1. Does the paper address relevant scientific questions within the scope of HESS?

YES

2. Does the paper present novel concepts, ideas, tools, or data?

YES

3. Are substantial conclusions reached?

YES

4. Are the scientific methods and assumptions valid and clearly outlined?

YES, but the authors need to consider information criteria as an alternative to LOOCV for model selection. I think the data set could potentially support models with more than two predictors. Information criteria could identify if that is the case, but the authors chose to search only for two-parameter models. This potentially limits predictive ability and scientific understanding. And, the diagnostics of the final model need to be presented for the readers to fully assess its utility.

5. Are the results sufficient to support the interpretations and conclusions?

YES

6. Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)?

YES

7. Do the authors give proper credit to related work and clearly indicate their own new/original contribution?

YES

8. Does the title clearly reflect the contents of the paper?

YES

9. Does the abstract provide a concise and complete summary?

YES

10. Is the overall presentation well structured and clear?

NO. It is a chapter from a dissertation that has not been sufficiently reformatted to stand on its own and conform to standards of a journal article. The figures are not numbered in the

order in which they are referred to in the text. As a result, the methods and logic are harder to follow. Fewer figures with better organization would improve readability.

11. Is the language fluent and precise?

YES.

12. Are mathematical formulae, symbols, abbreviations, and units correctly defined and used?

YES, but they are a little clunky. Shorter abbreviations and variable names would make the paper easier to read and understand.

13. Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated?

YES. The number of figures could be reduced. In particular, figures 9-14 could be put into supplementary material or an appendix.

14. Are the number and quality of references appropriate?

YES, except for references to alternative model selection procedures.

15. Is the amount and quality of supplementary material appropriate?

I didn't see references to any supplementary material, but I suggest that figures 9-14 could be put into a supplement.

General Observations

Manuscript hess-2023-41 "Seasonal prediction of end-of-dry season watershed behavior in a highly interconnected alluvial watershed, northern California" by Kouba and Harter provides an important contribution to the literature on behavior of alluvial river systems in Mediterranean climates. These types of river systems are ubiquitous around the globe and often have high ecological significance. In California and southern Oregon alone, these rivers support imperiled runs of anadromous fish, and the end-of-dry season behavior is critical to the migration and spawning success of these fish. The authors identify and propose predictors of timing of reconnection of surface flow in the Scott River, Klamath Basin, using key hydrologic indicators known five months in advance. The final predictive model uses peak snow water equivalent and total October-April precipitation as predictors. The authors use leave-one-out-cross validation (LOOCV) to select the best model, but they restricted model selection to a candidate set that consisted only of models with one or two predictors. This choice of methodology eliminates more highly parameterized models that could be better by some criteria and also restricts the amount of scientific insight than can be gained through the model-selection procedure. Further, presentation of additional model diagnostics, goodness-of-fit, and measures of predictive uncertainty beyond those depicted in figures 9-14 would improve the reader's interpretation of how the final model

will perform when used in a predictive capacity among a potentially larger set of models that could be considered. I highly recommend that the authors use an information criterion such as AIC for model selection. Although AIC and LOOCV are asymptotically equivalent for selection of regression models—and in this case may well end up producing the same model—application of AIC or other information criteria to a wider candidate set would offer much more scientific insight, especially if presented in the context of multi-model inference. I recommend acceptance upon major revision, with that revision consisting primarily of use of an information criterion for model selection. The manuscript would also benefit from some restructuring and from moving figures 9-14 (and any additional diagnostic figures) to a supplement.

Specific Comments.

1. The manuscript is obviously a chapter from a dissertation, and it needs quite a bit of structural rearrangement to be suitable for a stand-alone peer-reviewed publication. First, there are numerous references such as "See Chapter 1 of this dissertation" that need to be either replaced with formal citations of the dissertation or of other peer-reviewed papers or eliminated. Second, the figures are not numbered in the sequence in which they are first referenced in the text, which is contrary to most journal style guides and makes it really hard to follow the logic of the paper. Third, there are several acronyms used that are not defined. Most likely these were defined earlier in the dissertation, but they need to stand alone in this paper. Finally, again with respect to figures, figures 9-14 serve as model diagnostics and should be moved to a supplement.

2. Mathematical quantities have very cumbersome notation. I appreciate that the nomenclature is complete enough to describe the quantity (e.g., $V_{min,30\ days}$ as 30-day minimum dry season streamflow volume), but after an initial definition, a much more concise symbol would make the manuscript easier to read. I think V_{min} would be sufficient. Once selected, make sure the variables are consistently presented in the text, tables, figures, and figure captions. That is not currently the case. Additionally, the two equations in the manuscript, which are currently not numbered, are simply algebraic linear equations, which do not need explicit formulaic listing in the manuscript. The one on line 301 explicitly lists the predictor variables, which also have very cumbersome notation. A table listing the final model coefficients and their standard errors would be more informative and easier to follow.

3. Although leave-one-out-cross validation (LOOCV) is a widely accepted and defensible model-selection method, the value of this research would be much greater if an information criterion (IC) such as AIC, BIC, or FIC were used as the model-selection technique. At the very least, the authors should acknowledge that IC are widely used and provide statistical justification for why LOOCV is used instead of an IC. Two good references on IC are Burnham and Anderson (2002) "Model Selection and Multimodel Inference" and Claeskens and Hjort (2008) "Model selection and model averaging." In defense of LOOCV, both of these books indicate that for simple regression models like the ones the authors fit, LOOCV and Akaike's IC (AIC) are asymptotically equivalent, meaning that they will select the same optimal model from a candidate set given a

large enough sample size. So, it's possible that use of AIC will produce the same optimal model as LOOCV in this case. Further, although the authors cite a reference that models be restricted to one or two predictors with a sample of size 80 to avoid over-fitting, this rule of thumb is not a good guideline to follow in the era of fast computing and default calculation of AIC and BIC in all of R's model-fitting functions. The modern scientific literature is full of examples in which top models selected with an IC had three or more predictors, even with sample sizes smaller than 80. If this data set will only support a two-predictor model without overfitting, the IC will identify that, but without even entertaining the possibility in the candidate set, there is no way of knowing whether there may be a better model that has more than two predictors. Lastly, if an IC is used and the results are presented in tabular form showing model likelihoods, relative IC weights, and parameters included in each model, the scientific value of this research would be much higher, even if the IC selected the same optimal model as presented in the current version of the paper. Figures 9-10 and 12-13 in the current version present the different one- and two-parameter models, but it is difficult to tease out the same kind of information that would be readily apparent from an IC table. For example, figures 12 and 13 show that the best two-parameter model includes maximum SWE and Oct-Apr precipitation as predictors, with LOOCV error of 461. Looking at the single-parameter models, it is apparent that of these two predictors, Oct-Apr precipitation is by far the stronger predictors, with LOOCV error of 496. Addition of peak SWE improve the model relatively little. But, this observation would be much easier to glean and much more strongly quantified if the two models appeared in an IC table.

4. Regardless of model selection technique, more model diagnostics are needed. Currently, LOOCV error and the scatter plots in figures 9-10 and 12-13 are the only reported diagnostics, making it somewhat difficult to assess whether all of the assumptions of linear model fits have been met. The scatterplots general indicate that assumptions have been met, but R's standard four diagnostic plots would be a much better way to confirm that assumptions have been met. These should go in a supplement, but they should be included. In addition, although RMSE is the model selection criterion in LOOCV and hence should be reported, it does not provide the reader with easily interpretable information about how good the model is in absolute terms. This is true of ICs as well. In either case, some other *relative* measures like MAR, R^2, or Wald's z or t statistics on the estimated parameters (parameter value divided by standard error) or predicted values provide much more information about model performance in prediction. The best model by IC or LOOCV is the best model in the candidate set, but it may not be a very good predictive model. In this case, the best model for P_{spill} has an average RMSE of 20.7 mm, relative to observed mean P_{spill} of around 60 mm. This means that a 95% prediction interval around the estimated value of P_{spill} is roughly 2*20 = 40 mm on either side of a quantity with a mean value of 60 mm. Further, the "strong" correlations you refer to (e.g., top of page 17) are not very strong in reality. R values of 0.5 to 0.73 are equivalent to model R^2 values of 0.25 to 0.53, which are low to moderate at best for predictive models.

   Some mention of attention to model assumptions should be indicated in the text. Related to this, the authors did assess the potential lags in some predictors, which is a good idea in any system with strong groundwater influence. However, using ARIMA models with appropriate lags included is a much more statistical defensible way to do this than explicitly lagging the

predictors. For the purposes of correlation with other predictors (Figure 7), explicit lagging is fine, but in model selection, ARIMA models with different AR components can be included in the candidate set and ranked with the IC right along with all other models. Again, use of an IC and multi-model table would provide a lot more information about the role of antecedent watershed conditions on the response variables of interest.

5. Some additional explanation of the characteristics of the three time periods would be helpful to provide context for the results. I suggest doing that in the introduction, rather than waiting until the results (line 240) to present that information. I agree with the authors' delineation of the three time periods, but I suggest expanding a little more on climatic differences among the time periods. I agree that 1977 coincided with widespread implementation of groundwater irrigation, but some large-scale climate indicators such as the Pacific Decadal Oscillation also changed around 1977. In other parts of the West, most hydrologists would come up with the same time-period delineations as the authors have for the Scott River, and these would be based solely on climatic factors. Climate was fairly stable prior to the late 1970s, it was highly variable from the late 1970s through 1999 and included some very wet and very dry years occurring in close succession, and very dry from 2000-present. Also across the West these climate periods were coincident with changes to water use and irrigation practices—such as the increase use of groundwater in Scott Valley starting in the late 1970s or the widespread conversion from flood to sprinkler irrigation in other parts of the west prior to the late 1990s. The intersection of climate and water use has made it very challenging to unravel the relative contributions of climate change and water use/management to observed streamflow changes. The authors have done a good job of presenting quantitative analysis that is useful for predicting important hydrologic parameters regardless of how those parameters have changed over time and regardless of reasons for the change. That is a strength of this paper. But, my two cents is that the paper could be even stronger with definition and more discussion of the three time periods right up front in the paper's introduction.

Line-by-line comments

These are in addition to and generally do not duplicate those made above, e.g., I don't identify each instance of a reference to the original dissertation here; the authors can find those with global find and replace.

Line 30. Suggest also citing use of the Surface Water Supply Index (SWSI) in other states like Idaho.

Lines 55-60. Provide quantities of agricultural and domestic water use relative to supply, e.g., "annual withdrawal of water for irrigation is XX Mm^3 relative to a total annual basin supply of YY Mm^3."

Figure 1 caption. Change "low-to-medium storage" to "medium-to-low storage" for consistency with Table 1.

Line 177: Define CDEC

Line 188: Define CIMIS. If defined here, you can use the acronym only in Figure 3 and its caption.

Figure 4. Nice figure! The reference line for Qspill is referred to in a legend but is not shown in panel A. Add the reference line.

Line 242: See comment above about moving this information to the introduction and expanding the climate discussion a little. Hence, I suggest “is coincident with” rather than “corresponds to”.

Line 262: change “or or near the valley floor” to “on or near…”.

Line 278: R values of -0.11 to -0.24 provide little to no evidence for a lagged effect, not “moderate” evidence as the authors suggest. These R values are equivalent to model R^2 values of 0.01-0.06, which nobody would consider useful in a predictive capacity.

Figure 8. The text indicates that there are 74 wells in the groundwater basin. This map shows far fewer, presumably because many are so close to on another the symbols overlap. If this is the case, either explain that in the caption or make the symbols smaller.

Line 310: Provide context for these RMSE values. How large is the error relative to the typical (mean) observation?

Lines 315-318. Standard residual plots, including leverage plots, would be much more useful than the existing figures used to illustrate model diagnostics.