Water Temperature Modeling Platform Mid-Term Review

Incidental Comments

Delta Science Program
DELTA STEWARDSHIP COUNCIL

September 30, 2022

Additional Thoughts

- These suggestions are for intended for the authors' use.
- No responses to these comments are required.

Documents reviewed

- 2. Task 4 TM_Data Management (DRAFT)_052322(v1-clean)_POB(VI_508)_061022.pdf
- 3. cvp-wtmp-tech-memo-5-system-model-framework-selection-2021-12-10.pdf
- 4. cvp-wtmp-tech-memo-6-model-selection-2021-12-10.pdf
- 5. Tech Memo_ Data Development_DRAFT_06-1-22_V1_POB061422_clean.pdf
- 6. TM8_Model Development_DRAFT (v1) 6-6-22_MLD(v2)_POB(clean).pdf

Documents read - no separate review comments

See comments for 6. TM8_Model Development

- 6a. App_A_Shasta Lake Model Results and Model Performance Statistics.pdf
- 6b. App_B_Keswick Model Results and Model Performance Statistics.pdf
- 6c. App C_2000-17_WTMP_report_draft-2022.06.17-1737_(POB-format).pdf

Document not read

See comments for 5. Tech Memo_ Data Development

- 7. DRAFT_Data_Inventory_06-20-22(v1).xlsx

Document read for background - no review comments

- 8. USBR_TM_SelectiveWithdrawal_FINAL_REPORT_9-28.pdf

In the following, each section starts with "General comments", which includes major comments and comments that pertain to the whole document, followed by "Specific comments", which refer to a specific chapter, page, figure, or table.

Some of the general comments (especially does related to the Model Development document) have already been included in the document: Q5_CalibrationEvaluation,

Water Temperature Modeling Platform: Model Development, Calibration, Validation, and Sensitivity Analysis (INTERIM DRAFT).

Specific comments

Page 2-5. Please note that the epilimnion is defined here as “the upper, warmest layer” of a thermally stratified reservoir. Yet, this definition is not entirely correct because the epilimnion can be colder than underlying layers in a stratified reservoir during the winter (e.g., Tranmer et al., 2018). A better definition would be “the lower-density surface layer of a thermally stratified reservoir that is warmer during the summer – and cooler during the winter in colder climates – than underlying layers”.

Citations

Tranmer AW, D Weigel C Marti, D Vidergar, R Benjankar, D Tonina, P Goodwing J Imberger (2020) Coupled reservoir-river systems: Lessons from an integrated aquatic ecosystem assessment. J Environmental Management 260(110107). https://doi.org/10.1016/j.jenvman.2020.110107

Data Management

Filename: Task 4 TM_Data Management (DRAFT)_052322(v1-clean)_POB(VI_508)_061022.pdf

General comments

a. Include information on backups / failovers. At least discuss the requirements for these.
b. Discuss anticipated future data volumes and make sure the databases are compatible with those.
c. No actual metadata standards or data types are discussed as part of this document. Consider appending those to the document.

Specific comments

a. p.1-1: Have you estimated data volumes for tasks anticipated in the future to make sure your Data Management System (DMS) can adapt, for example:
    I. gridded data
    II. ensemble forecasts
    III. hindcasts (single-trace or ensemble) used to assess forecast accuracy
b. p.1-1: Real-time data is often provisional until some later date when QA/QC is performed but it has to be used in the provisional form as part of real-time forecasting and operations. Is there a way to indicate that a simulation is based on provisional data rather than QA/QC'd information?
c. p.2-1: Does your time series datatype allow for gridded time series data sets or is it limited to point data?
d. p.2-1: Can operational data be modified, for example, to do scenario evaluation to evaluate alternative management strategies?
e. p.2-2: Perhaps avoid the "verification" terminology (if only to avoid discussion about the term. Model "evaluation" may be easier.
f. p.3-1: "Document, as feasible," - why the qualifier on documentation? Under what conditions would it not be feasible to document data types and data ranges?
g. Figure 3-2: The time series in panels a and b appear to differ for reasons other than just gap filling. The short-term variability is much greater in panel a) than in panel b) (different time step perhaps?). To avoid confusion, it may be useful to update the caption to mention what else was done or replace the figure.

h. p.4-1: What are the expected data volumes? How quickly will these volumes grow? Will any data in the database be retired?
i. p.4-1: What are the performance requirements for the database?
j. p.4-1: How will the database be backed up or mirrored? Is there a fail-over requirement (automatically connect to a different system when operational system is out of commission)?
k. p.4-1: Data streams inevitably fail. How do you backfill the database to recover from outages?
l. p.4-2: The data throughput rate of 5,000 data points / day will rapidly be exceeded when you start storing model output, which is one of the expressed uses of the database.

Model Framework Selection

Filename: 3. cvp-wtmp-tech-memo-5-system-model-framework-selection-2021-12-10.pdf

General comments:

a) Ongoing model evaluation can be a very useful activity to learn about model performance. That would mean that model output is compared to new observations on a continual basis, both in near real-time and in forecast modes. Given the proposed DMS and modeling framework, it should be possible to do that for this project as well. This requires the development of dedicated reports / tools that show how recent observations compare with past model simulations and past forecasts. Making these updated forecast evaluations available to outside users as well, can play an important part in building trust (or in highlighting areas for improvement).
b) While most temperature models produce similar outputs (p.3-3), the reports and analysis can be quite different when single-trace versus ensemble techniques are used for forecast generation or for Monte Carlo simulation. These techniques are mentioned in Tables 4-3 and 5-3, but the implications for the DMS and report generation are not discussed in any detail. For example, do you store every forecast trace in an ensemble or only ensemble statistics? How do you display ensemble model output? These methods can also create additional requirements for model execution times, since it may not be possible to run an "expensive" model for a sufficiently large number of iterations.

Specific comments:

a) Table 2-1: IT support should also include system updates, backups (and recovery), maintenance of system libraries, user account management, and permission management.
b) p.2-5: "(see for example Bass, et al. (2003))" suggested change "(see for example Bass et al., 2003)"
c) p.2-5: "will be related to the whether the system" suggested change "will be related to whether the system"
d) p.2-5: "distributed modeling workstations". In this case the modeling is not necessarily distributed, but the workstations are. Perhaps just remove the 'modeling'. You can also remove the entire sentence since the following paragraph addresses the same topic in more detail.

e) p.3-3: Are there regulatory or legal mandates that require the ability to track versions and reproduce earlier model simulations (in addition to this being good practice and being desirable)?

f) Table 4.1: The first column does not really contain 'model types' but lists specific 'model codes'. The first column reads as something that is a requirement and it would be good to explain why these specific model codes are required. Do they have specific capabilities that make them suitable (in other words, why are they 'expected to be used by Reclamation')? Even though the selection has already been made, it may still be useful to document why 'CEQUAL-W2' is a 'must' and why 'HEC-5Q', 'HEC-ResSim', and 'HEC-RAS' are preferred.

g) Table 4-1: last row: 'may be advantages' should read 'may be advantageous'

h) Table 4-2: Motivate why 'tight coupling' is unnecessary. I am not arguing that this is not true, but it means you already made certain decisions about model selection that are not well documented and explained.

i) Table 4-4: Does "Data Acquisition" fit inside the model framework, or would that be done outside the model framework (e.g., as part of the DMS)?

j) Table 4-4: Why is "Forecasting support" preferred rather than a "must". After all, from page 1-1 of this document: "A primary development goal of the WTMP is to provide realistic predictions of downstream water temperatures with sufficient confidence to carry out the necessary planning for seasonal, real-time, and long-term study applications while also describing situational risk and uncertainty."

k) Chapter 5: While the tables are useful, it is not always clear why a particular modeling framework was given a particular score or why a particular modeling framework was selected as a candidate. A lead-in paragraph describing each of the model frameworks, their main purpose, and why they were selected, would be useful (perhaps copy some of that text from Appendix A). For example, ESMF was developed for "developing high-performance, multi-component Earth science modeling applications" (ESMF online documentation), which is somewhat different than the application envisioned here. It is basically a library of component interfaces, which by design requires further code development for any particular application. There is no particular "aptitude" for water resources projects, unlike some of the other modeling frameworks.

l) Chapter 5: Address the shortcomings of the selected framework "HEC-WAT". Which particular requirements are not as well developed in HEC-WAT as in some of the other modeling frameworks?

m) p.6-2, last paragraph: How will the data extracted from the DMS be aggregated / disaggregated to match the model time step? For example, if you have daily values for a particular value in your DMS, but the model timestep is monthly, how will the data be aggregated? Will that be done in the DMS or in model framework (which component takes responsibility for this task)?

n) p.6-3: Is there an expectation that a user can check out an earlier version of the model framework (and component models) to reproduce earlier model results? If so, how will that capability be managed (will also require access to older configuration files, etc.)? It may require older libraries, etc. to make these simulations possible, which can be handled through containerized virtual environments like Docker, but which can be difficult to manage.

o) p.6-4: What metadata will accompany the base simulation configuration files that will be copied into a new subdirectory hierarchy? Is this subdirectory only available locally on the workstation where the simulation was performed? How do users share these configurations?

Model Selection

Filename: 4. cvp-wtmp-tech-memo-6-model-selection-2021-12-10.pdf

General comments

a) Since the component models played a big role in choosing the Model Framework, it may be better to reverse the other of these two documents.
b) The model requirements (or model selection criteria) do not make a strong connection with the resource management requirements imposed on Reclamation, for example, temperature requirements imposed by the State Water Resources Control Board and various Biological Opinions at specific locations and times (compliance). It would be good to confirm that the model requirements are consistent with these management requirements.
c) It is not entirely clear how CE-QUAL-W2 (the selected detailed reservoir model) is expected to interact with the selected system model HecRes (for example to optimize releases for downstream temperature targets). Please elaborate. For example, is the execution of CE-QUAL-W2 triggered by HecRes (as the system model), by HecWat as the framework model, or is that done differently?

Specific comments

a) p.2-2: "Such a rigid approach was not undertaken herein due to system specific conditions, [...]" What are these system-specific conditions? I do not object to a categorical "high, medium, low" rating rather than a numerical score, mostly because I think that a numerical rating will simply be more precise without being more accurate and will provide little additional information. But if there are system specific conditions for using a categorical rating, then you should list those.
b) Table 2.1: Predicting the daily maximum stream temperature does not necessarily require a sub-daily timestep (e.g., one could use a data-driven model).
c) Table 2.1: If "computation time is important", then you should be more specific in stating what execution lengths are acceptable for what types of simulations on what hardware. Should a short-range forecast be run in less than 10 seconds, less than 10 minutes, or less than 10 hours? Is that for a single forecast trace or for a forecast ensemble? What about multi-decadal planning simulations? Can you run simulations in parallel (or is that perhaps a requirement)?

d) Table 2.2: "Modeling framework compatible": Can information exchange between models occur by passing files (or database queries) or does the information exchange need to happen through shared memory? I assume the former, but it would be good to be more explicit.
e) Table 2.5: The ability to represent specific current and planned features (such as the Shasta Dam Temperature Control Device) is rather high, but in the existing implementation, the behavior of this device is parameterized rather than represented explicitly.
f) Table 2.6: "Model is relatively easily operated" is rather vague without specifying at the very least by whom and for what purpose.
g) Table 2.6: "Model can readily assess uncertainty" -- unclear what that means. How would you define "easy" for this application?
h) Table 2.6: "Model can readily assess uncertainty with the ability to modify inputs to assess uncertainty preferred over internal logic to assess uncertainty." Not entirely clear why the preference of one or the other. I am not sure what is meant by "internal logic" in this case or whether this is an either/or situation (that is, are the two types of uncertainty equivalent?).
i) p.2-4: "Type" Section -- Rather than formulating the two bullets as questions I would reformulate them as requirements.
j) p.2-5: "Dimensionality" Section -- Be specific about the requirements of the Reclamation application. The requirements section speaks too much in generalities.
k) p.2-5: "Dimensionality" Section -- Indent the two bullets starting with "represent" so that they are sub-bullets of "two-dimensional models may be represented differently:"
l) p.2-6: "System geometric" Section -- Be specific about the requirements when you say that "model spatial resolution should capture vertical temperature gradients and outflow temperatures in reservoirs and longitudinal temperature gradients in rivers sufficient to support temperature management operations and activities." To make evaluation meaningful you should state what is "sufficient". Is it a vertical resolution of 0.01 m, 0.1 m, 1 m?
m) p.2-8: "Modeling Framework Compatible" Section -- Seems to me that the main requirement here is that model configuration and execution can be scripted (batch mode) or, in other words, can be programmatically controlled either through command-line arguments or through an API. As long as that is the case, it does not matter whether

the model has a GUI or whether the model is accessed through a command-line interface (note that models with a GUI still have an executable).

n) p3.1-3.2: I appreciate the explanation of the rationale for excluding certain (categories of) models from review.

o) Table 3.5: The distinction between 'discrete' and 'system' model types is rather confusing in the context of numerical models, since 'discrete' in a numerical context would typically be used as the antonym of 'continuous'. Please define 'discrete' in this context or rephrase. I prefer the term 'component model' as was used in the presentations to the review panel.

p) Table 3.5: 'Model type (River/Reservoir)' is defined as whether the model is 'designed for predicting vertical distributions and release-water temperatures in a reservoir reach' and it is stated that all the models are able to do this. However, further in the table, most of the models are classified as one-dimensional, and some of those have 'longitudinal' as their principal dimension. Doesn't that mean that you have a vertically averaged temperature and that it is therefore not possible to predict a vertical temperature distribution? Please clarify and/or correct.

q) Table 3.5: What does the '*' mean on the 'Yes' in the row marked 'Long-term planning'?

r) Table 3.5/3.11: CE-QUAL-W2 is marked as 'Yes*' for long-term planning in Table 3.5 (reservoir) and 'No' in Table 3.11 (river). What is the difference?

Data Development

Filename: 5. Tech Memo_ Data Development_DRAFT_06-1-22_V1_POB061422_clean.pdf

General comments

a) See comments on Data Management System document as well.
b) It would be good to be explicit about time stamps for observations and model data and to indicate in the metadata whether values are instantaneous, period-averages, highest or lowest value over the period, etc. If values are calculated over an interval (means, maxima, minima, etc.), be clear whether the time stamps are period-ending or period-starting. Similarly, be explicit in how daylight savings time will be handled (one option is to store all data in UTC or Pacific Standard Time).
c) Chapter 4: How will gap-filled records be indicated in the DMS (metadata)? Will there be a way to indicate what algorithms were used for gap-filling and whether gap-filling was automated or manual?
d) The document mentions a lot of times how things "can" be done but is often unclear on how things "will" be done. As such, it often reads as a smorgasbord of data-filling and data-analysis methods, without providing a clear idea of how this will be implemented in practice. I would suggest that you indicate which methods are at the very least considered and what the default, minimally acceptable method will be for gap-filling for each variable.
e) There is a rich literature of methods for estimating atmospheric variables. See below under specific comments. There are also other sources of information (gridded datasets, model outputs) that may at times be better sources for filling missing data than interpolation or a linear regression against neighboring sites.

Specific comments

a) p.1-1: Suggest replacing 'validate' with 'evaluate'.
b) Table 3-1 and Table 3-2: How are updates to geometry data managed? For example, changes in river cross-section due to scour and deposition or changes in the stage-volume curve due to sedimentation? Are geometry data time-stamped or versioned?
c) p.3-3: How will you use meteorological point measurements to represent the atmospheric conditions over your study domains?
d) p.3-4, solar radiation: Unclear what you mean with "normal" in "follow a daily normal curve". This is not a Gaussian curve (normal distribution). If you mean a curve that can

be calculated, then that is only true for clear-sky radiation (provided that you know the atmospheric transmissivity, which itself is a function of humidity, aerosols, etc.).

e) p.3-4 lapse rate: Lapse rates depend on environmental conditions (humidity in particular). It would be useful to use local resources/measurements to determine the lapse rate that is appropriate for the region, rather than use a standard -6ºC/km lapse rate. This lapse rate may vary seasonally (for example, because of seasonal changes in humidity).

f) p.4.1: "which are not are missing" should read "which are not missing"

g) p.4-1, flow and storage data: This paragraph is not particularly informative on what will actually be done. Will filling of flow and stage data be an automated or manual process or a combination of the two? Flow and stage data are often missing during extreme high flows because of equipment failures (e.g., gauge washed away) and these events are often of particular interest. How will extreme events be handled as part of the gap-filling process?

h) p.4-2, error in equation for flow-weighted temperature: The denominator should be the sum of the flows rather than the sum of the temperatures. As written, this is the temperature-weighted flow rather than the flow-weighted temperature (just check units). As written, the equation would "blow up" if the temperatures are 0 degrees at all penstocks.

i) p.4-2: equilibrium temperature equation: Is "S" the sum of all sources and sinks (or the net source) rather than just "sources and sinks"? Note that A/V is essentially 1/D, where D is some equivalent depth.

j) p.4-3: "Net het flux" should read "Net heat flux"

k) p.4-3: "several days several weeks" needs some punctuation.

l) p.4-3: Gap-filling multiple missing years with data from a similar year seems problematic since you would lose any correlation with events from neighboring sites (unless you replace data at all sites). If possible, using a relationship with neighboring sites is likely to be less problematic or use data from gridded historic datasets (which may need some form of correction to match the local record).

m) p.4-3, meteorological data: I would disagree with the statement that "However, these data sets are typically based on a daily frequency or longer, and though they can provide useful insight, they are not sufficient for sub-daily boundary conditions necessary to model temperature." There are a number of techniques for disaggregating daily meteorological data into sub-daily values which have been used routinely and widely for

hydrological modeling and stream temperature modeling. See for example Bohn et al. (2013) and Bennett et al. (2020) and the references therein. So, while the datasets may not be sufficient without further processing, they are definitely better than using "a similar year" as model inputs.

n) p.4-3, solar radiation: See previous comments. Pretty good methods exist for estimating solar radiation exist based on other quantities such as the daily temperature range and humidity. These methods have been widely tested and are pretty robust. They could be calibrated to local observations as needed. Note that using a similar year does not account for variations in things like smoke or fog either (at least not on the right days), without significant extra effort. The above algorithms (see e.g., Bohn et al. 2013) can also be used for quality-checking observations for items such as fouled sensors.

o) p.4-4, cloud cover: Clouds tend to be poorly resolved in even high-resolution model outputs (such as NARM). The above-mentioned methods for estimating solar radiation may be more robust. Alternatively, data products based on satellite data may be available that provide cloud cover estimates from geo-stationary satellites at a sub-daily time step. However, this may be more trouble than it is worth at this stage.

p) p.4-4, air temperature: See earlier comment about lapse rates.

q) p.4-4, dew point temperature, and wet bulb air temperatures: There are methods to estimate dew point temperature and wet bulb temperature based on air temperature. For example, assume that the dew point temperature equals the daily minimum air temperature. This works well in humid environments and over open water, but not as well in very dry environments. See for example Bohn et al. (2013) and Kimball et al (1997).

r) p.4-4, wind speed and direction: Surface wind speeds and directions are difficult to transfer from observations at nearby sites. For this variable it may be better to look at outputs from atmospheric models and perhaps develop statistical relationships between surface winds and winds modeled at a higher level in the atmosphere. Quite a lot of work has been done in this area in the context of developing resource maps and wind records in support of wind energy projects.

s) p.5-2: Unclear what the 02/21/2014 date refers to. Was that the date of the 1000ft elevation or the date of the 940ft elevation (surely the lake was not at both these levels on the same date)?

t) p.5-2: How is measured hourly data fused to develop the stage-volume relationship? I doubt that the storage volume is measured on an hourly basis (I don't know how that

would be done), so isn't the stage-volume relationship simply a function of the terrain geometry? Please clarify what "Measured" means in this context.

u) p.5-4, figure 5-3: The text states that the spillway has a capacity of 186,000 cfs at water surface elevation of 1,065 ft, but figure 5-3 shows the bottom of the spillway at a level greater than 1,065 ft (at 1,067 ft). Perhaps the 1,067 ft in the figure should be 1,037 ft? Please reconcile (or explain).

v) Table 5.2: Would be much easier to read as a graph with a time axis along the bottom.

w) Table 5.3: "Sq___C ab Shasta" should read "Sq___C at Shasta". Also consider using "Sq___C" per the directive of the Department of the Interior (Secretarial Orders 3404 and 3405 and resulting Task Force decisions: https://www.doi.gov/pressreleases/interior-department-announces-next-steps-remove-sq-federal-lands).

x) p.5-10: "with more frequent measurements taken during summer and under certain conditions" - what are those conditions?

y) p.5-12: How is "light intensity" used in the models? Do you mean shortwave radiation or is this something that is used as a measure of turbidity? Please clarify.

z) Table 5.6: Can you provide a map of the data sources in Table 5.6 (or perhaps plot them in Figure 5.4?)

aa) Figure 5-6: As before -- unclear what the "measured" refers to. I understand that the stage is measured, but I don't understand how the storage would be measured (I assume that the volume estimate is simply based on the existing stage-volume relationship).

bb) Tables 5-10/11/13/14/15/17/18/19/20/23/24/25/26/28/29/30/31/33/34/35/36 and Table 6-1: Maps would be helpful (or combine with other maps).

cc) p.5-41: The text refers to Figure 5-14, but I think it should be Figure 5-18 (first instance) and Figure 5-19 (following instances). Similarly, I think Figure 5-17 should read Figure 5-18.

dd) p.5-48: What is the source of the in-tunnel heating?

Citations

Bohn, T.J., B. Livneh, J.W. Oyler, S.W. Running, B. Nijssen, and D.P. Lettenmaier, 2013: Global evaluation of MTCLIM and related algorithms for forcing of ecological and hydrological models. Agricultural and Forest Meteorology, doi:10.1016/j.agrformet.2013.03.003.

Bennett, A., J. Hamman, and B. Nijssen, 2020: MetSim: A Python package for estimation and disaggregation of meteorological data. Journal of Open Source Software, doi:10.21105/joss.02042.

Kimball, J.S., S.W. Running, R.R. Nemani, 1997: An improved method for estimating surface humidity from daily minimum temperature. Agric. For. Meteorol, doi:10.1016/S0168-1923(96)02366-0.

Model Development

Filename: 6. TM8_Model Development_DRAFT (v1) 6-6-22_MLD(v2)_POB(clean).pdf

General comments

a) As noted in the primary document, we suggest moving away from the 'validation' terminology and thinking more broadly about model evaluation. There is a strong argument that models cannot be 'validated'. For example, Oreskes et al. (1994) argue that "[...] modelers misleadingly imply that validation and verification are synonymous, and that validation establishes the veracity of the model. In other cases, the term validation is used even more misleadingly to suggest that the model is an accurate representation of physical reality. The implication is that validated models tell us how the world really is. [...] But the agreement between any of these measures and numerical output in no way demonstrates that the model that produced the output is an accurate representation of the system". This is not merely a philosophical discussion; it matters because one of the goals of the WTMP project is to build trust with other stakeholders.

It also leads to further confusion, because the use of the term 'validation' leaves the mistaken impression that a model is either good (if it is 'validated') or bad (if it fails to pass some *a priori* established performance criteria). Models are by necessity evaluated using a limited set of observations and satisfactory model performance for some variables, which does not guarantee that the model will perform well for other variables (see for example: Grayson et al., 1992). For example, calibrating a hydrological model to reproduce streamflow does not guarantee that internal moisture stores (such as snow, soil moisture, and groundwater) are well reproduced.

That is not to say that models are not useful tools in water resources management (including temperature management), but it may be more productive and realistic to carefully select models that include the important processes, calibrate those models with available observations, and then use any additional observations to

evaluate model performance and learn about the strengths and weaknesses of the model approach. In that case, you would not necessarily declare the model 'validated', but it would allow you to learn and then communicate when and under what conditions you have greater or less confidence in your model results. For example, knowing that your model has a high temperature or flow bias under certain conditions, may allow the user to account for that bias when deciding on the amount and withdrawal depth of flow releases.

b) During calibration and model evaluation, don't be seduced to give the same weight to every observation. Instead, design evaluation measures that capture the features that are important for your modeling application. For example, the depth of the thermocline relative to the depth of the intakes may be an important quantity and you could evaluate the model performance with respect to that particular quantity. McMillan (2020) summarizes some of these hydrologic signatures and their use and a similar approach could be applied to temperature modeling.

c) It is difficult to do this in hindsight, but it would be good to establish guidelines for model performance up front and motivate the need for a given level of performance considering the decisions that you need to make. Rather than just stating that a certain evaluation metric needs to be above or below a certain threshold, explain why that matters and what the consequence will be if that is not possible. After all, you may well find that the model will not perform at a certain level at all times and at all locations, but in most cases that may not require you to abandon your chosen modeling approach entirely. This goes back to building confidence with your partners, which includes being transparent about the situations for which your model does not perform well.

d) Be explicit about the timestep at which a model performance metric was calculated (hourly, daily, monthly, etc.). A Nash-Sutcliffe efficiency (NSE) or Kling-Gupta efficiency (KGE) calculated at a daily timestep will have a different value (and expectation) than one calculated at an hourly timestep.

e) Focus calibration and model evaluation on quantities that are important when making decisions and which you can calculate from observations, such as the size of the cold pool, the location of the thermocline relative to the intake / gate locations, the downstream temperature at selected target locations (e.g., Red Bluff and Bend Bridge), the amount of heating in the Spring Creek tunnel, etc.

f) Chapter 4: Because there is a strong diurnal and seasonal cycle for the water temperature as well as a strong seasonal cycle for streamflow, care should be taken in the choice of your model calibration and model performance metrics. For example, as

Schaefli and Gupta (2007) note "In the case of strongly seasonal time series, a model that only explains the seasonality but fails to repro- duce any smaller time scale fluctuations will report a good NSE value; for predictions at the daily time step, this (high) value will be misleading." As a result, it may be useful to remove the periodic signal (season or diurnal) before calculating the performance metric.

g) It is not clear how the actual model calibration was performed. Was the calibration done manually or in an automated manner? What was the objective function for the calibration? Since there are multiple performance metrics, how were they used in the calibration? Were they combined into a single objective function for the calibration or was this a multi-objective calibration (if so, what are the trade-offs)? How many simulations were run and how did the performance metrics change as a function of the number of iterations? How did you determine that the calibration was finished?

h) The model performance figures (e.g., Figure 4-1, 4-2, etc.) provide little insight into model performance. It would be more useful to plot the measured values and then on a separate scale (or in a separate panel) the difference between the measured and simulated values. Also refrain from plotting very high frequency values (e.g., hourly) for a long period (e.g., 1 year) in a single figure, because it provides no insight into model performance on the short time scales. It would be better to plot the daily values and then separately plot/assess how the model performs at the sub-daily time scale (for example by making a plot with the diurnal cycle by month or something like it). That way you can actually see whether there are structural problems at the shorter time scales (for example, whether the simulations lag the measurements or under/overestimate the diurnal maxima and minima). For an extreme example see Figure 4-18, which provides no information.

i) p.4-44: The challenge with the outflow temperatures from Shasta Lake in 2015 is of concern (the simulated cold pool was depleted too quickly), because it happened during the year with the highest outflow temperatures, which likely had the greatest impact on fish survival. Consequently, the statement that "This parameterization performs well for the majority of other simulated summer-fall periods" comes across as a bit too self-congratulatory. It would be good to discuss whether the inability of the ResSim parameterization to capture the thermal structure of Shasta Lake in 2015 prevents it from being useful during conditions seen in 2015 and whether this requires further development. As an analogy, a model that predicts sunny weather in the desert at all

times is likely to have excellent bias, mean-squared error (MSE), Nash-Sutcliffe efficiency (NSE), etc., but will not be useful for making decisions under flash flood conditions.

Specific comments

a) p.2-6: "a minimum of five gates must be open". Does the side gate count as 1 gate as part of this requirement?
b) p.3-1: "The current modeling effort uses Version 3.6". Please be as precise as possible about the source of the model version. Is this the version distributed by ERDC or Portland State (both of which provide more recent versions).
c) p.3-16: Is there any logic in the model that prevents frequent switching of open gates in the ResSim model (since that is not something that is operationally done/feasible)?
d) p.3-17: If you include equations, make sure to define all the terms, e.g., val1, etc. Also make sure to provide units (Q_max, H). For readability it may help to number the equations.
e) p.3-18 and following: Number equations, provide units, and perhaps move to appendix.
f) p.3-21: storage for Keswick: 2.936x107 m3 - the '7' and '3' should be superscripted.
g) Figure 3-5: Provide units for vertical axis label
h) p.3-26: I appreciate the separate header of "Assumptions and Considerations". However, the implications of the assumption that "the ResSim water quality model does not account for the lag or attenuation effects of routing flow through stream reaches" are not clear and should be made explicit (or at least discussed). Are these travel times not accounted for in the flow routing, in the temperature model, or both. If there is no lag, does that mean that water released from Keswick shows up immediately further downstream (e.g., at Red Bluff)? I suspect this is not the case and clarification would be helpful.
i) Chapter 4: Be specific about the time step (e.g., hourly, daily) used to calculate the root-mean squared error (RMSE) and Nash-Sutcliffe Efficiency (NSE).
j) Table 4-1: These metrics are only meaningful if you provide the time step at which they are calculated.
k) Table 4-1: How are the model performance metrics related to the decisions that will be made based on the model output? For example, will the decision be the affected if the RMSE of the water temperature is greater than 1.5ºC?
l) p.4-3: CEQUAL-W2 Calibration: What was so different about 2016 that the minimum model timestep (DLTMIN) needed to be set to 0.4 seconds.

m) Figure 4-1, Table 4-3: Shasta Lake Stage: It is not clear to me how much of the Shasta Lake stage is actually modeled and how much is determined by the boundary conditions and the stage-storage relationship. In chapter 3 (p.3-26), both inflows and outflows are listed as boundary conditions that are provided to the Shasta Lake CEQUAL-W2 and ResSim models. In that case, the only water balance terms for the reservoir that are not provided are the open water evaporation and net groundwater outflow. As a result, it's difficult to assess from this figure and this table whether the model is performing well or not. Yes, the stage matches the observed, but that may be prescribed by the boundary conditions. In addition, the figure would be more informative if you plotted the measured quantity and then the difference between the measured and simulated on a separate scale (or in a separate panel below).

n) Figure 4-2, Table 4-4: Shasta Dam Outflow: Same as previous comments. Shasta outflows are provided to the model as boundary conditions according to chapter 3. As a result, it is not surprising that there is no error. This may be a necessary sanity check to make sure that input files are read correctly, but it does not say anything about the model performance. If the flows are actually calculated by the model and the error is 0.0 cfs, then I would argue that something is wrong, since that is simply not believable.

o) Table 4-4: If these are flows then the units should be 'cfs', not 'ft'.

p) p.4-7: "temperature profiles tracked measured values closely in all years, except for short periods": Be more critical of your own results. This is where you can use model evaluation to learn more about your own model. What was the commonality among these short periods where the model did not perform well? Were temperature thresholds exceeded during these periods? If these short periods were the important periods for water resources management decisions, then your model is not performing well, even if you measured values track closely the rest of the time.

q) p.4-7: It seems that model evaluation criteria are calculated by month rather than over the full period. That may be OK, but it would be good to explain that earlier (when you introduce the model performance metrics). Make sure that you still have enough samples in each month. If that is not the case, the error or uncertainty in your model performance metric may show a large change from month to month.

r) p.4-12: Outflow temperature: Same comment as before (be more critical of your own results). The model performance evaluation should try to determine under what conditions the model performs well and perhaps more importantly, under what conditions it does not. No model will do well all of the time for all purposes. That does not mean that

a model is not useful, but you should take the opportunity to learn about your model. Why did the model not perform as well during "a few short periods"? Or at least what is the common element among these periods. Even if the errors are due to poor measurement (or imprecise boundary conditions) that would be good to document.

s) Figure 4-4: Interesting and I can understand that "The information contained in these figures was particularly useful to the analyst during model calibration", but part of the purpose of documentation is to convey this information to third parties (including stakeholders and reviewers). I do not know to what extent this figure succeeds in conveying information to and building trust with a larger community. I would encourage you to summarize this information in a way that captures and conveys the salient details (both strengths and shortcomings) to a broader audience.

t) Table 4-9: These output statistics are calculated on hourly values. See earlier comment about effect of strong periodicity on NSE values.

u) The same comments pertain to the remaining figures and tables in this chapter, including the sensitivity analysis, ResSim calibration, and performance evaluation.

v) p.4-43/44: I appreciate the more detailed discussion of the problems with the simulated outflow temperatures during August 2005 and November 2010. They provide a better insight into model performance and some ideas are offered for future improvement.

w) Figure 4-19 (and other temperature profile plots): You are trying to provide insight into the model performance with respect to the temperature profiles. All temperatures (modeled and observed) fall between 5 and 10ºC, yet the horizontal axis goes from 0-30ºC, which makes it look like you are hiding errors. Maximize the detail available in the plots by resetting the horizontal axis to 5-15ºC.

x) Figure 4-23: The modeled values miss all the high peaks (consistently).

y) Table 4-45: Why is the bias in the flows so much larger in 2000? Is this a result of model initialization? If so, perhaps you need to exclude the first year or you need to revisit how the model is initialized.

z) Table 4-70: I appreciate the summary of the sensitivity of each of the ResSim parameters, even if the discussion is mostly qualitative (this is still very useful).

aa) Chapter 5: Summary: In the final version of the report, it would be useful to briefly summarize the conditions under which the model performs well or not and whether any further model development / tuning is needed.

Citations

Oreskes, Naomi, Kristin Shrader-Frechette, and Kenneth Belitz. “Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences.” Science 263, no. 5147 (1994): 641–46. http://www.jstor.org/stable/2883078.

Grayson, R. B., I. D. Moore, and T. A. McMahon, 1992: Physically based hydrologic modeling: 2. Is the concept realistic? Water Resources Research, doi:10.1029/92WR01259.

Schaefli B. and H. V. Gupta. (2007). Do Nash values have value? Hydrological Processes, doi:10.1002/hyp.6825

Model Results

Filenames:

- 6a. App_A_Shasta Lake Model Results and Model Performance Statistics.pdf
- 6b. App_B_Keswick Model Results and Model Performance Statistics.pdf
- 6c. App C_2000-17_WTMP_report_draft-2022.06.17-1737_(POB-format).pdf

See comments for 6. TM8_Model Development_DRAFT (v1) 6-6-22_MLD(v2)_POB(clean).pdf)

No additional comments

Data Inventory

Filename: 7. DRAFT_Data_Inventory_06-20-22(v1).xlsx

See comments for 5. Tech Memo_ Data Development_DRAFT_06-1-22_V1_POB061422_clean.pdf

No additional comments

Selective Withdrawals

Filename: 8. USBR_TM_SelectiveWithdrawal_FINAL_REPORT_9-28.pdf

No comments. Only read as background