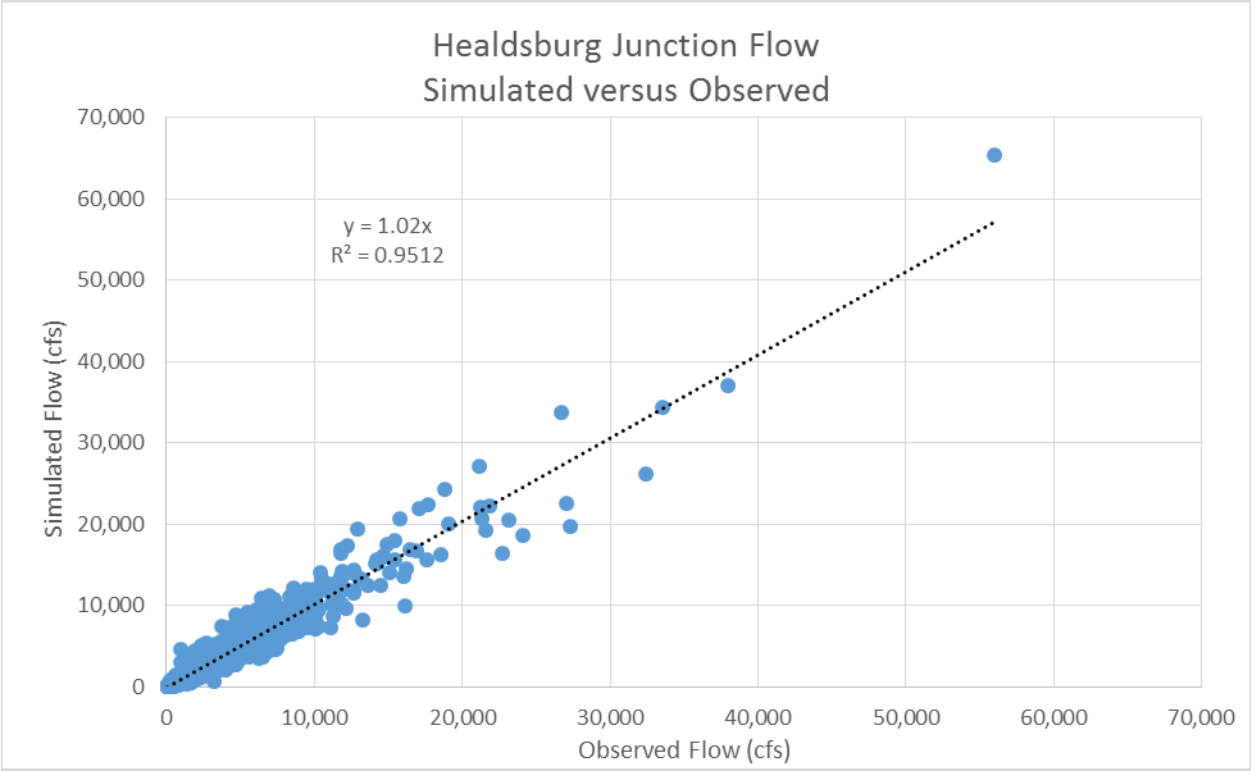

Figure 12. Observed versus simulated Healdsburg gage flow from 2000 to 2010.

The hydrograph of Healdsburg junction flow for the New Year's Day event is shown in Figure 13. Observed flows reach a peak flow of 56,000 cfs on New Year's Day. Simulated flows peak at a level above observed conditions for this event, reaching a peak flow of 65,400 cfs. As with the upstream Hopland junction, simulated flows trend very well with observed flows for the days leading up to and following the peak flow event.

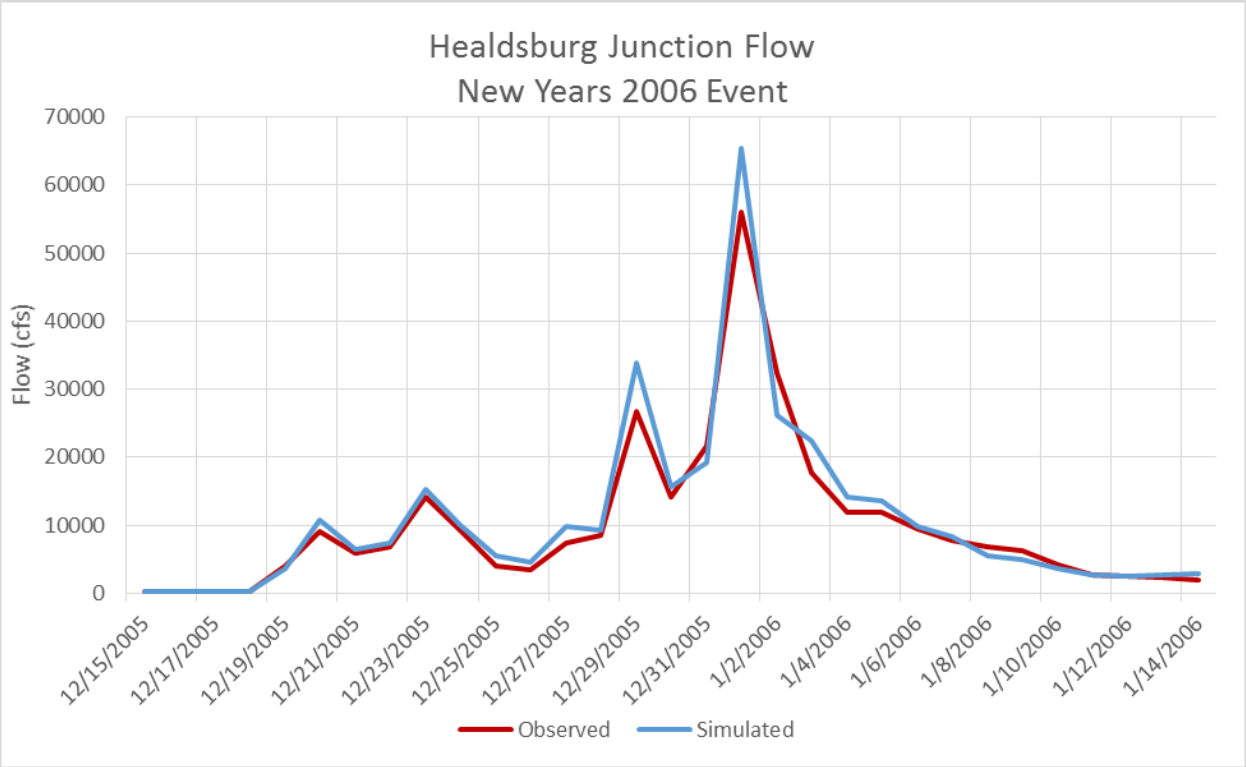

Figure 13. New Year's Day 2006 flood event Simulated and Observed flows for the Healdsburg model junction for the verification scenario.

Results of the test scenario were also evaluated against observed conditions using the Nash-Sutcliffe Efficiency (NSE) (Nash, 1970) and volumetric efficiency (VE) (Criss & Winston, 2008) performance metrics for all of the model junctions. Results of this assessment are summarized in Table 4. These results indicate good agreement between simulated and observed conditions for the test scenario period, 2000 to 2010.

Table 4. Model test performance metrics results.

Model Junction	NSE	VE
Lake Mendocino Storage	0.90	0.93
Hopland gage flow	0.89	0.76
Cloverdale gage flow	0.93	0.79
Healdsburg gage flow	0.95	0.81

Based on the results of the test scenario, we have high confidence that the model input data are of adequate quality to support the comparative analysis of alternatives completed for this study and to discriminate impacts between reservoir management alternatives.

4 Ensemble Forecast Operations Flowchart

A flowchart of the computational steps of the Ensemble Forecast Operations (EFO) alternative for calculating flood control releases can be seen in Figure 14.

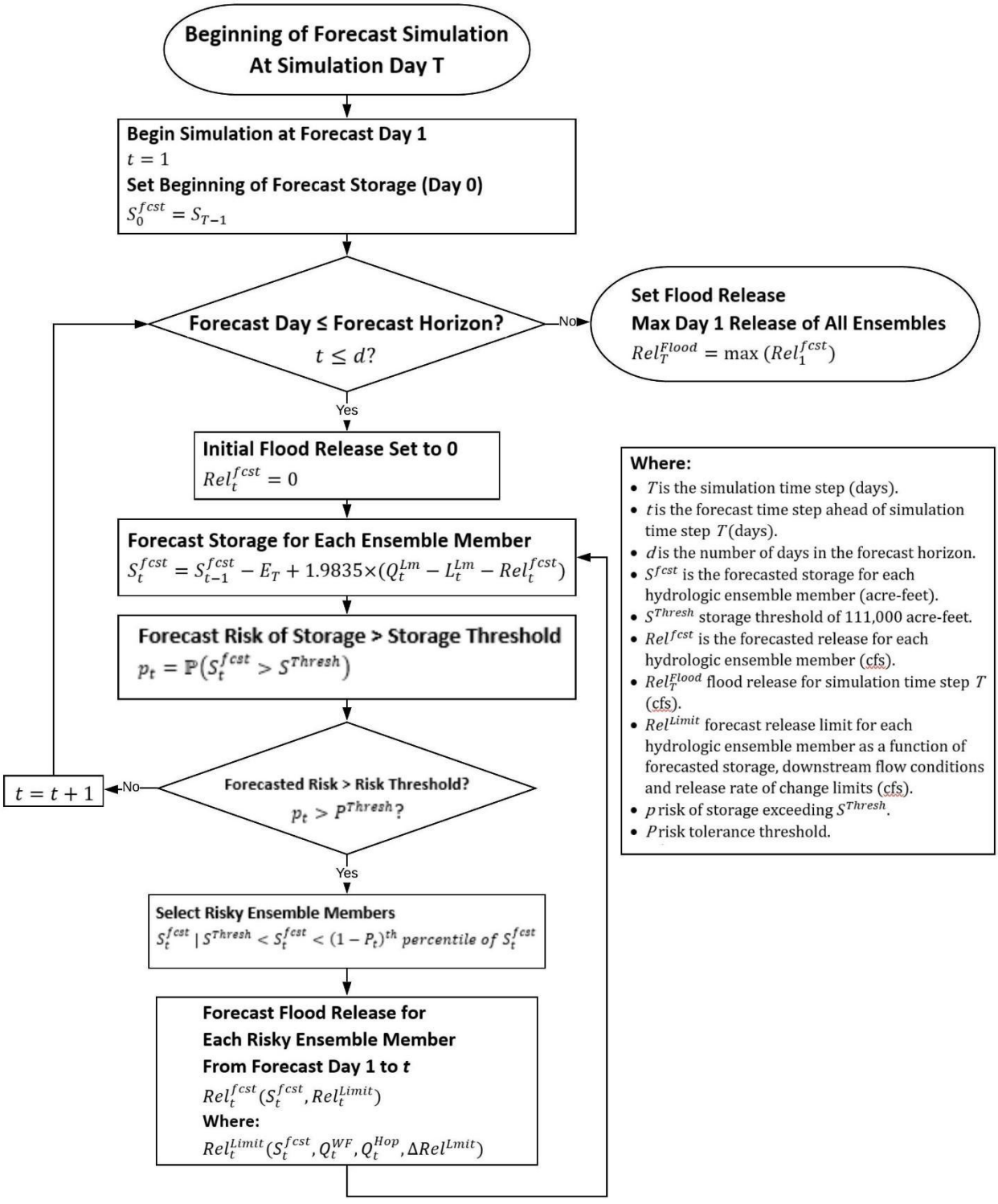

Figure 14. Flowchart for calculating flood control releases for EFO.

5 Development of the Risk Tolerance Curve

The EFO alternative utilizes ensemble streamflow predictions to forecast risk of exceeding a reservoir storage threshold. A central component of this methodology is a risk tolerance curve that is used to formulate forecasted release schedules for Lake Mendocino. The risk tolerance values for the 15 forecast time steps are not independent from each other. Therefore, given 61 possible risk tolerance values (61 ensemble members) at each time step, there are 15^{61} (approximately 5.5 x 1071) possible risk tolerance curves. Given the very large number of possibilities, a brute force method for optimization would be impractical. Additionally, many of the possible combinations of values that would be evaluated in a brute force analysis would not be plausible curve shapes, such as curves that vary erratically from one forecast timestep to the next. Therefore, for this study, a methodology was developed that evaluates a number of plausible curve shapes to select a curve that could be used to support a proof-of-concept simulation of the EFO methodology, but likely not the level of optimization that would be pursued for full implementation. The risk tolerance curve used for this study was developed by modeling numerous curves and selecting one that best meets project objectives of improving storage reliability for downstream water supply and environmental flows without increasing flood risk downstream.

Development of the risk tolerance curve involved the formulation of 5 trials where each trial divided the full simulation period (population [1985 to 2010]) into training and testing periods. Training periods were used to identify an optimal curve, and testing periods were used to evaluate that optimal curve under different hydrologic conditions. Each of the five trials, as shown in Figure 15, evaluated different periods of training and testing. The training periods consisted of 16 years (60% of the population) and the testing periods consisted of 10 years (40% of the population).

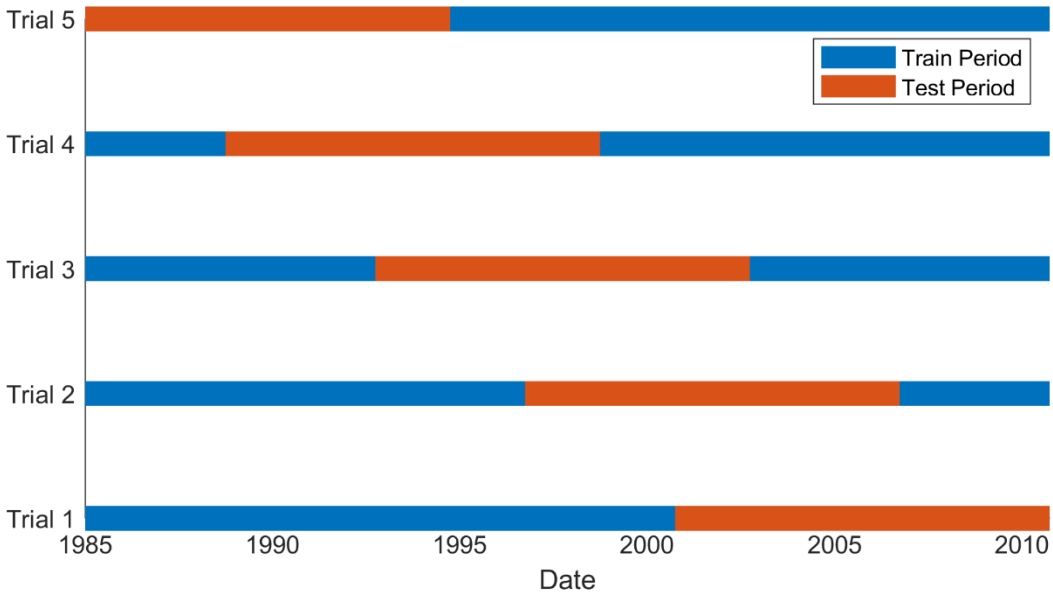

Figure 15. Delineation of testing and training periods for trials.

Simple statistics of standard deviation and mean 3-day inflow volumes of Lake Mendocino from December through March for each of the training and testing periods of each trial are presented in Figure 16. This figure also includes lines showing standard deviation and mean inflow volumes for the full simulation period (population). These statistics differ in standard deviation and/or mean flow volumes between training and testing periods of each trial as well as variability across trials. Notably, Trials 3 and 4 show similar standard deviation, but there are differences in mean values. Trials 1 and 5 also show a similar trend. Visual review of these statistics show that the training and testing sample statistics for each trial do not reflect long term increasing or decreasing trends, but rather reflect seemingly random differences. These results support an assumption of stationarity of the hydrology used for risk tolerance curve development, and differences for the various training periods are likely attributed to natural hydrologic variability within the population.

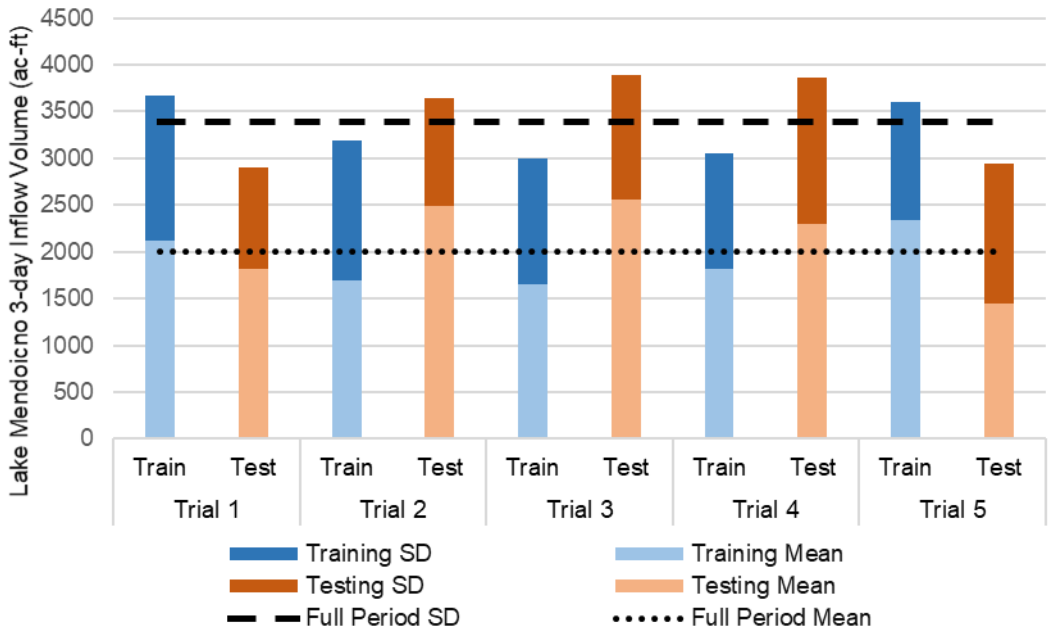

Figure 16. Standard deviation (SD) and mean Lake Mendocino 3-day inflow volumes from December through March for each trial.

Probability exceedance of Lake Mendocino 3-day inflow volumes for the months December through March for both training (left panel) and testing periods (right panel) of each trial is presented in Figure 17. Also included in each panel is the distribution of the full period (population) volumes shown as the dotted line. The training periods of Trials 1 through 4 show similar distributions from the 0 to 0.1% exceedance levels and also have the same maximum values (0%) as the full period volumes, which is from the 1986 flood. Training period for Trial 5, which does not include the 1986 flood, shows lower volumes from the 0 to 0.1% exceedance levels. The Trial 5 testing period includes the 1986 flood, and therefore has the highest maximum value of all of the testing samples, but also lower values for much of the distribution. Trials 3 and 4 show very similar values over the entire distribution for both training and testing periods. Comparing inflow volumes for training and testing periods for each trial show significant variability between testing hydrology and training hydrology. Additionally, variability from the population (full period analysis) is well demonstrated over most of the distribution for both training and testing samples.

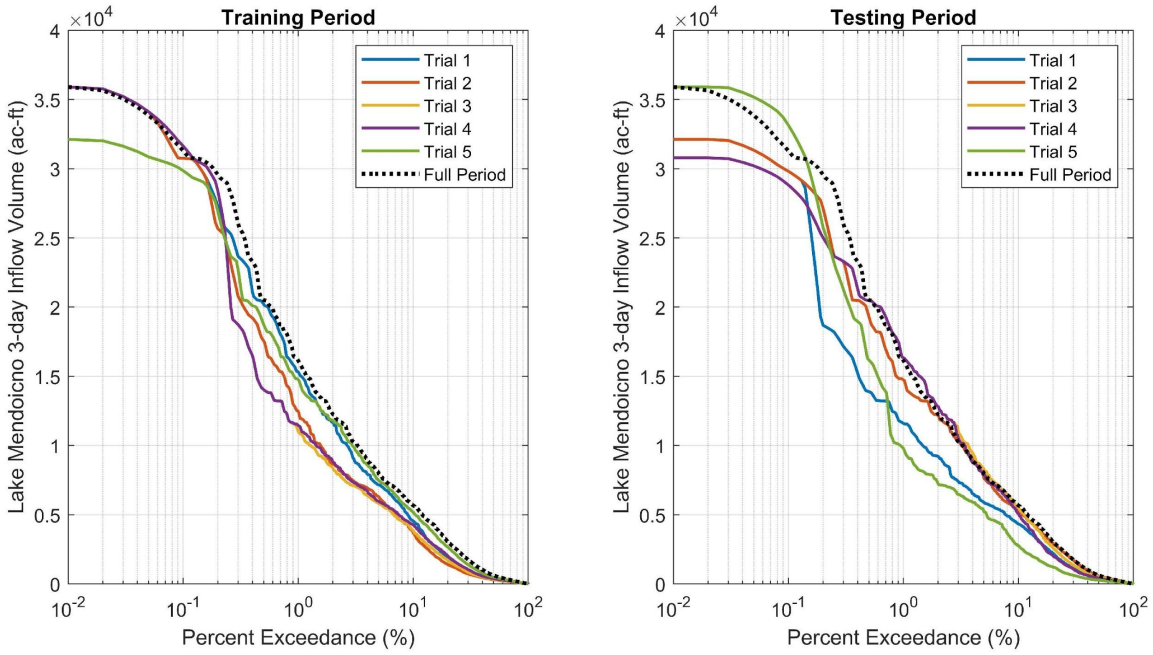

Figure 17. Percent exceedance of December through March Lake Mendocino 3-day inflow volumes for the robustness evaluation trials (solid lines) and the full period of analysis (dashed lines) for training (right panel) and testing (left panel) periods.

The training period of each trial was used to generate the candidate risk tolerance curves to be evaluated by first simulating perfect forecast operations (PFO) for each training period. Then for each day of the training period, forward looking modeled ensemble storage levels (assuming no reservoir releases) were generated using the HEFS hindcasts with the beginning storage determined by the PFO storage level at that time step. Forecasted probabilities or risk of exceeding the storage threshold of 111,000 ac-ft were calculated for each of the 15 days of the hindcast, producing one risk tolerance curve. This process is illustrated in Figure 18 for the training period of Trial 1 for simulation day February 9, 1986, where the risk tolerance curve provided in the inset plot is one of the candidate curves evaluated. This process generated 5,752 risk tolerance curves for Trial 1 (1 curve for each day of the training period), which was reduced to a set of 1,124 unique curves as shown in Figure 19. This process was repeated for the training period of each trial to generate a set of candidate risk tolerance curves for the respective trial.

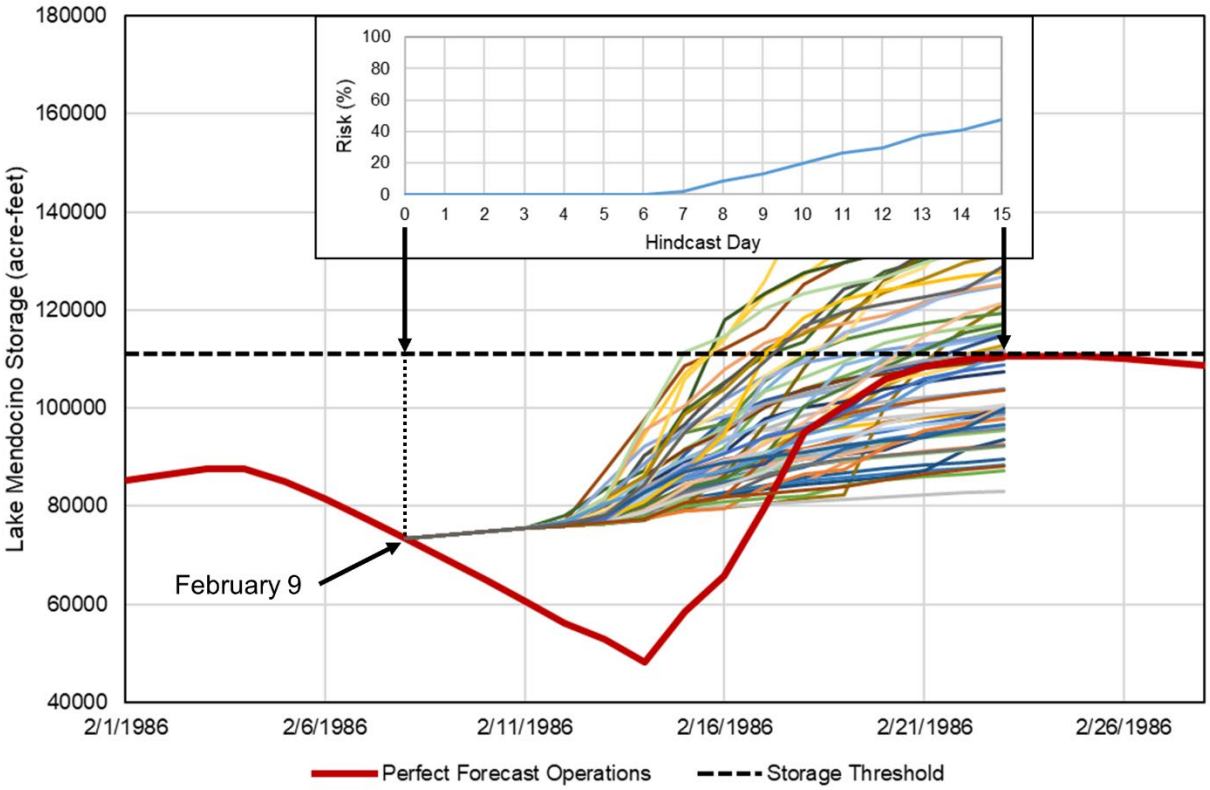

Figure 18. February 9, 1986 example of candidate risk tolerance curves using storage results from the PFO simulation of the training sample from Trial 1.

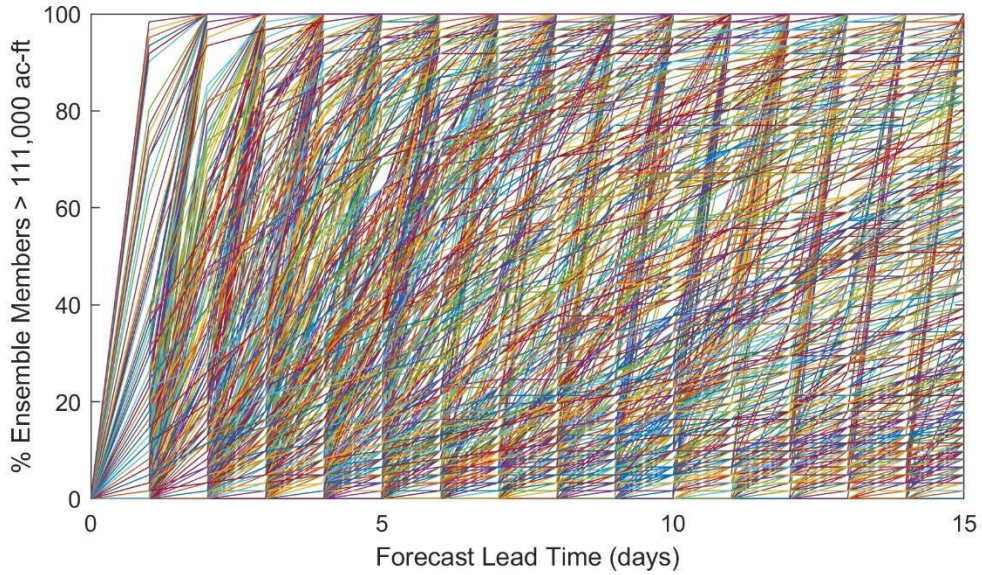

Figure 19. Candidate risk tolerance curves evaluated with LMO model for Trial 1.

Each of the candidate risk tolerance curves developed for each trial was simulated with the LMO model for the training period of each trial. For example, each of the 1,124

risk tolerance curves developed for Trial 1 was simulated for the entire training period of Trial 1. An objective function consisting of 3 decision variables (provided in the Equation 2) was developed to evaluate the simulation results of each training period. This objective function was formulated based on the project objectives to improve reservoir storage reliability while not increasing downstream flood risk.

$$Max\ J = (S + Q) \times NoSpill \quad (2)$$

The S decision variable is calculated as the mean May 10 storage ($\underline{s}_j^{May10}$ in ac-ft) over the entire training period for a candidate curve (j) greater than the minimum May 10 storage of all curves evaluated in the training period, divided by the range of mean May 10 storage results of all curves evaluated, which is summarized in Equation 3 and 4, given n is the number of years in a training period.

$$S_j = \frac{\bar{s}_j^{May10} - \min(\bar{s}^{May10})}{\max(\bar{s}^{May10}) - \min(\bar{s}^{May10})} \quad (3)$$

$$\bar{s}_j^{May10} = \frac{s_1^{May10} + s_2^{May10} + \cdots + s_n^{May10}}{n} \quad (4)$$

The S decision variable captures the objective to improve storage reliability, and, therefore, higher values favor curves with a higher risk tolerance.

The Q decision variable is calculated as the total flow volume (q_j^{vol} in ac-ft) above flood stage (where flows at Hopland [$q^{Hopland}$] exceed 8,000 cfs) for a candidate curve (j) less than the maximum flood volume for all curves evaluated, divided by the range of flood volumes of all curves evaluated, which is summarized in Equation 5 through 7, given 1 cfs for 24 hours is 1.9835 ac-ft.

$$Q_j = \frac{\max(q^{vol}) - q_j^{vol}}{\max(q^{vol}) - \min(q^{vol})} \quad (5)$$

$$q_j^{vol} = \sum q^{flood} \times 1.9835 \quad (6)$$

$$\{q^{flood} | q^{flood} \in \mathbb{R}, q^{Hopland} > 8{,}000\} \quad (7)$$

Due to the size of Lake Mendocino and the watershed area impounded by the lake, flood control capacity is limited, and during flood events most flooding is caused by natural flow downstream of the lake. This makes decreasing flood risk by an appreciable amount

challenging for this system. However, the risk tolerance curve is used in EFO to select the ensemble member for the calculated flood control release, which includes accounting for downstream conditions. Lower risk tolerance values result in the selection of more conservative (wetter) ensemble members for scheduling reservoir releases. The EFO model seeks to limit flows above flood stage when scheduling releases; therefore, a wetter ensemble member (a more conservative forecast) can reduce the contribution of releases to downstream flows. As a result, higher values of the Q decision variable favor curves that are more risk intolerant (lower risk levels).

During the process of information gathering for this study, reservoir operators expressed a strong aversion to spillway releases. Additionally, design documents of the reservoir indicate that storage levels were designed to reach the emergency spillway for a 1 in 50-year event (USACE, 1954). A frequency analysis completed by the USACE (USACE, 2020) of Lake Mendocino 3-day inflow volumes showed that the largest flood event within the simulation period (February 1986) was less than a 1 in 50-year event. Therefore, to capture the importance of this criteria to the stakeholder and stay within the original level of design, $NoSpill$ is defined as an indicator variable such that if there are any simulated spillway releases then $NoSpill$ is set to a value of 0, and if there are no spillway releases then $NoSpill$ is set to a value of 1.

Results of the objective function computations for each of the 1,124 simulations for the training period of Trial 1 are provided in Figure 20. It can be seen that many of the candidate risk tolerance curves resulted in a spillway release, and therefore have objective function values of 0.

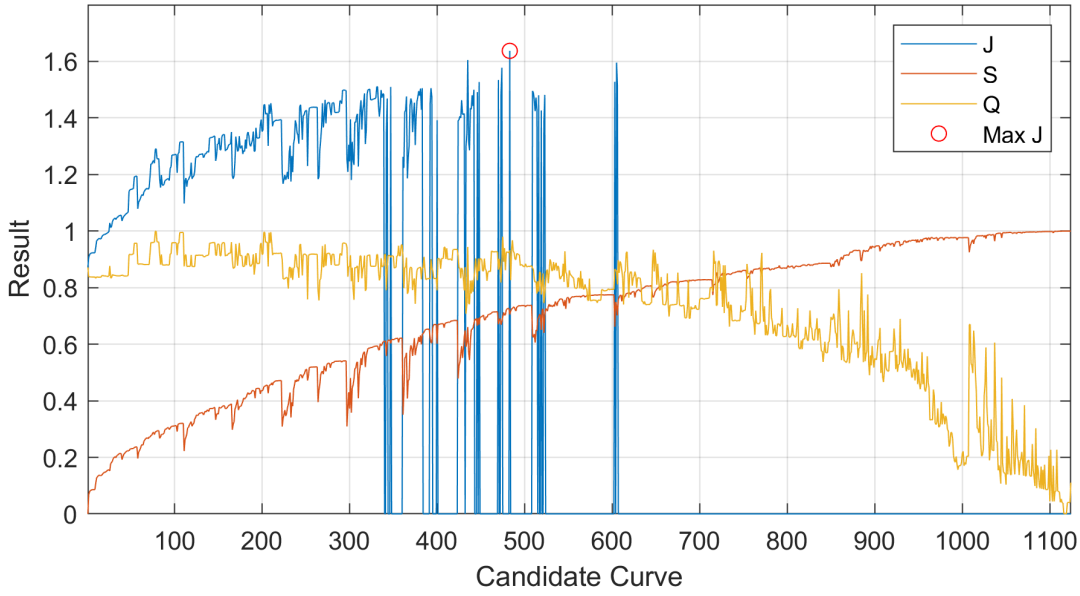

Figure 20. Objective function results (J) for candidate risk tolerance curves, and S and Q decision variable results for the training period of Trial 1.

The objective function result circled in red shows the highest value, and, therefore, the risk tolerance curve associated with that result was selected as the optimal for Trial 1. This process was repeated for each trial, and optimal risk tolerance curves were identified for each training period which are shown in Figure 21. Trials 1, 2 and 4 identified the same curve shape, while Trials 3 and 5 showed different optimal curves due to differences in hydrology of their training periods.

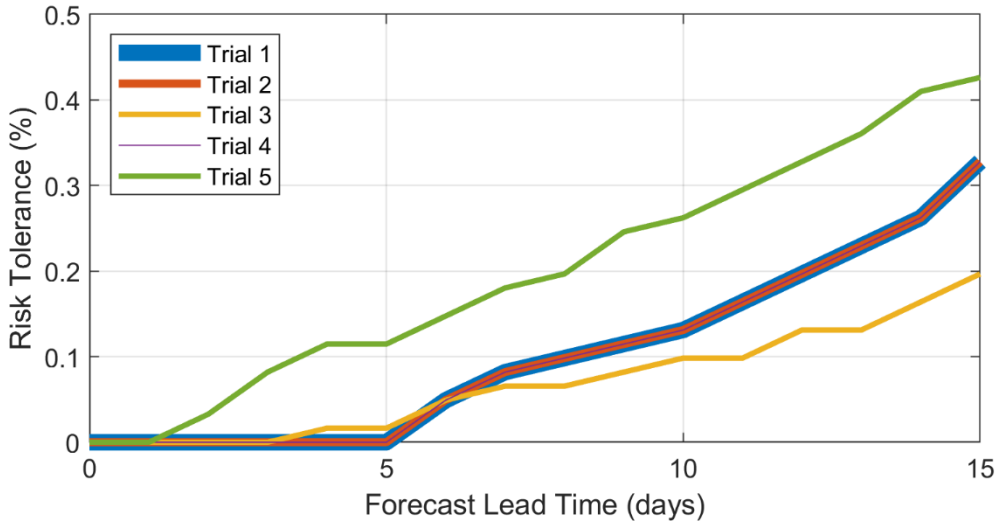

Figure 21. Optimal risk tolerance curves identified in training period trials.

The testing period of each trial was used to evaluate the robustness of the optimal curve derived from the training period of the trial. The robustness was evaluated by comparing the performance in meeting the objective function of the derived optimal curve to the candidate curves using the testing period. The curves were compared using rank position by ranking their objective function score in ascending order, and, since many objective function scores were 0 due to the presence of spills, curves were secondarily ranked by spill volume in descending order.

A summary of the results of the training and testing simulations is presented in Table 5, which shows the number of candidate risk tolerance curves evaluated for each trial as well as the decision variable and objective function results for the optimal curve identified in the training and testing period. Also included is the percent rank position of the testing period objective function result (J^{test}) for the optimal curve determined in the training period relative to all the curves derived from the training period and evaluated in the testing period.

As previously shown in Figure 21, the training periods of Trials 1, 2 and 4 resulted in the same optimal curves, and the results in Table 5 also show that curve shape provides robust results for the test period simulations with J^{test} values ranking greater than 92% of the candidate curves. Also as previously discussed and shown in Figures 16 and 17, the training period of Trials 3 and 4 had similar December through March standard deviation and 0 to 1% exceedance of 3-day Lake Mendocino inflow volumes, yet showed different mean values. Although these trials showed similarities in the very high flow ranges (0 to 1%), training resulted in different optimal curves. However, it should be noted that the optimal curve from Trials 1, 2 and 4 ranked 3rd from optimal in Trial 3 with a training objective function score within 2% of the optimal curve score for this trial. As shown in Figure 21, the training period of Trial 5 produced a very different optimal curve due to the large differences in hydrology for the 0 to 0.1% exceedance range (the highest flows) from the other trials. The training period optimal curve from Trial 5 resulted in a spillway release for the testing period, which led to an objective function result of 0 and a low rank (60%) relative to the other curves. However, the S and Q decision variables for the Trial 5

curve showed high values in the testing period compared to the optimal curves in other trials.

Table 5. Summary of robustness testing results for each trial.

		Trial				
		1	**2**	**3**	**4**	**5**
No.	**Risk Curves**	1124	964	1010	1080	1284
Training Periods	**S**	0.71	0.70	0.72	0.69	0.90
	Q	0.93	0.99	0.77	0.82	0.79
	NoSpill	1	1	1	1	1
	J^{train}	1.64	1.69	1.49	1.51	1.69
Testing Periods	**S**	0.70	0.71	0.69	0.73	0.90
	Q	0.72	0.74	0.80	0.91	0.96
	NoSpill	1	1	1	1	0
	J^{test}	1.42	1.45	1.49	1.64	0.00
	% Rank J^{test}	92%	93%	75%	93%	60%
	$J^{train} + J^{test}$	3.06	3.14	2.99	3.15	1.69
% Rank	**$J^{train} + J^{test}$**	100 %	100%	99%	100%	69%

Table 5 also includes the combined objective function results for the training and testing periods ($J^{test} + J^{train}$), and the percent rank position of the combined result for the optimal curve for a given trial within all of the curves evaluated. These combined results show that the risk tolerance curve of Trials 1, 2 and 4 was the top ranked curve for those trials. The curve from Trial 3 also showed good results in the 99% rank position (7th from the optimal curve), and the curve from Trial 5 showed to be suboptimal for the combined result.

Given the assumption of stationarity for the population of the hydrologic data as previously discussed, a final test was completed that included the full period of simulation (1985-2010) to train and test all candidate risk tolerance curves to fit the final optimal risk tolerance curve. Given the longer period of simulation than what was completed for the trials, the PFO simulation for the full period derived 1,772 candidate risk tolerance curves which included the 3 optimal curves (shown in Figure 21) identified in the 5 trials. Results of this analysis are presented in Figure 22, which shows objective function (J) plotted in ascending order of value, and, since many have a result of 0 due to spills, secondarily by

descending order of spill volume. S and Q decision variables are also included in this plot with plotting positions mapped to the objective function values. Additionally, markers are included showing the position of the results from the optimal risk tolerance curves derived from the 5 trials with marker colors set to match the colors of the variables.

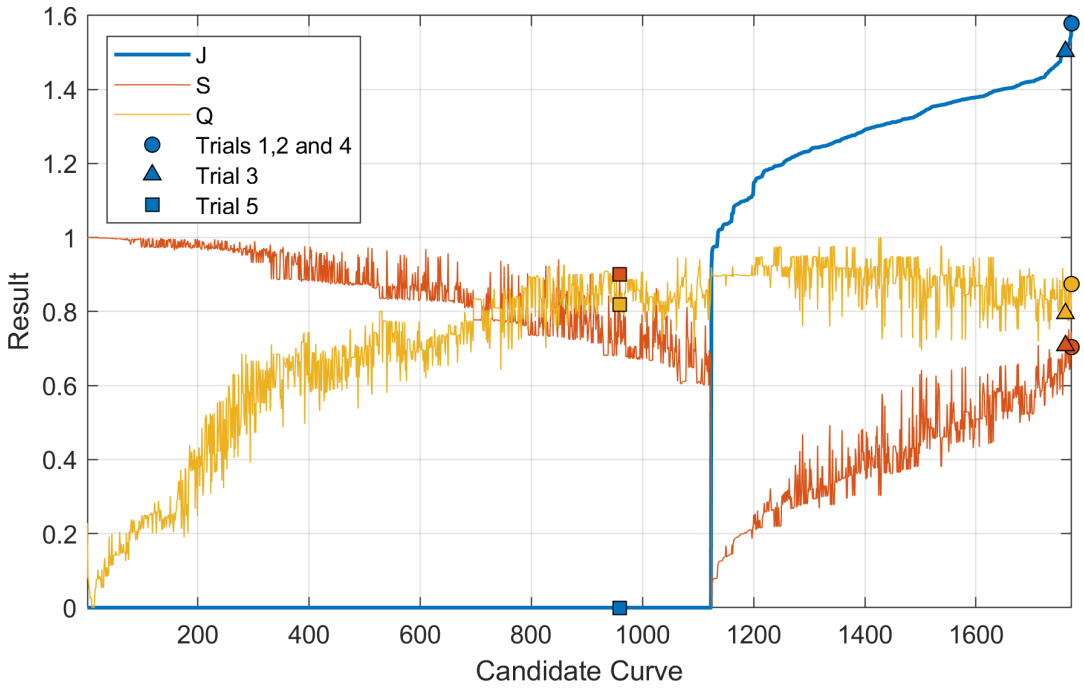

Figure 22. Objective function results (J) for the full period of simulation in ascending order of values, and secondarily by descending order of spill volume; S and Q decision variable results are mapped to the objective function results; markers show positions of risk tolerance curves identified in the trials with colors matching the line colors of J, S and Q.

Results of the trials and the full period analysis showed strongest consensus for the optimal curve identified in Trials 1, 2 and 4, which was selected and used for this study to test EFO for Lake Mendocino. However, based on the testing samples from Trial 3 and 5, results show high sensitivity to the hydrology used for optimization. This indicates that the risk tolerance curve used for this study is likely overfitted due to lack of hydrologic variability available to train the risk tolerance curve, and future efforts to develop a risk tolerance curve could produce different results.

The methodology presented here is a heuristic approach to optimize a risk tolerance curve that is adequate for proof-of-concept testing and evaluating EFO for Lake Mendocino. The novelty of this methodology is in using the PFO simulation to derive a

realistic number of plausible, candidate risk tolerance curves to be evaluated with an objective function, which significantly reduces the scope of the problem compared to a brute force approach. Based on the results of the trials, this methodology showed sensitivity to the hydrology used for training the risk tolerance curve; therefore, future research could be completed that incorporates more extreme hydrology both for flooding and drought. Incorporation of more extreme hydrology would also likely require a different objective function formulation. The current objective function is heavily weighted against spills by incorporating the *NoSpill* indicator variable as a Boolean multiplier. The incorporation of more extreme hydrologic events in the training and testing of risk tolerance curves would likely increase the frequency of spills, and the current objective function would likely favor very risk averse curves that are not consistent with management objectives. Future formulations of the objective function could seek to minimize spill volumes while also balancing the storage reliability and flood risk reduction objectives.

Future efforts of sampling training and testing hydrology for cross validation should be informed by an evaluation of stationarity of observed and hindcast hydrology using more sophisticated methods such as spectral analysis (Fleming et al., 2018). Additionally, further research could be completed to cross validate the approach presented here against more established optimization methods such as stochastic dynamic programming (Stedinger et al., 1984).

References

Bellier, J., Zin, I., & Bontron, G. (2017). Sample Stratification in Verification of Ensemble Forecasts of Continuous Scalar Variables: Potential Benefits and Pitfalls. *American Meteorological Society*, Vol. 145, 3529-3544.

Brown, J. D., Wu, L., He, M., Regonda, S., Lee, H., & Seo, D. (2014). Verification of temperature, precipitation, and streamflow forecasts from NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 2. Streamflow verification. *Journal of Hydrology*, Vol. 519, Part D, 2847-2868.

Brown, J. D., Wu, L., He, M., Regonda, S., Lee, H., & Seo, D. (2014). Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 1. Experimental design and forcing verification. *Journal of Hydrology*, Vol. 519, Part D, 2869-2889.

Criss, R. E., & Winston, W. (2008). Do Nash Values Have Value? Discussions and Alternate Proposals. *Hydrological Processes*, 22, 2723-2725.

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., & Searight, K. (2013). Probabilistic Weather Prediction with an Analog Ensemble. *Monthly Weather Review*, Vol. 141, 3498-3516.

Demargne, J., Wu, L., Regond, S. K., Brown, J. D., Lee, H., M., H., . . . Zhu, Y. (2014). The Science of NOAA's Operational Hydrologic Ensemble Forecast Service. *American Meteorological Society*, 79-98.

Elmore, K. L. (2005). Alternatives to the Chi-Square Test for Evaluating Rank Histograms from Ensemble Forecasts. *Monthly Weather Review*, Vol. 20, 789-795.

FERC. (2004). *Order Amending License (106 FERC 61,065).* U.S. Federal Regulatory Commission.

Fleming, S., Lavenue, A., Aly, A., & Adams, A. (2002). Practical Applications of Spectral Analysis to Hydrologic Time Series. *Hydrological Processes.*, 16. 565 - 574. 10.1002/hyp.523. .

Hamill, T. M. (1998). Evaluation of Eta-RSM Ensemble Probabilistic Precipitation Forecasts. *Monthly Weather Review*, Vol. 126, 711-724.

Hamill, T. M. (2000). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *American Meteorological Society*, Vol. 129, 550-560.

Hamill, T. M., & Colucci, S. J. (1997). Verification of Eta-RSM Short-Range Ensemble Forecasts. *Monthly Weather Review*, Vol. 125, 1312-1327.

Hamill, T. M., Bates, T. M., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau, T. J., . . . Lapenta, W. (2013). NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. *American Meteorological Society*.

Jolliffe, I. T., & Primo, C. (2007). Evaluating Rank Histograms Using Decompositions of the Chi-Square Test Statistic. *Monthly Weather Review*, Vol. 136, 2133-2139.

Nash, J. E. (1970). River flow forecasting through conceptual models part I -A discussion of principles. *Journal of Hydrology*, 10 (3), 282-290.

NMFS. (2008). *Biological Opinion for Water Supply, Flood Control Operations, and Channel Maintenance conducted by the USACE, the SCWA, and the MCRRFCWCID in the Russian River Watershed.* Santa Rosa, California: National Marine Fisheries Service.

Sonoma Water. (2016). *Fish Habitat Flows and Water Rights Project Draft Environmental Impact Report.* Santa Rosa (CA): Sonoma County Water Agency.

Stedinger, J. R., Sule, B. F., & Loucks, D. P. (1984). Stochastic Dynamic Programming Model for Reservoir Operation. *Water Resources Research*, Vol. 20, Issue 11, p 1499.

U.S. Army Corps of Engineers. (2020, February). E-mail communication. *Frequency Analysis of Lake Mendocino Inflows*. Davis, CA.

USACE. (1954). *United States Army Corps of Engineers, Design Memorandum No. 2, Hydrology and Hydraulic Analysis, Russian River Reservoir (Coyote Valley), California.* San Francisco: U.S. Army Corps of Engineers San Francisco District.

USACE. (2003). *United State Army Corps of Engineers, Coyote Valley Dam and Lake Mendocino, Russian River, California, Water Control Manual: Appendix I to master water control manual Russian River basin, California.* Sacramento (CA): U.S. Army Corps of Engineers, Sacramento District.