Water Temperature Modeling Platform Mid-Term Review

September 2022

Report to the Delta Science Program

Delta Science Program

DELTA STEWARDSHIP COUNCIL

TABLE OF CONTENTS

Panel Information **3**

Panel Members *3*

Water Temperature Model Platform *3*

USBR Technical Memoranda (Draft) *3*

Panel Charge *4*

General Findings **6**

Question 1. Does the modeling design (e.g., model selection, framework) include the necessary processes and resolution (spatial and temporal) to represent the short-term and long-term temperature dynamics expected in the reservoir and river environments throughout the CVP project area? ***8***

Suggestion 1. High-Level Overview *8*

Suggestion 2. Risk-Informed Analysis *9*

Suggestion 3. Alternative Management Plans *10*

Citations *10*

Question 2. Are the models adequate for describing water temperature during extreme hydrologic/storage conditions (e.g., droughts/low storage)? ***12***

Citations *14*

Question 3: Are unique features (i.e., selective withdrawal devices, thermal curtains, and submerged structures) adequately represented? ***15***

Citations *16*

Question 4. Are available data sufficient for the development of the selected models and intended uses? Where data gaps have been identified, are the assumptions and methodologies used to address them suitable? ***17***

Citations *18*

Question 5: Are testing methods (calibration and evaluation) adequate to demonstrate confidence in model performance for the historic period? ***19***

Citations *23*

Question 6. Does the modeling documentation include adequate information, assumptions, and detail to allow for transparency and replication of model results? ***25***

Citations *25*

Panel Information

Panel Members

- Todd E Steissberg, US Army Corps of Engineers (Panel Chair)
- Todd C Rasmussen, University of Georgia (Lead Author)
- Bart Nijssen, University of Washington
- Laurel Stratton Garvin, US Geological Survey
- Daniele Tonina, University of Idaho

Water Temperature Model Platform

- Scope of Work
- Panel Charge (see below for excerpt)

USBR Presentation (Randi C Field, 29 June 2022)

- CVP Temperature Management Overview

USBR Technical Memoranda (Draft)

1. Project Workplan
2. Data Management Plan
3. Model Selection Framework and Design
4. Model Selection
5. Data Development – Sacramento, Trinity, and American Rivers Systems
6. Model Development, Calibration, Validation, and Sensitivity Analysis
    a. Shasta Lake Model Results and Model Performance Statistics (Years 2000-2019)
    b. Keswick Reservoir Model Results and Model Performance Statistics (Years 2000-2019)
    c. Temperature Model Validation Summary, 2000-2017
7. Data Inventory (20 June 2022 v1)
8. Water Temperature Management in Reservoir-River Systems through Selective Withdrawal

Panel Charge

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

Reclamation requests the Panel identify both:

1. Modeling elements that are appropriately represented and consistent with the project objectives; as well as
2. Critical input and associated direction and recommendations to improve the model development and application.

Specific questions include:

1. Does the modeling design (e.g., model selection, framework) include the necessary processes and resolution (spatial and temporal) to represent the short-term and long-term temperature dynamics expected in the reservoir and river environments throughout the Central Valley Project, CVP, area?
2. Are the models adequate for describing water temperature during extreme hydrologic/storage conditions (e.g., droughts/low storage)?
3. Are unique features (i.e., selective withdrawal devices, thermal curtains, and submerged structures) adequately represented?
4. Are available data sufficient for the development of the selected models and intended uses? Where data gaps have been identified, are the assumptions and methodologies used to address them suitable?
5. Are testing methods (calibration and validation) adequate to demonstrate confidence in model performance for the historic period?
6. Does the modeling documentation include adequate information, assumptions, and detail to allow for transparency and replication of model results?

Abbreviations

Abbreviation	Meaning (w/ link)	Agency
ARG	American River Group	USBR
BDO	Bay-Delta Office	USBR
BO (BiOp)	Biological Opinion	NOAA
CMP	Comprehensive Program Manuals	USBR
CVP	Central Valley Project	USBR
DOI	Department of the Interior	US
DSP	Delta Science Program	CA
FWS (the Service)	Fish and Wildlife Service	US
HISA	Highly Influential Scientific Assessments	GAO
ISI	Influential Scientific Information	GAO
LTO	Long Term Operations (CVP/SWP)	USBR
MTC	Modeling Technical Committee	USBR
NMFS	National Marine Fisheries Service	NOAA
OMB	Office of Management and Budget	US
PRP	Peer Review Panel	CA
RISE	Reclamation Information Sharing Environment	USBR
RPM	Reasonable and Prudent Measures	NOAA
SRTTG	Sacramento River Temperature Task Group	USBR
SWP	State Water Project	CA
SWRCB	State Water Resources Control Board	CA
TCD	Temperature Control Device	USBR
TM	Temperature Modeling	USBR
USBR (Reclamation)	Bureau of Reclamation	USBR
WRF	Weather Research and Forecasting Model	UCAR
WTMP	Water Temperature Modeling Platform	USBR

General Findings

The panel commend the Bureau of Reclamation (USBR) for developing a project guided by several key features:

- Transparency along with open software, data, and meta data using the USBR Reclamation Information Sharing Environment (RISE) platform.
- Engagement with interested parties through open science.
- Dissemination of models and data to build community capability both in-house and within the community of interested parties.
- A systems framework with data flow through the modeling elements, automation of routine tasks, and standardized reporting.
- Vision for a framework that accommodates running the systems at different spatial-temporal scales and for different purposes.
- Ability to analyze model behavior at the element scale.

Developing a modeling framework to address all of these elements is challenging. However, it provides valuable benefits to develop water management operational scenarios that are environmentally sustainable.

The selected modeling framework is based on a modular approach that subdivides the system into key elements (e.g., rivers, tunnels, reservoirs) with information (e.g., inputs, outputs) exchanged among elements. Each element function is simulated with a selected model informed by external forcing and/or other element outputs.

The framework has been designed to allow swapping model elements so that an element can be replaced to accommodate new model development and/or a model of different spatial and temporal resolutions and scales to better address specific questions. Thus, calibration and evaluation of model performance needs to be able to account for different model elements.

Also, the framework model is expected to be used under a wide range of climatic conditions from dry to wet water years. It is important to consider that model performance varies with the climatic conditions, especially with those climatic conditions that differ the most from those used during model calibration.

Overall, the panel found that the guiding features and modeling framework are appropriate, and the selected numerical models are adequate for most intended applications. However, the panel

also identified features that may limit the model framework to fully reach the guiding features. The panel suggests that USBR:

(1) Further assist engagement with interested parties by:
    a. Providing a high-level overview that describes model features (i.e., your conceptual model showing entities and relationships), model events (i.e., dynamic conditions or scenarios that affect system behavior), and model processes (i.e., the physical, chemical, and biological relationships that describe model features).
    b. Providing a description of the dynamic characteristics of the target performance of the modeling, as it depends on water-year climatic conditions and/or water availability.
    c. Focusing on, documenting, and explaining model performance rather than model validation. This would help strengthen confidence and trust among the interested parties that use the model to develop scenarios and operational scenarios.
(2) Improve visualization and presentation of the model calibration and performance. Improve the description on how model calibration was performed and under what scenarios. Discuss model performance during critical periods and when it performs the least accurately.
(3) Provide model performance also at the framework scale. Most of the calibration and evaluation has been presented at the element model scale.

We address these issues in the following sections, where we provide additional context and information about these overarching suggestions.

Question 1. *Does the modeling design (e.g., model selection, framework) include the necessary processes and resolution (spatial and temporal) to represent the short-term and long-term temperature dynamics expected in the reservoir and river environments throughout the CVP project area?*

The document and presentation would benefit from a high-level overview of the project and its goals. This overview would provide the Panel, as well as interested parties, with the motivation and challenges associated with water-temperature modeling. We suggest clarifying the short- and long-term goals of the program to facilitate communication and awareness of the problems facing USBR.

Suggestion 1. High-Level Overview

To assist external reviewers and interested parties, we suggest that USBR start with a high-level overview that describes model features (i.e., your conceptual model showing entities and relationships), model events (i.e., dynamic conditions or scenarios that affect system behavior), and model processes (i.e., the physical, chemical, and biological relationships that describe model features).

- What are the Model Features? You may wish to show a conceptual model.
- What are the Model Events? You may wish to show the dynamic conditions that affect the model.
- What are the Model Processes? You may wish to show the physical, chemical, and biological relationships that describe model operations.
- What is the regulatory and operational context for the model framework and how will it be used by USBR?
- What temporal, geographic, and biologic metrics are critical to understand and target accurately?

Suggest development of clear maps and model diagrams, including:

- Locations, times, and thresholds that determine system operation.
- System-wide (e.g., illustrating connections between different model elements, key control points, etc.).
- Individual model elements (e.g., a diagram of the Lake Shasta CE-QUAL-W2 model)
- Map of all data sources, potentially linked to a table providing the period of record for each source.

- Suggest replacing use of the word "Sq___" for all references to Branch 2 in the Shasta CE-QUAL-W2 model or other usage with the appropriate updated place name.[1] [2]

Specifically, this orientation section would clarify whether the following hydrologic components are important, and if so, how they would be conceptualized in the model:

- Groundwater inflows, hyporheic exchanges.
- Point and nonpoint source inflows.
- Riparian and hyporheic exchanges (e.g., Turtle Bay).
- Solar insolation and riparian shading.
- Thermal mass of benthic sediments (both in reservoirs and channels).
- Linkages to other factors affecting fish recruitment (e.g., benthic-sediment composition due to peak flow scour and deposition).
- Is incidental leakage observed through multilevel discharge structures, and if so, how are they modeled?
- How is the thermal curtain feature modeled in the 2D and in the 1D?

Suggestion 2. Risk-Informed Analysis

Specific goals should be identified for proposed management activities that describe desired outcomes, as well as undesired outcomes. The ability to achieve these goals should be quantified using specified Performance Measures (e.g., regulatory thresholds or operational outcomes). The model should then be used to determine the frequency distribution of the Performance Measures.

From this, scenarios that either achieve or fail-to-meet the required performance measures can be identified. The overall risk can then be quantified using the frequency of each outcome weighted by the consequences of failure to meet performance measures.

We suggest that USBR provide a Risk-Informed Performance Assessment

- What are the Performance Measures (regulatory and operational)?
- What scenarios are expected to cause the inability to meet Performance Measures?
- What are the consequences of failure to meet Performance Measures?
- What metrics are suggested by the US National Marine Fisheries Service (NMFS) and US Fish and Wildlife Service (FWS).

[1] https://www.doi.gov/pressreleases/interior-department-completes-removal-sq-federal-use
[2] https://edits.nationalmap.gov/apps/gaz-domestic/public/all-official-sq-names

Figure 1 provides an example flowchart for conducting a risk-informed analysis developed for commercial nuclear power plants (ANS/ANSI 2.17, 2010). Note that this analysis includes characterization, monitoring, and performance modeling to assist in scenario evaluation.

For example, two management activities could be evaluated, one management action that fails to meet performance measures 80% of the time, but with inconsequential biological effects, while a second management action fails 10% of the time with highly consequential effects.

Risks (i.e., consequences weighted by their frequency) should be evaluated over long periods to allow for ergodic climatic conditions (e.g., multiple realizations that include variable flood and drought sequences). An important component of the risk-informed approach is to focus on collecting datasets that represent conditions where the greatest risk is anticipated, as well as calibration and evaluation of model performance during these conditions. Additionally, the modeling should be used to identify additional information needs (hard and soft data) to better evaluate model performance.

Suggestion 3. Alternative Management Plans

Models should be designed to evaluate alternative management activities, including alternative timing, duration, and magnitudes of discharges, as well as their sources (e.g., Trinity, Shasta, Whiskeytown). A possible framework for considering these factors is illustrated in the Colorado Water Conservation Plans (2015, 2023) that articulate the legal and institutional setting, inventory of resources, as well as scenario planning. Plans should also include defined metrics for use in evaluating the performance of implemented plans.

Citations

ANS/ANSI 2.17 (2010) Evaluation of subsurface radionuclide transport at commercial nuclear power plants. American Nuclear Society, 35 pp.

Colorado Water Conservation Board (2015) Colorado Water Plan. Colorado Department of Natural Resources.

Colorado Water Conservation Board (2023) Colorado Water Plan – Draft. Colorado Department of Natural Resources.

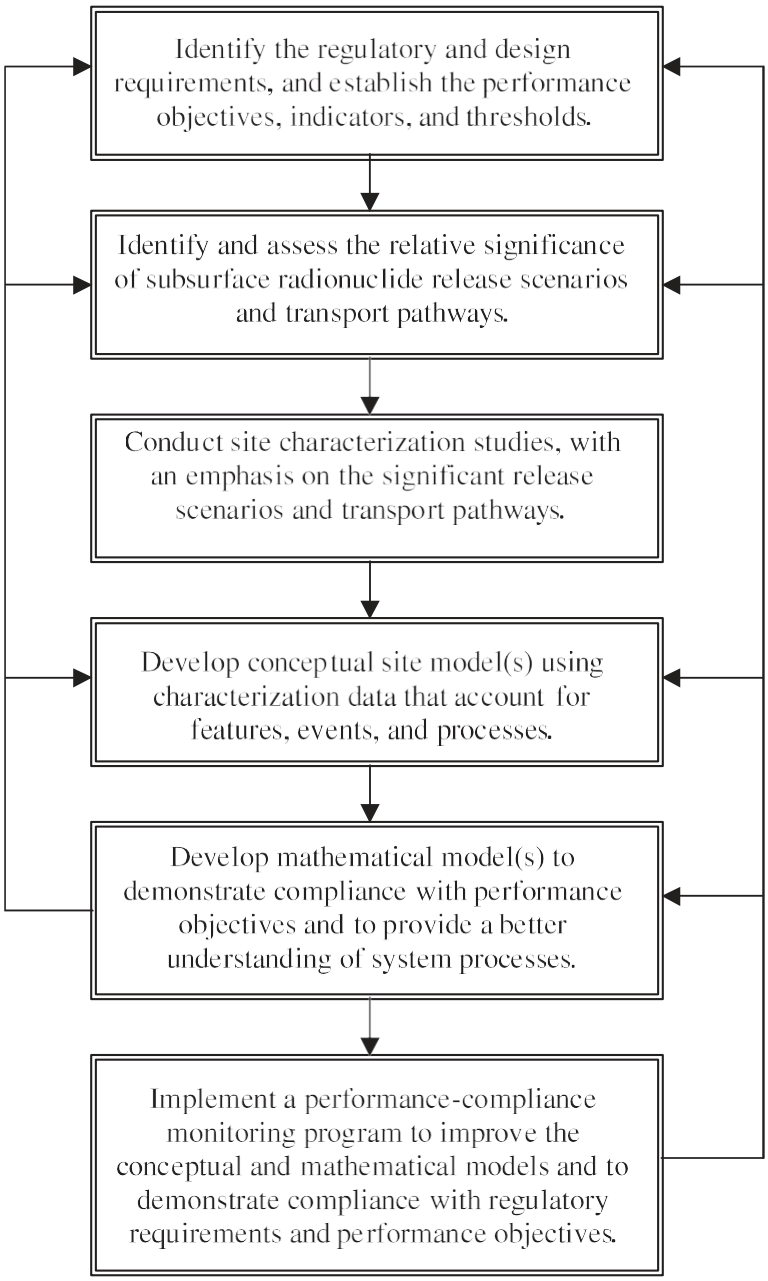

Figure 1 – Flowchart describing performance assessment activities and the relationships among these activities

(Source: ANS/ANSI 2.17, 2010)

Question 2. *Are the models adequate for describing water temperature during extreme hydrologic/storage conditions (e.g., droughts/low storage)?*

Models calibrated and validated thus far appear reasonable and their performance appears to be generally within the 1°C mean absolute error (MAE) standard often applied to CE-QUAL-W2 model implementations to indicate a "good" model (Sullivan et al., 2007). Exceptions to this broad metric, however, indicate that further refinement may be warranted.

For example, the Shasta Lake CE-QUAL-W2 model performance declines in Fall 2015 (a year of exceptional drought), reporting MAE of as high as 2.04°C through November. The reported profile comparison suggests that CE-QUAL-W2 simulates the thermocline depth to be shallower than measured, which suggests that either a) the wind-sheltering coefficient in low-storage conditions requires adjustment, b) inaccuracies in the width of the withdrawal zone from the temperature control device, or c) other considerations.

In general, however, the panel feels more information is required to determine the "adequacy" of model performance in extreme hydrologic/storage conditions. The panel suggests two approaches for improving the evaluation of model performance under drought conditions.

First, additional context for and model evaluation under drought and other low-pool conditions. While goodness-of-fit statistics and comparison figures are provided for the calibrated and verified years, the panel suggests additional effort to interpret these statistics and compare them to a range of hydro-climatologic conditions. For example,

- How does low-pool drought performance compare to other hydroclimatic conditions (e.g., a cool and wet year)?
- Of the calibrated/verified model years, which are considered to be droughts/low storage?
- In what ways does performance degrade during low-flow conditions?

With considerable effort, a reviewer or model user could obtain this information from the provided plots and statistics, but it is not easily parsed and would likely be onerous for many users. Both for the purposes of better understanding model performance and increasing public trust in model performance, the panel suggests providing an explicit comparison and discussion of model results in low-flow years, including a discussion of how and where errors increase under low-flow conditions, for example:

- How does the model bias change?

- How far downstream are errors in the thermal structure of the reservoir propagated before they are lost to environmental heat exchange, tributary inflow, etc.?
- Under what conditions and during what seasons are model results most robust?
- Under what conditions and during what seasons are they least robust?

Second, the panel suggests additional thought and discussion into what “adequate” means and how it can be communicated. While a range of performance statistics and metrics have been applied, no context is provided for their selection, and it is difficult to evaluate their meaning. Specifically, indices and metrics could be better focused on key components, including, for example, the position of the thermocline (which provides information on the pool-water volume), performance at specific control points downstream, performance relative to individual biological thresholds, etc. What matters, to whom, and when? Such an effort would allow modelers to develop better intuitive understanding of the model performance and also better frame discussion around the potential and limitations of the model capabilities for other users and the public.

Finally, the panel suggests providing additional information on the CE-QUAL-W2 models to better evaluate their performance. How will the use of average wind-sheltering coefficients for the CE-QUAL-W2 models affect their ability to forecast extreme events, which will have a significant impact on reservoir stratification? Additionally, because both outflow and reservoir elevation are specified as boundary conditions, goodness-of-fit statistics and measured vs. model plots of these parameters have limited utility in evaluating model fit. What is the size of the distributed tributaries applied in the models relative to total inflow and outflow? Are they explainable?

It would also be helpful if USBR further clarifies how the reservoir model will be used in long-term planning for critical conditions (e.g., drought conditions with low flows, small cold pool).

Specific clarifications would address:

- Have (or could) alternative flow and reservoir-storage scenarios beyond those in the historical record been developed and evaluated?
- Have (or could) Monte-Carlo methods using a distribution of flow and reservoir-storage scenarios been used to evaluate ensemble forecasts based on past and potential future modeling scenarios?
- Have (or could) long-term precipitation predictions be used in a Bayesian framework to identify potential scenarios?

- Have (or could) long-term monitoring data been used to improve model predictions at short- and long-term forecast horizons?

Citations

Sullivan AB, SA Rounds, S Sobieszczyk, HM Bragg (2007) Modeling hydrodynamics, water temperature, and suspended sediment in Detroit Lake, Oregon: U.S. Geological Survey Scientific Investigations Report 2007-5008, 40 p. https://pubs.usgs.gov/sir/2007/5008/pdf/sir20075008.pdf

Question 3: *Are unique features (i.e., selective withdrawal devices, thermal curtains, and submerged structures) adequately represented?*

The modeling framework approach separates the system into key elements, e.g., rivers, tunnels, and reservoirs. Each element function is then simulated with a selected model informed by external forcing and other element outputs. This results in the integration and use of multiple element models, which allows dynamic flow of information among models.

This modular approach has the advantage of using models of different spatial and temporal resolutions and scales for selected elements, e.g., a reservoir. The CVP Water Temperature Modeling Platform (CVP-WTMP) has several elements that include reservoirs, rivers, lakes, tunnels, and canals. Rivers have been modeled with a 1D longitudinal model, whereas reservoirs are modeled with a 1D vertical model or 2D transversally averaged model depending on the simulation objectives.

This approach provides great flexibility to address objectives with different temporal scales. However, it also presents challenges, because unique features within 1D and 2D models need to be modeled or parameterized differently due to averaging from 2D to 1D which results in loss of information, which then needs to be parameterized.

The river element is proposed to be modeled via HEC ResSim supported by tabulated information on flow hydraulics previously modeled in HEC-RAS 1D. This provides a fast and robust means to inform the temperature model of longitudinal and temporal hydraulic changes.

The 1D modeling is adequate for the main objective to quantify reach-scale stream water temperatures, which mainly change longitudinally due to advection and dispersion and several heat fluxes, which may include (1) latent heat flux, (2) sensible each flux, (3) long-wave radiation from the water body, (4) atmospheric and short-wave radiation from the sun and (5) bed heat conduction.

At this spatial resolution, local-scale effects (which may form spatio-thermal variability and/or thermal refugia formed by hyporheic exchange, river confluences and surface transient zones) such as expansion and local recirculating are typically accounted for by other processes within the one-dimensional heat transport equation.

Hyporheic exchange is water pumped in and out of the streambed sediment and may provide advective heat exchange between water column and pore-water. For the size of the Sacramento river its effect can be also lumped within other parameters (Marzadri et al., 2014; Tranmer et al., 2018), but could be important in small streams (e.g., King & Neilson, 2019).

Reservoirs within the CVP-WTMP will be modeled with HEC ResSim as a 1D vertical model and CE-QUAL-W2 as the 2D transversal model. These elements show several unique features with critical effects on water temperature releases. Those features include the a) temperature control device (TCD) at Shasta Dam, b) selective water withdrawal shutter at Folsom Dam, c) outlet facility at Whiskeytown Lake, d) thermal curtains in Whiskeytown Reservoir, and e) submerged dams in the New Melones Lake. All these features present challenges in modeling.

The TCD has several challenges, including a) leakages at the panels, b) large panels that withdraw water from broad vertical bands with potentially different water temperatures and c) lateral locations. Thus, their effects on water temperature in both 1D and 2D modeling has been parameterized depending on reservoir water stage and TCD operation.

This approach resulted in good match between predicted and measured temperatures both in calibration and validation stages. The provided documentation and presentation focused mainly on the unique characteristics of Lake Shasta. The other unique features were identified but their effects were not presented during modeling or parameterization within the modeling element but will be described in Phase II.

Citations

King, T. V., & Neilson, B. T. (2019). Quantifying Reach-Average Effects of Hyporheic Exchange on Arctic River Temperatures in an Area of Continuous Permafrost. *Water Resources Research*, *55*(3), 1951–1971. https://doi.org/10.1029/2018WR023463

Marzadri, A., Tonina, D., McKean, J. A., Tiedemann, M. G., & Benjankar, R. M. (2014). Multi-scale streambed topographic and discharge effects on hyporheic exchange at the stream network scale in confined streams. *Journal of Hydrology*, *519*(PB), 1997–2011. https://doi.org/10.1016/j.jhydrol.2014.09.076

Tranmer, A. W., Marti, C. L., Tonina, D., Benjankar, R., Weigel, D. E., Vilhena, L. C., McGrath, C. L., Goodwin, P., Tiedemann, M. G., Mckean, J. A. J., & Imberger, J. J. (2018). A hierarchical modelling framework for assessing physical and biochemical characteristics of a regulated river. *Ecological Modelling*, *368*, 78–93. https://doi.org/10.1016/j.ecolmodel.2017.11.010

Question 4. *Are available data sufficient for the development of the selected models and intended uses? Where data gaps have been identified, are the assumptions and methodologies used to address them suitable?*

While the data are extensive, both spatially and temporally, we have several questions and suggestions:

- From a risk-informed perspective, it would be helpful to identify those data that are most consequential in terms of evaluating system performance. That is, what data are needed to show success and/or failure, and which data are unimportant.
- While USBR is limited to existing data and operational constraints, can the modeling framework be used to identify additional resources that improve model forecasts, as illustrated in the lowermost box in Figure 1, "Performance Assessment Compliance Monitoring"?
- Is it correct that one weather station and depth-profiling sensor (temperature, dissolved oxygen etc.) are used for the model domain?
- Gap filling in the time-domain may not be as accurate or efficient as using frequency-domain methods (e.g., Fourier, HALS; Dilmaghani et al., 2007; Bessenbacher et al., 2021; Schweizer et al., 2021) that use the entire data record to reproduce the long-term behavior during periods with missing data.
- Cross-correlations between stations (e.g., Sq____Creek) using zero-lag correlations may not be as accurate or efficient as using transfer functions (i.e., convolution). Rather than fitting a regression equation for missing data of the form $y(t) = a + b\ x(t)$ that suggests that nearby flows are synchronous with the desired site, it would be more accurate to assume that there is a lag time between peak flows so that the response function is lag-dependent:

  $$y(t) = y_o + b_o\ x(t) + b_1\ x(t-1) + b_2\ x(t-2) + \ldots$$

  where lags are generally positive when low-order streams are used to predict flows in higher-order (peaks in lower order streams precede peaks in higher order streams) and negative when higher order streams are used to predict flow in lower order streams.
- Could existing watershed models be integrated to provide inflows from ungaged streams?
- How are long-term (decadal) data (e.g., snowpack, precipitation) considered for forecasting?

Citations

Bessenbacher V, SI Seneviratne, L Gudmundsson (2021) CLIMFILL: A framework for intelligently gap-filling Earth observations. EGU Geoscientific Model Development, 37 pp. https://doi.org/10.5194/gmd-2021-164

Dilmaghani S, IC Henry, P Soonthornnonda, ER Christensen, RC Henry (2007) Harmonic analysis of environmental time series with missing data or irregular sample spacing. Environmental Science & Technology 41(20):7030-7038. https://doi.org/10.1021/es0700247

Schweizer D, V Ried, GC Rau, JE Tuck, P Stoica (2021) Comparing methods and defining practical requirements for extracting harmonic tidal components from groundwater level measurements. Mathematical Geosciences 53:1147-1169, https://doi.org/10.1007/s11004-020-09915-9

Question 5: *Are testing methods (calibration and evaluation) adequate to demonstrate confidence in model performance for the historic period?*

Most of the suggestions in response to this question rely on review of the Model Development document and the associated appendices that show model performance statistics (in figures and tables).

The panel suggests moving away from the 'validation' terminology and thinking more broadly about model evaluation. There is a strong argument that models cannot be 'validated'. For example, Oreskes et al. (1994) argue that

> "[...] modelers misleadingly imply that validation and verification are synonymous, and that validation establishes the veracity of the model. In other cases, the term validation is used even more misleadingly to suggest that the model is an accurate representation of physical reality. The implication is that validated models tell us how the world really is.
>
> [...] But the agreement between any of these measures and numerical output in no way demonstrates that the model that produced the output is an accurate representation of the system".

This is not merely a philosophical discussion; it matters because one of the goals of the WTMP project is to build trust with other interested parties.

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

That is not to say that models are not useful tools in water resources management (including temperature management), but it may be more productive and realistic to carefully select models that include the important processes, calibrate those models with available observations, and then use any additional observations to evaluate model performance with the goal of learning about the strengths and weaknesses of the model approach.

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

Different characteristics of the modeling system matter for the different model modes (real-time, seasonal, long-term) in which the WTMP will be used. Although question 5 is focused on the model performance for the historic period, it may be useful to consider which model performance characteristics will be most important for the different model modes.

The panel suggests focusing the calibration and model evaluation on quantities that relate more directly to the decisions that the model framework needs to support. At the same time, the panel suggests re-evaluating how to weight individual measurements during the calibration and model evaluation process.

Not all observations have the same weight in assessing model performance. For example, the panel suggests that model performance could be evaluated in terms of the size of the cold pool, the location of the thermocline relative to the intake / gate locations, the downstream temperature at selected target locations (e.g., Red Bluff and Bend Bridge), the amount of heating in the Spring Creek tunnel, etc. McMillan (2020) in a review paper about model evaluation in hydrology refers to such derived quantities as ‘hydrologic signatures’ and similar signatures could be defined or used for stream temperature.

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

The current calibration and model performance evaluation have been done at the element scale but not at the modeling framework scale. The panel suggests to extend this evaluation to a full prediction of temperature in the river system accounting for all operations upstream (Shasta,

Trinity, Whiskeytown, Lewiston, and Keswick). This will demonstrate the performance of the model framework and may highlight any challenges with uncertainty and error propagation. It also raises the question how to assess model performance when run as a system (as opposed to individual elements).

The panel suggests providing more detail about the model calibration process. The document provides little information about the calibration process.

- Was the calibration done manually or in an automated manner? What was the objective function for the calibration?
- Since there are multiple performance metrics, how were they used in the calibration?
- Were they combined into a single objective function for the calibration or was this a multi-objective calibration (if so, what are the trade-offs)?
- How many simulations were run and how did the performance metrics change as a function of the number of iterations?
- How did you determine that the calibration was finished?

In addition, it would be helpful to distinguish between

- **Code Verification:** Determine whether the code is functioning as desired using analytic solutions or by comparing results between multiple modeling systems
- **Model Accuracy:** Distinguish between parameter calibration errors and model prediction errors considering parameter covariance and equifinality.
- **Model Resiliency:** Identify the limits of model utility considering uncertainties in future conditions

Additionally, we:

a) Suggest using Mean and Standard deviation of residuals instead of RMSE.

   The documents use the following index for model performance: Mean Bias, MAE – mean absolute error, RMSE – root mean squared error, NSE – Nash Sutcliffe efficiency. The panel would suggest also reporting the mean and standard deviation of the error. This is because the mean and the standard deviation report two different types of error. The mean provides the bias, systematic error, whereas the standard deviation reports the random error. The RMSE is a lumped index that reports both types of errors, bias and random. Substituting RMSE with the standard deviation will not change your conclusions but strengthen them.

Note that these statistics assume homoscedastic residuals (i.e., the magnitude of the error is constant as a function of the magnitude of the observation). Yet many hydrologic variables are heteroscedastic, such as discharge where both observational and fitting errors increases with the magnitude of the observation. In these cases, a logarithmic transform is required to obtain homoscedastic residuals.

The temporal resolution of the temperature modeling should be reported. In the presentation, the resolution is hourly, which yields good threshold (Table 4-1).

b) Limit the number of significant digits (to two) of the correlation coefficient in regression equation (e.g., Water Temperature Modeling Platform: Data Development, page 5-8). It would also be helpful to include the standard error of regression coefficients along with their significance. Comparisons between observed vs predicted values may be better in a graph showing observed vs predicted so a good fit is along the 1:1 line. This also shows the range of variability of the quantity.
c) The model performance figures in the Model Development document and its appendices provide little insight into model performance. Rather than plotting measured and simulated curves on top of each, it would be more useful to plot the measured values on one axis (or in one panel) and then plot the difference between the measured and simulated values on a separate scale (or in a separate panel).
d) The panel also suggests refraining from plotting very high frequency values (e.g., hourly) for a long period (e.g., 1 year) in a single figure, because it provides no insight into model performance on the short time scales. It would be better to plot the daily values and then separately plot / assess how the model performs at the sub-daily time scale (for example by making a plot with the diurnal cycle by month or something like it).

   That way you can actually see whether there are structural problems at the shorter time scales (for example, whether the simulations lag the measurements or under/overestimate the diurnal maxima and minima). For an extreme example see Figure 4-18 in the Model Development document, which provides no information about model performance relative to the measurements.

e) The panel suggests that the authors be explicit about the timestep at which a model performance metric was calculated (hourly, daily, monthly, etc.). An NSE or KGE calculated at a daily timestep will have a different value (and expectation) than one calculated at an hourly timestep.

f) Because there is a strong diurnal and seasonal cycle for the water temperature as well as a strong seasonal cycle for streamflow, care should be taken in the choice of your model calibration and model performance metrics. For example, as Schaefli and Gupta (2007) note "In the case of strongly seasonal time series, a model that only explains the seasonality but fails to reproduce any smaller time scale fluctuations will report a good NSE value; for predictions at the daily time step, this (high) value will be misleading." As a result, it may be necessary to remove the periodic signal (season or diurnal) before calculating the performance metric.

g) Suggest discussing situations when the model does not perform well in more detail. These are the events from which you can learn the most about your model. For example, in the Model Development document (p.4-44), the challenge with the outflow temperatures from Shasta Lake in 2015 is of concern (the simulated cold pool was depleted too quickly), because it happened during the year with the highest outflow temperatures, which likely had the greatest impact on fish survival.

   Consequently, the statement that "This parameterization performs well for the majority of other simulated summer-fall periods" ignores what could potentially be a serious problem. It would be good to discuss whether the inability of the ResSim parameterization to capture the thermal structure of Shasta Lake in 2015 prevents it from being useful during conditions seen in 2015 and whether this requires further development.

   As an analogy, a model that predicts sunny weather in the desert at all times is likely to have excellent bias, MSE, NSE, etc., but will not be useful for making decisions under flash-flood conditions.

Citations

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

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

McMillan HK (2020) Linking hydrologic signatures to hydrologic processes: A review. Hydrological Processes. 34:13931409. https://doi.org/10.1002/hyp.13632

McMillan HK, SJ Gnann, R Araki (2022) Large scale evaluation of relationships between hydrologic signatures and processes. Water Resources Research 58(6):e2021WR031751. https://doi.org/10.1029/2021WR031751

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

Question 6. *Does the modeling documentation include adequate information, assumptions, and detail to allow for transparency and replication of model results?*

The ability to comprehensively document model algorithms, components, datasets, and performance is an industry-wide challenge. While the provided documentation is excellent, it appears to be *ad hoc*, and does not reference standard methods of model documentation. Like many other modeling efforts, this study does not provide a summary documentation of all modeling elements. Without a formal approach for standardizing documentation, it is difficult to identify the important elements that require documentation. Industry efforts to establish standardized model-documentation guidelines that specify the essential components of the model-documentation effort are still in their infancy.

A more rigorous and formal method for documenting modeling efforts has been developed in another hydrologic discipline (i.e., groundwater modeling). Within the groundwater profession, a wide range of ASTM guidance documents provide standard methods for proper model documentation, shown in the citations. Development of model documentation “best-practices” would substantially improve the robustness and adequacy of all modeling efforts. Not incorporating these practices compromises the utility and confidence in model predictions. Every effort should be made to adhere to professional standards using published guidance documents.

The panel would be most appreciative of information regarding:

- Is documentation available for other model elements beyond what has been provided (e.g., Trinity Basin models, etc.)?
- Will documentation of the model system be produced, including documentation of the Keswick W2 model performance using modeled inflow and temperature?

Citations

Standard guide for application of groundwater flow model to a site-specific problem (D 5447)

Standard guide for comparing groundwater flow model simulations to site-specific information (D 5490)

Standard guide for defining boundary conditions in groundwater flow modeling (D 5609)

Standard guide for defining initial conditions in groundwater flow modeling (D 5610)

Standard guide for conducting a sensitivity analysis for a groundwater flow model application (D 5611)

Standard guide for documenting a groundwater flow model application (D 5718)

Standard guide for subsurface flow and transport modeling (D 5880)

Standard guide for calibrating a groundwater flow model application (D 5981)

Standard practice for evaluating mathematical models for the environmental fate of chemicals (E 978)

Standard guide for developing conceptual site models for contaminated sites (E 1689)