Peer Review Summary

Peer Review for 24-Month Study Year 1 and Year 2 Analysis and Recommendations

Date

May 28, 2026

Originating Office

Research and Modeling Group, Upper Colorado Basin Region, Bureau of Reclamation

Reclamation Roles

Director or Delegated Manager: Valerie Deppe, Projects, Operations, and Modeling Division Manager, Upper Colorado Basin Region, Bureau of Reclamation

Peer Review Lead: Paige Becker, Research Hydrologist, Upper Colorado Basin Region, Bureau of Reclamation

Peer Review Scope

Reclamation publishes a 24-Month Study (24MS) each month projecting monthly reservoir operations for the Colorado River. The 24MS uses an unregulated inflow forecast that is derived through a combination of methods, which includes water supply forecasts from the Colorado Basin River Forecast Center (CBRFC), that are disaggregated monthly in the 24MS's first year and coupled with historical streamflow statistics for the 24MS's second year. The analysis that was peer reviewed implements different methods to calculate April through July forecast exceedances, uses different forecast methods for the 24MS than currently used by Bureau of Reclamation, and recommends changing the exceedance trace language presently shared by Reclamation.

The subject of this review is to determine if the analysis and recommendations are reasonable, and should focus on if the methodology, metrics, and recommendations are aligned with the goal of improving forecasts in the Colorado River. Specifically, the reviewer will respond to the following questions:

Question 1: Is the analysis clearly explained and results properly interpreted?

Question 2: Are the recommendations supported by the results of the analysis?

Peer Reviewers

Peer reviewers were selected who have a familiarity with operations in both the Upper and Lower Colorado regions. The two selected reviewers are from the Bureau of Reclamation and listed below:

Shana Tighi

Hydrologist, Team Lead, River Operations Group, Boulder Canyon Operations Office, Bureau of Reclamation, Boulder City, NV.

Alex Walker, Civil Engineer, Water Management Group, Upper Colorado Basin, Bureau of Reclamation, Salt Lake City, UT

Summary of Reviewer Comments

Responses to the specified questions in the peer review scope:

1. Is the analysis clearly explained and results properly interpreted?
   Response: The concept of reliability is not well explained, particularly in the Year 1 analysis. Because of this, it is hard to say that the results are properly interpreted. Once reliability is better explained, it will be clearer if the results are properly interpreted.
2. Are the recommendations supported by the results of the analysis?
   Response: In the current drafts, the recommendations are not explicitly stated, they are only implied. Suggest stating the results more directly.

Key Issues Identified

In addition to the two questions above, a few key issues were identified by the reviewers that have been addressed in the revised analysis reports. Summaries of these key issues are listed below, as well as all the comments and responses in the tables attached below.

1. Explanation of reliability is unclear.
   Summary of Comments: Explanation of reliability is confusing, as well as why it is an important metric to use and how to interpret the results.
   Response: Further description of reliability was included in the reports as well as why it is a useful metric, and how to interpret the results of this metric. Revisions were provided to peer reviewers to confirm explanations were clearer and addressed their concerns.
2. Explanation for why the 10^{th} and 90^{th} percentiles and Probable Minimum and Probable Maximum errors/ biases should be more negative or more positive with greater lead times.

Response: Further explanation was provided on what the expected results are for bias and error when forecasting the dry and wet conditions, and how to interpret the results. Revised text should more clearly explain the expected results and why the high positive bias or error are a good thing for the wet and dry scenarios.

3. It is not clear what the priorities are – more accuracy (lower error) or greater uncertainty (greater absolute error) – when forecasting in the out year, and why that should be the priority.
   Response: Explanation for why greater uncertainty and better reliability for the dry and wet conditions in the out year should be prioritized have been added to the reports for clarity.
4. The recommendations are not explicitly stated in either report and are only implied.
   Response: Recommendation sections have been added to both analysis reports, adding clarity to how Reclamation should proceed based on the findings.
5. All of the comments from the reviewers as well as their responses can be found in the tables below.

Table 1: Peer Review comments and responses for 24MS_Year1 Report

Section	Text	Comment	Reviewer	Response
I. Comparing Streamflow Forecast Methods for Year 1 of the 24-Month Studies: Period versus Monthly Methods	Comparing Streamflow Forecast Methods for Year 1 of the 24-Month Studies: Period versus Monthly Methods	Overall, I have these general notes: 1. Defining the key takeaway and then focusing on the takeaway would be very helpful. Streamlining the analysis section will help with that. What is the reader trying to take away from this report. 2. Can you recommend a preferred method in the text? Hard to tell opinions on using one vs the other for 24MS analysis. 3. My hunch is that the report will be read and consumed by a wide variety of stakeholders in the basin, the majority of whom will be non-technical and not fully up to speed on forecasting. Considering the audience, I feel that this report cannot assume a ton of background knowledge for a reader and will need to have a lot of context, either in the report or in references. Otherwise it will add to confusion. My guess is that these findings will be interpreted entirely in the context of operations, so beginning the ability to run comparative analyses between methods in Riverware asap will be important as well.	Walker, Alexander E	A recommendations section is added to the summary so the reports can be read separately as well. As it stands, this report is intended for internal use and reading and so the background is intended to be limited and focus more on the analysis. Finally, the Year 2 Report includes the RiverWare run using monthly ESP 10s and 90s.
A. Background	Background	This section also needs 1-2 sentences covering Year One vs Year Two	Walker, Alexander E	This is unnecessary for the purposes of this specific report which is primarily about the CBRFC forecasts. It feeds into year 2 but is not necessarily the same product.
A. Background	two	I think it would be a more accurate characterization to say that they produce a broad suite of forecast products, including the two highlighted here.	Tighi, Shana G	Statement updated to say there are a suite of forecasts issues

Section	Text	Comment	Reviewer	Response
		The paragraph only references the Upper Basin products used in the monthly 24-month study, but does not include other key products such as the CRMMS traces, the LML forecast, or the peak inflow forecasts for flood control purposes.		
A. Background	The Colorado Basin River Forecast Center (CBRFC) issues two main forecast products for the Bureau of Reclamation (Reclamation). The first is the spring runoff water supply volume forecasts (for April – July), period, which are issued twice a month from two times each month January through June (Figure 1). The second is the monthly volume forecasts for the entire water year, issued once per month year-round. The forecasts are generated using 30 years of historical weather data (1991-2020 precipitation and temperature) at a 6-hour time step. The Weather data are then	I think that including a methods source is needed so that the reader can learn more about CBRFC process if they would like.	Walker, Alexander E	Link to CBRFC is included

Section	Text	Comment	Reviewer	Response
	input to the CBRFC's hydrologic model initialized with current basin conditions, which result in 30 possible hydrograph scenarios, or traces. These traces are used to create forecasts of the 10th, 50th, and 90th percentile water supply volumes, where the 10th percentile indicates that 10% of the values are likely to be equal to or less than that amount.			

Section	Text	Comment	Reviewer	Response
A. Background	There are two different methods for calculating the percentile volumes: the Period Method and Monthly Method. The Period Method calculates the percentile volumes based on the aggregated period volume (e.g. April-July) from each trace. $Period\ Method\ Percentile = n < Apr\text{–}July\ Inflow\ N\ periods * 100$ The Monthly Method calculates percentile volumes for each month then adds those together for each percentile.	General equations for both Period and Monthly method will be helpful here.	Walker, Alexander E	Equations for both methods have now been included

Section	Text	Comment	Reviewer	Response
A. Background	There are two different methods for calculating the percentile volumes: the Period Method and Monthly Method. The Period Method calculates the percentile volumes based on the aggregated period volume (e.g. April-July) from each trace. $Period\ Method\ Percentile = n < Apr{-}July\ Inflow\ N\ periods * 100$ The Monthly Method calculates percentile volumes for each month then adds those together for each percentile.	I tried to put them together based on my understanding of the math	Walker, Alexander E	See response to the comment above

Section	Text	Comment	Reviewer	Response
A. Background	volumes	Suggest appending "for use in hydrologic modeling at a monthly time step" or similar, since we also use the period totals for other purposes.	Tighi, Shana G	This statement was added as suggested
A. Background	Reclamation disaggregates these period volume forecasts into monthly volumes.	Is there a methodology here as well that can be cited or otherwise referenced? Also, why are we disaggregating? Because that's the only forecast product we get?	Walker, Alexander E	The CBRFC website and forecast discussions explain the methodology. The aggregation is only for the 10th and 90th due to those being supplementary forecasts for the Min and Max 24MS, as opposed to the Most Probable. But per the recommendations, Reclamation will receive monthly 10s and 90s as well in the future.
A. Background	The Monthly Method is then used by the CBRFC for the 50th percentile forecast for the remainder of the first water year	I get why you want to say first water year, but it's the only one being discussed here and it's just kind of pops up. I think define that in the "background" section.	Walker, Alexander E	Added a note in background section to clarify this is the first year of the 24 MS.
A. Background	July	Did you mean to start with August here? You included July at the beginning of this paragraph with the spring runoff period.	Tighi, Shana G	Fixed to say August.
A. Background	The Period Method is used by the CBRFC for the 10th and 90th percentile water years forecasts, which Reclamation disaggregate to monthly volumes based on the CBRFC Monthly Method 50th percentile forecast's monthly distribution.	What time frame here? I'm not 100% tracking what forecast is for what. Could there be a table, a conceptual figure or a simplified matrix shown here?	Walker, Alexander E	Clarified to explain this is for April through July

Section	Text	Comment	Reviewer	Response
A. Background	It is unclear which method, Period or Monthly, provides more reliable and accurate forecasts	This needs to be either beginning or ending a paragraph not hiding in the middle of a sentence. It's one of the driving factors of the research!	Walker, Alexander E	Moved to be at the beginning of the paragraph.
A. Background	The analysis includes an assessment of both spring runoff volume forecasts (for April-July) being issued starting in January 1, and water year forecasts issued (for October- through September), issued year-round.	Again, this terminology is confusing for someone who isn't steeped in the different forecasts. I think it requires a discussion of how this information wants to be conveyed, then referring to consistently throughout the document. I think that referring to it as the April-July forecast and the Water Year forecast is the easiest way to structure it.	Walker, Alexander E	Removed the Water Year analysis as that's not a forecast used by CBRFC or Reclamation
A. Background	Figure 1: Example of the spring runoff water supply forecast for Lake Powell showing raw and official monthly 10th, 50th, and 90th percentile volumes. This can be found at cbrfc.noaa.gov/wsup.	This will need greater explanation for the viewer of what they are seeing with particular care to emphasize that the data shown here are forecasts through time. In addition, caption should have the day the figure was accessed/downloaded from CBRFC website.	Walker, Alexander E	Caption updated to better explain the figure
A. Background	Figure 1: Example of the spring runoff water supply forecast for Lake Powell showing raw and official monthly 10th, 50th, and 90th percentile	I'm not 100% sure this need to be included.	Walker, Alexander E	Date accessed is not needed as it's an internal document

Section	Text	Comment	Reviewer	Response
	volumes. This can be found at cbrfc.noaa.gov/wsup.			
B. Approach and Data	forecasts	Question - are these forecasts you evaluated from the raw ESP? Or are the historical operational (adjusted) forecasts of the 10,50,90 (aka Min/Most/Max)?	Tighi, Shana G	Added sentence to clarify that they are the raw ESP forecasts without adjustments from forecasters
B. Approach and Data	twelve	There are only 11 entries in table 1	Tighi, Shana G	Edited to say 11
B. Approach and Data	Forecasts for years 1991-2020 (WY 1992-2020) were evaluated against the observed volumes	Do forecasts go back any farther? I would be curious if a larger sample size would produce different results.	Walker, Alexander E	The forecasts only go back to 91 due to the recalibration of the model to match with the 91-20 forcings
B. Approach and Data	Forecasts for years 1991-2020 (WY 1992-2020) were evaluated against the observed volumes.	Include total number of data points (n) here to give the read an idea of how many observations are being consider for each column or boxplot.	Walker, Alexander E	Clarified that it's 30 years of forecast observations for 7 issuance months.
B. Approach and Data	Reliability, bias, and root mean square error were calculated. A Mann- Whitney U testU was computed to understand if the errors were significantly different between the Period and Monthly methods.	Reasoning for each of these tests should be explained here in the methods section.	Walker, Alexander E	A sentence was added to state that further explanation of the metrics is included in a section below.
1. Monthly versus Period Method	Method	This paragraph could benefit from a bit of clarification about where you are talking about Apr-July period total forecasts, individual monthly forecasts, and WY period total forecasts. Also might help to clarify that	Tighi, Shana G	Clarified this paragraph and removed the WY portion.

Section	Text	Comment	Reviewer	Response
		are indeed talking about forecasts and not forecast errors. It took me a bit of scrolling back and forth between the graphic and re-reading the paragraph a couple of times for me to fully tease out what you are talking about here.		
1. Monthly versus Period Method	, blue bars for Monthly Method, and yellow bars for Period Method). This is due to the Monthly Method assuming that the driest and wettest volumes happen sequentially for each month.	Sequentially throughout the calendar year or through the month in the data set? Not sure which is supposed to be sequential. Can you more clearly define the Monthly Method expectations?	Walker, Alexander E	It doesn't matter really, just that it assumes a driest April, driest May, driest June, etc. has equal chance of happening but this section has been better clarified.
1. Monthly versus Period Method	Figure 2: Comparison of April -- July Spring runoff percentile volumes for the Monthly (blue) and Period (yellow) Methods. The black line is the median or 50th percentile (median), and the numbers on the top and bottom of each bar represent the 10th and 90th percentile volumes. This is for one site (BMDC2) with an April issuance month, but other sites and	This is a really good figure, in fact, I think that it and Figure 3 form the core of why this analysis was performed. I recommend that it be done for GLCA3 and clearly state the n for each plot is 12	Walker, Alexander E	Updated the figure to show all 30 years but kept it as BMDC2 as the analysis was done for all sites not just GLDA3.

Section	Text	Comment	Reviewer	Response
	issuance months show similar patterns.			
1. Reliability	Reliability	Since reliability is a specific term in this context , it needs to be defined, possibly with a reference.	Walker, Alexander E	Similar to Year 2 analysis, a section on metrics and expected outcomes will be added to this report.
1. Reliability	Reliability	Additionally, this is the first analysis point, so it could be considered to be the most important - is reliability the most important metric to assess here? Over bias?	Walker, Alexander E	Ordering is not a reflection of importance. Both metrics are equally important and provide different information.
1. Reliability	Reliability	The Wiki page for reliability states that in stats, it doesn't imply validity. Is this true here? If so, why do we know that the results here are valid for our purposes?	Walker, Alexander E	We can assume the results are valid because they overall behave as expected (high positive bias for the 90s, high negative bias for the 10s, and near 0 bias for the 50s). The bias measurement, or error, checks the validity of the results since reliability does not.
1. Reliability	then the forecasts are biased towards over-forecaststst.	Suggest putting the conclusion more directly in terms of reliability. As written, the section introduces reliability, already a somewhat ambiguous concept, but then shifts to discussing bias, which is addressed separately in a later section. Clarifying this distinction up front would help the reader follow the intent. It would also be helpful to provide a clear definition of "reliability," along with an explanation of why it matters. In many contexts, users are primarily concerned with forecast accuracy and minimizing forecast error. It may not be immediately clear to water managers why the distributional properties of the forecasts are important if the forecast magnitude itself is incorrect. While this is a technical paper, many users of inflow	Tighi, Shana G	Re-wrote this section to better explain what reliability is and focus on its interpretation more rather than the bias.

Section	Text	Comment	Reviewer	Response
		forecasts (and the 24MS more broadly) are not academics. They are focused on practical decision-making and on having forecasts that support operational needs. Providing some practical context on why forecast reliability has real-world implications and why managers, operators, and stakeholders should care about it would strengthen the discussion.		
1. Reliability	then the forecasts are biased towards over-forecaststst.	Regarding a more precise definition of reliability, I just returned from the Year 2 memo, where I added a comment on the definition of reliability that I'll copy here, in case it is useful for clarification: Reliability is the percentage of years the observation is <= the forecast. If reliability exceeds the method's expected percentile (e.g., >90% for a 90th percentile), the method is over-forecasting that percentile; if it is below, the method is under-forecasting. I like some of the language used with Reliability in the Year 2 paper. For example, this is helpful (from the Metrics section on page 8): "The reliability calculation determines whether the observations fall below the forecasts the expected number of times and is useful for assessing the performance of the wet and dry scenario". Also I like the language used in the Expected Results section on page 9.	Tighi, Shana G	Section re-written for better definition of reliability and explanation of expected results
1. Reliability	(red x)	We expect the observation (red x) to be above the 90th percentile 10% of the time and below the 10th percentile 10% of the time. It appears with this chart that you are arguing that the observation should ideally fall between the 10th and	Tighi, Shana G	Added more description to better clarify the expected results for reliability.

Section	Text	Comment	Reviewer	Response
		90th percentile forecasts 100% of the time? What am I misunderstanding? What clarification needs to be added so I can read this the way you intend?		
1. Reliability	Figure 3: April - July volume forecasts for each year for BMDC2 comparing monthly method (blue) to Period Method (yellow) against the observations (red x).	This is also a really good plot, showing the issue	Walker, Alexander E	Thanks!
1. Reliability	Figure 3: April - July volume forecasts for each year for BMDC2 comparing monthly method (blue) to Period Method (yellow) against the observations (red x).	Recommend moving to GLDA3 and also renaming Y axis to “thousand acre-feet (kaf)”	Walker, Alexander E	Changed the figure to say thousand-acre ft
(1) 10th Percentile	left panel	This is where calling it figure 4a would come in handy	Walker, Alexander E	Added labels a, b, and c to the plot
(1) 10th Percentile	Across all sites, from the corresponding range for the period method is 20 to 57	Pretty significantly lower range for all sites vs Lake Powell - meaningful?	Walker, Alexander E	Because this is unregulated flow (no upstream reservoirs present) and because it's a larger body of water that is the integration of all the upstream sites, the errors from the other sites are also integrated.
(1) 10th Percentile	As forecast lead time decreases, the reliability of both methods decreases, which is expected	So they get less consistent? Going to need more discussion here. I would also say try to eliminate the sentence structure of “X is happening, which is expected because Y” and instead just state the expectations going into the analysis section based on	Walker, Alexander E	Added explanation of why results did or did not match the expectations.

Section	Text	Comment	Reviewer	Response
	given that the forecasted range of volumes narrows through the end of the spring runoff	the forecast structure, then these paragraphs can discuss why it diverged or matched expected		
(1) 10th Percentile	As forecast lead time decreases, the reliability of both methods decreases, which is expected given that the forecasted range of volumes narrows through the end of the spring runoff.	I think this needs to go in the general reliability paragraph and also needs to be explained why this is expected.	Walker, Alexander E	Stayed in the analysis section but with improved clarification.
(1) 10th Percentile	Both methods exhibit a high bias, especially at Lake Powell, indicating that both models tend to over-forecast low values, though the Period Method over-forecasts more frequently.	Should this go in the Bias discussion?	Walker, Alexander E	Reliability also shows bias, just not the magnitude of it.
(2) 50th Percentile	For Lake Powell, the Monthly Method indicates that 83% to 97% of observations fall below the forecasted 50th percentile (Figure 4, middle panel, blue bars). Across all sites,	What would the expectation be here? 50% of observations? These numbers are hard to parse without knowing what is to be expected.	Walker, Alexander E	Added explanation of what is expected.

Section	Text	Comment	Reviewer	Response
	the observations fall below 50th percentile 49 to 69% of the time. The Period Method has roughly the same reliability, with observations falling below the 50th percentile 73% to 97% of the time for Lake Powell (Figure 4, middle panel, yellow bars). Across all sites, the observations fall below the 50th percentile volume 54% to 67%.			
(3) 90th Percentile	At Lake Powell, both methods tend to over-forecast, whereas across all sites, the Monthly Method over-forecasts and the Period Method under-forecasts the higher volumes.	I need to come back and digest this a bit more It seems to me that both methods tend too high (Monthly method has 93-100% under vs 90% under and Period method has 97-100% under vs 90% under). Although Figure 3 seems to indicate that the actual falls under the 90th percentile 100% of the time for the Monthly method and exceeds the 90th percentile twice (1999 & 2019) for the periodic method. Doesn't that Figure indicate that the Periodic method is more reliable than the Monthly method for the 90th percentile? What am I misunderstanding? Why does this seem to be a contradiction?	Tighi, Shana G	This section was re-written to provide more clarity.
(3) 90th Percentile	At Lake Powell, both methods tend to over-forecast, whereas across all sites, the	Is this a reliability metric? I think it would be best to have a discussion of where the method over and under forecast, then follow that with the Reliability/Bias/RMSE discussion	Walker, Alexander E	Kept the explanation in the same place.

Section	Text	Comment	Reviewer	Response
	Monthly Method over-forecasts and the Period Method under-forecasts the higher volumes.			
(3) 90th Percentile	Figure 4:	For all 3-panel plots, please split them in a), b) and c) - it'll help when referencing them	Walker, Alexander E	Added a, b, and, c labels
(1) 10th Percentile	For the 10th percentile at Lake Powell, the Monthly Method indicates 27% to 87% of observations fall below the forecasted 10th percentile (Figure 5, left panel, pink bars).	Does this mean that up to 87% of observations fall below the 10th percentile for a given month? That seems pretty bad! Is there a reason why that is occurring? Is that reason significant for forecasts or operations? Need more analysis of what we are seeing.	Walker, Alexander E	Removed this section (Water Year analysis), however, it is still problematic which is why further exploration of the bias is ongoing work. Additionally, this is important because it suggests that raw ESPs are currently "missing" any low values that may happen and limit long term planning for operations.
(2) 50th Percentile	50th Percentile	I would try to make it very clear that this means the tendency for Lake Powell is to over-forecast, like a lot, in both methods.	Walker, Alexander E	Added statement that there is a tendency to over-forecast and mentioned that addressing the bias is part of future and ongoing work. Additionally, Powell is exceptionally worse than the other sites because it is integrating the error from the upstream sites. By addressing the bias in the underlying model, reliability should be improved in the future.

Section	Text	Comment	Reviewer	Response
(3) 90th percentile	the Period Method tends to under-forecast high values with observations being below the forecasted 90th percentile 84% to 91% of the time.	I think there is something backwards in this 90th percentile (10% probability of exceedance) approach/concept. Or maybe there are just a few typos confusing me. The observation should fall BELOW the 90th percentile 90% of the time and above the 90th percentile 10% of the time. If the observation falls Below the 90th percentile 84-91% of the time, that 91% is actually pretty decent, and the 84% is actually a tiny bit low since it implies that the observation exceeds the 90th percentile 16% of the time rather than 10% of the time. Also, you also wrote above that the forecasts are "greater than" the observations 87-97% of the time. Did you mean to write less than instead of greater than? Do you also have typos in the numbers in this paragraph? You wrote that both the Monthly method and Period Method have observations below (greater than?!?!?) the 90th percentile 87-97% of the time. That implies that the two methods are equally reliable.	Tighi, Shana G	Removed water year analysis section but also clarified the reliability sections for better understanding. However, this was a typo as pointed out by the reviewer. But the two methods were similarly reliable for the 90^{th} percentile in the water year.
2. Bias	Bias	Does this section need to be included? What does it tell us about the outcomes? I'm having trouble tracking how it's needed to fully explain the story.	Walker, Alexander E	Yes, Bias is an important indicator of the direction and magnitude of the errors the forecasts have compared to the observed values. Knowing what direction and how much the error is lets us better understand how accurate the models are. Explanation of direction and magnitude also was added to the reports for explanation. Additionally, a sentence was added in the Next Steps section that says the high bias will be addressed in future work.

Section	Text	Comment	Reviewer	Response
2. Bias	Bias for the 10th percentile forecasts are expected to be negative since the forecast represents low volume conditions. Similarly, bias for the 90th percentile methods are expected to be positive since the forecast represents high volume conditions.	I recommend a detailed discussion of what are the expected outcome at the 10th percentile and the 90th percentile, then place those discussions into the "Approach and Data" section. The reader needs to be prepared for which outcomes are to be expected. Then, this section can focus on biases that are notable or where the magnitude is greater.	Walker, Alexander E	Added more explanation in the data and methods section explaining what is expected from the different metrics.
a) April – July	has a positive median bias at shorter leads (Figure 6, left panel, yellow bars).	Is this bad? Previous paragraph says it is expected to be negative - what does it mean to be positive? What is the reader supposed to take away from these plots?	Walker, Alexander E	Better explained the results.
a) April – July	Figure 6	If I am reading this right, the median bias value is very similar for May in both 10th and 50th. Are these data sets using the same observed volumes? Does this mean that the forecast tends to overestimate for both? Seems like that the bias flipping positive is an undesirable outcome	Walker, Alexander E	The biases are different and are using the same observations. Also it was explained that the 50th are supposed to be similar. The 10th will get more similar for both methods as the range in values decreases. An explanation about the decreased ranges is added to the text throughout

Section	Text	Comment	Reviewer	Response
a) April – July	kaf) for Lake Powell (GLDA3) volume by lead month of forecasted percentile volume minus the observed (actual) volume for April through July. Blue is for Monthly Method volumes; gold is for Period Method volumes. The left column is 10th percentile, middle is 50th percentile, and right is the 90th percentile.	These plots also seem to show that the forecasting tools are good at the 90th percentile (the high end), but far worse on the low end and they don't improve as dramatically with reduced lead time. This has significant implications for interpreting the forecasts since this data suggests that for the 10th percentile forecast, you cannot expect the level of reduced bias with a decreased lead time that is seen at the 90th percentile. That, coupled with the reliability metrics, implies to me that theres a large amount of uncertainty in lower range forecasts and we are missing that lower end quite a bit. Do I have that right? If so, it should be discussed more intensely. For the 50th percentile, the persistent positive bias is also concerning, esp since the IQR does not cover the 0 bias value. Finally, the Bias is listed in KAF, but 2500*1000 is 2.5 million acre-feet. Is that the intended range, that forecasts in the 50th percentile have biases >1.0 MAF? Or, does the scale and labels need to be adjusted?	Walker, Alexander E	Better elaborated on how the methods are overall more positively biased as shown by the 50th percentile and elaborated that this isn't the case for all sites, but worse at Lake Powell. The >1 MAF is difference is entirely possible. Just this year (WY 2026) the forecasts for the 50th have already changed by more than 2 MAF. That is the intended range, and the scale and labels do not need adjustment.
b) Water Year	Figure 7	For this, it feels like it would be good to focus some discussion on May of the 50th percentile. Considering the known information (snowpack, etc), seems like its' the biggest takeaway that it has a tight spread (high precision), but overestimates in all data points (low accuracy)	Walker, Alexander E	Removed water year analysis
b) Water Year	Figure 7:	This WY forecast is a lot better, but the same comments apply here as in Figure 6.	Walker, Alexander E	removed water year analysis
3. Root Mean	Root Mean Square Error	I think the discussion in this section is explained by the Bias section, since the RMSE seems to track Bias. I	Walker, Alexander E	RMSE section has been removed

Section	Text	Comment	Reviewer	Response
Square Error		think this is a candidate for elimination unless the RMSE section is illuminating something that isn't shown by the Bias analysis.		
3. Root Mean Square Error	Root Mean Square Error	RMSE and BIAS could be collapsed to a single section as well with all these values displayed in a table.	Walker, Alexander E	RMSE section has been removed
3. Root Mean Square Error	Root mean square error (RMSE) compares the accuracy of the two methods to observations. RMSE was calculated for the 10th, 50th, and 90th percentile forecast for each lead time and location.	Helpful to explain that Bias is a direction of error and RMSE is a magnitude of error, in that a value can have low bias and a higher RMSE, vice versa, etc.	Walker, Alexander E	RMSE section has been removed
a) April – July	Figure 8:	Use of the Month on the X-axis: Is this data a time series or, is it the lead time of the forecast? If lead time, I think it would be better to Have the X-Axis be "Lead Time" and the x values be month of lead time to clearly delineate when you are talking about forecast lead time. The plot would also benefit from Capitalizing all the legend items and putting a box around the plot area. Also, would points work better than lines?	Walker, Alexander E	RMSE section has been removed
b) Water Year	Figure 9	See comments on Figure 8	Walker, Alexander E	RMSE section has been removed
4. Mann-Whitney U test	two independent populations.	Citation here would be ideal. Here's a possible one: https://pmc.ncbi.nlm.nih.gov/articles/PMC12366075/	Walker, Alexander E	Internal document so not needed.

Section	Text	Comment	Reviewer	Response
4. Mann-Whitney U test	two independent populations.	https://www.jstor.org/stable/2236101%20	Walker, Alexander E	Internal document so not needed.
4. Mann-Whitney U test	two independent populations.	I think if you want a detailed explanation of p-value, include it, but the test is in service of determining if the biases are different and as such, you can focus on the results - someone who wants to know more can look up the citation.	Walker, Alexander E	I kept the explanation but moved into the Data and Methods section instead.
4. Mann-Whitney U test	A Mann-Whitney U test was used to determine if the biases for either method were significantly different from each other, we employed the. The Mann-Whitney U test, a nonparametric test that checks the difference between two independent populations. is a nonparametric statistical test used to determine if two populations have the same distribution.	Would be helpful to say why this test was performed. What does it tell us in practical terms and how will it help us come up with a conclusion or recommendation?	Tighi, Shana G	Added more explanation as to why this metric was used.
4. Mann-Whitney U test	BAs shown in Table 2, the biases for the 10th percentile are more likely to be significantly different than the other percentiles. Additionally, the biases	Need to explain why these results are important, beyond simple describing what is in the table. Is something unexpected happening? This seems to be showing that the methods operated differently at the extremes. Is that to be expected based on how the methods aggregate data? Seems to show that the choice of method matters.	Walker, Alexander E	Added and explanation.

Section	Text	Comment	Reviewer	Response
	at longer lead times tend to be more different than shorter lead times, but not in a consistent manner. The 90th percentile bias is significantly different between both methods for the months of February, April, May, and June. The 50th percentile bias distributions are significantly different.			
4. Mann-Whitney U test	The 90th percentile bias is significantly different between both methods for the months of February, April, May, and June. The 50th percentile bias distributions are significantly different.	Seems notable the methods differ significantly during the peak runoff months - is these an explanation why? Is it due to the handling of extremes in different methods?	Walker, Alexander E	Further elaborated this.
4. Mann-Whitney U test	Table 2:	Helpful also to explain why the p-values would be different between A-J and WY volumes - different sample populations?	Walker, Alexander E	Removed WY results so no longer relevant.
4. Mann-Whitney U test	0.21	Explanation for why this is the case?	Walker, Alexander E	I don't have a good explanation as why this is the case.
4. Mann-Whitney U test	0.01	Should this value also be highlighted?	Tighi, Shana G	Removed WY results so no longer relevant.

Section	Text	Comment	Reviewer	Response
E. Summary	accurate	However, they are both still inaccurate, looking at the Bias plots? Do I have that correct?	Walker, Alexander E	There is some inaccuracy as the error is biased in the positive direction. Because the error is presented as a boxplot with multiple error values there is not a way to definitively say all values are accurate or inaccurate.
E. Summary	but the Monthly Method is more reliable and accurate for forecasting volumes in dry and wet years.	Based on the graphics presented, and the background information provided, the connection to the conclusion is not entirely clear. The figures show that, at the 90th percentile, the Monthly method exhibits higher bias and higher RMSE, which might intuitively suggest it is less accurate for forecasting volumes. The RMSE at the 10th percentile is also higher for the Monthly method until May. This is all very unintuitive because typically we want to minimize errors and bias. I think it would make your conclusions a lot more intuitive and tell a clearer story if you explained the following (earlier in the doc when explaining the analysis results is appropriate, or here to tie the conclusions to the concepts): Larger errors in the 10-90 are not a sign of poor accuracy in longer lead times. They reflect uncertainty increasing as lead time increases. When you forecast further into the future (ex, January for the A-J, or October for the WY), the spread between the upper and lower percentiles should widen. So it is desirable and expected that the errors at the 10-90 extremes should get bigger. Bias tends to be a concern in the error metrics for the 50th percentile, but it is a different kind of concern for the 10-90. When discussing whether the model is biased in the 10th/90th percentiles, we're not looking	Tighi, Shana G	In the methods section added writing to explain what the expectations are for the results and elaborated in the analysis section why the models are or are not performing as expected and added explanation in the summary why the Monthly Method is recommended over the Period method to better address these points.

Section	Text	Comment	Reviewer	Response
		for small errors close to 0, but rather for whether observations fall below or above those percentiles with the correct frequency - thus we use the reliability analysis to help us understand bias. It appears that the underlying objective is to use a method that produces a wider range of uncertainty (higher highs and lower lows), and particularly lower forecasts for the 10th percentile, which the Monthly method achieves. Choosing a method that results in lower lows is a reasonable and defensible choice, especially if the resulting forecasts are shown to be less biased and more accurate. That being said, I remain uncertain (pun intended) that the Monthly method is demonstrably more reliable for forecasting inflow volumes during wet years (90th percentile). Once the inconsistencies between the figures and text (particularly Figure 3 vs the Reliability section) above are addressed, I will be better able to assess that conclusion. At this stage, I am concerned that the Monthly method may produce 90th percentile inflow forecasts that are unrealistically high, to the point that they offer limited value for operational decision-making.		
E. Summary	Additionally, as lead times decrease, the accuracy of the forecasts increases for both water year and spring runoff volumes as the range in possible volumes decreases.	I'm not sure this is fully explained in the analysis section - I certainly felt like it wasn't supported by any of the earlier text.	Walker, Alexander E	Added explanation for this in the Analysis Section

Section	Text	Comment	Reviewer	Response
E. Summary	However, the reliability of both methods decreased as lead times decreased for the 10th percentile forecasts (Figures 4 and 5, left panels).	Really really need some interpretation here as I don't have the background to assess the import of this occurring	Walker, Alexander E	Added explanation for this in the Analysis section
E. Summary	over-forecast all percentile values (Figures 4 and 5)	This seems way more notable to stakeholders and operators than reliability. Is this not a big deal because it's well known? If it is considered to be well known, then it should be covered in the background section.	Walker, Alexander E	It's a little unclear what the reviewer means here but the rewritten analysis and summary sections better explain why reliability is important. Additionally, a sentence was added in the next steps section that says the high bias will be addressed in future work as it also is important along with reliability.

Section	Text	Comment	Reviewer	Response
E. Summary	but the Monthly Method is more reliable and accurate for forecasting volumes in dry and wet years.	Based on the graphics presented, and the background information provided, the connection to the conclusion is not entirely clear. The figures show that, at the 90th percentile, the Monthly method exhibits higher bias and higher RMSE, which might intuitively suggest it is less accurate for forecasting volumes. The RMSE at the 10th percentile is also higher for the Monthly method until May. This is all very unintuitive because typically we want to minimize errors and bias. I think it would make your conclusions a lot more intuitive and tell a clearer story if you explained the following (earlier in the doc when explaining the analysis results is appropriate, or here to tie the conclusions to the concepts): Larger errors in the 10-90 are not a sign of poor accuracy in longer lead times. They reflect uncertainty increasing as lead time increases. When you forecast further into the future (ex, January for the A-J, or October for the WY), the spread between the upper and lower percentiles should widen. So it is desirable and expected that the errors at the 10-90 extremes should get bigger. Bias tends to be a concern in the error metrics for the 50th percentile, but it is a different kind of concern for the 10-90. When discussing whether the model is biased in the 10th/90th percentiles, we're not looking for small errors close to 0, but rather for whether observations fall below or above those percentiles with the correct frequency - thus we use the reliability analysis to help us understand bias.	Tighi, Shana G	Addressed the inconsistencies pointed out by the reviewer and cleaned up the writing to better explain the expected results, why the models are or are not matching those expectations, and what the interpretations are.

Section	Text	Comment	Reviewer	Response
		It appears that the underlying objective is to use a method that produces a wider range of uncertainty (higher highs and lower lows), and particularly lower forecasts for the 10th percentile, which the Monthly method achieves. Choosing a method that results in lower lows is a reasonable and defensible choice, especially if the resulting forecasts are shown to be less biased and more accurate. That being said, I remain uncertain (pun intended) that the Monthly method is demonstrably more reliable for forecasting inflow volumes during wet years (90th percentile). Once the inconsistencies between the figures and text (particularly Figure 3 vs the Reliability section) above are addressed, I will be better able to assess that conclusion. At this stage, I am concerned that the Monthly method may produce 90th percentile inflow forecasts that are unrealistically high, to the point that they offer limited value for operational decision-making.		
1. Next Steps	Additionally, both methods have a higher than expected bias when forecasting dry conditions and tend to over-forecast total volumes in both the spring runoff season volume and water year total volume. Given that both methods use the same inputs, this	This reads more like a summary conclusion than a next step.	Tighi, Shana G	Separated the Next Steps and Summary sections and made clearer what the actual next steps are.

Section	Text	Comment	Reviewer	Response
	suggests that the model used by CBRFC water supply model might be biased towards over-forecasting volumes.			
1. Next Steps	This could be explored in a handful of ways: including looking at impacts of initial soil moisture and snowpack conditions used as model inputs, as well as comparisons of comparing historical weather, and sensitivity analyses of the driest and wettest years to each other to understand how sensitive of historical climatology and initial conditions are toto outputs in the model.	Is the objective of this "exploration" and "understanding" to remove the bias towards higher volume forecasts? This section would be strengthened by clarifying the operational goals/objectives of the next steps. Outlining operational goals also helps ensure that the work remains grounded in real-world application rather than academic framing. Also, what happens if the CBRFC forecast is "fixed" so it is no longer biased towards higher forecasts? Would you repeat this analysis to make sure the Monthly Method is still reliable, or if it vastly over/under forecasts the 10-90 tails of the distribution? Might be worth considering.	Tighi, Shana G	Removed this analysis as it did not contribute to the overall analysis but rather provided a possible explanation for the bias the model is seeing. It was more of an academic exercise and not directly related to operations.
1. Next Steps	Table 3: 30-year average and median April through July volumes, as well as Skewness for all sites. The positive skew indicates that forecasts have more outliers that are large	This is analysis and should go in the analysis section.	Walker, Alexander E	Removed this analysis

Section	Text	Comment	Reviewer	Response
	than small, confirming the forecasts are biased towards wet volumes.			

Table 2: Peer Review comments and responses to 24MS_Year2 Report

Section	Text	Comment	Reviewer	Response to reviewers
I. Comparing Streamflow Forecast Methods for Year 2 of the 24-Month Studies	Comparing Streamflow Forecast Methods for Year 2 of the 24-Month Studies	I feel like this could be one report split into Year One and Year Two sections.	Walker, Alexander E	The reports will remain split as they include different analyses and years for the 24MS, and Year 1 report is more directly related to CBRFC forecasts and methodology, while Year 2 is directly related to Reclamation forecasts and methodology.
A. Background and Objectives	Background and Objectives	I like the background and objective here much better than the Year One draft, but this draft has very similar issues to the Year One draft, where it's not fully clear to the reader what they are looking at, what the primary findings are, and what the recommendations resulting from this work. I recommend determining the primary conclusion on the report and working to make the text of the document and the analyses support that primary conclusion as clear as possible. Additionally, some of the language varies between the two documents. Please make sure to align language between the two sections. Particularly the use of terms "error" and "bias"	Walker, Alexander E	Edits have been made to make it clear what the primary findings are and more clearly state the recommendations. Additionally, addressed the language differences throughout both documents.

Section	Text	Comment	Reviewer	Response to reviewers
A. Background and Objectives	Future conditions for streamflow and reservoir operations are forecasted using separate modeling systems – the CBRFC's streamflow forecast model and Reclamation's CRMMS reservoir operations. The CBRFC's modeling system uses temperature, precipitation, soil moisture, and snowpack conditions to forecast streamflow. Outputs from this model are then used in CRMMS, which simulates reservoir operations in accordance with the "Law of the River," a framework of laws, agreements, and treaties governing management of the Colorado River.	Some sort of flow chart or diagram would be helpful for conceptualizing this relationship here.	Walker, Alexander E	Authors feel that the paragraph is clear enough as it's explaining that hydrology forecasts are developed by CBRFC and those outputs are then used in CRMMS model for operations and that a diagram would not add to the comprehension.
A. Background	greater correlation coefficient	Perhaps explain what this is? What is correlated with what in this statement?	Tighi, Shana G	Explained that it's correlated with observations

Section	Text	Comment	Reviewer	Response to reviewers
and Objectives				
A. Background and Objectives	A study done by S. Baker in 2019 revealed that the 50th percentile ESP method outperformed the Most Probable method and had a greater correlation coefficient for longer lead times compared to using the historical 50th percentile from climatology.	Need reference	Walker, Alexander E	Internal report that was not publicly published doesn't need a reference.
A. Background and Objectives	Recently	In 2021? Can you put year of updating here?	Walker, Alexander E	Updated with the year
B. Approach and Data	2020	Was it recalibrated in 2020? I thought the recalibration occurred in 2021 including 2020 data, and was implemented in time for WY 2022. Might need to double check on this.	Tighi, Shana G	Edited this statement to make clear it includes 2020 data and was implemented in WY 2022
B. Approach and Data	greater uncertainty	I don't intuitively follow why this is good. Is there a different way to phrase this? Perhaps, "better encompasses uncertainty"	Walker, Alexander E	Changed the language to say it encompasses a wider range of possible conditions
B. Approach and Data	This study will be divided into an analysis of methods that represent wet scenarios, moderate scenarios, and dry	Language here seems different than in the year one report. I don't remember Year One talking about wet, moderate and dry. Please make the language are consistent across both.	Walker, Alexander E	Spring runoff specifically looks at 10th, 50th, and 90th percentiles. However, the Probable Maximum and Probable Minimum do not use just a single percentile, so the language needs to be slightly different for the purposes of these reports

Section	Text	Comment	Reviewer	Response to reviewers
	scenarios, which correspond to the Maximum, Most, and Minimum Probable 24MSs, respectively			
1. Ensemble Streamflow Prediction Percentile Hindcasts	then forced	Unclear terminology here for those of us less familiar with modeling	Walker, Alexander E	Changed to say "then fed with a set of time series"
1. Ensemble Streamflow Prediction Percentile Hindcasts	The 10th, 50th, and 90th percentiles are then calculated and analyzed here as ESP_10, ESP_50, and ESP_90, respectively. The ESP traces are developed by calculating the 10th, 50th, and 90th percentiles for each month for the 24 months of interest, based on a previous study analyzing the water supply forecasts issued by the CBRFC.	The sentences appear to be talking about the same percentiles with different methods? Please edit for clarity.	Walker, Alexander E	Rewrote the sentences for clarity.
3. Probable Minimum Hindcasts	For the hindcast formulation, ESP 50th percentiles are used in place of the RFC official forecasts (blue	Is this using the official forecast or the ESP 50th percentile? I'm a little confused here. I think for hindcasts, best thing here would be to change RFC to 50th %ile ESP on the charts in this report	Walker, Alexander E	These are the official matrices used by Reclamation so that's why they are here with the "official" RFC forecasts. However, the text has made it clear for the purpose of the study we have to use the raw hindcasts. But

Section	Text	Comment	Reviewer	Response to reviewers
	boxes labeled as "RFC").			the official forecasts should be very similar to the raw
3. Probable Minimum Hindcasts	volumes	The word "volumes" is repeated from earlier in the sentence: "to get the volumes for the Probable Minimum August and September volumes"	Tighi, Shana G	Fixed
5. Forecast Locations	Forecast Locations	Like in the year one report, the split between analyzing all spots and the Lake Powell focus makes the report feel somewhat incomplete. I think that if it were a journal article, I'd recommend just focusing on Lake Powell. Since it's a report I think that the results need to be referenced, in a table, appendix or elsewhere in the report, but the main body can be streamlined to just look at Lake Powell.	Walker, Alexander E	The analysis needed to be done for all but since it's just an internal document it can stay focused on Lake Powell. However, all sites will be part of the changes and results are available upon request. Results and reports will be stored for archival purposes as well.
5. Forecast Locations	12 forecast locations are listed below (Table 2)	You listed 11 locations in table 2 below.	Tighi, Shana G	Fixed to list all 12
5. Forecast Locations	Table 2: Forecast location names used for this study. Note that only Lake Powell is being presented but the streamflow input traces were created for all the locations listed	Like the year one draft, there is only 11 locations presented here	Walker, Alexander E	Fixed to list all 12

Section	Text	Comment	Reviewer	Response to reviewers
6. ESP vs 24MS in the out-year	historical streamflow	Natural, unregulated or simply observed (ie, including reservoir operations)? I think that's a good distinction to make here.	Walker, Alexander E	Historical unregulated observations (not gage flow). This is now clarified in the document.
C. Metrics	Correlation assesses how strongly the forecasts align with the observations, which we expect to align only in moderate scenario methods and are therefore omitted from the other scenarios.	Not run in year one analyses? Why?	Walker, Alexander E	Because that was just for A-J spring runoff which add the months together. Additionally, the correlation coefficient was calculated for all 24 months so year 1 was done and shows up in this analysis.
C. Metrics	The reliability calculation determines whether the observations fall below the forecasts the expected number of times and is useful for assessing the performance of the wet and dry scenario.	It would be helpful to have a more precise definition of reliability in this paper. For example: Reliability is the percentage of years the observation is <= the forecast. If reliability exceeds the method's expected percentile (ex. >90% for the 90th percentile), the method is over-forecasting that percentile. If it is below, the method is under-forecasting. You have some of this in the Expected Results section, but it still leaves out the "what does it mean if it falls above or below the percentile/expected value?" and more importantly, "Why do we care? Why is this important?"	Tighi, Shana G	Used the same wording in the updated Year 1 report here to talk about reliability and explain why it's useful for the 10th and 90th percentiles, especially.
2. Expected Results	Error can be interpreted based on scenario-specific expectations. Under moderate scenario methods, we expect little to no bias	The reference to Bias occurs immediately without discussion - I would use it instead of "Error" in the metrics section. Make it clear for our less-statistically inclined readers!	Walker, Alexander E	in the Metrics section it says "error quantifies systematic bias" so it is discussed but elaborated to state that Bias gives the direction for error (positive or negative). This is described in the reports.

Section	Text	Comment	Reviewer	Response to reviewers
2. Expected Results	For example, in wet scenarios, the 24MS Probable Maximum method uses a 75th percentile of historical observations for the out-year. Thus, we expect the observation to fall below the forecast 75% of the time. For the wet scenarios for CRMMS-ESP and the ESP percentile (i.e., CRMMS_90 and ESP_90), we expect the observation to fall below the forecast 90% of the time.	This also helps to explain why a positive bias is expected in wet and negative expected in dry. I think a modified version of this should go in the error paragraph immediately above	Walker, Alexander E	Added an explanation in the error paragraph above explaining why a positive or negative bias is expected within the error calculations.
2. Expected Results	For example, one method may better forecast extreme conditions but exhibits a higher bias and lower accuracy compared to a method that doesn't capture extreme values as frequently.	In such a case, what is the priority? How do you choose a forecast method where one is better at some metrics and another is better at others?	Tighi, Shana G	Added a paragraph to explain why reliability should be the priority.
2. Expected Results	For example, one method may better forecast extreme conditions but exhibits	Perhaps discussed below, but what is the Reclamation priority? If there isn't one, good thing to discuss in the analysis section.	Walker, Alexander E	See comment above.

Section	Text	Comment	Reviewer	Response to reviewers
	a higher bias and lower accuracy compared to a method that doesn't capture extreme values as frequently.			
D. Analysis: Powell Unregulated Inflow	Analysis: Powell Unregulated Inflow	Different analyses here vs Year One analysis - is there a reason why? I would think they could be consistent across both documents	Walker, Alexander E	They serve slightly different purposes and the A-J forecast was just looking at streamflow. But the A-J gets disaggregated and thus is analyzed through RiverWare.
1. Moderate Scenarios	the deterministic streamflow forecasts	Not clear based on text if this refers to CRMMS, ESP or 24MS	Walker, Alexander E	All three and text to the report is updated to clarify.
1. Moderate Scenarios	quantile	"percentile"?	Tighi, Shana G	changed to percentile
1. Moderate Scenarios	quantile	"percentile"?	Tighi, Shana G	changed to percentile
1. Moderate Scenarios	out-year,	Any interest in being more specific? Months 13-16 look spot on, just months 17-24 are a little off. Not a critical change, but might be a "nice to" item.	Tighi, Shana G	Edited to be more specific.

Section	Text	Comment	Reviewer	Response to reviewers
1. Moderate Scenarios	indicating the slight under-forecasting of all three methods.	Because the concept of reliability can be somewhat abstract, it may be helpful to break it down. For example: The chart indicates that we expect the actual to fall below the 50th percentile 50% of the time, but for lead time 17-24 months, it falls below the 50th percentile ~45% of the time and falls above the 50th percentile forecast ~55% of the time. This aligns with earlier findings that all three methods tend to produce forecasts that are biased low, or under-forecast, at those lead times. Is it necessary to provide this level of clarifying detail in each section? While this breakdown is helpful for my own understanding, the overall approach of the paper seems to favor summarizing rather than presenting specific numbers. If this is indeed the preferred style, readers who wish to examine the details can perform their own mental breakdown as needed. I confess my own personal preference to add clarification and details in an attempt to avoid confusion.	Tighi, Shana G	Wording has been edited to be clearer about reliability and more consistent across results and reports.

Section	Text	Comment	Reviewer	Response to reviewers
2. Wet Scenarios	This suggests that the historical climatology likely under-forecasts wet conditions.	This is such a weird and unintuitive concept. Typically we want to minimize errors. But you are saying for the upper and lower percentiles we want the highest errors! We want to be more wrong in the out-year! We want to be more biased! We want to be more inaccurate! Perhaps it would help to explain this a bit more. You kind of touched on it a tiny bit in the Expected Results section, but I think your write-up would do a much better job with "analysis clearly explained and results properly interpreted" if you clarified explicitly the following concepts: Percentile forecasts are supposed to miss by design. A 90th percentile forecast should be too high about 90% of the time. A 10th percentile forecast should be too low about 90% of the time. In other words, "being wrong" is exactly how these bounds demonstrate that they're calibrated correctly. Larger absolute errors in the 10-90 are not a sign of poor accuracy in longer lead times. They reflect uncertainty increasing as lead time increases. When you forecast further into the future (i.e., into the out-year), the spread between the upper and lower percentiles should widen. So it is desirable and expected that the errors at the extremes should get bigger. "Bias" tends to be a concern for the 50th percentile, but it is a different kind of concern for the 10-90. When discussing whether the model is "biased low," we're typically talking about the median. But for the 10-90th	Tighi, Shana G	Added this to the expected results section and will be sure to elaborate more in the analysis section as well to help readers better grasp

Section	Text	Comment	Reviewer	Response to reviewers
		percentiles, we're not looking for "small errors", but rather for whether observations fall below or above those percentiles with the correct frequency - thus the reliability analysis.		

Section	Text	Comment	Reviewer	Response to reviewers
2. Wet Scenarios	A comparison of absolute errors in Figure 10 shows that ESP_90 has the widest range, and therefore, the greatest uncertainty in modeling wet conditions. The Probable Max method has the narrowest range and lowest median error, suggesting that it has the lowest uncertainty of the methods.	Similar to my comment above, I think it would make the write up so much more clear if you explained Why is one good/desirable and the other bad/undesirable. The concern is that the forecasts in the out year are too tight and do not fully represent what we might see in the out-year, so we want a method that better reflects the uncertainty and a more full range of realistic possible future conditions. Again, this may or may not be the best place for this content. Perhaps it belongs in the background sections. Regardless of where it goes, I think it is helpful to link this very general summary of results to why we care and what we are looking for in this analysis.	Tighi, Shana G	Elaborated more on expected results in the expected results section and in the summary and recommendations section.
2. Wet Scenarios	/	I wonder if it would be more clear to add a 2nd dotted line at the 75 marker for the Probable Max comparison? The description of the Figure below is not completely accurate, as the grey dashed line shows the expected reliability of only 2 of the 3 forecasts.	Tighi, Shana G	Will leave it as it is explained in the text it should be 75 for the second year.
2. Wet Scenarios	/	Although it switches to 90 for year 1 (<= 12 months) so maybe that would be confusing? Maybe just have the dotted line go from lead time 24 to 13?	Tighi, Shana G	See comment above.
3. Dry Scenarios	90	Should this be 10?	Tighi, Shana G	Fixed to be 10.
3. Dry Scenarios	Figure 14	Can you make sure to make it explicit in this figure caption which direction is positive and which direction is negative?	Walker, Alexander E	The figure gives percentages from 0 to 100 so there is no negative. As explained in the text, under the dashed line is under-forecasting. Over the dashed line is over-

Section	Text	Comment	Reviewer	Response to reviewers
				forecasting which is explained in the caption.
3. Dry Scenarios	/	As above for the wet conditions, perhaps it makes sense to add a dotted line at the 25 mark to indicate what the Probable Min ought to be projecting for the outyear?	Tighi, Shana G	Will leave it as it is explained in the text it should be 25 for the second year.
E. Analysis: Powell Pool Elevation	which is heavily impacted by Lake Powell release decisions governed by the 2007 Interim Guidelines.	It might be helpful to explain this in just a little more detail. For example: "Small differences in inflow forecasts can translate into larger elevation differences when tier thresholds trigger different release volumes." or something similar.	Tighi, Shana G	Elaborated on this
2. Wet Scenarios	However, the Probable Maximum method is much higher than its expected value compared to the other two methods, suggesting it is the least reliable for predicting wet conditions.	This would be particularly well illustrated if you added the horizonal dotted line at 75, since the Probable Max method is higher than even 90.	Tighi, Shana G	Leaving the figure as is but elaborated that the current 24 MS methods are at a different reliability.
3. Dry Scenarios	the Probable Minimum method under-forecasts (less than 25%) for the out-year.	Would be more clearly illustrated with a horizontal line at 25 in your figure	Tighi, Shana G	See comment above.

Section	Text	Comment	Reviewer	Response to reviewers
F. Summary	Summary	Question 2 of the Peer Review Plan is "Are the recommendations supported by the results of the analysis?" In the current draft, the recommendation is not explicitly stated. While I understand the intended recommendation based on our in-person discussions and presentations, it is only implied in the written document through references to the preferred results. Suggest stating the recommendation more directly.	Tighi, Shana G	Added a recommendations section at the end.
F. Summary	The ESP percentile-based forecast would provide a 10th and 90th percentile range that is more accurately reflects the 10th and 90th percentile risk given our 30-year historical record as shown in the reliability analyses.	The reliability plots seem to indicate that the ESP percentile functions more like a 5-100 percent interval rather than a reliable 10-90 percent interval. Similar to the Year 1 analysis, the Summary Results suggest an expectation/desire that the observed values should fall within the 10–90 forecast range nearly 100% of the time, rather than within that range 80% of the time as would typically be expected. If that's what you want, that's a valid choice, but it's hard to claim statistical reliability as a justification for that choice. In effect, the reliability analysis is being used to support adopting a forecast that is statistically miscalibrated/unreliable, but aligns with the preference for having observations fall within the 10–90 percentile bounds nearly 100% of the time.	Tighi, Shana G	When looking at Reliability metrics for the inflow at Lake Powell, we see that the ESP traces do the best for reliability. Additionally, these are raw hindcasts without any review by forecasters who may understand and adjust the forecasts as needed based on known biases. Further, when addressing reliability for Lake Powell Pool elevation (post operations), we see that the ESP_50 does the best over the other methods. This shows that the shift is actually a reflection of the operations model, not the forecasts themselves. Therefore, there is future research needing to be done to address the shift that happens in the operations model. We appreciate you pointing out this discrepancy as a future issue to address!