JUNE 2025

HEC ResSim Water Temperature Model Platform (WTMP): Independent Peer Review

A report to the Delta Science Program

Prepared by

Scott A Wells, PhD – Portland State University (Panel Lead)

Todd C Rasmussen, PhD – University of Georgia (Lead Author)

Sarah E Null, PhD – Utah State University

Laurel Stratton Garvin, PhD – Garvin HydroGeo, LLC

Delta Science Program

DELTA STEWARDSHIP COUNCIL

Table of Contents

Table of Contents 2
Background 3
- Review Objectives 4
- Panel Membership 5
- Panel Format 5
- Panel Letter 5
- Review Materials 5
- Panel Charge 6
- Review Questions 6
- Schedule 7

Panel Findings 8
- Executive Summary 8
- Review Goal 11
- Documentation 12
- Theoretical Basis 14
- Review of Mass Balance Analyses 17
- Review of Energy Balance Analyses 19
- Case Test Studies 23
- References 29

Appendix: Review Comments 32
- HEC-ResSim Water Quality User's Manual 32
- Water Quality Engine Technical Manual 38
- Russian River System: Sonoma Case Study 42
- American River System: Appendix J 54
- Upper Sacramento River System: Appendix F 59
- Stanislaus River System: Appendix M 62

Background

The U.S. Bureau of Reclamation (Reclamation) is developing a Water Temperature Model Platform (WTMP), which provides a data modeling framework that combines data collection, a data management system, and physically based modeling tools.

A successful application of WTMP using the U.S. Army Corps of Engineers' Hydrologic Engineering Center's (HEC) Reservoir Simulation System (ResSim) would provide short and long-term water temperature predictions to assist resource managers of major Central Valley Project reservoirs with balancing water resources and downstream temperature needs.

WTMP relies on the Water Temperature Simulation Module (WTSM) to simulate temperature dynamics within lakes, reservoirs, and stream reaches. Note that this report refers to both WTMP and WTSM, as WTSM is the temperature simulation component within the broader WTMP temperature modeling platform that also includes data management.

As noted in the HEC ResSim Water Quality User's Manual:

- Pg 20: "WTSM computes the kinetics of water temperature, sediment temperature and heat exchange between the water column and bed sediment."

  "Kinetics for the water column and active sediment layer are fully coupled in the Water Temperature Simulation Module (WTSM)."
- Pg 21: "Water temperature is a critical physical characteristic of aquatic environments. Besides its direct effects, temperature also influences numerous biological and chemical reactions in water quality models. In fact, virtually all kinetic rates are temperature-dependent. In WTSM, water temperature is calculated using an energy balance approach based on the laws of conservation of energy (heat)."

  "This energy balance accounts for heat inputs and outputs from the forcing functions, as well as heat exchange at the water surface and at the sediment–water interface. Schematic of sources and sinks of heat exchange in the multi-layer water column illustrates the heat exchange processes at both the air-water and sediment-water interfaces."

"The primary sources of heat at the water surface include short-wave solar radiation, long-wave atmospheric radiation, heat conduction from the atmosphere to the water, and direct heat inputs. The main sinks of heat are long-wave radiation emitted by the water, evaporation, and heat conduction from the water to the atmosphere."

- Pg 74: "Meteorological inputs (i.e., atmospheric pressure, air temperature, humidity, shortwave radiation, cloudiness, wind speed) must be provided for at least one location"
- Pg 324: "Details are provided in Appendix A, §16.1"

Review Objectives

The intent of this peer review is to provide Reclamation with recommendations to improve [HEC-ResSim website] (ResSim) as a reservoir temperature simulation model. Although the use of ResSim for reservoir flow modeling is well established in the literature (e.g., USACE, 2019), reservoir temperature modeling using [WTMP] was only recently developed.

This peer review of ResSim for reservoir temperature modeling is conducted to further the model's acceptance by Reclamation and associated interested parties. The development effort included in this review covers the following river systems:

- Russian (Lakes Mendocino and Sonoma, East-West Junction, Hopland, Cloverdale, Di____ Bend[1], Dry Creek, Hacienda Bridge)
- Upper Sacramento (Shasta Lake, Keswick Reservoir, Trinity Lake, Lewiston Reservoir, Upper Trinity River, Upper Sacramento River, Whiskeytown Reservoir, and Clear Creek)
- American (Folsom Lake, Lake Natoma, and three locations on the Lower American River)
- Stanislaus (New Melones, Tulloch, and Goodwin Reservoirs, and six locations on the Stanislaus River)

---

1 Note that "Di____" is a pejorative word that is derogatory to Native Americans in California

Panel Membership

The Panel consists of four members whose expertise includes hydrology, modeling, and water quality.

- Scott A Wells, PhD – Portland State University (Panel Lead)
- Todd C Rasmussen, PhD – University of Georgia (Lead Author)
- Sarah E Null, PhD – Utah State University
- Laurel Stratton Garvin, RG, PhD – Garvin HydroGeo, LLC

The Panel addresses questions based on their expertise and provides comments solely based on the scientific information being reviewed.

Panel Format

The panel convened virtually for closed teleconference meetings on five occasions prior to submitting their Panel Letter to the Delta Science Program (DSP). DSP coordinated all correspondence between the Panel and Reclamation.

Panel Letter

The deliverable of the final review includes a Panel Letter (Panel Findings) developed by the entire Panel that addresses Review Questions based on their expertise. For the letter format, the Panel uses a Delta Science Program template, with an executive summary and table of contents.

At the conclusion of receiving peer review comments, the Peer Review Panel Lead submits a final Peer Review Report to Reclamation's peer review website[2], which includes the Panel Letter along with reviewer comments.

Reclamation's responses to the comments, including actions the agency will undertake regarding the comment, and reasons the agency believes those actions will satisfy any key concerns or recommendations will be included in the publicly posted materials.

Review Materials

Materials consistent with the focus of the peer review provided to the Peer Review Panel include:

---

[2] http://www.usbr.gov/main/qoi/peeragenda.html

Review Documents

- ResSim Water Quality User Manual (DRAFT)
- Water Quality Engine Technical Reference (DRAFT)
- HEC-ResSim Simulation of Water Quality in the Russian River Basin: 2024 Update (DRAFT)
- Water Temperature Model Platform (WTMP). Appendix J: American System ResSim Results (DRAFT)

Supplemental WTMP Material (Optional)

- Water Temperature Modeling Platform: Model Development Calibration, Validation, and Sensitivity Analysis
- Water Temperature Model Platform (WTMP). Appendix F: Upper Sacramento System ResSim Results
- Water Temperature Model Platform (WTMP). Appendix M: Stanislaus System ResSim Results

Panel Charge

Specific questions are identified below to guide the Panel for the review. The Panel is encouraged to review each question carefully and clarify, refine, or otherwise modify questions as appropriate.

Review Questions

1. Documentation
    a. How can the documentation of the model be more complete, comprehensive, clear, and readily accessible for users?
    b. Does the documentation contain appropriate references?
2. Theoretical Basis
    a. Evaluate the first principles, scientific studies, and empirical measurements used to develop the fundamental components.
3. Review mass balance analysis for each of the test cases.
    a. What criteria have been used to validate the model mass balance?

    b. What types of studies were used to validate the mass balance? What are the spatial and temporal scales? What are the flow regimes?
    c. Are there any evident biases—spatial or temporal—in the model?
4. Review energy balance analysis for each of the test cases.
    a. What criteria have been used to validate the model energy balance?
    b. What types of studies were used to validate the energy balance? What are the spatial and temporal scales? What are the flow regimes?
    c. Are there any evident biases—spatial or temporal—in the model?
5. Case Test Studies
    a. Were case test studies effectively used to validate the model during the development phase, and were they appropriate and thoroughly tested?
    b. Was sensitivity analysis included in the case test studies? How does ResSim respond to sensitivity analysis? How could it be improved?

Schedule

- Panel Review: Commenced March 2025.
- Final Panel Report: Completed June 2025.

Panel Findings

Executive Summary

The panel appreciates the exhaustive coupling of field data with physically based processes to understand and characterize both fluid flow (discharge) and temperature for the four river systems included in the review (Upper Sacramento-Trinity, American, Stanislaus, and Russian). Panel findings are supplemented with more granular comments, found in the Appendix Review Comments.

Technical and User's Manual

We note that documentation would be more useful by focusing more on higher-level content and less on user functionality. For example, guidance regarding the systems for which the model is appropriate vs. inappropriate should be included. In what systems and/or conditions (e.g., drought vs. flood, summer vs. winter) is it appropriate or inappropriate to use ResSim, the WQ Module, and WTSM? What considerations dictate appropriate use of the various models?

A section on model calibration in the User's Manual is absent. In particular, it would be helpful to discuss parameters that could be used for calibration, their default values and typical ranges. What parameters should be used to calibrate models? What are the default parameters and the recommended range of all parameters? The Panel suggests that a rigorous sensitivity analysis be performed to identify key parameters influencing the model results and better understand its capabilities and limitations for use as a scenario or forecast model.

References (citations) in both the Users and Technical Manuals are limited with many references dating to before 2000. Also many citations are absent in the references. Providing links to cited references would allow users to locate older references. References to ResSim applications in other systems – along with analyses of modeling results in these systems—would provide reviewers with a better understanding of model use as well as greater confidence in model utility and performance.

The theoretical basis should provide guidance on where to place the boundary between the 1-D (one-dimensional) river and the vertical 1-D lake or reservoir models and the proper setup for connecting the longitudinal to vertical portions of the model. This is especially critical for those systems that have a large water-level variation. The process of entrainment is confusing because the conceptual model describes entrainment as a 2-D (two-dimensional) process, but then only describes a 1-D model. We suggest that figures should only show the 1-D vertical framework if that is the actual formulation used by the model.

It is also unclear why the water-quality time step needs to be shorter than the time step for the river hydraulics, unless the model uses steady-state hydraulics. There are also concerns about how sediment layers, shortwave radiation and dust effects, water temperature mixing from dam releases, sediment-groundwater interactions, sub-watershed integration, and biogeochemical processes are incorporated.

Case Studies

Mass and energy balances were evaluated using thresholds based on four performance metrics

- Mean bias: ±1.35 ºF; ±0.5 ft; ±150 cfs
- Mean absolute error (MAE): 1.8 ºF; 1 ft; 300 cfs
- Root mean-squared error (RMSE); 2.7 ºF; 1.5 ft; 500 cfs
- Nash-Sutcliffe Efficiency (NSE): 0.65

We suggest that the rationale for establishing each performance limit be provided, particularly why an absolute discharge is used instead of a percentage (in that discharges range over many orders of magnitude while temperature and elevation do not), as well as why the Nash-Sutcliffe Efficiency is used instead of the Kling-Gupta Efficiency.

An evaluation using these four metrics was performed for three river systems (Upper Sacramento, American, Stanislaus) but not for the Russian. Model performance showed seasonality behavior, with some months having better fits than others. The Panel also notes that an evaluation of thermal masses and fluxes is not provided; the total heat content (cold water pools) of reservoir and stream segments should be included.

While an energy balance was included in a model test for Lake Folsom in the Water Quality Engine Technical Manual (§12.2), none of the case studies showed a computation of whether mass or energy was conserved by the computational scheme. One can assess mass and energy balances by checking model data agreement between flow, water level and temperature, but allowing the model user to compute it internally to demonstrate mass and energy conservation of the computational scheme (as is done in CE-QUAL-W2 and was demonstrated in §12.2) would be useful, even though it is rarely provided in models.

It is important to note that the panel has substituted the term "evaluate" for "validate" to convey the scientific use of hypothesis testing to determine the suitability of model predictions for management purposes.

It is unclear whether WTSM simulations will be used for providing hindcasts—as the case studies suggest—or will they be used for "providing short and long-term water temperature predictions", as Reclamation desires. If forecasting is an objective, the Panel suggests conducting forecasts where model performance is evaluated without updated meteorological and hydrological information.

Users need to be cautioned that the current model is only suitable for annual (or shorter term) simulations that start with stratified reservoir conditions in the spring, and is not suitable for multi-year temperature simulations. This may be an important limitation of the current model framework for forecasting. According to the documentation, reservoir models should be initialized with spring stratification conditions to prevent temperature errors in the wintertime period affecting the onset of stratification.

Case studies suggest that considerable calibration and input data analysis is required to develop useful water temperature models. We recommend providing guidance for how and what the ResSim and WTMP models should be used for, and necessary steps to make them useful for the range of anticipated modeling objectives to meet Reclamation goals.

Additional discussion of the WTMP and associated models—along with suggestions for the conditions under which each should or should not be used (e.g., CE-QUAL-W2 vs. ResSim, etc.)—would provide helpful context for evaluating the limits of ResSim to model reservoir and stream temperature in this and other systems.

.

It is often difficult to evaluate model fit over long time periods due to small temporal offsets between observations and predictions. Crossplots between observed and modeled results, such as temperature, discharge, and thermal content (temperature$\times$discharge, temperature$\times$volume) would allow visualization over these longer periods. Both P-P (frequency) and Q-Q (actual values) plots could demonstrate model performance over the full range of conditions. Hysteresis in plotted values can also be used to evaluate time offsets between observed and modeled values.

And finally, the Panel recommends that a greater effort should be devoted toward predicting cold-water pool volumes within reservoirs, in that real-time allocation of the thermal content is important for meeting downstream objectives. The use of thermal content with depth-dependent volumes is expected to provide tools for optimally meeting thermal objectives.

Review Goal

The goal of this review is to provide constructive feedback on the use of the ResSim Water Temperature Simulation Module (WTSM) within the Water Temperature Modeling Platform (WTMP), which extends the ResSim hydrologic modeling platform to simulate reservoir and downstream water temperatures. As noted in the Panel Charge,

Reclamation will use the WTMP for “providing short and long-term water temperature predictions to assist resource managers of major Central Valley Project reservoirs with balancing water resources and downstream temperature needs”.

Panel members with extensive experience in hydrology, water quality, and modeling were tasked with reviewing questions in five areas, provided below, with the objective of evaluating and improving the utility and accuracy of the ResSim water temperature modeling platform.

As requested in the Panel Charge, the panel was provided six WTMP documents for their review:

- HEC ResSim Water Quality User’s Manual
- Water Quality Engine Technical Manual
- Russian River System: Sonoma Case Study

- American River System: Appendix J
- Upper Sacramento River System: Appendix F (supplementary)
- Stanislaus River System: Appendix M (supplementary)

Documentation

How can the documentation of the model be more complete, comprehensive, clear, and readily accessible for users?

The WTSM Users Manual is thorough and detailed. However, it is written at a basic level to provide non-technical users with instructions for model execution. The intended audience appears to be users who are generally unfamiliar with computers or other software. Most guidance is devoted to functions that are common to other software (e.g., + and – nodes expand or collapse subitems in a list/menu, clicking "OK" in a GUI window). The emphasis on basic GUI interactions–such as clicking buttons–rather than explaining the model does not provide the panel with sufficient information of the inner-workings of the model. Also, the text is redundant because many of the menus and functions are arranged similarly. This limits its utility as a searchable reference.

The focus on providing elementary instructions results in challenges for advanced users who wish to understand the assumptions and limitations of model functions. As a result, it is difficult to absorb the overall strategy and information needed without applying the model to a problem. As opposed to the extensive “click here” approach, documentation would benefit from a discussion on how to evaluate various datasets and how to prepare them for model use. This discussion could include tables of required (and optional) data, perhaps grouped by model function. The Panel suggests that a better venue for evaluating this manual would be using a workshop or online tutorial where participants could work through an application and provide feedback.

The WTSM Technical Manual is easier to navigate, and the use of hyperlinking is commended. The Technical Manual would benefit from a discussion of how to apply the model, as well as the appropriate uses of the model. Specific guidance on selecting the model that best meets user needs, and when a model may not be suitable for a specific application would improve the manual. Some discussion that helps potential users understand when a 1-D model is appropriate vs when a 2-D model is required would be valuable. And specifically, in what systems and/or conditions is it appropriate or inappropriate to use ResSim, the WQ Module, and the WTSM?

For the WTSM, comparison to other river and reservoir water temperature models (e.g., CE-QUAL-W2) would be helpful. What does WTSM provide that other models do not? How is this model easier to couple with Reclamation models?

Specific warnings are needed to inform users that the WTSM runs each year independently and should not be used for multi-year temperature simulations. Reservoir models should begin with spring stratification initial conditions so that reservoirs do not remain isothermal, and that forecasting for future conditions is currently out of the scope of intended model uses.

A section on model calibration is absent for the Russian River study. In particular, missing is a discussion of parameters that could be used for calibration, their default values and typical ranges. What parameters should be used to calibrate models? What are the default parameters and the recommended range of all parameters?

Finally, the Panel suggests removing material that's duplicative, both within and between the User's and Technical Manuals, such as the WTSM in Appendix A of the User's Manual and §11.1 in the Technical Manual.

Does the documentation contain appropriate references?

Both the User's Manual and Technical Manual require better referencing. The reference section in both manuals is missing many of the in-line citations. Additionally, the citations that are included could be greatly improved by including more, and more current, references to the primary literature. See "References" for examples of literature the panel suggests considering for incorporation into the documentation. Without additional referencing, the scientific basis for—and description of—physical, chemical, and biological processes embedded within the model are generally lacking. The User's Manual includes only three references for HEC reports on DSS, HEC-RAS, and HEC-ResSim, and five reports by Zhang and Johnson. Most references provided within the Technical Manual are older, with few citations after 2000.

Reference to original studies, or other documents that summarize these studies, would provide a useful resource for evaluation of model assumptions. In-line citations are generally good, but they are not often included in the references section; e.g., Water Resources Engineers, Inc. (1967), Brown and Barnwell (1987) and Deas and Lowney (2000) are referenced on page 325 but are not included in the references section.

Also helpful would be links to provided references, in that some citations are difficult to locate using online searches. References to ResSim applications in other systems--along with analyses of modeling results in these systems--would provide reviewers with a better understanding of model use and confidence in model utility and performance.

The Technical Manual is more useful than the User's Manual because it includes pertinent equations and references. While governing equations are detailed, more thorough referencing should be provided; it is difficult to track the development and the reason for inclusion and evolution of each equation. Providing references to older versions of the model could help, but referencing primary literature would be better.

Theoretical Basis

The theoretical basis is provided in the Water Quality Engine Technical Manual, which summarizes both temperature and other water-quality processes. The principles (governing equations), measurements, and fundamental components of this model are well described; governing equations are presented and align with other 1-D water temperature models, such as GLM.

Note that page numbers are not included in the Technical Manual, nor are enumerated captions provided for figures and tables.

Evaluate the first principles, scientific studies, and empirical measurements used to develop the fundamental components.

We recommend guidance on where to set the junction between 1-D river segments and lake or reservoir segments. This is a serious issue in lakes and reservoirs that have large water-level changes during the year. We assume that riverine segments are not added or subtracted as reservoir water levels change. This has implications for model applicability and calibration but is not discussed.

It is unclear whether junction elements (between reservoirs and river reaches) can vary spatially depending on reservoir level. Static junctions can create errors in reservoir heating. A section addressing this and providing examples and guidance is necessary for this model. Further discussion on routing between the 1-D longitudinal to the 1-D vertical model (i.e., how flow and water quality parameters move from the riverine to the lake/reservoir environment) would also be beneficial.

The discussion of entrainment is confusing because the model is introduced as if it were a 2-D model. This discussion would be more clear if the figures and explanation showed only the 1-D vertical framework. The entrainment rate is mentioned as being a constant over time at the end of the entrainment section. Also, references to the equations for this section are incorrect; they refer to un-numbered equations, not the numbered Equations 1 and 3, which are for extra-terrestrial solar radiation and transport, respectively.

It is unclear why the water-quality time step must be shorter than the river-hydraulics time step. This must mean that the river discharge predictions are not fully hydrodynamic, but are step-wise steady-state. This distinction should be clarified because the numerical stability criterion for hydrodynamics is always more constraining than water quality for a dynamic simulation. As noted below, calibrating a model using the system's hydrodynamic behavior provides an estimate of the kinematic velocity (celerity), not the fluid velocity.

Sediment layers are indicated for each vertical layer of the reservoir model. Are they also included in the sediment heat transfer to each layer? If so, then arrows on the figure shown in §11.1 should indicate this, otherwise they should be removed because they would not be calculated.

Is the thermal flux associated with advective flux across the water-sediment boundary considered? The thermal influence of groundwater inflows may or may not be important. The magnitude and direction of groundwater inflows is a function of the hydraulic gradient and the hydraulic conductivity of benthic and shoreline media. This influence is likely to be greatest when groundwater levels are higher than lake levels, resulting in a hydraulic gradient toward the lake. Groundwater inflows may be constant, or vary over time, such as when lake water levels are declining rapidly.

Is there a reference in the literature to a dust coefficient being used as a calibration parameter for the short-wave solar? This appears to be just a method of reducing short wave radiation directly. Using the approach of the short-wave solar radiation models of Meeus (1999) and Bird and Hulstrom (1981) may provide a better theoretical foundation for the calculation of short-wave solar. These models include the following coefficients, which were adjustable to calibrate solar radiation (rather than one): ratio of forward scattering, aerosol absorbance, and two atmospheric turbidity terms. These coefficients were summarized in Annear and Wells (2007) and were shown not to be a linear reduction of short-wave solar as in the "dust coefficient" used in WTMP.

Another possible issue arises if one uses measured shortwave solar. In that case the dust coefficient should not be used because short-wave solar is measured. Does the model prevent the model user from adjusting the measured short-wave solar by the dust coefficient?

Water temperature mixing from outflow/dam releases is pertinent for Reclamation. Adding this source of mixing to the model should be prioritized. The optimization of temperature control device (TCD) gates to a downstream target temperature is a nice feature. But what happens when thermal layers are depleted as the cold pool is depleted - does this force mixing of the reservoir? Clarification is needed.

How are data passed when models are separated into sub-watersheds to ease the computational burden. Are outflows, temperatures, constituents automatically formatted for boundary conditions for the downstream model, or must this be done manually by users? Manually reformatting data for input to the next model presents an opportunity for human error.

Biogeochemical processes that affect lake and river temperatures are discussed in both the Water Quality Engine Technical Manual and the HEC-ResSim Water Quality User's Manual as an addition to the light extinction coefficient for pure water. But the relationships of light extinction to suspended solids and organic solids, like algae, are not consistent between the documents. The importance of these processes lies in their relationships to water temperature in that they influence thermal absorption by benthic and pelagic algae within the photic zone as well as vertical mixing (e.g., convective cooling, §8.4 in the Technical Manual). Also note that the equation for light extinction in the HEC-ResSim Manual in §16.4 Algae, Eq 32, uses POM for computing light extinction but the subsequent paragraph discussing light extinction uses POC.

While pH and dissolved oxygen concentrations do not directly affect water temperatures, they are indicators of biogeochemical reactions that alter water clarity, which does affect water temperatures within the photic zone. Monitoring and explicitly modeling these variables can provide a better understanding and prediction of biogeochemical processes as they relate to water temperature.

Also lacking is the incorporation of nitrogen fixation within the water column by cyanobacteria, which influences light absorption and phytoplankton mass. Harmful algal blooms are often related to cyanobacteria, and are most likely observed during warm periods when other forms of biologically available nitrogen (e.g., nitrate, ammonium) are absent.

Biogeochemical processes affect many water-quality parameters, including pH, specific conductance, P/R ratios, Secchi depth, turbidity, N fixation, P sorption (Fe/DOC), and benthic/pelagic algae. Photosynthetic processes are most important with the photic zone, including both epi- and hypo-limnetic systems, while respiration dominates within the deeper, aphotic zone.

Of particular interest for water temperatures are the appearance of thermal absorption by organic and mineral matter that results in greater opacity. Near-surface temperature is affected by reduced water transparency due to stormwater inflows of fine sediments (clays, humic substances) along with pelagic algal blooms. Also, HAB (harmful algal blooms) create health concerns (Van Dyke and Fadness, 2021) along with rapid water-column heating.

Review of Mass Balance Analyses

What criteria have been used to evaluate model mass balances?

Four metrics are used to evaluate model performance: the mean bias, mean absolute error (MAE), root mean-squared error (RMSE), and the Nash-Sutcliffe Efficiency (NSE). In addition, specific performance criteria[3] were applied to determine acceptable model accuracy for temperature, elevation, and discharge:

- Mean bias: ±1.35 ºF; ±0.5 ft; ±150 cfs
- Mean absolute error (MAE): 1.8 ºF; 1 ft; 300 cfs
- Root mean-squared error (RMSE); 2.7 ºF; 1.5 ft; 500 cfs
- Nash-Sutcliffe Efficiency (NSE): 0.65

[3] The Delta Stewardship Council (2023) recommends the use of performance metrics in model evaluation

An evaluation using these four metrics is performed for all four river systems (Upper Sacramento, Russian, American, Stanislaus). Comparison of each metric was compared against performance criteria in three of the four rivers systems (i.e., all except the Russian River System).

Model performance showed seasonality behavior, with some months showing better fits than others, A discussion of why some months are easier to represent than others would provide insight into model behavior.

We suggest that the rationale for establishing each performance limit be provided (c.f. Moriasi et al., 2007; Moriasi et al., 2015). The panel also recommends that percentage errors (e.g., 5%, 10%, 15%) be used instead of explicit values (e.g., 150, 300, 500 cfs) for discharge because discharges range over many orders of magnitude while temperature and elevation do not. It would also be helpful to provide the rationale for using the Nash-Sutcliffe Efficiency instead of the Kling-Gupta Efficiency.

Also, it is unclear whether WTMP simulations will be used for providing hindcasts—as the case studies suggest—or the intent is to use them as plug-and-play models that are reliable for forecasting diverse systems and scenarios. If the latter purpose is desired, the Panel suggests conducting blind forecasts where model performance is evaluated without updated meteorological and hydrological information.

For operational forecasting, real-time data will likely be available to improve model state variables (e.g., temperature, inflows, discharges, meteorological conditions). Ongoing real-time calibration could also be used to update model parameters that would likely improve model forecasts.

What types of studies were used to evaluate mass balances? What are the spatial and temporal scales? What are the flow regimes?

While principles and components are discussed conceptually, model verification (i.e., assuring that model functions are correctly incorporated) is not discussed in the Technical or User’s Manuals, so that it is not possible to identify coding or application errors from the materials that were provided.

Examples of recent studies where this or other models have been applied to other systems are lacking. These prior studies could provide useful information about model strengths and weaknesses, and lessons from these could be incorporated to improve model performance.

Mass balances in the case studies were evaluated over multiple years that include a range of flow conditions, as well in four river systems that span a wide geographic area. Model performance varied across reservoir and riverine systems, with larger reservoirs and riverine systems with greater flows showing better performance.

Are there any evident biases—spatial or temporal—in the model?

Evaluation criteria are used to evaluate model fitting to historic data in the case studies. While some spatial and temporal biases are noted, they tend to lie within acceptable ranges. We note that using arithmetic discharge leads to greater errors due to large discharge ranges, something that a logarithmic transform would mitigate.

The Technical Manual includes a section (§12.1) that evaluates mass conservation for a conservative constituent in the American River System, including a zero-concentration test, constant-concentration test, and tracer-pulse test. The latter includes first- and second-order transport with null routing that preserves concentration or mass. These tests are shown at four-hour intervals, over twenty river miles, with modeled flow ranging between 400 and 500 cfs.

While calculated travel times were accurate in mass-balance evaluations, choosing to “preserve concentration” in the tracer pulse test resulted in increased mass bias. Choosing “conserve mass” in the tracer pulse test correctly conserved mass but resulted in decreased travel times.

Care must be taken when using calibrated kinematic velocity for predicting fluid velocity, in that hydraulic models based on Muskingum Routing and similar methods may be inappropriate for predicting fluid velocities (e.g., tracer transport) due to differences between hydraulic models that are calibrated using kinematic (hydrodynamic) data rather than mass transport. That is, kinematic velocity (or celerity), **c**, seldom equals the fluid velocity, **v**, where the ratio between the two velocities is the kinematic ratio, which equals **k = c/v = 5/3** when using Manning’s Equation in a wide, rectangular channel (Fread, 1993, Eqn. 10.3.5).

Review of Energy Balance Analyses

What criteria have been used to evaluate model energy balances?

Four metrics are again used to evaluate model performance; the mean bias, mean absolute error (MAE), root mean-squared error (RMSE), and the Nash-Sutcliffe Efficiency (NSE).

It would be helpful to reviewers if histograms of model performance were plotted along with their acceptable range. As an example, data from Table 27 in Appendix F were used by the panel to compare the mean monthly bias of hourly outflow water temperatures from Lake Shasta with the performance criteria of ±1.35 ºF, shown in *Figure 1*.

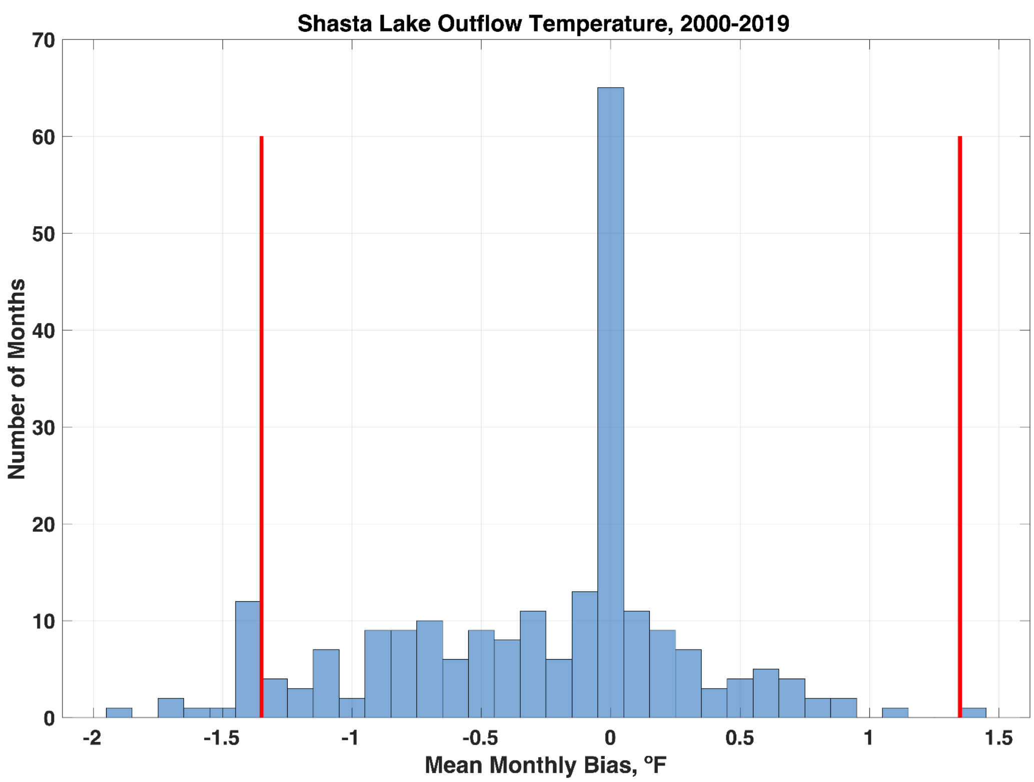

Figure 1. Example showing mean monthly bias of hourly outflow water temperatures from Lake Shasta obtained from Table 27 in Appendix F. Also shown (in red) are the performance limits of ±1.35 ºF. Note that most months fall within acceptable performance limits.

Another way to illustrate the inability to meet a performance measure is shown in *Figure 2*, which provides the number of years that the monthly mean bias is not achieved. This diagnostic information could be used to provide an explanation for the poorer performance in January, as well as for the excellent performance in March through June and December.

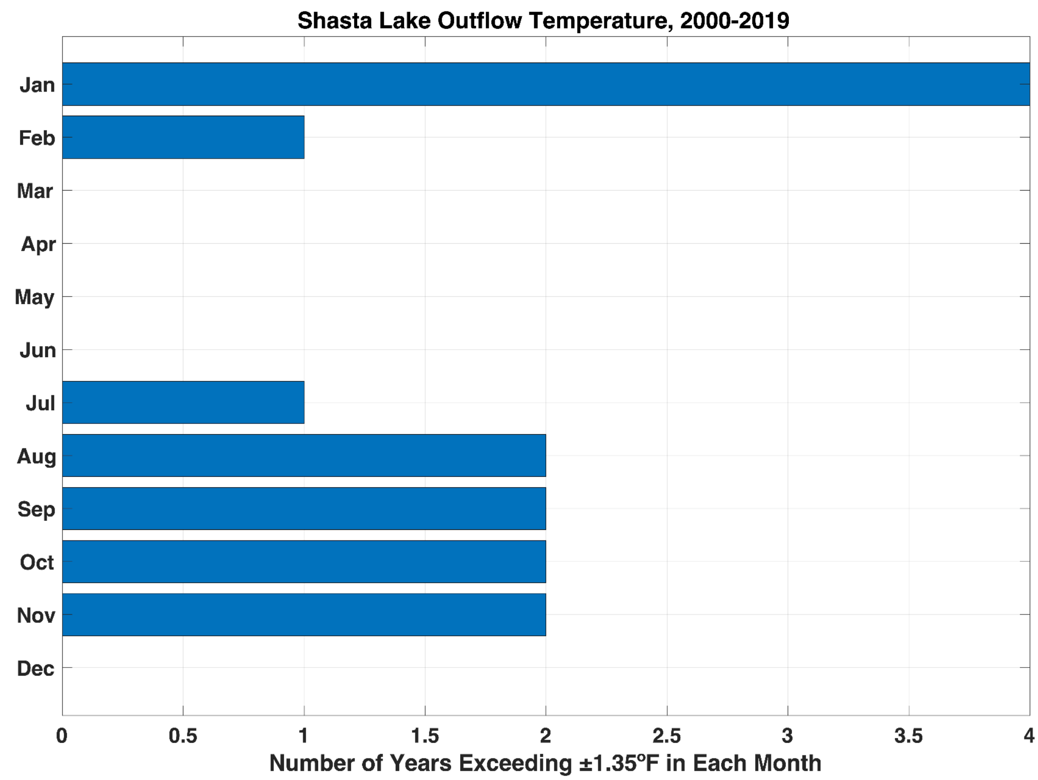

Figure 2. Example showing number of years in each month that the mean monthly bias fails to meet the ±1.35 °F performance measure. Note that January has the poorest performance, failing to meet the performance measure in four out of twenty years, while August, September, October, and November fail in two years each, February and July fail in only one year each, and March, April, May, June, and December meet the performance measure in all twenty years.

The Technical Manual includes a section testing the energy balance using atmospheric heat fluxes over nine days in July, with model residuals close to zero. Also, changes in heat content were compared to the net boundary heat fluxes for Folsom Reservoir. Again, residuals centered around zero.

While water temperatures are used for model evaluation, the energy balance should also include the thermal content, as measured using Joules or calories. These represent the total energy of the system, and are found by integrating temperatures over a known volume, or by finding the product of the fluid flux and the temperature. Note that the temperature-dependent fluid density as well as the heat capacity are also important, but can be problematic when suspended and dissolved solids are present.

A greater effort should be devoted toward predicting cold-water pool volumes within reservoirs, in that real-time allocation of the thermal content is important for meeting downstream objectives. The use of thermal content with depth-dependent volumes is expected to provide tools for optimally meeting thermal objectives.

What types of studies were used to evaluate energy balances? What are the spatial and temporal scales? What are the flow regimes?

Analyses use an hourly timestep. Inflow and outflow heat fluxes are included in Figure 2.5 in the Russian River Case Study, as well as in other case studies, but it is difficult to interpret the fits given overlapping time series as plotted. It is also difficult to evaluate model fit over long time periods due to potential temporal offsets between observations and predictions.

Crossplots between observed and modeled results, such as temperature, discharge, and thermal content (temperature multiplied by discharge) would allow visualization over these longer periods. Both P-P (frequency) and Q-Q (actual values) plots could demonstrate model performance over the full range of conditions. Hysteresis in plotted values can also be used to evaluate time offsets between observed and modeled values.

The case studies provided select random years for calibration and validation. Although commonly used, this approach is not consistent with what true validation actually is (Chapra and Reckhow, 1983; Oreskes et al., 1994). The panel recommends an approach where the entire time series of data, such as 2015-2023, be used in a long-term calibration. This approach is outlined in Shen et al. (2022).

Are there any evident biases—spatial or temporal—in the model?

There do not appear to be temporal biases in the Folsom Reservoir energy balance evaluation (at the end of the Technical Manual), and a spatial bias could not be observed using the energy balance verification provided using only Folsom

Reservoir. The spatial resolution of the Folsom Reservoir model is not provided in this section.

The extensive calibration effort is admirable (but verging on overfit, in our opinion). We feel that calibrating each year independently with a new temperature profile as an initial condition is a major limitation. How does the model perform when run for the entire Period of Record? What drift do you see? If the model is unable to accurately simulate establishment and breakdown of seasonal stratification without the control of an initial condition, it introduces concerns about the accuracy of any long model runs and the model's predictive capability. For an operational predictive model this seems limited; it limits confidence in its ability to simulate different scenarios or add model years.

Russian River Case Study Pg 3-20: "Calibration of internal dispersion at Lake Mendocino was difficult and involved box models and other methods of analyzing temperature budgets to understand whether internal mixing or bathymetry might influence the outflow temperature from Lake Mendocino. Discrepancies between the outflow temperature, which was at varying times colder and warmer than what was available in the bottom of the lake according to observed profiles, were not completely resolved in the final calibration."

More information is needed to understand this issue. Does this mean that the model determined that the (vertical?) dispersion was not accurate because the temperature profiles were incorrect? In cases like this, it is often helpful to remove all the calibration knobs that have been applied and see if the model is over-calibrated, which tends to fix some issues but cause others because these calibration adjustments are not based on physical phenomena.

Case Test Studies

Were case test studies effectively used to evaluate models during the development phase, and were they appropriate and thoroughly tested?

While four case studies were used to evaluate the model, we focus our comments on the Russian River Case Study because it provides the most comprehensive documentation. All references in this section refer to the Russian River Case Study.

The Russian River Case Study provides an excellent example of using modeling to better understand this system. The study is to be commended for the transparency of calibration steps and procedures, as well as thoughtful recommendations for future modeling efforts.

However, the Russian River Case Study was heavily tuned, including changing latent heat flux, changing reservoir release outlet depths to match observed temperatures, changing channel geometry, restarting the model each year, and starting the model with spring stratification initial conditions. Such a highly tuned model is subject to over-fitting, where improved fit comes at the cost of increased uncertainties due to parameter uncertainties. In effect, the model was fit to the data instead of the model being used to simulate the system.

The Russian River model is appropriate for hindcasting and has been effectively utilized to better understand this system. However, given that the model was so closely fit to observed data, its utility for forecasting is not proven. It would be helpful to forecast behavior without current inputs (e.g., meteorological, hydrologic, temperature) to evaluate forecast uncertainties.

Model runs are performed a single year at a time because the reservoir component of the model could not be initiated during the winter due to the unique timing and evolution of spring stratification in each year. This implies that the heat balance and mixing among reservoir layers does not function correctly. As a result, this model cannot represent reservoir (and downstream) water conditions over multiple years. Even stream temperature runs were re-started each year. There was limited forecasting potential constrained to conditions like 2013-2023. Users cannot use this model to simulate future releases through the temperature control gates to meet downstream temperature targets or to represent potential upstream infrastructure changes.

As noted above, the Russian River model appears to have required extensive calibration in order to fit observed data. Ideally, field-estimated parameters would have been used to parameterize the model to avoid extensive calibration efforts. While additional python scripts were developed to assist with model runs, these may not be suitable for other watersheds. Could this model be used "out of the box" for other systems? How were the python scripts used and are they unique to this system?

As documented in the report, specific calibration activities that were needed include:

- Changing the channel geometry cross sections (0.75 for Dry Creek & and 0.5 for Russian River downstream of SCWA Diversion). The effect of reducing cross sections is to increase the travel time.
- Pg 2-12: A major calibration effort was devoted to assigning meteorological stations to water quality regions.
- Pg 2-18: Curated Lake Mendocino inflow temperatures were fit to observed lake temperatures. It is unclear why 2024 ResSim calculated Lake Mendocino BC inflow temperatures were used instead of measured East Fork Russian River water temperatures near Calpella?
- Pg 2-18: "Ambient boundary conditions for inflow water temperature were used at Mark West Creek and Hacienda Bridge, as a better calibration was achieved with this setting than using calculated local inflow temperatures."
- Pg 2-20: "Entrainment in both Lake Mendocino and Lake Sonoma was necessary to correctly handle heating of the cold pool across summer and fall months, but the parameterizations were reduced from defaults to achieve the proper amount of mixing."
- Pg 3-2: "Each calibration year was simulated independently".
- Pg 3-20: "Calibration of internal dispersion at Lake Mendocino was difficult and involved box models and other methods of analyzing temperature budgets to understand whether internal mixing or bathymetry might influence the outflow temperature from Lake Mendocino."
- Lake Sonoma Power Plant, Controlled Outlet, and New Outlet structure elevations (inverts) were changed to better match outflow temperatures. This ignores the actual separation of flow between Power Plant and Controlled Outlet.
- Independent calibration of river reaches was needed, in that a new boundary condition was needed from Coyote Dam (Lake Mendocino) for the downstream river reach.
- Temperatures are off by Di____ Bend to the extent that simulations couldn't be relied upon for maintaining temperature targets. But it is worth noting that Di____ Bend is probably at least sixty river miles from Lake Mendocino, which is a challenging simulation for a stream temperature model.

- As noted previously, Lake Mendocino calibration suggests fundamental water temperature / mixing errors in the model: “Discrepancies between the outflow temperature, which was at varying times colder and warmer than what was available in the bottom of the lake according to observed profiles, were not completely resolved in the final calibration” (Pg 3-20),
- The model introduces three new parameters. The first two of which have been integrated into ResSimWQ, while the third is in the process of being integrated.
    1. Longwave radiation shading to limit excessive cooling in reaches with riparian canopy.
    2. COARE 3.6 [Coupled Ocean-Atmosphere Response Experiment], a state-of-the-art surface bulk heat flux algorithm, was incorporated into ResSim. COARE includes features such as wind gustiness, permitting cooling under low wind speeds, and non-equilibrium water skin temperature that is understood to affect heat fluxes.
    3. Latent heat flux limitation at cold temps (reservoirs have high humidity near their surface when Ta<Tw.

We recommend providing guidance for what and how ResSim and the WTMP models should be used, and necessary steps to make them useful for the desired modeling objectives. The Russian River case study suggests that considerable calibration and input data analysis is required to develop useful water temperature models. This study also suggests that users should only conduct annual (or shorter) models that start with stratified reservoir conditions in the spring.

It is difficult to evaluate water temperature performance without also examining streamflow and reservoir stage. This is because it is important to consider the total thermal load (temperature$\times$discharge) and mass (temperature$\times$volume) in order to correctly perform the energy balance.

The panel notes that errors in monthly vs. annual Nash-Sutcliffe Efficiency values need to be evaluated and corrected. Verification of scripts to calculate NSE values is recommended.

Was sensitivity analysis included in the case test studies? How does ResSim respond to sensitivity analysis? How could it be improved?

While parameter sensitivity analyses were included in the Sacramento-Trinity, American, and Stanislaus River System Models in Chapters 5, 7, and 9 of USBR (2024), respectively, sensitivity testing appears limited in the Russian River Study. The only mention of “sensitivity” is found in relation to the temperature at Di____ Bend.

> Pg 3-54: “This section of river contains variable conditions, including depths, shaping and microclimates that are difficult to capture with the single parameterization, and resulted in high sensitivity of changes. The calibration at Di____ Bend was a compromise: by allowing greater heating, the reach temperatures swung much too high, and by allowing more cooling temperature swings were mitigated but the net cold biases grew too large.”

The Panel suggests that the Russian River Model examine parameter sensitivities to model fits, state variables, as well as other parameters. This analysis could provide an improved understanding of model performance over a range of conditions. The Panel appreciates the extensive effort at calibration that went into the model, but has concerns about their applicability to modeling scenarios or future forecasts.

As noted by Loney (2023), sensitivity analysis is commonly used to evaluate which parameters have the greatest influence on model fits. These sensitivities are often a function of ambient conditions, reflected using state variables such as discharge, time-of-year, etc. The linear response is commonly referred to as the Jacobian, with non-linear (second-order) effects characterized using the Hessian. The Jacobian matrix can be used to not only find the sensitivity of model outputs to model parameters, but also to find the covariance between model parameters that arises whenever changing one parameter affects other parameters.

In many situations, multiple sets of calibration parameters provide the same level of model performance, leading to a non-unique solution called “equifinality” (Bevin, 2006). Having multiple calibration parameter sets results in uncertainties in model performance for conditions outside the historic range of observations. For linear systems, model parameter uncertainties are a function of the Jacobian matrix along with the residual errors (RMSE).

While the goal of calibration is often to reduce residual errors, this may come at the expense of increasing parameter and forecast uncertainties, which is especially problematic when forecasts extend beyond the range of calibration data. Prediction errors are generally smallest for conditions similar to historic, but increase for conditions not previously observed.

Model over-fitting is often quantified using the Akaike Information Criterion (AIC), which balances the goodness-of-fit with the number of model parameters that can be adjusted. The concept of parsimony refers to a model that requires the least number of model parameters that provides an acceptable model fit.

The Panel cannot assess the reservoir temperature optimization functionality given the case studies that were reviewed. Those case studies were hindcasts intended to match observed data. While forecasting seems more of a need for Reclamation operations, this part of the model remains untested.

The modeling effort should note that there may be multiple parameters that can be individually or simultaneously adjusted to yield a better match. Model calibration often requires adjusting these parameters to better match model outputs to observations. Note that a table of parameters that could be used for calibration, their default values and typical ranges should be added to guide users.

References

The panel suggests reviewing—and potentially incorporating concepts from—these references:

Annear R, SA Wells (2007) A comparison of five models for estimating clear-sky solar radiation. *Water Resources Research*, 43:W10415. https://doi.org/10.1029/2006WR005055

Beven K (2006) A manifesto for the equifinality thesis. *Journal of Hydrology*, 320(1-2):18-36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Bird RE, RL Hulstrom (1981) A simplified clear sky model for direct and diffuse insolation on horizontal surfaces. *Solar Energy Research Institute*, TR-642-761, 33 pp,. https://www.osti.gov/servlets/purl/6510849

Chapra SC, KH Reckhow (1983) Engineering approaches for lake management. Volume 2, Mechanistic modeling. *Butterworth / Ann Arbor Science*. ISBN 978-0250403929

Delta Stewardship Council (2023) Delta Plan Performance Measures Guidebook: Understanding the expectations and metrics for measures adopted by the Delta Stewardship Council. Version 2.0. *Delta Stewardship Council*, 80 pp. https://deltacouncil.ca.gov/pdf/delta-plan/2023-08-31-delta-council-performance-measures-guidebook.pdf

Etiope G, N Samardžić, F Grassa, H Hrvatović, N Miošić, F Skopljak (2017) Methane and hydrogen in hyperalkaline groundwaters of the serpentinized Dinaride ophiolite belt, Bosnia and Herzegovina. *Applied Geochemistry*, 84:286-296. https://doi.org/10.1016/j.apgeochem.2017.07.006

Fread DL (1993) Chapter 10: Flow Routing. In DR Maidment (Editor-In Chief) *Handbook of Hydrology*. McGraw-Hill. ISBN 978-0071711777

Henderson-Sellers B (1984) A new formula for latent heat of vaporization of water as a function of temperature. *Quarterly Journal of the Royal Meteorological Society*, 110(466):1186-1190. https://doi.org/10.1002/qj.49711046626

Loney DA (2023) Calibration, validation, and sensitivity approach. *Reclamation, Technical Service Center.* https://deltacouncil.ca.gov/pdf/science-program/presentations/2023-09-12-wtmp-final-peer-review-cal-val-sens-approach.pdf

Moriasi DN, JG Arnold, MW Van Liew, RL Bingner, RD Harmel, TL Veith (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the American Society of Agricultural and Biological Engineers*, 50(3):885-900. https://doi.org/10.13031/2013.23153

Moriasi DN, MW Gitau, P Daggupati, N Pai (2015) Hydrologic and water quality models: Performance measures and evaluation criteria. *Transactions of the American Society of Agricultural and Biological Engineers*, 58(6):1763-1785. https://doi.org/10.13031/trans.58.10715

Meeus J (1999) Astronomical algorithms. Second Edition. *Willmann-Bell, Inc.*, Richmond, VA , 477 pp. ISBN 978-0943396613

Nordstrom DK, CN Alpers (1999) Negative pH, efflorescent mineralogy, and consequences for environmental restoration at the Iron Mountain Superfund Site, California. *Proceedings, National Academy of Sciences*, 96(7):3455-3462, https://doi.org/10.1073/pnas.96.7.3455

Oreskes N, K Shrader-Frechette, K Belitz (1994) Verification, validation, and confirmation of numerical models in the Earth sciences. *Science* 263(5147):641-646. https://doi.org/10.1126/science.263.5147.641

Rasmussen TC, M Foroughi, D Markewitz (2021) Chapter 12: Carbon and nutrient fluxes within Southeastern Piedmont critical zones. In: A Hunt, M Egli, B Faybishenko [editors] (2021) *Hydrogeology, Chemical Weathering, and Soil Formation*. https://doi.org/10.1002/9781119563952.ch12

Shen H, BA Tolson, J Mai (2022) Time to update the split-sample approach in hydrological model calibration. *Water Resources Research*, 58:e2021WR031523. https://doi.org/10.1029/2021WR031523

USACE (2019) Willamette Basin Review Feasibility Study: Integrated Feasibility Report and Environmental Assessment. *U.S. Army Corps of Engineers*, Available at Willamette Basin Review.

USBR (2024) Water Temperature Modeling Platform: Model Development, Calibration, Validation, and Sensitivity Analysis. *U.S. Bureau of Reclamation, Technical Memorandum*, 707 pp.

Van Dyke M, R Fadness (2021) Recreational water users urged to be cautious about toxic algae in the Russian River. *Sonoma County Gazette*, June 22, 2021.

https://www.sonomacountygazette.com/sonoma-county-news/recreational-water-users-urged-to-be-cautious-about-toxic-algae-in-russian/

Wells SA (2021) Modeling the effectiveness of cooling trenches for stormwater temperature mitigation. Water 2021, 13(3):373. https://doi.org/10.3390/w13030373

Zeng XQ, TC Rasmussen (2005) Multivariate statistical characterization of water quality in Lake Lanier, Georgia, USA. *Journal of Environmental Quality*, 34:1980-1991. https://doi.org/10.2134/jeq2004.0337

Zeng XQ, TC Rasmussen, MB Beck, AK Parker, ZL Lin (2006) A biogeochemical model for metabolism and nutrient cycling in a Southeastern Piedmont impoundment. *Environmental Modelling & Software*, 21(8):1073-1095. https://doi.org/10.1016/j.envsoft.2005.05.009

Appendix: Review Comments

HEC-ResSim Water Quality User's Manual

General Comments

- Will ResSim WTMP software be open-source?
- The User's Manual is written at a basic level and is redundant, making it not highly useful. The User's Manual is written at a fairly low level, for people who are generally unfamiliar with computers or other software. Most space is devoted to functions that are common to other software (e.g., + and – nodes expand or collapse subitems in a list/menu, clicking "OK" in a GUI window).
- Very thorough and detailed but truly a user's manual – hard to glean much information from this or to evaluate it without poking around in the model itself.
    - Maybe a better venue to get feedback on this manual would be from a workshop where participants try to work their way through an example.
    - Would appreciate more discussion of how to evaluate various properties (datasets etc.) and prepare them for use as opposed to the extensive "click here" approach (which is useful, but insufficient for evaluating the suitability of the data for use in the model)
    - A table of required and optional data would be useful, perhaps grouped by function in the model.
- Very little discussion of how to apply the models and when is appropriate. It's a gap generally in water quality and hydraulic modeling: how to select what model will best meet your needs and when a model should be excluded from

further consideration. Documenting the decision making and helping potential users understand when a 1-D model is appropriate vs when a 2-D might be required would be valuable.

- Focusing less on clicking buttons and more on higher level content would improve the manual. For example:
    - In what systems is it appropriate/inappropriate to use ResSim, the WQ Module, and the WTSM?
    - For the WTSM specifically, comparison to other water temperature models would be helpful. What does this one do that others don't? How is this one easier to couple with Reclamation models?
    - What parameters should users use to calibrate models? What are the default parameters and the recommended range of all parameters?
- Extensive hyperlinking is useful.
- Thanks for including bookmarks when the PDF is exported – so useful and not done enough.

Specific Comments

- §2.1. The model is set-up to support "1D longitudinal water quality modeling in rivers and 1D vertical water quality modeling in reservoirs."

  There needs to be guidance on where to set the 1-D river from the 1-D lake or reservoir. This is a serious issue in lakes and reservoirs that have large water level changes during the year. We assume that as the water level drops in the reservoir riverine segments are not added and conversely subtracted as the water level rises. This has implications for model applicability and calibration. A section on addressing this and providing examples and guidance is necessary for this model.
- §2.1.1. "For example, if very small water quality cells are surrounded by larger ones, this could lead to computational challenges and instability in the water quality model. A single small water quality cell forces the model to select a correspondingly small time step to satisfy the Courant and Peclet conditions."

  It is unclear why the water quality time step would be more constraining than the river hydraulics. This must mean that the river predictions are not dynamic, but are steady-state? This distinction should be clarified because the numerical stability criterion for hydrodynamics is always more constraining than water quality for a dynamic simulation.

- §4.5.2 Atmospheric Pressure: It would be a useful sensitivity test to examine atmospheric pressure using just elevation compared to atmospheric pressure measurements. In many models of temperature, atmospheric pressure has only a small impact on surface heat transfer processes unless there is a drastic change during storm events.
- §4.5.5.1 Lat-Long: "The dust coefficient estimates the attenuation of solar radiation due to scattering and absorption by dust particles. It ranges from 0.0 to 1.0 (the larger the coefficient, the greater the attenuation). In general, urban areas tend to have higher attenuation, while rural areas have lower attenuation."

  Is there a reference in the literature to a dust coefficient being used in this way as a calibration knob? Even measurements of short wave solar are often problematic, but this is just a method of reducing short wave radiation.

  The short-wave solar radiation models of Meeus (1999) and Bird and Hulstrom (1981) include the following coefficients which were adjustable to calibrate solar radiation (rather than one): ratio of forward scattering, aerosol absorptance, and 2 atmospheric turbidity terms.

  These were summarized in Annear and Wells (2007) and were shown not to be just a linear reduction of short-wave solar as in this "dust coefficient". Using the approach of these models may put their short-wave solar estimates on a more theoretical foundation.

  Another issue is if one has measured short wave solar. In that case the dust coefficient should not be used because short-wave solar is measured. Is there a way to prevent use of this term for calibrating for the wrong reasons?
- Appendix A. Technical Background for Water Quality Modules: Equation 1.

  The dust coefficient should also be shown in this equation that is multiplied by the short wave solar.

- §16.1: Is the thermal flux associated with advective flux across the water-sediment boundary considered? The thermal influence of groundwater inflows may or may not be important.

  The magnitude and direction of groundwater inflows is a function of the hydraulic gradient and the hydraulic conductivity of benthic and shoreline media. This influence is likely to be greatest when groundwater levels are higher than lake levels, resulting in a hydraulic gradient toward the lake. Groundwater inflows may be constant, or vary over time, such as when lake water levels are declining rapidly.
- §16.1: "The pH scale [missing verb] from 0 to 14, with 7 representing neutrality."
  - The stated range (0 to 14) is widely cited but incorrect; being a logarithmic scale, pH is not restricted to these limits. For example, the pH of 10 M HCl is pH = $-\log_{10}[10] = -1$, and the pH of 10 M NaOH = $-\log_{10}[10^{-15}] = 15$
  - The lowest pH (-3.6) ever measured in a geologic setting occurs in massive sulfide lenses at the Richmond Mine, a Superfund Site on Iron Mountain near Keswick Dam. (Nordstrom et al., 2000).
  - The highest pH (12.8) ever measured in a geologic setting occurs in hyperalkaline groundwaters within the serpentinized Dinaride ophiolite belt at Lješljani, Bosnia and Herzegovina. (Etiope, 2017).

Grammar, Syntax, Redundancy

- There are typographic errors throughout the document.
- Required meteorological data are described repeatedly, sometimes including solar radiation and sometimes without solar radiation.
- Please define all acronyms on first use.
- There are two versions of Table of Contents (starting Pg 2 & 11), e.g., there are two §2 1.
- There is considerable redundancy, so the same steps and functions are repeated in each section. As a result, it is difficult for users to search for and locate needed sections.

- §3.3, first paragraph: are the "regular', "existing", and "standard" ResSim models all the same? Suggest consistent wording to clarify.
- §3.4.1, first paragraph. This paragraph could be replaced with a sentence that says: "In the Water Quality module, a water quality menu is added to the existing File, View, Alternative, Tools, and Help menus.
- §3. Meteorological Dataset – revise header numbers.
- §4.1: "However, if this data is not available" should be changed to: "However, if these data are not available". ["Data" is plural, "Datum" is singular]
- §4, Meteorological Data: Marked this chapter to remove redundancy and clarify writing.
- §16 Water Temperature Simulation Model (WTSM) – Pg 324
    - Provides standard equations for the energy balance. Did not see red flags here.
    - Define all terms and definitions (a list would help). Some are repeated in each equation, some are listed only once so one has to scroll back and forth, some are undefined.
    - Numbers starting each new section in the TOC are incorrect and potentially duplicated incorrectly.
    - Pg 30 – Table has mismatched chapters numbers and titles.

Tables

- §4.3: Meteorological Dataset Editor Input Data Source Options and Units by Variable Type. The Panel is impressed with the number of units to convert to or from. But this is also potentially dangerous for users.
- §4.5.6 Cloudiness Table Guideline for Cloudiness:

  Why are overcast skies 0.9 fraction rather than 1.0?
- §7.1.4 Referred to as Table 1 in the text, but it has no caption nor table number. "Input parameters required for WTSM are listed in Table 1":

  The units for sediment specific heat capacity are incorrect. They should be J/kg/^{o}C (see e.g., Wells 2021).

Figures

- §2.1. Figure captions are missing. While the Panel recognizes that this is a draft report, the captions would have assisted in our review.
- §2.1.2: Figure "Schematic of sources and sinks of heat exchange"

  Sediment layers are indicated for each vertical layer. Are they also included in the sediment heat transfer to each layer? If so, then arrows on the figure should show that, otherwise they should be removed because they would be omitted from any calculations.
- Plots included within the model to compare observed and simulated output are useful.
- §3.1ff: GUI figures are useful.
- §2.1: Figure has general arrows to show flow; would appreciate a more detailed look at the interface between longitudinal 1-D and vertical 1-D grid. Note that this is nicely described in the Engine Technical Manual (§6.1.1) so maybe we missed it; otherwise, consider importing this information here.
- §2.1.3: doesn't show that CBOD affects DO. Also, Tw should affect DO, but the connection between the Water Temperature Simulation Model and the Nutrient Simulation Model isn't discussed.
- §11 10. Mixing Coefficients figure after §11.1 Mixing Set Editor.

  The units of the mixing coefficients are shown in ft^2/s. Are those units adjustable? Metric units m^2/s are more commonly used.

  Based on the Note that says "units for the coefficients will adjust according to the watershed units of measure and the display units set for the watershed", so it appears that they do change.

References

- References are sparse, not recent, and missing. A thorough vetting using references in peer-reviewed literature should be provided.
- Many citations are not provided in the references.

Water Quality Engine Technical Manual

General Comments

- Nice format – easy to navigate. Appreciate all of the hyperlinking.
- Equations look standard, and appreciate the generally detailed inclusion of governing equations, but would greatly appreciate much more thorough referencing for all of them. It's hard to track the development and evolution of each and the decisions for inclusion that were made. At the very least, provide references to older versions of the model, but more extensive referencing to the primary literature would be better (note that this is actually included in many in-line citations, but I couldn't find the citation itself anywhere).
- The ability to vary parameters spatially by regions is a nice feature.
- Pg 43: Time steps and (substeps) explained clearly.
- Suggest including documentation of the WTSM in the technical reference and omitting the duplicate version as Appendix A of the User's Manual.
- We appreciate the list of limitations in §5.5 – it would be helpful to see these limitations developed in more detail in the general manual, possibly with some discussion of the implications of each.
- §5.6 Background Section: This section is also useful; for those who aren't intimately familiar with the HEC pantheon of modeling programs, a diagram or description of the various options and how they relate (or can relate) to each other would be helpful.
- We appreciate the verification applications in §12.

Specific Comments

- We suggest breaking out §5.5 "Model Capabilities and Limitations" into two – much of the introductory material there would more properly be labeled under a section named something like "Model Setup Overview"
- Do we understand correctly that units are flexible in ResSim but not in the Water Quality Engine (e.g. temperature limited to °C)?

- The WQ Engine and the WTSM retain some dependencies, where reservoir operations and WQ are simulated iteratively. The process is less labor intensive, but flow and WQ are still not simulated jointly.
- The balance flows described in §6.4 as "Since their primary role is to adjust the reservoir volume, they should not significantly influence water quality" is a big statement that deserves caveats. In our experience these can end up being quite large if you are not paying close attention. The Panel would have appreciated seeing the balance flows plotted in the example case study, for example, rather than the fit itself.
- §6.4: "Ambient" BC allocation across all layers could use more consideration; we would like to see an example of it in use vs other approaches; does it predicate starting the model during isothermal conditions?
- §6.1.3: Do junction elements (between reservoirs and river reaches) vary spatially depending on reservoir level? This is important for stream-reservoir modeling, but is never discussed. Static junctions can create errors in reservoir heating.
- §6.3: Initial conditions "typically wash out within the first few days of the simulation". "Reservoir initial conditions—particularly for large reservoirs with low flushing times—may persist and influence the water quality profile for months or even years." Reword as number of timesteps to give more information.
- §7.2.1: The equilibrium temperature equation (source-sink term for water temperature) ignores temperature changes from sediment and the bed.
- §8.3.2: Entrainment is confusing because the idea is introduced as if this were a 2-D model, then the changes made for a 1-D model are described. Then at the end of the section, it says the entertainment rate is constant. References to the equations for this section are not correct. They refer to un-numbered equations, not the numbered Equations 1 and 3, which are for extra-terrestrial solar radiation and transport, respectively.
- §6.6: Water temperature mixing from outflow/dam releases is pertinent for Reclamation. Adding this source of mixing to the model should be prioritized.

- §10.1: The optimization of TCD gates for a downstream target temperature is a nice feature. What happens when layers are depleted, say, when the cold pool is depleted. Does it force mixing of the reservoir? The Panel assumes that it is not based on the previous section, but it should. Clarification is needed.
- §11.1.3: We can not reconcile the reservoir temperature optimization methods with the case studies that we reviewed. Latent heat flux representation does not match that of Sonoma Case Study. Changing latent heat flux for a case study was a bold step (because latent heat flux equations should be standard, even if meteorological conditions that lead to evaporation should be better represented).
- §11.1.6: Is the f_{shade} parameter specified spatially/regionally? Or is it the same value for the entire modeled extent? Representing (riparian or topographical) shade that varies by river reach or segment would be a nice feature.
- §6.4 and 8.1. Both TSS and temperature affect density currents, with inflows forming layers at depths consistent with neutral buoyancy.

  For example, inflows may be neutrally buoyant at the thermocline whenever inflows are denser than epilimnetic waters but lighter than hypolimnetic waters. Warmer, sediment-laden inflows may be lighter or denser than cooler water lacking sediments.

  Note that lakes may be polymictic in the fall when atmospheric temperatures cool surface layers that then mix with underlying layers. We often observe step-wise mixing downward, with each cold period eroding the epilimnion until the entire reservoir is mixed.
- §11.2: Why are the P/R ratio, pH, redox, and hardness not included? What are the roles of Fe, Mn, S (sulfite/sulfide/sulfate) in biogeochemical processes? These influence the opacity within the water column, resulting in increased solar absorption and thermal heating.
- §7.2.3: Reservoir seepage should be bi-directional. On Pg 43 figure, seepage is shown as outflow only. Pg 44 says only the top reservoir layer changes volume, but any layer with bed could have seepage in or out and should change volume. This could lead to errors in conservation of mass.

- §7.3.1: How are data passed when models are separated into sub-watersheds to ease the computational burden. Are outflows, temperatures, constituents automatically formatted for boundary conditions for the downstream model or must this be done manually by users? Manually reformatting data for input to the next model presents an opportunity for human error.

Tables

- Recommend a table of parameters, their default values, and recommended range to give users an idea of when parameters should be adjusted in model calibration.

Figures

- Would appreciate legends in the figures as opposed to in-line descriptions, as well as figure captions. Hard to find what you're looking for when after specific information.
- A figure of how data/output from HEC-RAS, HEC-ResSim, and the WQ Engine are passed would be helpful.

References

- The references section is incomplete. In-line citations are generally pretty good but they are not often included in the references section. Example: Water Resources Engineers, Inc. (1967), Brown and Barnwell (1987) and Deas and Lowney (2000) are referenced on page 325 but we did not find them in the references section. Another example is references on page 336.
- As with the User's Manual, this could be referenced more thoroughly. It is often difficult to trace the theoretical thread to primary literature for many of the equations. A thorough vetting using references in peer-reviewed literature should be provided.

Russian River System: Sonoma Case Study

General Comments

- Excellent example of using modeling to better understand this system.
- Transparency of calibration steps and procedures.
- Thoughtful recommendations for the future.
- Not stated explicitly but it appears that the point of the update is water quality; i.e temperature. Please clarify to help readers.
- Calibration and validation are used in the wording, somewhat interchangeably. Need to be explicit and rigorous in use of either. The Panel suggests "evaluation" might be a better option.
- Appreciate the thorough detail.
- Hard to track the various scenario names and what exactly the updates were as written out in the text – panel suggests that a table would be helpful.
- Lake Sonoma (Warm Springs Dam; Dry Creek)
    - "Blending activities were not understood until late in the calibration process, and without knowing the fractions of outflow coming from the different inlets, calibration of Lake Sonoma water quality cannot be accomplished with precision". Not sure exactly what this means – presumably there is no data?
    - Not able to follow what happened with the invert elevations well as described, given the number of knobs that were apparently turned.

      Note that it can work (and it's been done), but the Panel objects to changing the invert elevations to match outflows without a good and well-articulated reason.
    - Appreciate that without outflow data Resource Management Associates (RMA) is in a tough spot. But our takeaway here is that the calibration for Lake Sonoma is of limited value if the record was reverse engineered – forecasts are likely to have large uncertainties if there are no data. This should be re-evaluated after a few years of outflow data collection (as is acknowledged, to their credit).

    - The idea of a scripted temperature control device is a good one, but this approach needs to test the model first with real data.
- Evaluation and cleaning of meteorological data is extensive – well done.
- Nice job filling the faulty temp record at Coyote Dam.
- Nice job adding so much additional data compared to the previous version.
- General comment: next time, separating model development (new code or modules) from the development of the specific watershed model or its components would make for an easier and clearer read, as well as aid future users of ResSim or the Russian River model.
- Calibration vs validation approach: better to evaluate a calibrated model, and then estimate model forecasts where observed input data are unavailable.
- A summary section of model strengths, model weaknesses, potential next steps, and data limitations or suspected errors would be a useful addition. Something like a spatially organized discussion of bias as you move downstream would be helpful in understanding and using results.
- The model fit is heavily tuned. It does not truly represent and simulate this system. In general, the temperature fits are impressive, and it's clear that modelers spent a huge amount of thought, time, and effort to get there. Well done, but we do worry that it is overfit, and so is not effective for predictions and forecasts.
- This is not a plug and play type model. Additional python scripts were written and relied upon for existing runs. It is our understanding that these are available only for this watershed, but not other watersheds. Could this model be used "out of the box" for other systems?
- Limited forecasting potential constrained to conditions like 2013-2023.
- Users could not use this model to simulate future releases through the temperature control gates to meet downstream temperature targets or potential upstream infrastructure changes.
- Recommend providing guidance for what and how ResSim and WTMP should be used for, and necessary steps to make them useful for modeling objectives. This case study suggests that considerable calibration and input data analysis is required to develop useful water temperature models. Users

should only conduct annual (or shorter) models that start with stratified reservoir conditions in spring.

- Tables providing all the information on meteorological stations etc. are useful.

Specific Comments

- Pg 2-6: Where did the USGS model come into this? Not totally following description of alternative V4WQ. Okay to point at and summarize earlier reports, but need to provide a bit more context here (and references, ideally).

  Alternatives are hard to follow. Table 2-1 is useful; something similar would help a lot.

- Pg 2-7: Would appreciate more discussion of the decision to alter effective cross section; likely a reasonable approach but it would be good to offer some discussion (limited data? Other reasons to suggest this approach other than velocities?).

- Pg 2-10: Not convinced about the approach for scaling the latent heat flux. It would be nice to see the investigation and work that is briefly summarized on the part of the RMA team (is there an accompanying paper or other documentation in the works?). Are there any data sets available that can help confirm this approach and hypothesis?

  The use of a latent heat flux scaling coefficient results in a black box model that would need recalibration every time a meteorological dataset is switched out, and thus difficult to generalize. The effort to formalize it is admirable, but perhaps using this approach to bias-correct the meteorological datasets directly would be a more theoretically justifiable approach.

  While this approach is discussed later on, the Panel is confused as to whether the meteorological record was bias-corrected and the new parameter in the model was used, and why that would be necessary.

  This is discussed in the section on meteorological data, but is worthy of more explanation in terms of its implications for model use and predictive capacity. Some reorganization would help, too, as it would be easy to miss the discussion if you weren't reading straight through. The graphs are useful (Figures 2-2 and 2-3).

It would be helpful to see more corroboration of this approach than Moore et al. 2005 "and references therein".

- Pg 2-11: Would appreciate more discussion of the COARE method – "the COARE algorithm resulted in immediate improvement in water temperatures, without the need for extreme parameterizations" needs more explanation to understand and justify. I'm not doubting the results, but it's a bit surprising, and deserves more explanation.
- Pg 2-17: The difference between the 5Q estimate of inflow temperature to Lake Mendocino and the new estimates based on regression with data (Figure 2-4) seems worthy of further investigation or discussion – the 5Q estimate looks like a profile of a low-elevation reservoir release, while the new data indicates the temperature has more-or-less equilibrated to environmental conditions by the time it reaches the location in question. The regression looks to overestimate temperatures in the winter but spring and summer look pretty good at the scale of the figure.

  Well done on replacing this inflow – we're sure it has a huge impact on the reservoir modeling. That said, this suggests that there is (was?) a fundamental misconception in how temperature evolves downstream in the older modeling that points to a) the need for more data at other locations where this large gap in understanding might not be recognized, and b) a reevaluation of the conceptual model for temperature evolution in the system.
- Pg 2-20: Restrictions to the mixing envelope vs manipulation of outlet elevations not well discussed or explained – I'm having trouble following exactly what was done. Mixing envelope restriction makes sense to the Panel given the limitations in model discretization; moving outlet elevations seems less justifiable. The Panel recognizes how it's done but have never found it satisfying and should be avoided if possible.

  Need a better understanding of the limitations in outlet configuration data and use in general – this is key for controlling the model and also (presumably) key for using it to predict.

- Pg 2-22: Plunge point calculations – provide units for equations. Interesting that this was not a focus for calibration.

Executive Summary

- “The main project tasks included curating, cleaning, and transforming input data”. What does "curating" and "cleaning" mean? Reword for clarity.
- “process of data exploration and ResSim watershed calibration was an entwined process”. What does an "entwined process" mean? Reword for clarity.

Chapter 2: Model Description

- Pg 2-1 Observed Data Stations – Need to reference a figure to show locations of stations (perhaps Figure 2-1 but with station locations shown).
- Pg 2-3: “The outlets were then changed, in the simulations, to invert elevations…”. The selective withdrawal algorithm used to estimate the outflow distribution uses the centerline elevation not the invert elevation of the withdrawal.
- Pg 2-3: “…in ResSim in the future by using a scripted TCD…”. Please provide a definition and explanation for "scripted".
- Pg 2-6: “Specified flows, diversions, reservoir elevations and storage used in the calibration and validation of water quality were developed by the USGS (SCWA communication), except for the period 2018- 2023 where USGS model results were not yet available.” It is not clear if you are discussing USGS field data or USGS model results. Did the USGS have model results of reservoir elevations? What model were they using?
- Pg2-6: “All balance flows used ambient boundary conditions for water quality tracers.” Is this discussing the use of conservative tracers in the inflow water quality for the balance flows? What are "ambient" boundary conditions? Inflows from upstream rivers? Did they have conservative tracers in them?
- Also the term "upwind solver" for the solution scheme if used in the vertical for stratification is notoriously prone to having too much numerical diffusion. Does a better numerical scheme really add that much computational time to a simulation? What is an "uncoupled simulation"? Does that mean that a reservoir is simulated independently of a river section?

- Pg 2-7: Water Quality Geometry: Does this mean that the numerical computational grid for hydrodynamics is not the same as for water quality? Seems odd that they would not be the same.
- Pg 2-7: "The file name reflects that the effective cross section of the Dry Creek segments multiplied by 0.75, and the effective cross sections for the Russian River below the SCWA diversion dam were multiplied by 0.5. By decreasing the effective cross section, travel times decreased, as these sections of river were found to move water faster than the RAS model results indicated during calibration."

  Is this a calibration approach to reduce the cross-sectional area of the model grid? Does this mean the original grid was incorrect, yet it was the same as the HEC-RAS model? This does not seem like an appropriate approach to model calibration if the channel geometry was correct unless the discussion in the text was misunderstood.
- Pg 2-10: "New parameter 1: Latent heat flux limitation at cold air temperatures".

  Why adjust a physical constant that has known properties as a function of temperature? Do you have any studies to support this change? If the issue is meteorological data, then correct that.

  For example, Henderson-Sellers (1984) clearly shows the variation of Latent heat of vaporization with temperature. Any adjustments beyond what has been published is not physically sound and becomes just a calibration knob for fitting data.
- Pg 2-11: "New parameter 2: Longwave radiation shading". The reference for this change, Rutherford et al. (1997) is missing from the references. This can occur in heavily shaded rivers and streams but should not be applied to lakes and reservoirs which are not shaded. Adding more calibration knobs that can be used by modelers when there is little basis for using them such as in a lake/reservoir is problematic.
- P. 2-12: "New parameter 3: COARE surface heat flux model".

The use of this surface heat exchange model for computing long-wave radiation from the water, evaporation and conductive heat losses is a useful tool in the arsenal of heat flux formulae.

But in most cases when these are used with other heat flux formulae, they give similar results. Hence, the fact that the use of this model resulted in “immediate improvement in water temperature” suggests that there were bugs in the other formulations that also needed revising. Or could the use of other calibration knobs obscure what is really happening in the surface heat flux calculations?

- Pg 2-12. Assigning meteorological stations to WQ regions. Significant portion of calibration effort.
- Calibration of the river reaches independent of the modeled reservoirs; e.g., new boundary condition BC from Coyote Dam (Lake Mendocino) for downstream river reach.
- Pg 2-18. Curated Lake Mendocino inflow temperatures to fit observed lake temperatures. Why did they develop 2024 ResSim calculated Lake Mendocino BC inflow temperatures instead of using measured east fork Russian River water temperature near Calpella?
- Pg 2-18. “Ambient boundary conditions for inflow water temperature were used at Mark West Creek and Hacienda Bridge, as a better calibration was achieved with this setting than using calculated local inflow temperatures.”
- Pg 2-20. “Entrainment in both Lake Mendocino and Lake Sonoma was necessary to correctly handle heating of the cold pool across summer and fall months, but the parameterizations were reduced from defaults to achieve the proper amount of mixing.”
- Table 2-7 seems to show that the longwave shading factor was used quite heavily in some reaches (is it a simple multiplicative coefficient?); please explain further and provide some local justification for each.
- Table 2.2: The model uses a 2-hr time step, but the water quality time step is 1 hr. Is the model time step for hydrodynamics? Does this mean you are using steady-state hydrodynamics and updating the steady-state condition every 2 hrs? [It is too large for a dynamic simulation.] What is "preserve

concentration'"? Does this apply also to the concentration of heat, related to temperature?

Chapter 3: Model Calibration and Validation

- The Panel suggests replacing "Validation" with "Evaluation", in that validation implies an exact representation of reality. No model can ever provide an exact representation of environmental systems; there is always a residual error. The question is whether it provides an adequate fit to meet end-user objectives with errors that do not introduce excess uncertainty. This question can be answered by evaluating the model fit using evaluation criteria that are defined based on management needs (Delta Stewardship Council, 2023).
- Model was fit to the data instead of model simulates the system: Calibrated by changing channel geometry cross sections (0.75 for Dry Creek & and 0.5 for Russian River downstream of SCWA Diversion). Reducing cross sections increases travel time.
- Pg 3-2: We are curious as to the random approach for choosing years. Given limits in data it would not be more robust to either use the entire record or select years that bracket the full range of expected conditions to use for calibration?
- This is a seasonal, single-year model only. One cannot start a reservoir model in the winter because stratification does not set up; wintertime errors in the cold-water pool persist for the rest of the simulation. This implies that the heat balance and mixing among reservoir layers does not function correctly. As a result, this model cannot represent reservoir (and downstream) water conditions over multiple years. Even stream temperature runs were re-started each year.
- Calibrating each year independently with a new temperature profile as an initial condition is a big limitation. How does the model perform when run for the entire period of record? What drift do you see? If the model is unable to accurately simulate establishment and breakdown of seasonal stratification without the control of an initial condition, the Panel has concerns about the accuracy of any long model runs or its predictive capability.
- For an operational predictive model this seems limited given the extensive calibration – admirable effort but verging on over-fit, in the Panel's opinion,

which hurts confidence in its ability to simulate different scenarios or add model years P. 3-1: "Calibration was divided into a calibration and validation phase to avoid calibration errors related to overfitting of parameters."

What does this mean, "calibration errors related to overfitting'? By overfitting are you saying having too many calibration knobs available to the modeler to match field data? This raises the question whether more calibration knobs were introduced in Chapter 2 that may lead to overfitting. This concept of a calibration being overfit needs to be explained.

- Pg 3-2: "Based on the overlap of critical boundary, meteorological, and observational data, years 2015-2023 were selected for use in calibration and validation of ResSim water quality on the Russian River. The years 2013, 2014, 2017, 2018, 2021, and 2022 were chosen randomly as calibration years, but which included both wet (2017) and dry (2014, 2021) years. The years 2015, 2016, 2019, 2020, and 2023 were evaluated as validation."

  This approach goes against the basic idea of calibration and validation that the validation model is statistically independent of the calibration model. There were no tests to determine statistical independence because they were chosen randomly. Many others also mistakenly just use a different data set to validate their model.

  Also, the concept of validation is suspect (Oreskes et al., 1994). Chapra and Reckhow (1983) state that "comparison of model predictions with observations that represent conditions distinct from those represented by the calibration data—is often omitted or done in a perfunctory manner." The proper approach is outlined in Shen et al. (2022): "calibrating to the full historical data and skipping model validation entirely is the most robust choice".

  Hence, the model should just be applied to the entire time series from 2015-2023. The indication that the model is validated by picking random years is not an accepted practice even though it is used not only in this study but others.

- Pg. 3-2: "Each calibration year was simulated independently, beginning in late winter or early spring on the date where a water quality profile in Lake Mendocino was observed. The start date was set to the same date as the profile observation. This ensured that wintertime errors in heat flux while the reservoir was isothermal did not remain in the cold pool for the rest of the simulation and systematic bias was avoided in the calibration process. Profiles were not available for initial conditions at Lake Sonoma; instead the temperature of Lake Sonoma was set to the observed late-winter outflow temperature at the start of calibration/validation simulations in an attempt to reduce cold-pool bias."

This is not a way of building confidence in the model for scenario development. By restarting the simulation initial condition for each year, one is setting the initial conditions based on data, not a model simulation. If the model simulation is biased or is not predicting overturn and the start of stratification each year, this means that the heat balance or other mechanisms are not being modeled properly.

- Pg 3-3: "Because biases in reservoir outflows are passed into the river reaches, the river reaches themselves were first calibrated independently."

  It is much better to delve into what the issues are with the model and how the model can overcome this bias. In its current state, the model cannot be run continuously from 2015-2023 which would be the true test of a model calibration. Only in doing this continuous simulation could the model be developed properly to perform future scenarios under different climate conditions for example.This is a big limitation and suggests that the model is not able to set up spring stratification, as opposed to the explanation of "avoiding systematic bias."

  Given that it's probably a real limitation in the abilities of the 1-D model, The Panel believes that this issue deserves serious consideration of the limitations in the future use of the model, in a way that is highlighted up front. If the model is not able to set up stratification, its use should be limited to individual years given an established profile, which is likely restrictive for an operational model. This is discussed in passing in the section on Lake Mendocino, but would benefit from more focus and discussion

- Pg. 3-19: "Calibrating in this manner was the best method of estimating parameters to not be influenced by biases in initial conditions, or by potential errors in the wintertime cold pool development that may be "locked in" after springtime stratification. This method of simulating (starting in spring with observations in Lake Mendocino, and if available, Lake Sonoma) also seems the most likely and accurate way of leveraging ResSim for water quality forecasting."

Similar to the comments above, this approach used in this document does not seem like "the most accurate way of leveraging ResSim for water quality forecasting". This approach is not an accurate way to build confidence in a model that will be used for predicting future scenarios. Also, the term "leveraging" – what does that mean?

- Pg. 3-20: "Calibration of internal dispersion at Lake Mendocino was difficult and involved box models and other methods of analyzing temperature budgets to understand whether internal mixing or bathymetry might influence the outflow temperature from Lake Mendocino. Discrepancies between the outflow temperature, which was at varying times colder and warmer than what was available in the bottom of the lake according to observed profiles, were not completely resolved in the final calibration."

  Problems with Lake Mendocino calibration implies fundamental water temperature / mixing errors in the model.

  More information is needed to understand the issue involved. Does this mean that the model determined dispersion (vertical?) was assumed to not be accurate because the temperatures with depth were not correct? In cases like this, it is often helpful to remove all the calibration knobs that have been applied and see if the model is over-calibrated – fixing some issues but causing others because they are often not based on physical phenomena.
- Lake Sonoma Power Plant, Controlled Outlet, and New Outlet structure elevations were changed (inverted) to better match outflow temperatures. This ignores the actual separation of flow between Power Plant and Controlled Outlet.
- Three new parameters. Two have been integrated into ResSimWQ and the third is in the process of being integrated:

  1) Latent heat flux limitation at cold temps (reservoirs have high humidity near their surface when $T_a < T_w$.

2) Longwave radiation shading to limit excessive cooling in reaches with riparian canopy.

3) COARE 3.6 [Coupled Ocean-Atmosphere Response Experiment], a state-of-the-art surface bulk heat flux algorithm, was incorporated into ResSim. COARE includes features such as wind gustiness, permitting cooling under low wind speeds, and non-equilibrium water skin temperature that is understood to affect heat fluxes.

- It's hard to evaluate water temperature without also seeing streamflow and reservoir stage.
- Errors in monthly vs annual Nash-Sutcliffe Efficiency values. Check scripts to calculate values, they are not updating/calculating correctly.
- By Di____ Bend, temperatures get off to the extent that simulations couldn't be relied upon for modeling supporting maintaining temperature targets. Di____ Bend is probably at least 60 river miles from Lake Mendocino, which is challenging for a stream temperature model.
- Water temperatures can also be affected by biogeochemical processes (pH, Sp Cond, P/R ratio, Secchi depth, turbidity, N fixation, P sorption (Fe/DOC), benthic/pelagic algae, photic zone), in that incident solar radiation is absorbed within opaque water columns. Both organic and inorganic sediments reduce the clarity of water, leading to heat absorption within the water column. This is especially true during algal blooms, such as one noted by Van Dyke and Fadness (2021).
- Vertical mixing (convective cooling, S8.4 in Tech Manual) needs to incorporate stratification due to thermal absorption within the photic zone, which is a function of water clarity.
- There is a need to better describe thermal content and fluxes, as well as cold water pool volumes. The thermal content is the product of the volume (or flux) and the water temperature.
- It would be helpful to provide crossplots between observed and modeled data. This should better illustrate how offsets in times occur. Also, the linear range and prediction hysteresis would become evident.

- A table of available calibration data for each key model site would be useful to supplement or largely replace the "Calibration and Validation Time Periods" section, which is a tough read. It would be particularly useful to separate the different types of data (temp, DO, etc.) by location and time period.

American River System: Appendix J

General Comments

- "Calibration" and "validation" appear to be used interchangeably. We assume that calibration refers to adjusting model parameters to better match observed with model-predicted values, and validation refers to the goodness-of-fit between observed and modeled values. But these terms are ambiguously used.
- The Panel generally discourages the idea of "independent" model validation, but if that is what was done here, that needs to be made explicit.
- The results presented cover all model years for the three sections (Folsom, Natoma, and the American River downstream), while there are limited data for some years (e.g., profiles in Natoma are only available in 2001, 2002, and 2003). But, was calibration performed on a subset of model years to yield the results shown? Not clear from the writing.
- The appendix is admirably thorough but would benefit greatly from some effort at summarization to help readers understand the strengths and weaknesses of the model fit. As a stand-alone document it is nearly impossible to evaluate the model fit and its calibration (maybe there is more explanation in other documents, such as from our previous review? If so, the Panel was unable to find them).
    - Suggest providing a map of the system and all relevant locations, including collection points for measured data.
    - Suggest providing a table that helps readers understand the years that data were available from each location and how these compare to the model runs, as well as more information on model setup.
    - Suggest providing summary statistics and figures that allow readers to interpret the relative fits and understand where the model is best at

providing data and where it struggles. Ideas could be along the lines of summarizing by year, year type (wet, dry, warm, cool, or combinations thereof) or seasonality.

- What are the strengths and weaknesses of each individual model vs the system? Where does the model struggle?

- Along the lines of greater summarization, it would be beneficial to have some introductory information for each model. We understand this is an appendix but without understanding what and where we're looking at it's difficult to derive meaning. Some of this could be accomplished by providing model-specific introductory information in each section rather than a repeat of the canned text for each.
    - Canned introductory text needs some proofreading.
    - Was the model run continuously or restarted every year?
    - Are the statistics and plots provided for each model run independently or as a system?
    - What were the boundary conditions provided, particularly for upstream inflows? Assuming that dam outflows from Folsom are difficult to measure (leakage looks large); how was this accounted for in calibration? Are records complete for every year? How sensitive to Folsom outflow are the downstream models? (from the plots, the answer appears to be "very" – this is consistent with what we would expect, but would benefit from greater discussion
- Information on how these fits were achieved or what parameters were used for calibration is missing. It's difficult to evaluate the model fit without understanding how this fit was achieved.
- We assume that matching flow was a key component of the calibration steps. If such is the case, then showing the model fits is not particularly meaningful. If there is a reason that they are shown, consider editing the graphs to make them more readable
- We are not convinced that it is effective to plot model output for years in which there is no observed data for comparison in the context of a "validation" appendix. It's somewhat useful to see that the general patterns

and the range of magnitudes look reasonable, but summarization is needed to make this a useful presentation of the data.

- In general, model fits appear to be excellent. It would be helpful to understand more how this was achieved.
  - Temperature profiles in Folsom generally match the shape and magnitude of stratification well.
    - What are "Loc A", "Loc B", "Loc C", "Loc D", and "Loc E"? They aren't explained anywhere that we could find.
    - RMSE is occasionally on the high side.
    - Bias looks pretty balanced across the seasons/years.
  - Powerhouse temperatures seem to have a tendency toward negative bias (although multiyear plot this doesn't clearly show – is there a mistake in the plotting or is the scale so small it's hard to see?). We would really like to see this discussed or explained somewhere, particularly given the context of the model being used to manage temps for coldwater fish (what do you expect a negative model bias to mean operationally?).
  - Unclear why buzz plots are included under Folsom Lake section – consider moving to a new "system" section or similar, because they include data from downstream observations and downstream simulations

    Scaling of buzz plots occasionally is incorrect– need to expand y axis.
  - Lake Natoma results at USGS gauge show similar negative bias as Folsom powerhouse.
  - Label the location of the USGS gauge. Also, does this indicate anything about the Lake Natoma model or is it a better measure of how well total outflows from Folsom are captured?
  - Natoma observed data has only occasional weak stratification; this generally appears to be captured well.
  - Natoma elevation plot is unreadable.

- Temperature results from the American River site look good! General trend and magnitude recreated well. We would still like to see the negative bias discussed, particularly because it appears to propagate downstream at all observation sites provided.
- This document is challenging to review because no information was provided about how the model was calibrated. This document is titled a validation although the model names, figures, and the brief text indicate the document is to aid calibration. After reviewing the document, we do not know the extent to which parameters, geometry, input data or other components have been tuned for model fit.
- These seem to be hindcasts using observed flow data. If ResSim and the WTMP will be used for forecasting, functionality has not been demonstrated by this case study.
- The buzz plots are great figures. They show simulated stream temperatures are consistently colder than observed. This is problematic for using ResSim to manage cold water pools to meet temperature standards (e.g., temperature standard at Watt Ave).
- From William Pond to Watt Bridge – modeled stream flow doesn't heat as much as observed in this reach. William Pond modeled stream temperatures trend cold (especially during summer), while Watt Bridge trends warm in the summers.
- Table captions provide statistics and performance metrics for mean bias, MAE, RMSE, and NSE, but do not include how that range was identified. No citations provided.
- Some tables and figures are labelled as CE-QUAL-W2 results (e.g., Table 27, Pg 131). Is this correct, or is it a typographic error? If correct, discussion about the relative fits should be provided.
- All flows match observed data nearly exactly (Folsom outflow, Folsom storage, Natoma outflow, American River at Fair Oaks). We don't think they're truly showing results of a calibration, but rather these outflows are input data to the model or the model uses change in storage (so simulated outflows must match observed).

Perhaps these indicate that modeled conservation of water mass functions properly in the model (?), but that is hard to ascertain because nothing was explained about the modeling process.

We suggest showing:

- Time series of thermal volumes, inputs, and outputs as line plots and/or colored stacked bars.
- Calibration/simulation differences vs day-of-year (DOY) and temperature.
- Details of outlet structures (e.g., depths, size, shape, position). These could affect the "sweep" of water and subsequent mixing.
- How average daily temperatures compare with hourly data that varies due to daily peaking.
- Plot discharges on logarithmic scale. We also suggest using logarithmic transform of discharge to evaluate model performance. This is equivalent to evaluating percentage differences, as opposed to absolute differences.

  That is, a 100 cfs difference between observed and simulated discharges when flows are 100 cfs are likely to be more impactful than a 100 cfs difference when discharges are 100,000 cfs. Rather than a performance metric of 150 cfs (likely to fail at high flows, and pass at low flows), an error metric of 10% would balance errors across the full discharge range.

Folsom Lake

- Modeled temps start stratifying earlier than observed (e.g., March and April). Although by May, observed and modeled are well- aligned.
- Modeled temperature at depth tends to be about 1-2 F colder than observed (consistently).
- The powerplant temperatures (through the TCD) are important and generally have a good fit, except in late fall (mid Oct), when simulation temperatures decrease more quickly than observed.
- Daily simulated temperatures tend to hover 1-2 ºF colder than observed (e.g., Table 21).

Lake Natoma

- Modeling water temperature through Lake Natoma is challenging.

- Modeled temperatures are persistently colder than observed and the modeled diurnal range much larger and colder than observed.
- Given the smoothness of diurnal minima, it looks like a minimum threshold might be used in the model. We wish that was explained.
- Modeled temperatures reproduce step changes in temperature well.

Lower American River

American River at Fair Oaks

- Daily temperatures match fairly well.
- Modeled diurnal range is greater than observed.

American River at Warm Pond Park

- Many years have nearly perfect fit with a good match in diurnal variability.
- However, in some years, simulated temperatures are 2-3° F colder than simulated (with good match in the range of diurnal variability).
- If this model will be used to guide operations to meet temperature standards in the American River, these errors are important. According to Table 76, the model tends to be colder every year.

American River at Watt Bridge nr Carmichael

- The model matches observed temperatures well, or trends warm. This implies that the reach from William Pond to Watt Bridge heats too much.

Upper Sacramento River System: Appendix F

Shasta Lake

Near Dam Temperature Profiles

- It is stated (Pg 62) "In the winter or in shallower reservoirs with substantial flow through, vertical temperature profiles can be uniform from the surface to the bottom (Isothermal conditions)."

  While not relevant for the reservoirs simulated in the WTMP project, we routinely observe stratification in shallow systems (even < 2 m) with substantial flow through in Southeastern US impoundments due to thermal heating of surface layers with shallow Secchi depths (high algal concentrations, turbidity, and humic substances) and low benthic sediment temperatures due to conduction from underlying sediments and groundwater inflows.

- Summertime baseflow and stormwater inflows are cooler than surface layers, and warmer than benthic layers, so inflows migrate along the thermocline. In top-discharge impoundments, inflows displace overlying layers (increasing downstream temperatures) while inflows bottom-discharge reservoirs displace underlying layers (decreasing downstream temperatures).
- Simulated temperatures closely matched observations, with mean bias rarely exceeding ±1 °F and only a single value exceeding 2 °F (28-Oct-2015).
- It would be helpful to summarize evaluation metrics by plotting them as histograms grouped in relevant intervals (e.g., quarter-degree).
- It would also be helpful to summarize the thermal pools within the reservoir over time, i.e., volume within each integer temperature interval (e.g., 44-45 °F). This illustration could assist in understanding the remaining availability of cold water pools for future discharge.

Outflow Temperature, Flow, Elevation, and Gates

- While complete, the gate (top panel) and penstock (third panel) discharges are difficult to interpret due to overlapping lines. In these cases, it might be helpful to show cumulative discharges through all gates and penstocks using a shaded stacked chart, or some other graphical format.

Outflow Temperature

- Differences between data lines in figures are difficult to resolve, in part due to exaggerated y-axis scale. Suggest rescaling water temperatures to plot between 44 and 66 °F (min and max values, respectively). A cross-plot would also be helpful to illustrate prediction bias as a function of temperature.
- Performance, as measured using the bias between observed and simulated temperatures, are excellent; within ±1 °F for mean annual and ±2 °F for mean monthly intervals. But it might be helpful to provide histograms of differences between observed and predicted values for both hourly and daily intervals.

Reservoir Discharge (Outflow)

- It is stated (Pg 155) "Four lines are included in the figure: observed (light blue) and simulated (red-dashed) hourly water temperature and observed (dark blue) and simulated (dark red dashed) daily average outflow." Yet, there are only two lines shown (both discharge) on each plot.
- It would be helpful to provide cross-plots between observed and predicted flows, along with histograms of their differences.
- Also, discharge should be plotted on a logarithmic scale to allow inspection and understanding of lower flows.
- Mean bias in hourly discharge is routinely negligible, except in Jan-2017.

Elevation (Water Levels, Stage)

- It is again stated (Pg 202) "Four lines are included in the figure: observed (light blue) and simulated (red-dashed) hourly water temperature and observed (dark blue) and simulated (dark red dashed) daily average elevation." Yet, there are only two lines shown (both elevation) on each plot.
- Suggest shading the ranges of individual gates to more clearly identify the depth range of each gate.
- It would be helpful to provide a cross-plot between observed and simulated water levels, along with a histogram of observed differences.
- Mean biases appear large (>2'). How far ahead are elevations forecast? Shorter term (1-day ahead) forecasts should be small, with increasing errors for longer forecast horizons (e.g., 1-yr ahead).
- Could the cold water pool volume be predicted?

Keswick Reservoir

- Similar comments to Shasta Lake comments (above).
- It would be helpful to show the change in outflow discharges and temperatures between Shasta and Keswick Dams. That is, how much do discharges and temperatures change based on DOY, reservoir elevation, discharges, and temperatures.

Trinity Lake

- Similar comments to Shasta Lake comments (above).

- Model performance appears to be inferior to Shasta Lake, perhaps due to:

  Missing data during critical years and periods.

  Unique physical setting. Need to collect additional site-specific reservoir information.

Stanislaus River System: Appendix M

Similar comments to Upper Sacramento River System: Appendix F comments (above). It would be helpful to combine annual reports for all years (2005-2013) into a single document.