Water Resources Research

RESEARCH ARTICLE
10.1029/2023WR035703

Special Collection:
Quantifying human interferences in hydrologic process modeling

Key Points:

- Generic reservoir models with transparent structures are fully coupled with rainfall-runoff models for streamflow simulations
- Fully coupled models may be reliably used in large-scale hydrological models without losing physical significance
- Coupled models are evaluated based on ability to represent distributional properties of observed flows using state-of-the art metrics

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

X. Cai,
xmcai@illinois.edu

Citation:

Vora, A., Cai, X., Chen, Y., & Li, D. (2024). Coupling reservoir operation and rainfall-runoff processes for streamflow simulation in watersheds. *Water Resources Research*, *60*, e2023WR035703. https://doi.org/10.1029/2023WR035703

Received 29 JUN 2023
Accepted 30 MAY 2024

Coupling Reservoir Operation and Rainfall-Runoff Processes for Streamflow Simulation in Watersheds

Anav Vora[1], Ximing Cai[1], Yanan Chen[1], and Donghui Li[1]

[1]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign (UIUC), Urbana, IL, USA

Abstract We assess the overall watershed system representation via fully coupling a generic reservoir operation model with a conceptual rainfall-runoff model. The performance of the coupled model is evaluated comprehensively by examining watershed outflow simulations, model parameter values, and a key internal flux of the watershed model (here reservoir inflow). Five published generic reservoir operation models are coupled with a watershed rainfall-runoff model, and results are compared across the coupled models and one additional model called ResIgnore that ignores reservoir operation. Traditional loosely coupled watershed hydrologic models (where calibrated inflow is routed through reservoir operation models) are used as baselines to examine the differences in simulation performance and parameterization obtained from the fully coupled models. We find that fully coupling the Generic Data-Driven Reservoir Operation Model (GDROM) and the Dynamically Zoned Target Release (DZTR) reservoir operation models with the rainfall-runoff model obtains robust simulations of watershed outflow with realistic parameterization, suggesting that they can be reliably integrated into large-scale hydrological models for simulating streamflow in heavily dammed watersheds. Our results also show that compared to ResIgnore, the fully coupled watershed models more accurately simulate the entire distribution of watershed outflow, obtain more realistic values of model parameters, and simulate reservoir inflow with higher accuracy. Finally, we note that the prediction intervals of watershed outflow obtained from the GDROM- and DZTR-based fully coupled models consistently envelop observed watershed outflow across the study watersheds, indicating that GDROM and DZTR can be suitable reservoir components of large-scale hydrology models.

Plain Language Summary Reservoir operations greatly influence streamflow in heavily dammed watersheds, and hence incorporating a realistic reservoir component in watershed hydrological models to simulate the impacts is important. Recent efforts have greatly advanced generic reservoir operation model development. We couple various generic reservoir operation models with a rainfall-runoff model to develop watershed hydrological models. We use a comprehensive evaluation method with state-of-the-art metrics to examine the performance of the coupled watershed models in terms of simulated watershed outflows, model parameterization, and simulated internal variables. Fully coupled watershed models based on recently developed reservoir operation models obtain significantly improved representations of the watershed system (i.e., reservoir operation + natural rainfall-runoff processes) compared to models that ignore reservoir operations or use simplified representations of reservoirs.

1. Introduction

Reservoirs constructed for irrigation, flood control, water supply, hydropower generation, etc., have had a pronounced effect on river flow regimes and terrestrial hydrology (Ekka et al., 2022; Haddeland et al., 2014; Kåresdotter et al., 2022; Poff et al., 1997). Consequently, the development of reservoir operation models that can be integrated with rainfall-runoff process simulation in hydrologic models has received significant attention (Chen et al., 2022; Coerver et al., 2018; Dang et al., 2020; Döll et al., 2003; Hanasaki et al., 2006; Haddeland et al., 2006; Meigh et al., 1999; Solander et al., 2016; Turner, Steyaert, et al., 2021; Van Beek et al., 2011; Wisser et al., 2010; Yang et al., 2019; Yassin et al., 2019; G. Zhao et al., 2016). Developing a realistic reservoir component for hydrological models is paramount for hydrologic process simulation accuracy and further for water resources planning and management. However, the lack of details regarding reservoir operations, especially the impact of operators' behaviors that are usually intractable, has impeded the development of realistic generic watershed hydrologic models. In addition, the lack of a transparent easy-to-understand structure and heavy data requirements of certain complex reservoir operation models limit the potential for coupling with rainfall-runoff process simulation. Finally, watershed hydrologic models that couple unrealistic reservoir operation models could end with unrealistic model parametrization and biased simulations of internal fluxes (Dang et al., 2020; Hejazi, Cai, & Borah, 2008). In this paper, we examine the representation of the overall watershed system (i.e., reservoir operation + rainfall-runoff processes) obtained by coupling recently developed generic reservoir operation models with a rainfall-runoff model. The performance of the coupled models is studied in terms of watershed outflow, watershed model parameters, and a key internal flux of the watershed model (the reservoir inflow simulated by the rainfall-runoff model applied to the drainage area of the reservoir).

Lack of details on reservoir operation has led to some outstanding modeling issues, such as ignoring the existence of reservoirs (Abbaspour et al., 2015; De Paiva et al., 2013), assuming no human control on reservoir storage (e.g., a scheme implemented in the National Water Model which utilizes level pool routing) (Döll et al., 2003; Gochis et al., 2020; Meigh et al., 1999), and setting up simplified reservoir operation rules for routing flows through a reservoir (Hanasaki et al., 2006; Wisser et al., 2010). The simplified approaches for modeling reservoirs range from defining an empirical linear relationship between reservoir inflow and reservoir release (as in Wisser et al., 2010) to specifying total releases for the entire year based on storage at the beginning of the year (Hanasaki et al., 2006). Simplified approaches have found widespread applications in global modeling studies. For example, studies such as Hanasaki et al. (2008), Döll et al. (2009), Biemans et al. (2011), Pokhrel et al. (2012), and Yoshikawa et al. (2014) have adapted the simplified reservoir operation model developed by Hanasaki et al. (2006) into global hydrological models; similarly, the simple reservoir operation model proposed by Wisser et al. (2010) has been adopted by Fekete et al. (2010) for modeling global hydrology. Optimization schemes have also been proposed for simulating reservoir releases and adopted widely (Bierkens et al., 2019; Haddeland et al., 2006; Van Beek et al., 2011). Unfortunately, “optimized rules” may not directly correspond to real-world practices and such optimization models may be limited by imperfect objective(s) and missing or incorrect variables and constraints. In general, reservoir operation models, both optimization models and simulation models, have limited representation of the behaviors and actual operations of reservoir operators (Hejazi, Cai, & Borah, 2008; Solander et al., 2016). While operators normally follow regulations (usually embedded in the predesigned reservoir operation rule curve), they have flexibility to use their own judgment and behavior in response to hydrologic variability, change, and uncertainty associated with reservoir release forecasts. Hence, reservoir operation models that can justifiably capture reservoir operators' behavior are required for inclusion in modeling watershed hydrology. In particular, large-scale hydrologic models usually include both rainfall-runoff processes in a river basin (crossing multiple watersheds) and human interferences such as storage regulation, and an appropriate reservoir component will be needed for more realistic overall watershed system representation.

A few studies in literature have examined the problems arising from coupling unrealistic reservoir operation models with rainfall-runoff process simulations (Dang et al., 2020; Hejazi, Cai, & Borah, 2008). Hejazi, Cai, and Borah (2008) showed that a watershed hydrologic model could end with unreasonable calibrated parameter values (e.g., curve numbers) that are out of their physical ranges, if human controls on reservoir storage were ignored. The problem could be mitigated by accounting for the human interferences associated with human regulation of a reservoir located in the study watershed. Similarly, Dang et al. (2020) calibrated the variable infiltration capacity (VIC) model to the Upper Mekong basin with and without a reservoir component. They found that the model trained without reservoirs mimicked dry season releases from hydropower reservoirs through unrealistic model parameters that increased soil water storage capacity, baseflow and infiltration. These deficits could be resolved by the inclusion of an explicit reservoir component in hydrological models. In general, ignoring storage regulation including both physical effect (such as those captured in level pool routing), and human control, can lead to poor reproduction of the seasonal differences between reservoir inflows and releases for meeting operational demands such as irrigation, hydroelectricity, water supply or flood control.

Auspiciously, with the growing availability of historical reservoir operation data, it becomes feasible to derive reservoir operators' behaviors in a region or country via data mining and machine learning methods (Chen et al., 2022; Giuliani & Herman, 2018; Hejazi, Cai, & Ruddell, 2008; Q. Zhao & Cai, 2020). Recent efforts have attempted to incorporate reservoir operators' behaviors for developing generic reservoir operation models as a more realistic component of watershed hydrologic models. In particular, progress has been made to derive operation rules from observations of reservoir inflow, release, demand, and climate condition for many reservoirs serving various operational purposes (Chen et al., 2022; Coerver et al., 2018; Turner, Steyaert, et al., 2021; Yang et al., 2019; Yassin et al., 2019). These efforts recognize the fact that even if rule curves are available for a reservoir, directly implementing them into a model can ignore expected operational modification introduced by reservoir operators (Hejazi, Cai, & Borah, 2008; Solander et al., 2016). Chen et al. (2022) developed the generic data-driven reservoir operation model (GDROM) which can simulate daily reservoir operation dynamics. They used a data-driven approach to derive reservoir release rules from long-term daily operational records as a set of if-then conditions. Thus, GDROM benefits from having a transparent structure and can represent what reservoir operators actually did in response to intra-year seasonal variations, and inter-year changes in inflow and water demand. GDROM can also capture potentially different operational strategies used under drought or flooding conditions (Q. Zhao & Cai, 2020). Chen et al. (2022) successfully applied GDROM to over 450 reservoirs in the conterminous United States (CONUS). They also provided derived if-then rules for these reservoirs in an online repository in the form of text files for direct integration with watershed hydrological models (Li et al., 2023a). Another generic reservoir operations model, named dynamically zoned target release (DZTR) (Yassin et al., 2019), indirectly accounts for seasonally varying reservoir operators' behavior through model parameters that are derived from long-term observed records of reservoir storage and release. Yassin et al. (2019) have applied the DZTR model to 37 global reservoirs with reasonable performance. In addition, Turner, Steyaert, et al. (2021) derived a generic reservoir operation model called Inferred Storage Target and Release Functions (ISTARF) for 1,930 reservoirs in the CONUS, and provided model parameters for integrating ISTARF with hydrological models (Turner, Voisin, et al., 2021).

Compared to many other reservoir models that were developed for individual reservoirs (serving specific operational purposes such as hydropower, irrigation, flood control, etc.), GDROM, ISTARF, and DZTR models exhibit a generic and transparent structure describing reservoir operation, which makes it easy to incorporate the reservoir models into any rainfall-runoff model structure. In addition, many reservoir operation models require data such as reservoir bathymetry or downstream demands for determining release rules, and incorporating such models presents data requirement challenges. In contrast, GDROM, ISTARF, and DZTR models have relatively low and easily available data requirements, as described in the following section. Thus, GDROM, ISTARF, and DZTR have a strong potential for integration with generic rainfall-runoff process models and ultimately the development of generic hydrologic models of all spatial scales.

In this paper, we couple generic reservoir operation models like GDROM, ISTARF, and DZTR that have generic, transparent structures and can represent reservoir operators' behavior with a rainfall-runoff model. We then assess the overall representation of the watershed system obtained by the coupled watershed models. We also couple several other generic published reservoir operation models with the same rainfall-runoff model for comparing the coupled models with various reservoir operation components. A critical technical issue is the model performance assessment to illustrate the impact of a reservoir model coupled with the rainfall-runoff process on hydrologic process simulation accuracy. As discussed above, Hejazi, Cai, and Borah (2008) and Dang et al. (2020) proposed an approach focusing on the effect on model parameterization. Another option proposed by Khatami et al. (2019) is to examine the simulation performance of modeled internal fluxes of a watershed model, which has been adopted by previous studies focused on developing water quality models (Apostel et al., 2021; Wallington & Cai, 2023). We examine the overall watershed system representation comprehensively by studying watershed outflow simulations, model parameters, and internal fluxes of the watershed models. State-of-the-art model performance metrics are used to evaluate the coupled models in terms of both simulated watershed outflows and internal fluxes.

2. Methodology

Our analysis involves coupling a conceptual rainfall-runoff model, called F03+ here (D. Farmer et al., 2003), with five generic reservoir operation models for simulating streamflow in watersheds that drain into five reservoirs located in the CONUS (Section 2.1). We conduct a comprehensive evaluation of the coupled watershed models in terms of watershed outflow simulations, model parameters, and simulation of internal fluxes (Section 2.2). Aggregated goodness-of-fit (GoF) metrics and factors such as representation of the distributional properties of streamflow (Flow duration curves (FDC) and L-moments) are considered when evaluating watershed outflow and internal fluxes (Section 2.3).

2.1. Coupling Reservoir Operation Models With a Rainfall-Runoff Model

Generic reservoir models—GDROM (Chen et al., 2022), ISTARF (Turner, Steyaert, et al., 2021), DZTR (Yassin et al., 2019), HANA (Hanasaki et al., 2006), and WISS (Wisser et al., 2010) (Section 2.1.1) are coupled with

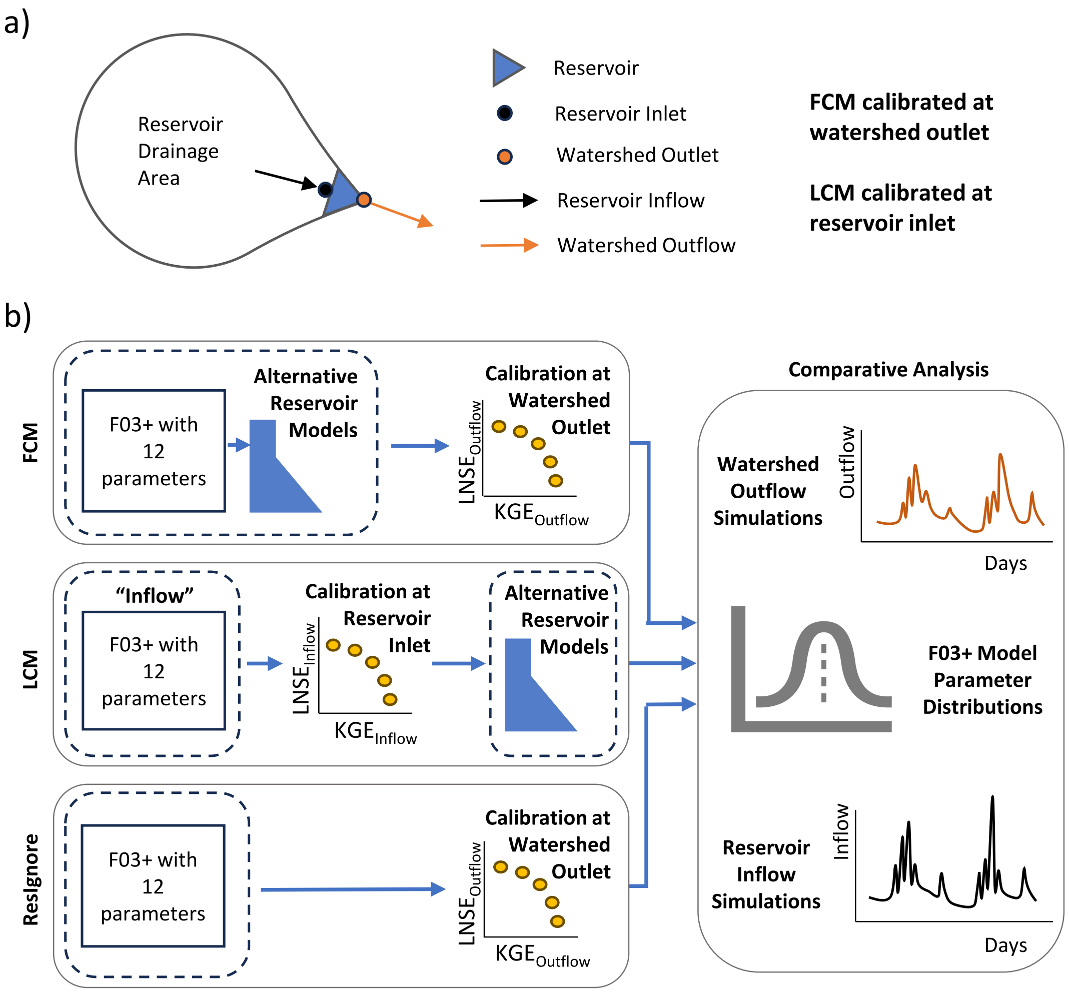

Figure 1. (a) A conceptual diagram showing the watershed system (reservoir + reservoir drainage area) and locations where FCMs and LCMs are calibrated; (b) the model evaluation technique used to examine the overall watershed system representation obtained by fully coupling alternative reservoir operation models (WISS, HANA, DZTR, ISTARF, GDROM) with F03+. For the FCMs, F03+'s parameters are adjusted jointly with the pre-trained reservoir operation models to maximize $KGE_{Outflow}$ and $LNSE_{Outflow}$, thus obtaining the best simulation of watershed outflow. For the LCMs, an "Inflow" model is first developed by adjusting F03+'s parameters to maximize KGE_{Inflow} and $LNSE_{Inflow}$. The calibrated inflow simulations from the "Inflow" model are then routed through the reservoir operation models. The ResIgnore approach ignores the existence of reservoirs and directly adjusts F03+'s parameters to maximize $KGE_{Outflow}$ and $LNSE_{Outflow}$. Watershed outflow simulations, model parameters, and reservoir inflow simulations obtained using all three approaches are compared.

F03+ (Section 2.1.2) to simulate daily flows in each watershed. Reservoirs in the coupled model are treated as nodes, a scheme implemented by G. Zhao et al. (2016). At every time step, the reservoir model receives simulated flows from F03+ as inputs. Subsequently, the reservoir model transforms the simulated inflows into reservoir releases; reservoir storage is updated according to the inflow and releases. Actually, the coupling procedure is identical to that used by Hanasaki et al. (2006) and Wisser et al. (2010).

Here, we make a distinction between fully coupled models (FCM) and loosely coupled models (LCM) (Figure 1a). In the context of our study, FCMs are defined as models in which F03+'s parameters are calibrated jointly with previously trained reservoir operation models to simulate watershed outflow. For LCMs, which have been used in most previous studies, reservoir inflow is simulated by F03+ applied to the drainage area of the reservoir, and the simulated inflow is routed through reservoir operation models to simulate reservoir release. Thus, the LCMs conduct data exchange between two independently calibrated models: "Inflow" models (F03+, calibrated to reservoir inflow) and reservoir operation models. For the watersheds we analyze, the reservoirs are

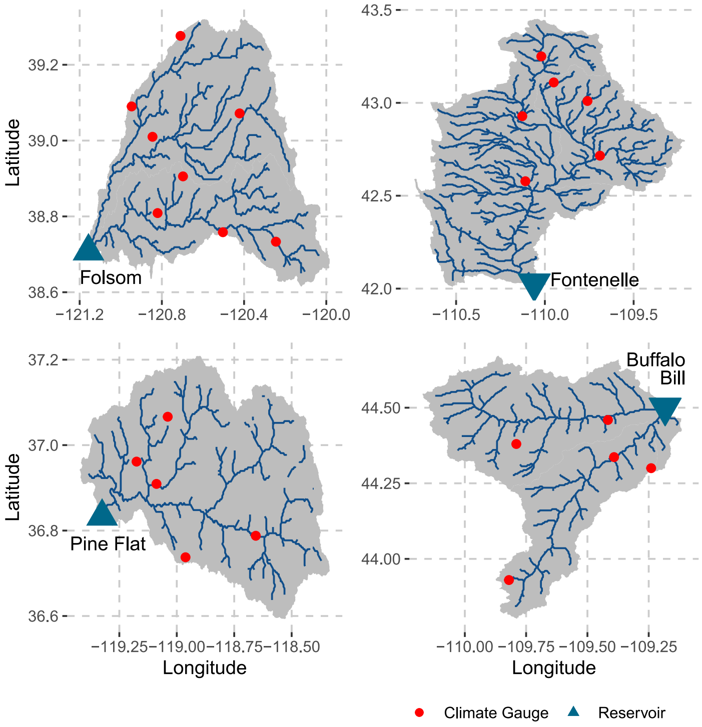

Figure 2. Watersheds modeled in our analysis. A reservoir is located at the outlet of each watershed. Hence, reservoir outflow is the same as watershed outflow in our analysis. Table 1 includes additional reservoir related details.

located at the watershed outlet. Hence, the simulated reservoir releases are the simulated watershed outflows from the coupled watershed model (Figure 2).

In this study, we use LCMs as baselines to evaluate differences in simulation performance and parameterization arising from the use of FCMs. We expect a trade-off between the LCMs and FCMs in terms of watershed outflow simulations and model parameterization. As the "Inflow" models used for LCMs are calibrated to the drainage area upstream of a reservoir, they are unaffected by reservoir operation, and thus obtain relatively realistic parameters representing the "natural condition" (i.e., without storage regulation). On the other hand, as the FCMs are calibrated using watershed outflow, they may perform better in simulating watershed outflows than LCMs. However, the parameters of FCMs and LCMs are likely different as the former are calibrated with watershed outflow and the latter are calibrated using reservoir inflow. We observe that FCMs and LCMs should not demonstrate extremely different parameterization, as this indicates unrealistic trade-offs made by FCMs to achieve improved watershed outflow simulations. Ideally, FCMs should improve watershed outflow simulations and demonstrate a parameterization resembling that obtained by the "Inflow" models, which are part of the LCMs. We argue that, in general, such a trade-off should be addressed in calibrated large-scale hydrological models that incorporate human interferences (especially storage regulation) as an internal component along with natural hydrologic processes.

2.1.1. Generic Reservoir Operation Models

GDROM derives reservoir release rules from operational records as a function of inflow, current storage, palmer drought severity index (PDSI) and the day of the year (DOY) (Chen et al., 2022). Its predictors such as DOY and PDSI implicitly account for factors such as seasonality in reservoir operation, and the effects of dryness and wetness conditions on reservoir operation strategies. Being derived from operational records, GDROM also incorporates reservoir operators' experience regarding decisions taken under specific inflow, storage, dryness, and wetness conditions at different times of the year. For a particular reservoir, GDROM provides a set of regression trees and a classification tree. The regression trees are termed operation modules. Each operation module predicts reservoir release based on specific inflow and storage conditions. The operation module to be applied for release prediction is selected using the classification tree based on inflow, storage, PDSI and DOY. As decision trees have a transparent structure, GDROM also does not suffer from the lack of model interpretability which are issues faced by other data-driven models such as neural networks.

Turner, Steyaert, et al. (2021) developed the ISTARF model by proposing that reservoir storage targets and releases can be modeled using harmonic functions. They fitted harmonic functions to observed reservoir storage records for defining a normal operating range (NOR) of storage for each reservoir, which varies weekly. For conditions where the reservoir storage falls within the NOR, Turner, Steyaert, et al. (2021) predicted a weekly value of reservoir release by fitting a harmonic function to observed reservoir releases. They also developed a linear adjustment equation to adjust the predicted weekly release depending on the actual reservoir inflow and current storage. Turner, Steyaert, et al. (2021) trained ISTARF for 595 data-rich reservoirs and extrapolated them to 1,335 data-scare reservoirs in the CONUS, creating a data set of 1,930 reservoirs with ready-to-apply inferred policies (Turner, Voisin, et al., 2021).

The generic DZTR model developed by Yassin et al. (2019) is based on the premise of dividing the reservoir storage into various operational zones based on long-term inflow and release data. Reservoir release is then related to reservoir storage using piecewise-linear functions for each of the defined zones. Yassin et al. (2019) improved upon promising previous models that were based on the same premise but had data requirement limitations or were applicable to reservoirs serving only specific operational purposes. For example, a previous model developed by G. Zhao et al. (2016) required bathymetric data to define operational zones and water demand data to determine releases; another model proposed by Dang et al. (2020) was applicable to hydropower purpose only. To address such issues, Yassin et al. (2019) proposed a generalized algorithm to derive monthly varying operational zones and target releases for each zone from long-term observed data of reservoir storage and releases. By virtue of deriving model parameters (operation zones and target releases) from operational records, the DZTR model indirectly accounts for reservoir operators' behaviors responding to various inflow and climate conditions. Yassin et al. (2019) also developed a version of DZTR with optimized model parameters, however, we adopt the generalized DZTR model in our analysis due to lower data requirements and a transparent structure. In this study, we determine the model parameters for the DZTR model, that is, operational zones and release targets for each month, using data for the same period used for training GDROM models.

The Hanasaki et al. (2006) model is generic and only requires information regarding reservoir storage capacity and inflow to simulate releases for reservoirs serving primarily non-irrigation purposes (hydropower, flood control, etc.). We adopt a modified version of the model proposed by Hanasaki et al. (2008), HANA henceforth. The HANA model determines total release in an operational year depending on the storage level at the beginning of that operational year, that is, reservoir storage in the first month when mean monthly inflow falls below mean annual inflow. If reservoir storage at the beginning of an operational year is greater than normal, releases are increased throughout the year and vice versa. The target reservoir release suggested in the HANA model for non-irrigation reservoirs is the mean annual inflow. This target release is adjusted depending on storage at the beginning of the operational year. For reservoirs with small storage capacities (Section 3), additional adjustments to the target release are made depending on actual daily inflow into the reservoir.

The Wisser et al. (2010) model, referred to as WISS in our analysis, proposes a piece-wise linear relationship between reservoir release and reservoir inflow applicable to all reservoirs. The constants of the equation were found empirically using operational data of 30 global reservoirs. Note that reservoir storage was not used as a predictor for reservoir release in the WISS model. We observe that the piece-wise relationship suggested by WISS would predict a reservoir release equal to the mean annual reservoir inflow if current reservoir inflow is much smaller than the mean annual inflow. Correspondingly, for large reservoirs, a constant release equaling the mean annual inflow may be simulated by WISS during the dry season.

We also consider a case, defined as “ResIgnore” in this study, that ignores reservoir operation and directly applies F03+ for simulating watershed outflow.

We note that the total reservoir release may include diversions for use in addition to streamflow releases to the downstream of the reservoir. We notice that not explicitly accounting for the diversions when training FCMs could result in systematic errors in storage simulation at each timestep (via mass balance) and their propagation crossing times. As watershed outflow simulations from the FCMs depend on simulated reservoir storage, systematic errors in storage simulations would lead to the accumulation of large biases in simulated watershed outflow. Hence, to account for diversions, we develop a two-step approach. In the first step, GDROM, ISTARF, and DZTR are used to identify a relationship between observed current conditions (storage, inflow, PDSI, DOY) and observed river releases. Similarly, in the second step a relationship between observed current conditions and observed total releases is determined. In the coupling, the relationship found from the first step is used to simulate river streamflow downstream of a reservoir depending on current conditions; while the relationship obtained from the second step is used to simulate the total reservoir release, which is used for updating reservoir storage for the next time step. We recognize that the above "two-step" approach can only be applied if data on diversions occurring from a reservoir are available. The WISS model directly predicts river release and does not require storage as an input for predicting release. Hence, no modifications are required for applying the WISS model to reservoirs where total release exceeds river release. On the other hand, the widely applied HANA model that was also developed to simulate river release requires reservoir storage as an input for predicting reservoir release. Unfortunately, the HANA model is theoretical, with the theoretical target river release being the mean annual inflow for non-irrigation reservoirs. The HANA model may be extended to determine theoretical diversions from the reservoir; however, this is beyond the scope of our paper. Hence, the HANA model is applied without any modifications following past literature that performs global hydrological modeling using the HANA model. We note that taking this approach can impact reservoir storage simulations, however, in our case, the issue only impacts one of the study areas we examine - Buffalo Bill Reservoir, Wyoming (Section 3).

2.1.2. Rainfall-Runoff Model

Our analysis adopts F03+, a modified form of the simple conceptual rainfall-runoff model developed by D. Farmer et al. (2003). The modifications include a multiple bucket formulation based on the Xinanjiang distribution introduced by Bai et al. (2009) to represent spatial variations in runoff generation and additional parameters to represent plant phenology suggested by Sawicz (2013). F03+ has 12 parameters spread across four modules representing snow, vegetation, near surface soil moisture accounting and deep recharge and routing. The snow module includes the parameters degree day factor (DDF), threshold temperature for snow formation (TTH), and base temperature for snow melt (TB). The vegetation module parameters are fractional cover of deep-rooted vegetation (M), and minimum and maximum leaf area index (LAImin and LAImax). The near surface soil moisture accounting module has the parameters depth of soil store (SB), shape factor (B), and field capacity threshold (FC). The soil storage is divided into saturated and unsaturated storages based on the parameter FC. The deep recharge and routing module comprises of the parameters deep recharge coefficient (KD) which controls the rate of percolation to the deep groundwater store from the saturated store, the recession coefficient for saturated soil (ASS) which controls the rate of subsurface runoff generation from the saturated store, and the baseflow recession coefficient (ABF) which controls the rate of drainage from the deep groundwater store as base flow. A detailed description of the F03+ model can be found in the supplementary section of Vora and Singh (2022).

The simple and parsimonious structure of F03+ allows easy examination of parameter differences arising from the use of different reservoir operation models for coupled model development. F03+ can be easily applied to multiple sub-watersheds to simulate flows in a large watershed with reservoir impacts. It is also benefitted from low data and computational requirements. Any other rainfall-runoff model may also be used to develop a coupled model and reproduce our results. A priori parameter value ranges for parameters DDF, TTH, TB, LAImax, LAImin, B, and KD of F03+ are obtained from literature (Bai et al., 2009; D. Farmer et al., 2003). Feasible parameter value ranges for M, FC, SB, ASS, and ABF are derived from observed land cover characteristics, soil properties, and recession curve analysis using the procedure described in Singh et al. (2014).

2.2. Evaluation of System Representation

Our comprehensive model performance evaluation method is focused on examining the overall representation of the watershed system (i.e., reservoir + rainfall-runoff components) obtained by fully coupling reservoir operation models with F03+. We consider three aspects to evaluate the system representation: watershed outflow, model parameters, and internal fluxes. The system representation evaluation is based on calibrated models, including FCMs, LCMs and ResIngore. We use the NSGA-II multi-objective optimization algorithm with the Kling Gupta Efficiency (KGE) (Gupta et al., 2009) and log-transformed Nash Sutcliffe Efficiency (LNSE) (Nash & Sutcliffe, 1970) objective functions for model calibration. The decision variables of the optimization algorithm are the 12 parameters of F03+ (Section 2.1.2). Using trial-and-error, a population size of 20 is selected and evolved over 250 generations to obtain Pareto-optimal solutions; the configuration resulted in consistent Pareto-optimal solutions over different runs. The optimization algorithm is run five times with different random seeds to account for effects of randomness in the initial population. Pareto optimal solutions that produce streamflow simulations better than the mean of the observed streamflow at every time step are considered acceptable and selected for further analysis. Knoben et al. (2019) found that using the mean flow as a benchmark yields a KGE value of −0.41 and an LNSE value of 0. Hence, we consider that Pareto optimal solutions with KGE > −0.41 and LNSE > 0 are acceptable. Note that KGE and LNSE values are calculated using daily observed and simulated streamflows. The GoF estimators KGE and LNSE are selected to capture high and low flow simulation performance, respectively, following Dang et al. (2020) who also considered high flow and low flow objectives for FCM development. Reasonable simulations of both high and low flows are important for riverine ecosystems downstream of a reservoir (Hoang et al., 2016; Poff et al., 1997). The KGE and LNSE estimators are selected also due to their lower variability across samples than other popular estimators such as the classical Nash Sutcliffe Efficiency (NSE) (Clark et al., 2021; Lamontagne et al., 2020). Finally, including LNSE for model calibration may also address the issues of heteroskedasticity in model residuals to some extent (Kuczera, 1983; McInerney et al., 2017).

For FCM and ResIgnore, the model parameters are optimized to maximize watershed outflow GoF estimators (i.e., $\mathrm{KGE}_{\mathrm{Outflow}}$ and $\mathrm{LNSE}_{\mathrm{Outflow}}$). For LCM, the "Inflow" model is first calibrated by maximizing the reservoir inflow GoF estimators (i.e., $\mathrm{KGE}_{\mathrm{Inflow}}$ and $\mathrm{LNSE}_{\mathrm{Inflow}}$) (Figure 1b). Calibrated inflow simulations are then routed through trained reservoir models. Note that the reservoir model parameters remain fixed in both FCM and LCM calibrations. Correspondingly, the FCM calibration procedure determines F03+'s parameters that work best in conjunction with trained reservoir operation models.

Our evaluation of the watershed system begins with the examination of the system output, that is, watershed outflow simulations. Ideally, the FCMs should achieve better simulations of watershed outflow than the LCMs in both calibration and validation periods. Simulation performance of watershed outflow is studied in terms of aggregated GoF estimators (i.e., $\mathrm{KGE}_{\mathrm{Outflow}}$ and $\mathrm{LNSE}_{\mathrm{Outflow}}$), as well as the ability to represent the distributional properties of observed streamflow through errors in estimated L-moment ratios and the FDC (Section 2.3).

We use kernel density estimation to visualize the acceptable Pareto optimal FCM, LCM, and ResIgnore parameters. The calibrated parameters are compared to understand the differences in model parameterization occurring with the use of a particular modeling approach. In particular, we note that ignoring reservoir operation, as with ResIgnore, introduces structural errors in the watershed model leading to parameter values that deviate from their true natural values (Dang et al., 2020; Hejazi, Cai, & Borah, 2008). We hypothesize that resolving structural errors by adding reservoir operation models for FCM development results in improved watershed model parameterization compared to ResIgnore; moreover as mentioned earlier, ideally, FCMs should have a parameterization resembling that obtained by the "Inflow" models, which are part of the LCMs. These hypotheses will be tested and validated by results of this study.

In addition, the effects of differences in parametrization arising from the different modeling choices can also be explored by studying internal fluxes of the watershed model. A convenient internal flux in our framework is the inflow to the reservoir. We thus examine the aggregated GoF estimators associated with reservoir inflow and metrics quantifying the distributional properties of reservoir inflow.

2.3. Evaluation of Model Ability to Represent Streamflow Distributional Properties

We examine the ability of the coupled watershed models to represent the distributional properties of observed streamflow. This analysis includes examining the watershed model's performance in reproducing the L-moments and FDC of observed streamflow.

We recognize that even after calibration, all hydrological models have residual errors, that is, differences in the observed and simulated responses. Consequently, during application, simulated responses from calibrated hydrological models have uncertainties due to unknown model errors (W. H. Farmer & Vogel, 2016). Application of coupled models for water management downstream of a reservoir warrants the addition of model errors back to the simulation by making a suitable assumption regarding the distribution of model errors (McInerney et al., 2017; Shabestanipour et al., 2023). Ignoring model errors can lead to unrealistic predictions of hydrologic extremes, and hence, for operational purposes, using a deterministic model (without adding model errors back to the simulation) is not recommended (W. H. Farmer & Vogel, 2016). We use a post-processing approach to add model error back to deterministic simulations and generate stochastic simulations of watershed outflow corresponding to each acceptable Pareto-optimal parameter set. The post-processing approach is based on the procedure developed by Shabestanipour et al. (2023) with only one difference; we assume that the log-transformed and differenced model errors have a Gaussian distribution with zero mean and constant variance. A detailed description of the post-processing approach we use is included in Supporting Information S1 Text S1.

To examine model performance in reproducing the L-moment ratios of observed streamflow, we compare the L-moment ratios of each streamflow simulation with the L-moment ratios of the observed streamflow (as simulated L-moment ratio minus observed L-moment ratio). The median difference in the observed and simulated L-moment ratios across all simulations is reported.

The FDC presents a relationship between specific values of discharge and the probability with which those discharge values may be equaled or exceeded, that is, the exceedance probability. Deterministic streamflow simulations are used to obtain the 50% confidence interval (75th minus 25th percentile of simulated flows) and 95% confidence interval (97.5th minus 2.5th percentile of simulated flows) of the FDC. Similarly, stochastic streamflow simulations are used to obtain the 50% and 95% prediction intervals of the FDC. We examine the extent to which the observed FDC is enveloped by the prediction or confidence intervals of the simulated FDCs obtained from various models, to understand whether using a specific modeling approach is associated with errors in the simulated flow at specific exceedance probabilities.

3. Study Watersheds and Data Sources

We simulate the hydrology of selected watersheds draining into five reservoirs in the CONUS - Folsom, Fontenelle, Pine Flat, Wappapello, and Buffalo Bill (Figure 2). The selected reservoirs differ in size and serve a range of operational purposes so that results may be applicable for large-scale hydrologic modeling (Table 1). Reservoir sizes are defined in relation to the mean total annual inflow into the reservoirs following Hanasaki et al. (2006). Reservoirs with storage capacities (SC) that are smaller than half of the mean total annual inflow (I_{mean}) are termed small reservoirs. For example, the ratio of SC to I_{mean} for Folsom reservoir is 0.33 (i.e., <0.5), and hence, it is termed a small reservoir. We include two small reservoirs and three large reservoirs in our analysis (Table 1). Literature suggests that releases from reservoirs with smaller values of SC/I_{mean} have a greater dependence on inflow in comparison to those with large values of SC/I_{mean} (>0.5). Releases from larger reservoirs are more dependent on reservoir storage (Coerver et al., 2018; Hanasaki et al., 2006; Yassin et al., 2019).

The climate data required for modeling, that is, daily precipitation, and daily maximum and minimum temperatures are obtained from the Global Historical Climatology Network daily (GHCNd) database. The Hargreaves method (Hargreaves & Samani, 1982) is used to estimate daily potential evapotranspiration. Climate data for all gauges operational between 10 October 1979 and 30 September 2021 within each watershed are downloaded to determine the period and gauges with maximum data availability. Geometric averaging across selected gauges is used to obtain lumped values of temperature and precipitation to be supplied as inputs to F03+. When unavoidable, we fill missing temperature data using linear interpolation for periods shorter than 7 days. Longer periods are filled using long-term average values (10 October 1979–30 September 2021) for that gauge for that period. We fill missing precipitation data using recorded rainfall at temporarily operational neighboring precipitation gauges. Observed daily flows are obtained from the United States Geological Survey (USGS) National Water Information System (NWIS). In case of unavailable streamflow gauges for measuring inflow into a reservoir, we use approximated inflow determined using the water balance equation. Li et al. (2023a) share information on approximated inflow along with GDROM models using the HydroShare platform developed by the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI). Information on the extent of vegetation coverage and soil data for defining a priori ranges of F03+ parameters is obtained from the USGS National Land Cover Database (NLCD) and the United State Department of Agriculture (USDA) Soil Survey Geographic Database (SSURGO), respectively.

4. Results

Here we examine the overall representation of a watershed system obtained by coupling reservoir operation models with a rainfall-runoff model; we compare the results obtained by using different reservoir operation

Table 1
Data on Selected Reservoirs and Watersheds

	Reservoir name	Operational purpose	SC (Acre-feet)	SC/I_{mean}	Calibration period	Validation period
a)	Folsom, California	IRR (ELE, WSP, FCON)	970,487	0.33	1 Oct 1999–30 Sept 2007	1 Oct 2010–30 Sept 2019
b)	Fontenelle, Wyoming	ELE (WSP)	350,698	0.35	1 Oct 1993–30 Sept 2001	1 Oct 2012–30 Sept 2019
c)	Wappapello, Missouri	FCON	758,643	0.52	1 Oct 2007–30 Sept 2012	1 Oct 2012–30 Sept 2015
d)	Pine Flat, California	FCON (IRR, ELE)	998,852	0.60	1 Oct 2011–30 Sept 2018	1 Oct 2003–30 Sept 2010
e)	Buffalo Bill, Wyoming	ELE (IRR, WSP)	646,647	0.75	1 Oct 1991–30 Sept 1999	1 Oct 2000–30 Sept 2008

Note. Reservoirs considered are sorted in increasing order of size as defined by SC/I_{mean}. For reservoirs serving multiple operational purposes, secondary operational purposes are mentioned in brackets. FCON: Flood control; ELE: Hydroelectricity; IRR: Irrigation; WSP: Water supply.

models to build watershed models via two calibration approaches (FCM and LCM) for five watersheds flowing to reservoirs of different sizes.

4.1. Simulation Performance of Watershed Outflow

Examining the GoF estimators $KGE_{Outflow}$ and $LNSE_{Outflow}$ in the calibration period, we note that the FCMs outperform the LCMs in simulating watershed outflow for all watersheds using any reservoir operation model (Figure S1 in Supporting Information S1). Higher values of $KGE_{Outflow}$ and $LNSE_{Outflow}$ are noted from the FCMs than the LCMs for all watersheds, irrespective of which reservoir operation model is used. Moreover, the simulations of watershed outflow from the FCMs result in L-moment ratios that are closer to the observed values than those from the LCMs (Figure S2 in Supporting Information S1). This result, however, is expected as the FCMs are designed to achieve the best simulation of watershed outflow (by maximizing $KGE_{Outflow}$ and $LNSE_{Outflow}$) in the calibration period. To indicate truly improved watershed outflow simulations from FCMs, higher values of the GoF estimators in both calibration and validation periods are necessary; higher values of GoF from FCMs than LCMs only in the calibration period can be indicative of unrealistic parameterization of the FCMs.

For the validation period, we observe a performance of the FCMs that is more dependent on reservoir operation models than the calibration period. We notice that GDROM- and DZTR-based FCMs obtain watershed outflow simulations that are comparable, if not better than LCMs (Figure 3). For example, the median $KGE_{Outflow}$ from the GDROM-based FCM for the Folsom reservoir watershed is 0.72, which is much higher than 0.60 from the LCM. On the other hand, comparable performance is seen for the Pine Flat reservoir watershed (median $KGE_{Outflow} = 0.80$ for both the GDROM-based FCM and LCM). Similarly, the DZTR-based FCM of the Fontenelle reservoir watershed achieves a higher median $KGE_{Outflow}$ value of 0.69, compared to 0.59 from the LCM. Comparable performance is also noted for the DZTR-based models of the Pine Flat reservoir watershed (median $KGE_{Outflow} = 0.75$ from the FCM and median $KGE_{Outflow} = 0.80$ from the LCM). GDROM-based FCMs and LCMs also obtain comparable values of $LNSE_{Outflow}$ for all the watersheds, though slightly improved median $LNSE_{Outflow}$ values from the FCMs are observed for three watersheds: Fontenelle, Pine Flat and Buffalo Bill. We observe that the GDROM-based FCMs obtain either equal or better median $KGE_{Outflow}$ values than the DZTR-based FCMs for all the watersheds we analyzed. Similarly, except for the Pine Flat reservoir watershed, GDROM-based FCMs either perform as well as or better than DZTR-based FCMs in terms of $LNSE_{Outflow}$. GDROM-based FCMs outperform DZTR-based LCMs particularly for watersheds with smaller reservoirs (i.e., Folsom and Fontenelle) (Figure 3).

The ISTARF-based FCMs of the Wappapello and Pine Flat reservoir watersheds show higher values of $KGE_{Outflow}$ and $LNSE_{Outflow}$ than the LCM alternatives. It is notable that the GDROM- and DZTR-based coupled models (both LCM and FCM) significantly outperform the ISTARF-based FCM of the Pine Flat reservoir watershed. Although the WISS-based FCMs outperform the LCMs in terms of median $KGE_{Outflow}$ and $LNSE_{Outflow}$ for Wappapello, Pine Flat, and Buffalo Bill reservoir watersheds, the overall performance of the WISS-based models is poor with negative values of $LNSE_{Outflow}$ for all the study watersheds in both FCMs and LCMs (Figure 3). The HANA-based models result in negative $LNSE_{Outflow}$ values for watersheds that drain into reservoirs with $SC/I_{mean} > 0.5$.

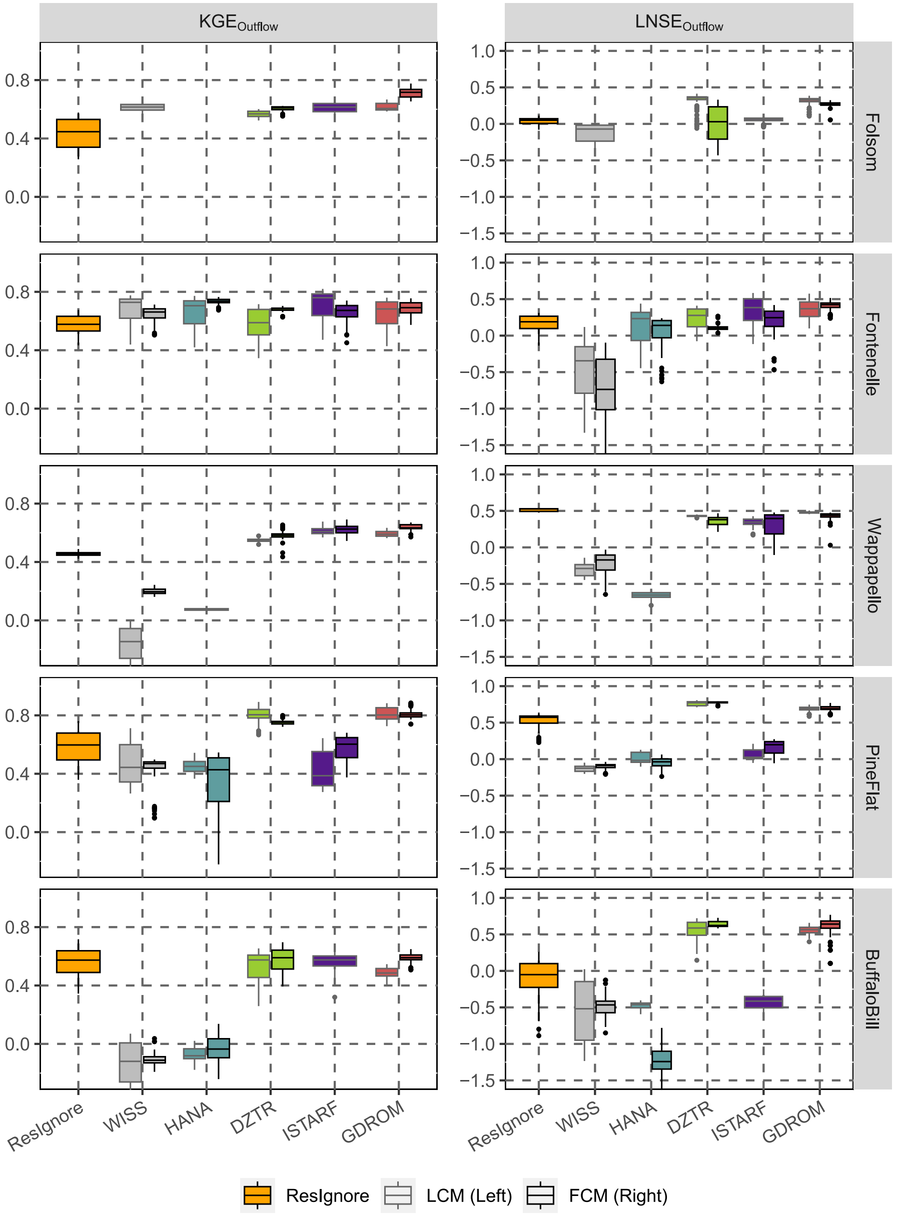

Figure 3. Validation period boxplots of KGE$_{Outflow}$ and LNSE$_{Outflow}$ (columns) obtained using acceptable Pareto-optimal parameter sets to simulate watershed outflow for different watersheds (rows). Each reservoir operation model (x-axis) is used to construct an FCM and an LCM; hence, two boxplots are plotted for each reservoir operation model. The boxplot on the left with the gray border corresponds to LCM results, while the boxplot on the right with the black border shows FCM results. A single boxplot with a gray border (i.e., LCM) is shown if none of the FCM Pareto optimal parameter sets yield KGE$_{Outflow}$ > −0.41 and LNSE$_{Outflow}$ > 0 in the calibration period (e.g., WISS and ISTARF for Folsom reservoir). ResIgnore results also have a single boxplot. No boxplot is shown corresponding to HANA for the Folsom reservoir watershed, as HANA cannot be applied to irrigation reservoir without additional data.

Examining the ability of coupled watershed models to capture the distributional properties of watershed outflow during the validation period shows that GDROM- and DZTR-based FCMs demonstrate lower errors in the estimates of L-moment ratios than corresponding LCMs. Moreover, the GDROM-, DZTR- and ISTARF-based models (both FCM and LCM) demonstrate lower errors in the estimates of L-moment ratios than those obtained from the HANA- and WISS-based coupled models, emphasizing the significance of using reservoir operation models that can capture operator behavior. In addition, significant errors in the L4 moment estimates are observed for the flood control reservoirs Wappapello and Pine Flat for all watersheds (for FCM and LCM). We

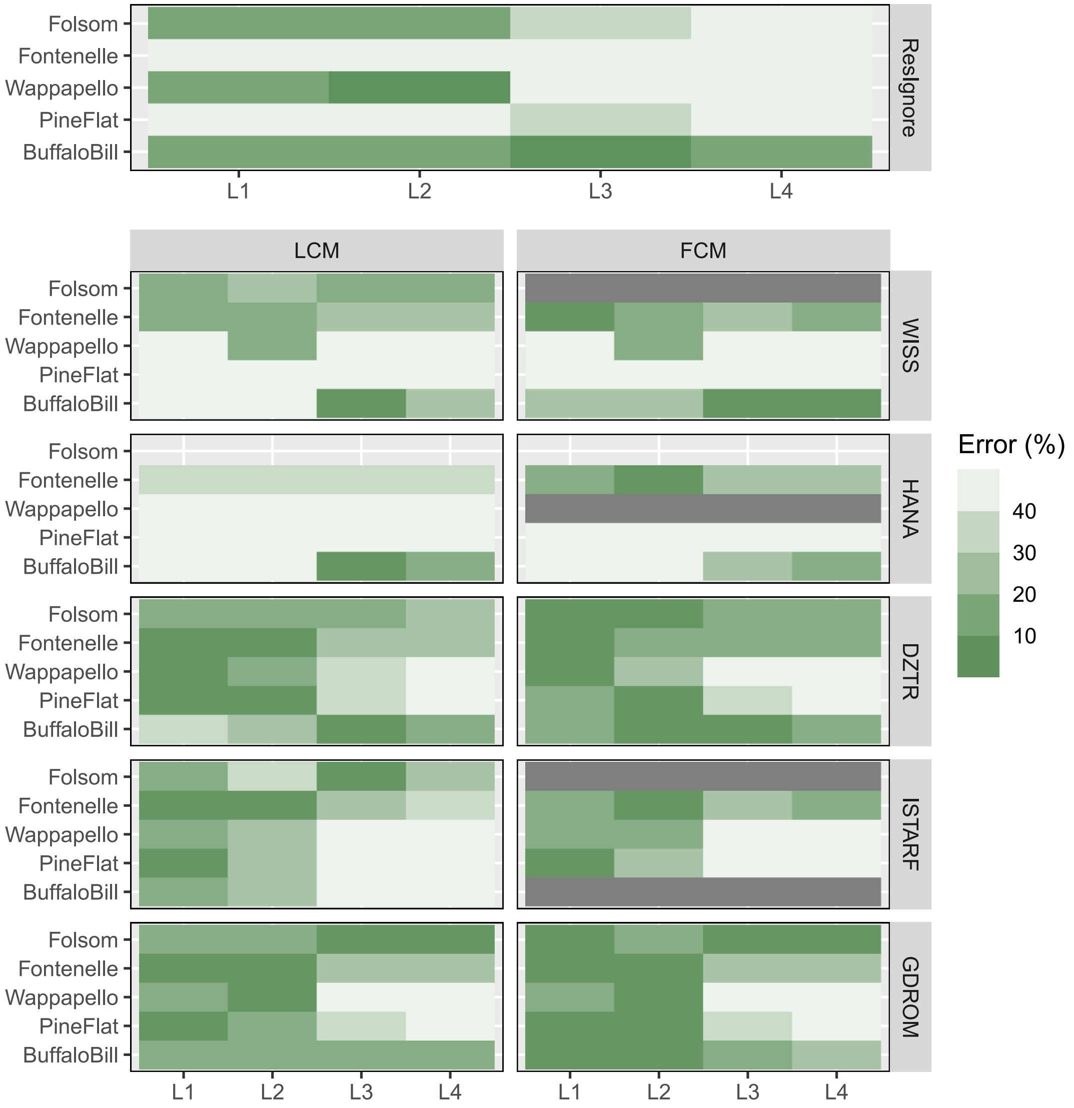

Figure 4. Median absolute percentage errors in the estimates of all four L-moment ratios (*x*-axis) from stochastic outflow simulations obtained using LCMs (left column) and FCMs (right column) based on different reservoir operation models (rows) in the validation period. ResIgnore presents the case where reservoir operation is ignored for watershed model development. Only acceptable Pareto optimal parameter sets are used to generate stochastic outflow simulations; tiles are grayed if none of the FCM Pareto optimal parameter sets yield $\text{KGE}_{\text{Outflow}} > -0.41$ and $\text{LNSE}_{\text{Outflow}} > 0$ in the calibration period. The *y*-axis corresponds to different reservoirs within each row. Lower errors correspond to more intense green shades of the tiles. No tiles shown for Folsom reservoir using the HANA reservoir operation model, as the HANA model cannot be applied to irrigation reservoirs without additional data.

attribute these errors in part to errors in the watershed outflow simulations and the procedure used to generate stochastic streamflow simulations, as described later in this section.

The results also show clear advantages of including a reservoir component in the watershed model; the GDROM-, DZTR- and ISTARF-based FCMs show lower errors in the estimates of the L-moment ratios than the ResIgnore models, especially for the Fontenelle and Pine Flat reservoir watersheds (Figure 4). The WISS- and HANA-based FCMs also obtain better estimates of the L-moment ratios than the ResIgnore models for the Fontenelle reservoir watershed. The large errors in the estimates of the L-moment ratios from the ResIgnore models suggests that humans alter streamflow through reservoir operations to the extent that a rainfall-runoff model that ignores reservoir operation cannot accurately simulate the entire distribution of observed streamflow. Examining the simulated deterministic FDCs of watershed outflow from the ResIgnore models confirms this; the ResIgnore models show large errors at multiple exceedance probabilities representing low, intermediate, and high flows

(Figures S3–S7 in Supporting Information S1). Furthermore, the stochastic FDCs of watershed outflow simulated by the ResIgnore models show extremely wide prediction intervals for Fontenelle, Pine Flat, and Buffalo Bill reservoirs, indicating that the structural error of ignoring reservoir operation generates simulations that are unsuitable for operational use (Figures S8–S12 in Supporting Information S1).

4.1.1. Watershed Outflow Flow Duration Curves

Examining validation period FDCs and hydrographs obtained using the different reservoir operation models reveals interesting insights regarding reservoir operation model functioning (Figures S3–S17 in Supporting Information S1). Our results show that the deterministic FDCs simulated by the WISS and HANA models for relatively larger reservoirs (Wappapello, Pine Flat, and Buffalo Bill) do not envelope large segments of the observed FDCs (Figures S5–S7 in Supporting Information S1). For large reservoirs, the WISS and HANA models tend to predict a constant average outflow with negligible variations resulting in large underpredictions at lower exceedance probabilities (<40%) (Figures S5–S7 and S15–S17 in Supporting Information S1). Large overpredictions are seen at higher exceedance probabilities from the WISS and HANA models of Pine Flat reservoir (Figure S6 in Supporting Information S1). For Wappapello reservoir, large underpredictions are noted even at higher exceedance probabilities (Figure S5 in Supporting Information S1). Stochastic simulations of HANA-based watershed outflows for Wappapello, Pine Flat, and Buffalo Bill reservoirs have wide prediction intervals as the error model adds large random errors to compensate for overall poor simulations (Figures S10–S12 in Supporting Information S1). Similarly, wide prediction intervals are also observed for WISS-based watershed outflows for the Pine Flat reservoir (Figure S11 in Supporting Information S1). Poor stochastic simulations of watershed outflow from WISS- and HANA-based models also explain the large errors in estimated L4 moments noted for Wappapello and Pine Flat reservoir watershed in Figure 4.

For the ISTARF model, the deterministic FDCs tend to demonstrate errors for low flows over the 75% exceedance probability, and high flows below the 5% exceedance probability (Figures S3–S7 in Supporting Information S1). A thresholding behavior is noted in the outflow FDCs for ISTARF, where an almost constant value of watershed outflow is simulated above 75% and below 5% exceedance probabilities. Large errors in the overall outflow FDC are noted for the ISTARF-based LCMs of the Folsom and Buffalo Bill reservoir watersheds (Figures S3 and S7 in Supporting Information S1). This is consistent with the poor $\mathrm{LNSE}_{\mathrm{Outflow}}$ values obtained using the ISTARF model for these reservoirs (Figure 3 and Figure S1 in Supporting Information S1). Using stochastic simulations to generate FDC prediction intervals for ISTARF-based model simulations removes the thresholding behavior for all reservoirs, however, the prediction intervals of the simulated stochastic FDCs only envelop the observed FDC reasonably for Folsom and Fontenelle reservoirs (Figures S8–S12 in Supporting Information S1). This explains the large errors in the L4 moment estimates as shown in Figure 4 for Wappapello and Pine Flat reservoirs.

The DZTR-based models reproduce watershed outflow reasonably for all watersheds; however, we observe a pattern of minor but consistent errors in the deterministic FDCs simulated by those models (Figures S3–S7 in Supporting Information S1). The ranges of exceedance probability showing errors vary by watershed. For example, for the Folsom reservoir watershed, slight overpredictions of watershed outflow are observed from the DZTR-based FCM above the 50% exceedance probability. Similarly, minor but consistent underpredictions are noted from the DZTR-based FCM of the Fontenelle reservoir watershed above the 50% exceedance probability. For the Wappapello reservoir watershed, slight overpredictions are noted between the 25%–75% exceedance probabilities, and slight underpredictions are seen above the 75% exceedance probability. Minor errors are also noted from the DZTR-based FCM of the Pine Flat reservoir watershed outside the 40%–90% exceedance probability range. Finally, for Buffalo Bill reservoir, the deterministic watershed outflow FDC from the DZTR-based FCM shows underpredictions beyond the 50% exceedance probability. The minor but consistent errors from DZTR-based FCMs explain the lower $\mathrm{KGE}_{\mathrm{Outflow}}$ and $\mathrm{LNSE}_{\mathrm{Outflow}}$ values those models achieve compared to GDROM-based FCMs (Figure 3).

The GDROM-based models tend to reproduce the high and intermediate flows reasonably across all watersheds. However, low flow simulations (above the 75% exceedance probability) demonstrate some errors (Figures S3–S7 in Supporting Information S1). A thresholding behavior like that noted for ISTARF is also seen with GDROM; however, the constant average release simulated by GDROM above the 75% exceedance probability is close to the observed mean of low flows. This explains the comparatively high $\mathrm{LNSE}_{\mathrm{Outflow}}$ values obtained from GDROM compared to ISTARF, despite a similar thresholding behavior in low flow simulations. Like ISTARF, adding model residuals back using stochastic simulations for GDROM removes the thresholding behavior.

Our procedure for generating stochastic streamflow simulations assumes that model error is identically distributed over time, which can lead to overprediction of peaks as the error model compensates for poor low flow simulations. We associate the assumption of identically distributed model error with errors in the L4 moments of watershed outflow from GDROM- and DZTR-based models of the Wappapello and Pine Flat reservoir watersheds. A more sophisticated approach for generating stochastic simulations that may explicitly account for the heteroskedasticity of model error should lead to better L4 moment estimates; however, this is beyond the scope of our study. Overall, the prediction intervals of the stochastic FDCs obtained from the GDROM- and DZTR-based models tend to consistently envelope the observed outflow FDC for most exceedance probabilities (Figures S8–S12 in Supporting Information S1). Hence, the GDROM and DZTR models may be suitable for operational use in hydrology models in their current form.

4.2. Parameterization of the Watershed Models

We compare acceptable Pareto optimal parameters of the watershed models to understand differences arising from the use of different reservoir operation models for developing FCMs (Figure 5 and Figures S18–S21 in Supporting Information S1). We observe that there are study area dependent differences in the parameters which are affected when using the FCMs and ResIgnore models. For example, FCMs developed for Pine Flat reservoir show deviations in parameters of the near surface soil moisture accounting module (SB, FC) from those obtained by calibrating F03+ to natural reservoir inflow (i.e., the "Inflow" model part of LCM) (Figure 5). On the other hand, parameters in the deep recharge and routing module (KD, ASS, ABF) demonstrate differences from those obtained by the "Inflow" model for FCMs of the Wappapello and Buffalo Bill reservoir watersheds (Figures S20 and S21 in Supporting Information S1). The extent of differences in parameters is found to depend on the reservoir operation model used for FCM development. For example, the GDROM-based FCM of the Pine Flat reservoir watershed shows smaller deviations in the SB and FC parameters compared to FCMs based on other reservoir operation models (Figure 5). Values of the ABF parameter obtained using the GDROM-based FCM for the Pine Flat reservoir watershed are also the closest to those obtained from the "Inflow" model which calibrates F03+ to reservoir inflow.

For the Folsom reservoir watershed, ignoring reservoir operation results in a model parameterization that mimics reservoir storage by increasing moisture storage across the watershed in various forms including snowpack, near surface soil moisture, and deep groundwater storage (Figure S18 in Supporting Information S1). TTH shows an increase compared to the natural state (i.e., "Inflow" model) when using ResIgnore resulting in greater snowpack accumulation because snowfall is simulated at warmer temperatures. ResIgnore also increases SB resulting in greater soil moisture storage capacity. Under ResIgnore, a larger portion of the precipitation reaching the soil surface enters the saturated soil moisture store and subsequently the deep groundwater store due to a reduction in FC; a reduction in FC reduces the moisture holding capacity of the unsaturated soil moisture store. Abstractions from the saturated store occur at slower rates under ResIgnore due to reduced ASS. The DZTR- and GDROM-based FCMs improve the watershed model parameterization in terms of FC and KD, indicating improvements achieved by the inclusion of a reservoir operation model. The DZTR-based FCM also obtains ASS values that are within ranges of those obtained by the "Inflow" model.

The ResIgnore model greatly reduces ABF for the Fontenelle reservoir watershed, reducing the rate of drainage from the deep groundwater store (Figure S19 in Supporting Information S1). All the FCMs obtain an ABF value closer to that obtained under natural conditions compared to ResIgnore. However, in general, the FCMs increase the value of FC. The DZTR-, WISS-, and ISTARF-based FCMs also show reduced values of KD compared to the natural state. A reduction in KD reduces the rate at which the deep groundwater store is recharged via drainage from the saturated soil moisture store. The FC values obtained using the GDROM-based FCM are closest to those obtained under natural conditions. The GDROM- and HANA-based FCMs also obtain KD values close to those obtained by calibrating F03+ to reservoir inflow.

For the Wappapello reservoir watershed, the ResIgnore model greatly increases KD and ABF (Figure S20 in Supporting Information S1). The ResIgnore model also demonstrates a large reduction in *B* compared to the natural state indicating differences introduced by ResIgnore in the near surface soil moisture accounting module. The WISS-, DZTR- and GDROM-based FCMs bring ABF values within natural ranges, although reductions in ABF

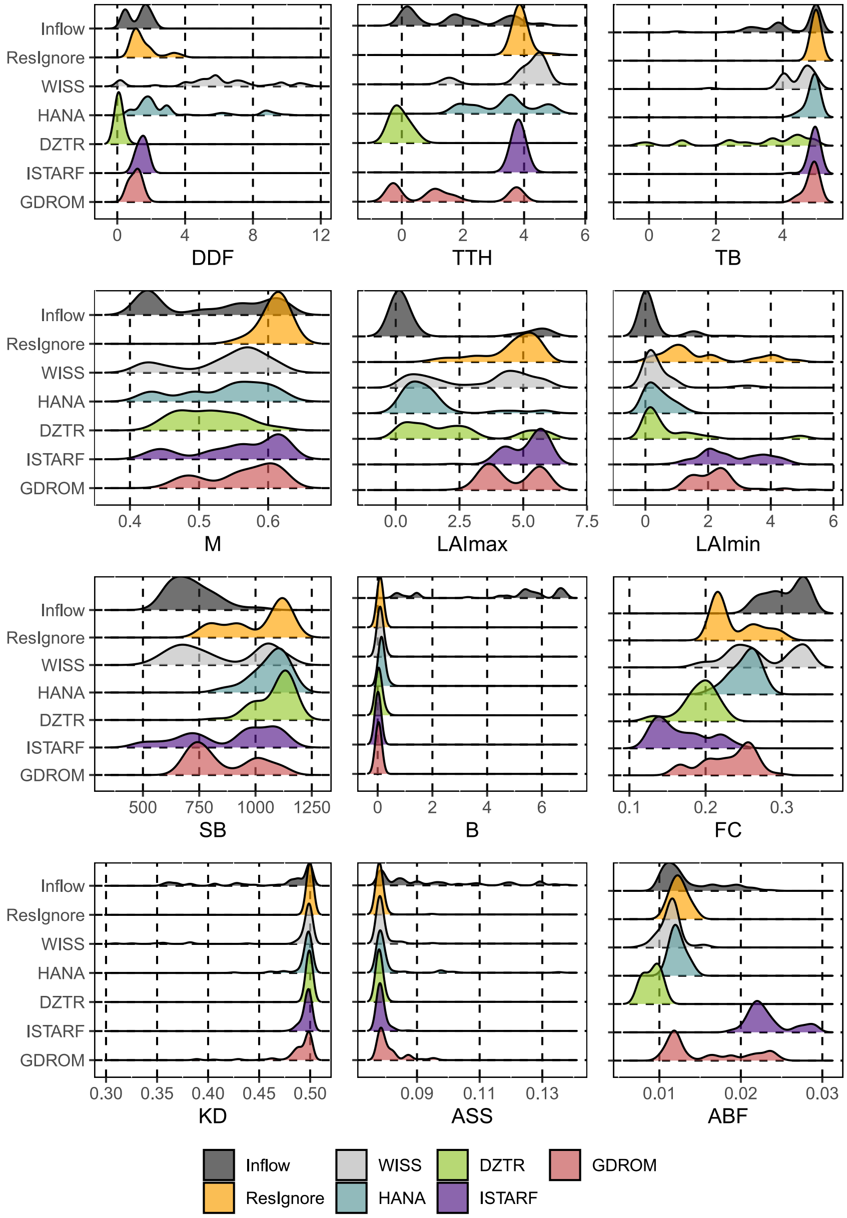

Figure 5. Visualizations of acceptable Pareto-optimal parameters (with KGE > −0.41; LNSE > 0) of watershed models developed for the Pine Flat reservoir watershed. Each grid corresponds to one of the 12 calibrated model parameters. In each grid, the *y*-axis states the name of the reservoir operation model used for FCM development. ResIgnore shows parameters found when reservoir operation was ignored (i.e., F03+ calibrated to streamflow downstream of the reservoir). Inflow corresponds to parameters found by calibrating F03+ to reservoir inflow (i.e., natural flow upstream of the reservoir). $\text{KGE}_{\text{Outflow}}$ and $\text{LNSE}_{\text{Outflow}}$ are used for calibrating FCMs and ResIgnore models, while $\text{KGE}_{\text{Inflow}}$ and $\text{LNSE}_{\text{Inflow}}$ are used for LCMs.

are noted. Similarly, although all the FCMs obtain KD values closer to those under the natural state compared to ResIgnore, small reductions in KD are observed. All the FCMs obtain improved parameterization of B. Finally, the WISS-based FCMs show large increases in the SB and FC parameters compared to the "Inflow" model.

The acceptable Pareto optimal parameters determined by the WISS- and HANA-based FCMs differ greatly from the natural state for the Buffalo Bill reservoir watershed (Figure S21 in Supporting Information S1). This result, however, is expected as HANA and WISS cannot model the diversions from Buffalo Bill reservoir that occur in addition to downstream river releases. The ResIgnore model decreases TTH and ASS values, while increasing SB, KD and DDF values compared to the natural state. The DZTR-based FCM improves upon ResIgnore,

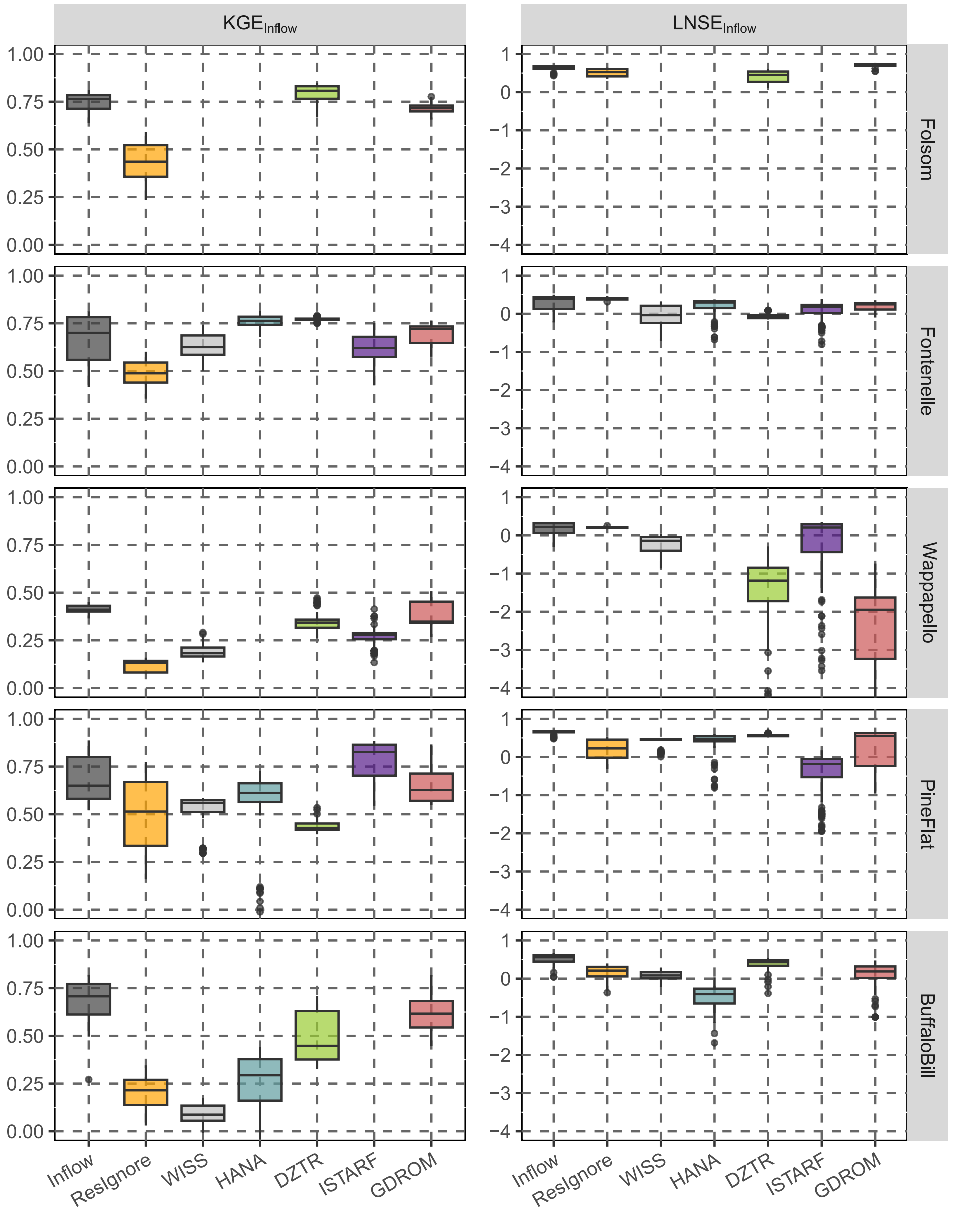

Figure 6. Validation period boxplots of KGE_{Inflow} and $LNSE_{Inflow}$ (columns) obtained using acceptable Pareto-optimal parameter sets to simulate reservoir inflow for different watersheds (rows). No boxplot is shown if none of the FCM Pareto optimal parameter sets yield $KGE_{Outflow} > -0.41$ and $LNSE_{Outflow} > 0$ in the calibration period (Ex., WISS and ISTARF for Folsom reservoir). No boxplot is shown corresponding to HANA for the Folsom reservoir watershed, as the HANA model cannot be applied to irrigation reservoirs without additional data. An identical plot for the calibration period is included in the as Figure S32 in Supporting Information S1.

obtaining values of SB that are close to those obtained by calibrating F03+ to reservoir inflow. The GDROM-based FCM obtains TTH and DDF values that are closer to the natural state than ResIgnore. The HANA-, DZTR- and GDROM-based FCMs show increases in ABF values compared to those obtained under a natural state.

4.3. Reservoir Inflow Simulations

In general, the inclusion of a reservoir component in the FCMs results in greatly improved simulations of reservoir inflow compared to ResIgnore. The FCMs obtain KGE_{Inflow} exceeding that obtained by ResIgnore for all cases except the WISS-based and DZTR-based FCMs of the Buffalo Bill and Pine Flat reservoir watersheds respectively (Figure 6). Furthermore, the KGE_{Inflow} values achieved by the FCMs are close to those obtained by

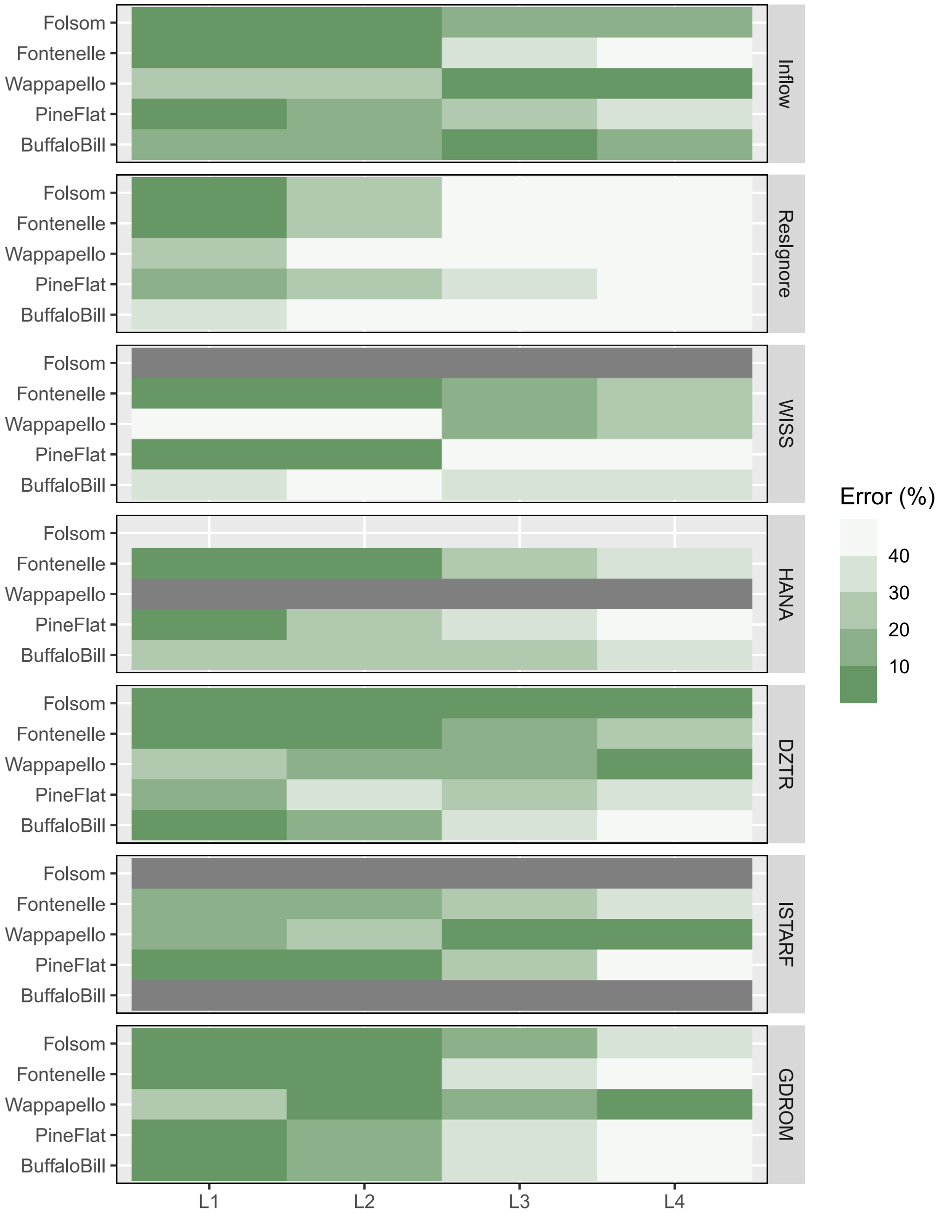

Figure 7. Median absolute percentage errors in the estimates of all four L-moment ratios (*x*-axis) from deterministic reservoir inflow simulations obtained using FCMs based on different reservoir operation models (grid rows) in the validation period. Only acceptable Pareto optimal parameter sets are used to generate deterministic reservoir inflow simulations; tiles are grayed if none of the FCM Pareto optimal parameter sets yield $KGE_{Outflow} > -0.41$ and $LNSE_{Outflow} > 0$ in the calibration period. The *y*-axis corresponds to different reservoirs within each row. Lower errors correspond to more intense green shades of the tiles. No tiles shown for Folsom reservoir using the HANA reservoir operation model, as the HANA model can only be applied to non-irrigation reservoirs. An identical plot for the calibration period is included in the as Figure S33 in Supporting Information S1.

the "Inflow" models from LCMs; this indicates reasonable trade-offs between FCMs and LCMs (Section 2.1). Although the values of $LNSE_{Inflow}$ obtained from ResIgnore and the FCMs are comparable, clear benefits of including a reservoir component can be noted by examining errors in the estimates of the L-moment ratios of reservoir inflow. The FCMs demonstrate lower errors in the estimates of the L-moment ratios of reservoir inflow compared to ResIgnore, especially FCMs developed using GDROM, DZTR and ISTARF (Figure 7). The specific effects of differences in optimal parameterization found by the various FCMs and ResIgnore models are examined by studying the deterministic FDCs of reservoir inflow.

The ResIgnore-based deterministic FDC of reservoir inflow for Folsom reservoir shows large underpredictions of high flows (below the 10% exceedance probability) and large overpredictions of intermediate flows between the 25%–90% exceedance probabilities (Figure S22 in Supporting Information S1). These errors can be explained by changes in ASS and FC introduced by ResIgnore for the Folsom reservoir watershed which result in streamflow simulations mimicking reservoir operation (Figures S18 and S23 in Supporting Information S1). Slowed drainage from the saturated soil moisture zone due to reduced ASS lowers streamflow peaks, while increased moisture accumulation in the deep groundwater store due to reduced FC combined with reduced ASS increases and sustains baseflow in the dry season, causing elevated intermediate flows. The DZTR- and GDROM-based FCM simulations of the deterministic FDC of reservoir inflow show significantly lower errors compared to ResIgnore till the 75% exceedance probability on account of their improved estimate of FC. The DZTR-based FCM also obtains better peak flow simulations compared to ResIgnore due to relatively more realistic values of ASS (Figures S18 and S22 in Supporting Information S1). Both the DZTR- and GDROM-based FCMs show significant underpredictions for low reservoir inflows above the 75% exceedance probability.

Like noted for the Folsom reservoir watershed, the deterministic FDC of reservoir inflow obtained using ResIgnore for the Fontenelle reservoir watershed also demonstrates severe underpredictions below the 10% exceedance probability and overpredictions between the 25%–90% exceedance probabilities (Figure S24 in Supporting Information S1). The reduction in ABF caused by ResIgnore reduces baseflow contributions to daily streamflow, lowering streamflow peaks (Figure S19 in Supporting Information S1). ABF also influences baseflow in the dry season; lower ABF drains the deep groundwater store slowly, leaving more moisture available for baseflow generation in the dry season. Thus, the ResIgnore model of the Fontenelle reservoir watershed mimics reservoir operation in terms of attenuated high flow releases and increased dry season releases to satisfy environmental and water supply demands (Figure S25 in Supporting Information S1). For the FCMs, increased FC results in reduced subsurface and baseflow generation, leading to underpredictions of low reservoir inflows (Figure S24 in Supporting Information S1). Further underpredictions of low reservoir inflow are noted for the DZTR-, WISS, and ISTARF-based FCMs due to reduced KD which reduces baseflow by lowering the rate of groundwater recharge. The simulated deterministic reservoir inflow FDCs from the GDROM- and HANA-based FCMs are very similar to that obtained by calibrating F03+ to reservoir inflow, due to comparable FC and KD values.

For the Wappapello reservoir watershed, increase in KD under ResIgnore mimics reservoir operation, resulting in underpredictions and overpredictions of reservoir inflow below and above the 10% exceedance probability respectively (Figures S20, S26 and S27 in Supporting Information S1). The higher rate of drainage to the deep groundwater store under ResIgnore, due to increased KD, reduces moisture available in the near surface saturated store for fast subsurface runoff generation, leading to underpredictions of high flows. At the same time, the increased moisture accumulated in the deep groundwater store results in increased baseflow, causing overpredictions of low flows. The GDROM-, DZTR- and ISTARF-based FCMs accurately simulate the reservoir inflow peaks and high flows below the 25% exceedance probability due to relatively better estimates of KD (Figures S20 and S26 in Supporting Information S1). However, we note large underpredictions of reservoir inflows beyond the 25% exceedance probability from the GDROM-, DZTR-, and WISS-based FCMs. This may be associated with the combined effects of reduced KD and ABF on the deep groundwater store (Figure S20 in Supporting Information S1). The deep groundwater store is recharged at a slower rate due to reduced KD and drains at a slower rate due to reduced ABF, leading to greatly reduced baseflow production.

The Pine Flat reservoir modifies incoming streamflow by attenuating and delaying the peaks, an effect that the ResIgnore model mimics by increasing SB (i.e., the overall moisture holding capacity of the near surface soil moisture zone), leading to underpredictions of high reservoir inflow (Figure 5; Figures S28 and S29 in Supporting Information S1). Increases in SB of differing extent noted for the FCMs (see Section 4.2) result in underpredictions of high reservoir inflow of differing magnitudes. The GDROM-based FCM shows relatively lower magnitudes of high flow underpredictions due to more realistic SB values. The ISTARF-based FCM also shows negligible underpredictions of high flows, however, this may partially be influenced by increases in ABF; increases in ABF can increase high flows and reduce low flows via mechanisms described previously in this section. Notably, the ISTARF-based FCM shows underpredictions of low reservoir inflows which we associate with increased ABF. Reservoir inflow simulations from the DZTR-based FCM are also impacted by changed ABF; DZTR-induced reductions in ABF lead to greater underpredictions of high flows below the 25% exceedance probability, and overpredictions at higher exceedance probabilities.

The ResIgnore-based deterministic FDC of reservoir inflow for the Buffalo Bill reservoir watershed shows large underpredictions below the 25% exceedance probability and above the 60% exceedance probability (Figure S30 in Supporting Information S1). The reductions in high flows can be explained by the increases in SB and decreases in ASS introduced by ResIgnore via mechanisms described previously in this section. Deterministic FDCs of reservoir inflow from the DZTR- and GDROM-based FCMs show smaller underpredictions of high flows due to improved estimates of SB. We note that improvements in high flow simulations from the GDROM- and DZTR-based FCMs may also be influenced by increases in ABF. Underpredictions of low inflows noted for the DZTR- and GDROM-based FCMs may also be ascribed to increased ABF values.

5. Discussion

5.1. Implications for Modeling Hydrology Over Large Spatial Scales

Overall, our results show that GDROM- and DZTR-based watershed models (both FCMs and LCMs) achieve a consistently reliable watershed outflow simulation performance (in terms of aggregated GoF estimators, L-moments and FDCs) across all five watersheds considered in our study. Recall that the reservoirs included in our study differ in size and serve a range of operational purposes including hydroelectricity, flood control, irrigation, and water supply. Correspondingly, the consistent reliable performance we find from the watershed scale GDROM- and DZTR-based models provides evidence that GDROM and DZTR may be coupled with hydrological models at the various larger scales such as river basin, national, continental, global, etc.

Large-scale hydrological models (such as national or global models) are applied in both calibrated and uncalibrated forms (Kauffeldt et al., 2016). The integration of GDROM or DZTR into uncalibrated large-scale hydrological models would yield the simulation performance that is noted for LCMs in this study, if the hydrological model reliably simulates reservoir inflow. On the other hand, for calibrated large-scale hydrological models, depending on whether the streamflow gauge used for hydrological model calibration is upstream or downstream of a reservoir, the simulation performance for either LCM or FCM will be attained. We find that GDROM and DZTR can be reliably used in both FCM and LCM modes for streamflow simulation. Moreover, the parameterization found by FCMs and LCMs based on GDROM and DZTR are comparable despite that LCMs are not affected by reservoir storage regulation (Figure 5 and Figures S18–S21 in Supporting Information S1). That is to say, the FCMs do not obtain "the right answers for the wrong reasons."

The ability of FCMs to obtain parameters that are comparable to LCMs, and simulate streamflow as well as, if not better than LCMs, demonstrates strong potential for using FCMs in large-scale multi-reservoir systems. For multi-reservoir systems, releases from upstream reservoirs impact inflows into downstream reservoirs. Applying LCMs to such systems would require performing "calibration in parts," that is, defining incremental drainage areas for each reservoir and calibrating each incremental drainage area separately; this becomes increasingly complicated as the calibration progresses downstream. Our results show that FCMs could be used to reliably represent all the reservoirs in a multi-reservoir system simultaneously, thus avoiding the "calibration in parts" issue and improving computational efficiency.

It should be noted that reservoirs in a multi-reservoir system are often operated in coordination. Ignoring the effects of coordinated reservoir management has been shown to misrepresent reservoir impacts during flood and drought conditions (Rougé et al., 2021). The reservoir operation models in our study are parameterized independently, and hence, there is no explicit consideration of coordinated reservoir management. However, as mentioned by Chen et al. (2022) and Rougé et al. (2021), inferring reservoir operation rules using machine learning techniques can implicitly capture coordinated reservoir operational behavior to some extent. We recognize that while the limited number of predictors used for GDROM may not comprehensively capture coordinated reservoir management, predictors such as DOY and PDSI can account for seasonality and basin dryness conditions, respectively, which are important factors affecting coordination. However, we would admit that the impacts of factors such as the institutional context, forecasts, etc., may need explicit consideration for modeling reservoirs operated in coordination, which in turn will affect the ease and generality of incorporating GDROMs into large-scale hydrological models.

An additional measure must be taken regarding flow diversions directly from reservoirs to meet water demands. Not accounting for the diversions would not only introduce errors in simulated reservoir releases at the reservoir where diversions occur, but also lead to incorrect inflow simulations for all reservoirs located downstream. In our analysis, we accounted for flow diversions from the Buffalo Bill reservoir using a two-step approach (Section 2.1.1). Future studies may analyze the resulting inconsistencies introduced in large-scale hydrological models when water management activities such as flow diversions are ignored.

GDROM and DZTR have been developed with the intention of coupling with large-scale hydrological models, and in our analyses, they result in better coupled model performance than with HANA and WISS which have already been integrated into global hydrological models (Chen et al., 2022; Fekete et al., 2010; Hanasaki et al., 2008; Wisser et al., 2010; Yassin et al., 2019). We notice that GDROM has been rigorously tested for simulating reservoir outflow using observed inflow across 467 reservoirs in the CONUS (Chen et al., 2022); while Yassin et al. (2019) have shown good performance of the DZTR model over only 37 reservoirs across the globe, of which 18 are located within the CONUS. Although we also obtain good performance using DZTR for reservoirs not tested by Yassin et al. (2019), we believe that GDROM may be more applicable for studies in the CONUS given its more in-depth validation by Chen et al. (2022) over the CONUS.

We hope that our analyses help inform researchers in coupling reservoir operation models with hydrological models at various spatial scales. To aid researchers in building coupled models, following the method of this study, we provide freely accessible code that can be used to derive operation rules for DZTR, HANA and WISS models from historical reservoir operation records, and convert the derived rules into reservoir components that can be easily integrated with rainfall-runoff models (Vora et al., 2024). Li et al. (2023b) have provided code to derive GDROM operation rules from historical reservoir operation records. Derived reservoir operation rules for the GDROM (Li et al., 2023a) and ISTARF (Turner, Voisin, et al., 2021) models are available publicly for reservoirs in the CONUS. The code we provide with our study can also read GDROM- and ISTARF-based operation rules and convert them into reservoir components.

5.2. Reasons for Differences in the FCM Performance

We associate the FCM performance with the incorporated reservoir operation model's ability to accurately simulate reservoir outflow, which in turn depends on the reservoir operation model's formulation. For example, the formulations of the HANA and WISS models simulate a constant reservoir outflow irrespective of variations in daily reservoir inflow for large reservoirs (see in Section 2.1.1 and Figures S15–S17 in Supporting Information S1). These near constant reservoir release simulations throughout the analysis period lead to low values of $\mathrm{LNSE}_{\mathrm{Outflow}}$ and $\mathrm{KGE}_{\mathrm{Outflow}}$, as well as poor estimates of the L-moments of watershed outflow from HANA- and WISS-based LCMs and FCMs (Figures 3 and 4). In fact, the HANA and WISS models obtain negative $\mathrm{LNSE}_{\mathrm{Outflow}}$ values even when routing calibrated inflow through the Wappapello, Pine Flat and Buffalo Bill reservoirs (Figure 3 and Figure S1 in Supporting Information S1).

For the ISTARF-based models, near constant values of watershed outflow are also simulated above the 75% and below the 5% exceedance probabilities, which may be explained by the formulation of ISTARF (Figures S3–S7 in Supporting Information S1). The NOR approach of ISTARF can result in the simulation of near constant releases when reservoir storage is above or below a particular value. If the reservoir storage is below the NOR for ISTARF, a constant minimum release is simulated. For reservoir storage above the NOR, the ISTARF model simulates outflows aimed at bringing the reservoir storage back into the NOR, subject to a maximum permissible outflow. Widening the NOR for the ISTARF models and choosing different values of minimum release (currently the 95th percentile of observed release is used) might improve outflow simulation performance of the ISTARF models. Furthermore, for certain reservoirs, the ISTARF model only predicts a seasonal pattern of reservoir outflow which is not affected by reservoir inflow or storage (when reservoir storage lies within the NOR); Buffalo Bill is one such reservoir. This disconnect between reservoir inflows and outflows may explain the negative $\mathrm{LNSE}_{\mathrm{Outflow}}$ values obtained even when calibrated inflow is routed through Buffalo Bill reservoir using the ISTARF model (as in the LCM; Figure S1 in Supporting Information S1).

The minor but consistent errors observed from DZTR-based FCMs may be explained by the parameterization chosen to define operational zones and target releases in the generalized DZTR reservoir operation model. Yassin et al. (2019) suggested that the operation zone and target release parameterization could be optimized using a bi-objective optimization approach, however, we adopted the generalized DZTR model for analysis due to lower computational and data requirements. This decision is in-line with the intention to couple generic reservoir models with simple and transparent structures and low data and computational requirements with rainfall-runoff models.

For GDROM, constant average releases above the 75% exceedance probability can be explained by the fact that GDROM was originally trained to maximize NSE values. Correspondingly, GDROM is inherently biased toward better simulations of relatively higher flows. For lower flows, GDROM tends to predict an average constant value that is close to the observed mean of the low flows; this leads to errors, albeit of smaller magnitudes than those observed for ISTARF. Improved low flow simulations may be obtained if the original GDROM model is retrained with an LNSE objective. Doing so should lead to even better representation of watershed systems from GDROM-based FCMs.

Reservoir operation model accuracy in simulating releases also impacts FCM parameterization, and thereby reservoir inflow simulations. For example, below a threshold level of reservoir inflow, the WISS and GDROM models of Wappapello reservoir simulate a constant value of reservoir release, irrespective of the value of reservoir inflow (Figure S5 in Supporting Information S1). Correspondingly, parameters controlling relatively lower levels of reservoir inflow are rendered insensitive to the calibration process which is designed with the sole objective of accurately simulating reservoir outflow (via maximizing $\mathrm{KGE}_{\mathrm{Outflow}}$ and $\mathrm{LNSE}_{\mathrm{Outflow}}$). For DZTR-based FCMs, the parameterization obtained compensates for deficiencies in reservoir outflow simulations from DZTR. For example, the DZTR-based LCM of Folsom reservoir shows minor but consistent overpredictions of watershed outflow beyond the 50% exceedance probability (Figure S3 in Supporting Information S1). To compensate for this, the DZTR-based FCM of Folsom reservoir reduces reservoir inflows beyond the 50% exceedance probability (Figure S22 in Supporting Information S1). Improvements in reservoir operation models should remedy parameterization related issues. Additionally, our results show that currently available reservoir operation models do obtain reasonable model parameterizations when used in FCMs. We acknowledge that the sensitivity of the rainfall-runoff model parameters should be accounted for when comparing the parameterization obtained by FCMs and LCMs. Ideally, more sensitive parameters should have consistent values between FCMs and LCMs, which can be explored by future studies.

6. Summary and Conclusions

We assess the overall representation of a watershed system (i.e., reservoir operation + rainfall-runoff processes) via fully coupling realistic reservoir operation models with a rainfall-runoff process simulation model. Full coupling entails obtaining rainfall-runoff model parameters that work best in conjunction with trained reservoir operation models for simulating watershed outflow. We couple five generic reservoir operation models—WISS, HANA, DZTR, ISTARF, and GDROM, with a 12-parameter conceptual rainfall-runoff model called F03+. The GDROM, ISTARF, and DZTR reservoir operation models are derived from long-term observed reservoir operation records, and to some extent, can represent reservoir operators' behavior in release decisions. The HANA and WISS models represent widely applied reservoir operation models that use simplified rules to route inflows through a reservoir. Our evaluation of the watershed system representation includes examining watershed outflow simulations, model parameters, and reservoir inflow (an internal flux of the FCMs) simulations.

Our results show that fully coupled watershed models based on GDROM and DZTR obtain parameters that are comparable to loosely coupled models, and watershed outflow simulations that are as good, if not better than loosely coupled models. Correspondingly, our results show that the fully coupled models can be used to model large multi-reservoir systems without loss of physical significance. We also find that the prediction intervals of watershed outflow FDC obtained by the GDROM- and DZTR-based watershed models (both fully coupled and loosely coupled) consistently envelope the observed watershed outflow FDC. Thus, the GDROM and DZTR models may be used for developing realistic large scale hydrology models for operational use. Finally, we note that simulations from the ResIgnore models cannot represent the distributional properties of watershed outflow. Large errors in the estimates of the L-moment ratios of watershed outflow are found when applying ResIgnore, indicating significant human impacts on watershed hydrology that cannot be captured without including a reservoir component in the watershed model.

The effects of improvements in watershed model parameterization are clearly seen by examining the reservoir inflow simulation performance. We find that the ResIgnore model introduces changes in parameters controlling near surface soil moisture storage and deep groundwater storage to mimic reservoir operation. Consequently, the FDCs of reservoir inflow simulated by the ResIgnore models show large errors. Including a reservoir component, as in the fully coupled approach, improves watershed model parameterization and reservoir inflow simulation. The fully coupled watershed models achieve significantly higher $\mathrm{KGE}_{\mathrm{Inflow}}$ values and obtain better estimates of the L-moment ratios of reservoir inflow compared to ResIgnore. The extent of improvements in parameterization achieved by the fully coupled models depends on the performance of the reservoir operation models in simulating reservoir outflow; limited improvements are achieved for parameters controlling reservoir inflow in the ranges where reservoir outflow simulations are not influenced by reservoir inflow. Improving the reservoir outflow simulation performance of the reservoir operation models across a wider range of values (low, intermediate, and high flows) would lead to fully coupled watershed models that achieve even better representations of the watershed systems.

Data Availability Statement

All codes required to carry out the analysis described in this study can be accessed from the Mendeley Data repository via Vora et al. (2024), https://doi.org/10.17632/t49cyrgtct.4, CC by 4.0. Parameters for the GDROM and ISTARF models can be obtained from Li et al. (2023a) (https://doi.org/10.4211/hs.63add4d5826a4b21a6546c571bdece10) and Turner, Voisin, et al. (2021) (https://doi.org/10.5281/zenodo.4602277) respectively. Information on reservoir characteristics and reservoir operation records can also be obtained from Li et al. (2023b). Climate data for running the watershed models was obtained from https://www.ncei.noaa.gov/cdo-web/. Streamflow data for watershed model calibration and validation was retrieved from https://maps.waterdata.usgs.gov/mapper/index.html.

Acknowledgments

The authors are grateful for Prof Richard Vogel at Tufts University for his help on the model evaluation performance. The authors also acknowledge and thank Kevin Wallington and Hari Dave (PhD students at University of Illinois Urbana Champaign) for helpful discussions during the formulation of the analysis's framework. Funding for this project was provided by the National Oceanic and Atmospheric Administration (NOAA), awarded to the Cooperative Institute for Research to Operations in Hydrology (CIROH) through the NOAA Cooperative Agreement with The University of Alabama, NA22NWS4320003.

References

Abbaspour, K. C., Rouholahnejad, E., Vaghefi, S., Srinivasan, R., Yang, H., & Kløve, B. (2015). A continental-scale hydrology and water quality model for Europe: Calibration and uncertainty of a high-resolution large-scale swat model. *Journal of Hydrology*, *524*, 733–752. https://doi.org/10.1016/j.jhydrol.2015.03.027

Apostel, A., Kalcic, M., Dagnew, A., Evenson, G., Kast, J., King, K., et al. (2021). Simulating internal watershed processes using multiple swat models. *Science of the Total Environment*, *759*, 143920. https://doi.org/10.1016/j.scitotenv.2020.143920

Bai, Y., Wagener, T., & Reed, P. (2009). A top-down framework for watershed model evaluation and selection under uncertainty. *Environmental Modelling & Software*, *24*(8), 901–916. https://doi.org/10.1016/j.envsoft.2008.12.012

Biemans, H., Haddeland, I., Kabat, P., Ludwig, F., Hutjes, R., Heinke, J., et al. (2011). Impact of reservoirs on river discharge and irrigation water supply during the 20th century. *Water Resources Research*, *47*(3), W03509. https://doi.org/10.1029/2009wr008929

Bierkens, M. F., Reinhard, S., de Bruijn, J. A., Veninga, W., & Wada, Y. (2019). The shadow price of irrigation water in major groundwater-depleting countries. *Water Resources Research*, *55*(5), 4266–4287. https://doi.org/10.1029/2018wr023086

Chen, Y., Li, D., Zhao, Q., & Cai, X. (2022). Developing a generic data-driven reservoir operation model. *Advances in Water Resources*, *167*, 104274. https://doi.org/10.1016/j.advwatres.2022.104274

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J., Tang, G., et al. (2021). The abuse of popular performance metrics in hydrologic modeling. *Water Resources Research*, *57*(9), e2020WR029001. https://doi.org/10.1029/2020wr029001

Coerver, H. M., Rutten, M. M., & Van De Giesen, N. C. (2018). Deduction of reservoir operating rules for application in global hydrological models. *Hydrology and Earth System Sciences*, *22*(1), 831–851. https://doi.org/10.5194/hess-22-831-2018

Dang, T. D., Chowdhury, A. K., & Galelli, S. (2020). On the representation of water reservoir storage and operations in large-scale hydrological models: Implications on model parameterization and climate change impact assessments. *Hydrology and Earth System Sciences*, *24*(1), 397–416. https://doi.org/10.5194/hess-24-397-2020

De Paiva, R. C. D., Buarque, D. C., Collischonn, W., Bonnet, M.-P., Frappart, F., Calmant, S., & Bulhões Mendes, C. A. (2013). Large-scale hydrologic and hydrodynamic modeling of the Amazon River basin. *Water Resources Research*, *49*(3), 1226–1243. https://doi.org/10.1002/wrcr.20067

Döll, P., Fiedler, K., & Zhang, J. (2009). Global-scale analysis of river flow alterations due to water withdrawals and reservoirs. *Hydrology and Earth System Sciences*, *13*(12), 2413–2432. https://doi.org/10.5194/hess-13-2413-2009

Döll, P., Kaspar, F., & Lehner, B. (2003). A global hydrological model for deriving water availability indicators: Model tuning and validation. *Journal of Hydrology*, *270*(1–2), 105–134. https://doi.org/10.1016/s0022-1694(02)00283-4

Ekka, A., Keshav, S., Pande, S., van der Zaag, P., & Jiang, Y. (2022). Dam-induced hydrological alterations in the upper Cauvery river Basin, India. *Journal of Hydrology: Regional Studies*, *44*, 101231. https://doi.org/10.1016/j.ejrh.2022.101231

Farmer, D., Sivapalan, M., & Jothityangkoon, C. (2003). Climate, soil, and vegetation controls upon the variability of water balance in temperate and semiarid landscapes: Downward approach to water balance analysis. *Water Resources Research*, *39*(2), 1035. https://doi.org/10.1029/2001wr000328

Farmer, W. H., & Vogel, R. M. (2016). On the deterministic and stochastic use of hydrologic models. *Water Resources Research*, *52*(7), 5619–5633. https://doi.org/10.1002/2016wr019129

Fekete, B. M., Wisser, D., Kroeze, C., Mayorga, E., Bouwman, L., Wollheim, W. M., & Vörösmarty, C. (2010). Millennium ecosystem assessment scenario drivers (1970–2050): Climate and hydrological alterations. *Global Biogeochemical Cycles*, *24*(4), GB0A12. https://doi.org/10.1029/2009gb003593

Giuliani, M., & Herman, J. D. (2018). Modeling the behavior of water reservoir operators via Eigenbehavior analysis. *Advances in Water Resources*, *122*, 228–237. https://doi.org/10.1016/j.advwatres.2018.10.021

Gochis, D., Barlage, M., Cabell, R., Casali, M., Dugger, A., FitzGerald, K., et al. (2020). *The WRF-Hydro® modeling system technical description, (version 5.2. 0)*. ncar technical note.

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Haddeland, I., Heinke, J., Biemans, H., Eisner, S., Flörke, M., Hanasaki, N., et al. (2014). Global water resources affected by human interventions and climate change. *Proceedings of the National Academy of Sciences*, *111*(9), 3251–3256. https://doi.org/10.1073/pnas.1222475110

Haddeland, I., Skaugen, T., & Lettenmaier, D. P. (2006). Anthropogenic impacts on continental surface water fluxes. *Geophysical Research Letters*, *33*(8), L08406. https://doi.org/10.1029/2006gl026047

Hanasaki, N., Kanae, S., & Oki, T. (2006). A reservoir operation scheme for global river routing models. *Journal of Hydrology*, *327*(1–2), 22–41. https://doi.org/10.1016/j.jhydrol.2005.11.011

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., et al. (2008). An integrated model for the assessment of global water resources–part 1: Model description and input meteorological forcing. *Hydrology and Earth System Sciences*, *12*(4), 1007–1025. https://doi.org/10.5194/hess-12-1007-2008

Hargreaves, G. H., & Samani, Z. A. (1982). Estimating potential evapotranspiration. *Journal of the Irrigation and Drainage Division*, *108*(3), 225–230. https://doi.org/10.1061/jrcea4.0001390

Hejazi, M. I., Cai, X., & Borah, D. K. (2008a). Calibrating a watershed simulation model involving human interference: An application of multi-objective genetic algorithms. *Journal of Hydroinformatics*, *10*(1), 97–111. https://doi.org/10.2166/hydro.2008.010

Hejazi, M. I., Cai, X., & Ruddell, B. L. (2008b). The role of hydrologic information in reservoir operation–learning from historical releases. *Advances in Water Resources*, *31*(12), 1636–1650. https://doi.org/10.1016/j.advwatres.2008.07.013

Hoang, L. P., Lauri, H., Kummu, M., Koponen, J., Van Vliet, M. T., Supit, I., et al. (2016). Mekong river flow and hydrological extremes under climate change. *Hydrology and Earth System Sciences*, *20*(7), 3027–3041. https://doi.org/10.5194/hess-20-3027-2016

Kåresdotter, E., Destouni, G., Ghajarnia, N., Lammers, R. B., & Kalantari, Z. (2022). Distinguishing direct human-driven effects on the global terrestrial water cycle. *Earth's Future*, *10*(8), e2022EF002848. https://doi.org/10.1029/2022ef002848

Kauffeldt, A., Wetterhall, F., Pappenberger, F., Salamon, P., & Thielen, J. (2016). Technical review of large-scale hydrological models for implementation in operational flood forecasting schemes on continental level. *Environmental Modelling & Software*, *75*, 68–76. https://doi.org/10.1016/j.envsoft.2015.09.009

Khatami, S., Peel, M. C., Peterson, T. J., & Western, A. W. (2019). Equifinality and flux mapping: A new approach to model evaluation and process representation under uncertainty. *Water Resources Research*, *55*(11), 8922–8941. https://doi.org/10.1029/2018wr023750

Knoben, W. J., Freer, J. E., & Woods, R. A. (2019). Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Kuczera, G. (1983). Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty. *Water Resources Research*, *19*(5), 1151–1162. https://doi.org/10.1029/wr019i005p01151

Lamontagne, J. R., Barber, C. A., & Vogel, R. M. (2020). Improved estimators of model performance efficiency for skewed hydrologic data. *Water Resources Research*, *56*(9), e2020WR027101. https://doi.org/10.1029/2020wr027101

Li, D., Chen, Y., Cai, X., & Zhao, Q. (2023a). Data-driven reservoir operation rules for 450+ reservoirs in contiguous United States [Dataset]. *HydroShare*. https://doi.org/10.4211/hs.63add4d5826a4b21a6546c571bdece10

Li, D., Chen, Y., Cai, X., & Zhao, Q. (2023b). *Generic data-driven reservoir operation model*. GitHub repository. Retrieved from https://github.com/lidh966/GDROM?tab=readme-ov-file

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., & Kuczera, G. (2017). Improving probabilistic prediction of daily streamflow by identifying P Areto optimal approaches for modeling heteroscedastic residual errors. *Water Resources Research*, *53*(3), 2199–2239. https://doi.org/10.1002/2016wr019168

Meigh, J., McKenzie, A., & Sene, K. (1999). A grid-based approach to water scarcity estimates for Eastern and Southern Africa. *Water Resources Management*, *13*(2), 85–115. https://doi.org/10.1023/a:1008025703712

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part i—a discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Poff, N. L., Allan, J. D., Bain, M. B., Karr, J. R., Prestegaard, K. L., Richter, B. D., et al. (1997). The natural flow regime. *BioScience*, *47*(11), 769–784. https://doi.org/10.2307/1313099

Pokhrel, Y., Hanasaki, N., Koirala, S., Cho, J., Yeh, P. J.-F., Kim, H., et al. (2012). Incorporating anthropogenic water regulation modules into a land surface model. *Journal of Hydrometeorology*, *13*(1), 255–269. https://doi.org/10.1175/jhm-d-11-013.1

Rougé, C., Reed, P. M., Grogan, D. S., Zuidema, S., Prusevich, A., Glidden, S., et al. (2021). Coordination and control–limits in standard representations of multi-reservoir operations in hydrological modeling. *Hydrology and Earth System Sciences*, *25*(3), 1365–1388. https://doi.org/10.5194/hess-25-1365-2021

Sawicz, K. A. (2013). *Catchment classification-understanding hydrologic similarity through catchment function*. The Pennsylvania State University.

Shabestanipour, G., Brodeur, Z., Farmer, W. H., Steinschneider, S., Vogel, R. M., & Lamontagne, J. R. (2023). Stochastic watershed model ensembles for long-range planning: Verification and validation. *Water Resources Research*, *59*(2), e2022WR032201. https://doi.org/10.1029/2022wr032201

Singh, R., Wagener, T., Crane, R., Mann, M., & Ning, L. (2014). A vulnerability driven approach to identify adverse climate and land use change combinations for critical hydrologic indicator thresholds: Application to a watershed in Pennsylvania, USA. *Water Resources Research*, *50*(4), 3409–3427. https://doi.org/10.1002/2013wr014988

Solander, K. C., Reager, J. T., Thomas, B. F., David, C. H., & Famiglietti, J. S. (2016). Simulating human water regulation: The development of an optimal complexity, climate-adaptive reservoir management model for an LSM. *Journal of Hydrometeorology*, *17*(3), 725–744. https://doi.org/10.1175/jhm-d-15-0056.1

Turner, S. W., Steyaert, J. C., Condon, L., & Voisin, N. (2021). Water storage and release policies for all large reservoirs of conterminous United States. *Journal of Hydrology*, *603*, 126843. https://doi.org/10.1016/j.jhydrol.2021.126843

Turner, S. W., Voisin, N., Steyaert, J. C., & Condon, L. (2021). Istarf-conus (0.0.1) [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.4602277

Van Beek, L., Wada, Y., & Bierkens, M. F. (2011). Global monthly water stress: 1. Water balance and water availability. *Water Resources Research*, *47*(7), W07517. https://doi.org/10.1029/2010wr009791

Vora, A., Cai, X., Chen, Y., & Li, D. (2024). Dataset for "coupling reservoir operation and rainfall-runoff processes for streamflow simulation in watersheds" [Dataset]. *Mendeley Data*. https://doi.org/10.17632/t49cyrgtct.4,V4

Vora, A., & Singh, R. (2022). Improving rainfall-runoff model reliability under nonstationarity of model parameters: A hypothesis testing based framework. *Water Resources Research*, *58*(11), e2022WR032273. https://doi.org/10.1029/2022wr032273

Wallington, K., & Cai, X. (2023). Updating SWAT+ to clarify understanding of in-stream phosphorus retention and remobilization: SWAT+ PR &R. *Water Resources Research*, *59*(3), e2022WR033283. https://doi.org/10.1029/2022wr033283

Wisser, D., Fekete, B. M., Vörösmarty, C., & Schumann, A. (2010). Reconstructing 20th century global hydrography: A contribution to the Global Terrestrial Network-Hydrology (GTN-H). *Hydrology and Earth System Sciences*, *14*(1), 1–24. https://doi.org/10.5194/hess-14-1-2010

Yang, S., Yang, D., Chen, J., & Zhao, B. (2019). Real-time reservoir operation using recurrent neural networks and inflow forecast from a distributed hydrological model. *Journal of Hydrology*, *579*, 124229. https://doi.org/10.1016/j.jhydrol.2019.124229

Yassin, F., Razavi, S., Elshamy, M., Davison, B., Sapriza-Azuri, G., & Wheater, H. (2019). Representation and improved parameterization of reservoir operation in hydrological and land-surface models. *Hydrology and Earth System Sciences*, *23*(9), 3735–3764. https://doi.org/10.5194/hess-23-3735-2019

Yoshikawa, S., Cho, J., Yamada, H., Hanasaki, N., & Kanae, S. (2014). An assessment of global net irrigation water requirements from various water supply sources to sustain irrigation: Rivers and reservoirs (1960–2050). *Hydrology and Earth System Sciences*, *18*(10), 4289–4310. https://doi.org/10.5194/hess-18-4289-2014

Zhao, G., Gao, H., Naz, B. S., Kao, S.-C., & Voisin, N. (2016). Integrating a reservoir regulation scheme into a spatially distributed hydrological model. *Advances in Water Resources*, *98*, 16–31. https://doi.org/10.1016/j.advwatres.2016.10.014

Zhao, Q., & Cai, X. (2020). Deriving representative reservoir operation rules using a hidden Markov-decision tree model. *Advances in Water Resources*, *146*, 103753. https://doi.org/10.1016/j.advwatres.2020.103753