General Response for both reviewers.

We would like to express our sincere gratitude to both reviewers for their valuable feedback, which greatly assisted in reorganizing the content and enhancing the presentation of the results and discussions. In response to the suggestions provided by both reviewers, we have made several significant revisions to the manuscript. We will first outline these changes in this general response to the reviewers and then provide a point-by-point response to each reviewer's comments. Additionally, we would like to note that, in accordance with HESS guidelines, the manuscript revisions are not yet finalized, and we are still in the process of refining them. However, we have included the changes in their current form here to illustrate the revisions made thus far. Below are the main changes made to the manuscript for the sections with major changes.

Introduction:

We improved the introduction based on the recommendations made by both reviewers, and added background information on the IWAA program, referring to the new USGS report, introduced the bias adjusted conus404 dataset and why we performed this WRF-Hydro application. Below is the introduction text after making multiple edits to it:

Water availability is crucial for sustaining life, supporting ecosystems, and driving economic development. However, the balance between water supply and demand is increasingly strained due to factors such as climate change, pollution, and over-extraction. Recognizing the critical importance of water availability, the U.S. Congress has mandated federal agencies to conduct regular, comprehensive assessments to monitor and evaluate water resources across the country. In response, the U.S. Geological Survey (USGS) published two preliminary reports (Alley et al., 2013; Evenson et al., 2018), conducting Focused Area Studies and laying the groundwork for comprehensive national initiative (Stets et al., 2025).

The USGS Integrated Water Availability Assessments (IWAAs) is a comprehensive national initiative designed to evaluate water availability in the United States (U.S.) on a recurring basis. The inaugural cycle of this national water availability assessment has two primary objectives: firstly, to provide a status assessment of water availability for the period 2010 to 2020 on a national scale, and secondly to conduct a historical trend analysis exploring multi-decadal changes over time for the period 1980 to 2020. Subsequent USGS IWAAs will expand the assessment scope to include projections and undertake more focused regional studies (Miller et al., 2020).

To enable continuous, nationwide analysis—even in regions with sparse observational data—two national-scale hydrological models were utilized in the IWAA framework. The first model is a national-scale implementation of the Precipitation Runoff Modeling System (PRMS, Regan et al.

(2018)), while the second is The Weather Research and Forecasting (WRF) model hydrological modeling extension package (WRF-Hydro, Gochis et al. (2020)) which is discussed in detail in this paper. Stets et al. (2025) provides comprehensive insights into the IWAA, including the results from its initial activity, which assessed water availability over the water years 2010 to 2020 (Gorski et al., 2025). While the IWAAs address various dimensions of water availability, including quantity, quality, use, and aquatic ecosystems (Stets et al., 2025), this paper specifically focuses on the water quantity aspect.

Errors in simulated hydrologic components such as streamflow are aggregated errors emerging from errors in initial states, deficiencies in model structure, model parameter, and atmospheric forcing. Errors in the forcing dataset nonlinearly contribute to streamflow errors (Rafieeinasab et al., 2015) and, therefore, it is of great importance to choose the right forcing dataset for the application at hand. Ideally, one would like to force (and calibrate) the model using a dataset with an appropriate temporal and spatial resolution, a long-term data record, and physically consistent variables. The modeling applications used to support the first cycle of the IWAAs are forced by the state-of-the-art, CONUS404 dataset, a regional hydroclimate dataset over the conterminous United States (CONUS) developed through a collaborative initiative between the USGS and the National Center for Atmospheric Research (NCAR) (Rasmussen et al., 2023a). CONUS404 provides 40+ years of data at a spatial resolution of 4-km across CONUS and hence called CONUS404.

Through better representation of fine-scale weather phenomena, such as mesoscale convective systems and orographic precipitation, CONUS404 is able to produce a relatively accurate distribution of rainfall and temperature over a large area and a long period. The CONUS404 dataset provides an opportunity to study water-budget components at a relatively high spatial and temporal scale, which is of importance to hydroclimate studies. There is also a future scenario of CONUS404 providing an opportunity for studying climate change impacts on water budget components, making CONUS404 an appealing candidate for this study. Initial assessment of the CONUS404 dataset revealed some notable regional biases that could introduce inaccuracy in the hydrologic modeling and the model calibration procedure. Hence, in this study, the CONUS404 dataset (Rasmussen et al., 2023a) air temperature and precipitation are bias adjusted. The bias-adjusted CONUS404 is used to force (and calibrate) both IWAAs model applications (PRMS and WRF-Hydro, Stets et al. (2025)).

As mentioned above, WRF-Hydro (Gochis et al., 2020) is one of the two hydrological model applications used in the first cycle of the IWAAs (Stets et al., 2025). WRF-Hydro has been widely used in research and operations in configurations coupled to the atmosphere (e.g. Yucel et al., 2015; Fredj et al., 2015; Senatore et al., 2015; Arnault et al., 2016; Givati et al., 2016; Kerandi et al., 2018; Naabil et al., 2017; Verri et al., 2017; Varlas et al., 2018) and uncoupled applications (e.g. Xiang et al., 2017; Yin et al., 2022, 2021; Mehboob et al., 2022; Lee et al., 2022; Bao et al., 2022) where the model is forced by reanalysis or observational atmospheric *data. One of the most prominent applications of WRF-Hydro is the National Oceanic and Atmospheric Administration (NOAA) National Water Model (NWM). A particular instance of WRF-Hydro has been running operationally as the NWM since August of 2016 (Cosgrove et al., 2024; Read et al., 2023). Covering the CONUS along with parts of Canada and Mexico, the NWM significantly enhanced both temporal and spatial simulation resolutions of operational hydrological forecasting across the CONUS. The number of features for which forecasts are generated has increased from approximately 3,700 River Forecast Center prediction locations to over 2.7 million stream reaches derived from the National Hydrography Dataset NHDPlus version 2.1 (McKay et al., 2012).*

The WRF-Hydro instance used in this study aligns with the hydrography specifications of the NWM (Cosgrove et al., 2024) and uses similar physics options to NWMv3.0, with the exception of waterbody treatment. Waterbodies and water use are being represented in the IWAAs as a post-process, so the hydrologic models are estimating "natural" stream and waterbody inflows only. The IWAA application utilizes the bias-adjusted CONUS404 dataset. Therefore, it is necessary to calibrate the model to the new atmospheric forcing dataset and adjust the parameters accordingly.

This paper focuses on providing an in-depth account of the WRF-Hydro modeling effort within the IWAAs, specifically delving into the details of the WRF-Hydro model configuration, describing calibration and regionalization procedures, and evaluating its performance. This paper offers model evaluations of not only streamflow, but also the evapotranspiration, soil moisture and snowpack that are key factors in assessing water availability. This study focuses on providing bulk statistics of model performance compared to the available observation or other widely used model estimates, while Gorski et al. (2025) offers in-depth analysis of water availability based on the model simulation produced in this study and compares WRF-Hydro and PRMS model simulations.

Model Calibration and regionalization

In an effort to reduce the text, and keep the manuscript focused we have only provided the essential information regarding the calibration and regionalization and moved the details to the main manuscript. To keep it consistent we also moved the first section, “Evaluation of Calibration Basins” including Figures 7 and 8 to the supplement. The reduced text in the paper us as follows:

Conducting regional calibration for distributed models like WRF-Hydro is computationally expensive. One strategy to minimize this cost is to calibrate a select subset of basins, then extrapolating parameters to non-calibrated locations through a parameter regionalization process. We employ this strategy and calibrate 1,522 basins (Figure 5) which have minimal human impacts and are generally considered mostly natural flow basins, consistent with the

WRF-Hydro IWAAs configuration's exclusion of reservoirs, diversions, and other management. The core optimization algorithm used is the Dynamically Dimensioned Search (DDS) algorithm introduced by Tolson and Shoemaker (2007). In total, 17 WRF-Hydro model parameters (Table S1) are calibrated for the IWAAs configuration informed by a combination of pertinent scientific literature (Cuntz et al., 2016; Cosgrove et al., 2024; RafieeiNasab et al., 2025) and expert opinion.

The optimization procedure exclusively employs streamflow observations, with the (minimized) calibration objective function defined as 1 minus the Kling-Gupta efficiency (KGE) of hourly streamflow, where KGE is as proposed by Gupta et al. (2009). KGE for daily streamflow is applied in instances where there are insufficient hourly flow measurements. The choice of the hourly streamflow calibration and also use of KGE as the objective function is based on previous WRF-Hydro applications (Cosgrove et al., 2024; RafieeiNasab et al., 2025). Due to time limitations of the project, we did not experiment with any other temporal scale (daily or coarser) or a different objective function that might be more suitable for the water availability assessment than the current choices. The number of iterations in the DDS algorithm is set to 400 except for large domains (> 5,000 km^2), where only 200 iterations are used for computational tractability.

Before initiating the calibration process, a model run for each basin from October 2010 to October 2021 was spun up using default parameters. Subsequently, the "warm" model states from October 2021 serve as initial conditions for the calibration model runs, commencing from October 2012. While it is recognized that conditions in 2021 may differ from those in 2012, we assume that the seasonality and regional climate are similar. In addition to the single spin-up run with the default parameter, each calibration cycle incorporates a distinct 1-year acclimation period (from October 2012 to October 2013) with updated model parameters. This is to mitigate instabilities that could arise from the parameter change. The calibration phase spans a total of five water years (from October 2013 to October 2018). Independent validation period includes 2 years preceding the calibration interval (October 2011 to October 2013) and 3 years succeeding the calibration period (October 2018 to October 2021). The error metrics of simulated streamflow for both calibration and validation periods are reported in the Supplement (Figure S2 and S3).

To successfully execute the model application with spatially varying parameters across the CONUS, it is imperative to assign appropriate parameters to each grid cell within the model domain through a parameter regionalization approach. The attributes of the cells in each calibration basin are summarized and compared to summaries of attributes of all (non-calibrated) cells in 200 each USGS 10-digit hydrologic unit code (HUC10) of the Watershed Boundary Dataset (Jones et al., 2022). For each HUC10, the parameters from the calibration basin with the most similar characteristics are assigned to the cells within the HUC10. Two different set of basins attributes are used here to define similarity, 1) the Hydrological Landscape Region (HLR) framework (Winter, 2001; Wolock et al., 2004; Liu et al., *2008) 2) the Catchment Attributes and MEteorology for Large-sample Studies (CAMELS) dataset (Addor et al., 2017).*

Finally, since neither the HLR- or CAMELS-based regionalization approach exhibits universal superiority across all spatial contexts, we optimize the performance on a national scale across the CONUS by employing a mix-and-match strategy to select the better-performing approach (HLR or CAMELS). To do this, USGS 8-digit hydrologic unit codes (HUC8) are chosen as the spatial unit. For each HUC8 basin, we select the regionalization scenario that yields the best KGE calculated based on daily streamflow across the HUC8. Following the implementation of the mix-and-match approach and the establishment of the final configuration of the IWAAs WRF-Hydro CONUS model application, we conduct model simulations spanning the period from October 2009 to October 2021, encompassing the entire 10-year timeframe of the IWAAs program. More details on description of the regionalization are provided in Supplement.

Result and Discussion:

Considering the comments from both reviewers, we made the following changes to the sections.

- Moved the first section, “Evaluation of Calibration Basins” including Figures 7 and 8 to the supplement.
- Made modifications to “Regionalized Streamflow Evaluation” subsection, for readability and also addressing raised concerns and comments. Moved Figure 9 (c) to the supplement and removed the NSE part of Figure 10.
- Snow, ET and SM verifications remained mostly as presented in the original manuscript version.
- We have added a new section titled "Discussion of Water Budget Components" to provide a more detailed analysis of the water budget components. However, as noted in the introduction, a comprehensive water budget analysis has already been conducted by the USGS. Therefore, in this section, we focus on explaining the interactions between the water budget components, particularly those discussed in the previous sections, and propose potential solutions to address the identified shortcomings. Below is the newly added text, Figure N1 is suggested to be added to the main text, while Figure NS1, NS2 and NS3 are newly suggested figures that will be added to the supplement.

Discussion of Water Budget Components:

In this section, we will discuss the model biases of SWE, ET, SM and streamflow and their interactions with each other. We will not perform detailed water budget analysis here as Gorski et al., 2025 provides a detailed analysis of all water budget components based on the simulations provided by this study and also compares the finding against the national-scale implementation of the Precipitation Runoff Modeling System (Regan et al., 2018) over the CONUS. Instead, we *focus on providing reasoning of model behaviour and offering potential solutions for different regions across the US.*

We recognize that the current configuration of the IWAA may not be fully suitable for all water budget components, particularly the groundwater component. The existing setup is more appropriate for surface water analysis due to its simplified representation of groundwater and baseflow. Rummler et al. (2022) and Felfelani et al. (2024a) also emphasize the need for a more accurate representation of groundwater in the WRF-Hydro model. Ongoing research is exploring the integration of the U.S. Geological Survey's modular finite-difference flow model (MODFLOW) with WRF-Hydro, a development that could lead to significant improvements in model performance (Felfelani et al., 2024b). Given the limitation of the current WRF-Hydro model in presenting groundwater, we do not evaluate this water budget component here. Gorski et al., 2025 also performed the groundwater analysis based on well observational data rather than model simulations, and highlighted the groundwater modeling as an area for improvement in future IWAA studies.

Figure N1 shows the seasonal biases of ET, surface SM, root zone SM, SWE as well as streamflow. The streamflow bias for each month is the median percent bias of the GAGES-II reference basins in a given RFC. Figure N1 provides the mean across the years as the solid line, and the shaded area shows one standard deviation of a given quantity for that month. We also provided the scatter plots of percent bias of streamflow against ET, SWE, surface and root-zone soil moisture biases for each individual month during simulations period in Figure NS1 (RFCs with snow) and NS2 (RFCs with little to no snow). Correlation coefficients between streamflow biases and other water budget components are presented at each subpanel. Below we start with discussion points for the northeast US, then west U.S. and finally the great plains and southeast us.

Despite very little to no biases in overall streamflow metrics in the east US, there is a strong seasonal streamflow pattern with overestimations of streamflow at the fall and winter followed by an underestimation in spring and summer. While snow biases don't always align directly with streamflow biases, they do share a common trend. Notably, in regions like the NERFC, OHRFC, and MARFC, there is a noticeable drop in streamflow estimates during the melt season, following underestimation of snow water equivalent (SWE) values. Previous studies, such as Naple (2011), have identified this SWE underestimation in the region, which is typically linked to negative precipitation biases, positive temperature biases, and errors in precipitation partitioning (Naple, 2011; Minder et al., 2015). In our study, the initial CONUS404 dataset also showed low precipitation biases in this area, but these biases have been somewhat corrected in the adjusted CONUS404 dataset. Consequently, errors in model simulations are likely attributable to model settings (e.g., precipitation partitioning algorithm) and parameterization (calibrated parameters). It is possible that the phase partitioning has misclassified certain events as rain instead of snow, potentially due to temperature biases. As shown in Figure NS1, biases in *streamflow for MARFC, OHRFC, and NERFC are negatively correlated with biases in ET. Specifically, low ET biases tend to occur when streamflow biases are high. This issue could potentially be addressed by adjusting the model parameter set to partition a larger portion of precipitation into ET during the fall and winter months. We recommend exploring a more granular calibration approach by calibrating each season individually, which could help identify the optimal partitioning and improve model performance or using a multiobjective function which takes into account the seasonal biases.*

The NCRFC also exhibits similar snow underestimation. In this region, streamflow biases are also strongly correlated with snow biases, leading to a drop in streamflow values and underestimation during the spring and summer months. To address these low streamflow biases, improving snow simulations—either through more precise atmospheric forcing bias adjustments or enhanced phase partitioning—could prove beneficial. Unlike the above-mentioned RFCs (MARFC, OHRFC and NERFC) streamflow biases in the NCRFC are positively correlated with ET, except during the fall season, where a similar pattern of positive streamflow and negative ET is observed. Throughout the season, soil moisture also exhibits a consistent low bias. The region as a whole could benefit from a more effective partitioning of available water between streamflow (both direct and indirect runoff) and other components, particularly during the fall season. Despite calibrating parameters, streamflow still shows an overall low bias, suggesting that calibration alone may not fully address the limitations of the atmospheric forcing or model deficiencies. One potential improvement for this region could be the inclusion of subsurface tile drainage, given the area's high agricultural water management density. Valayamkunnath et al. (2022) demonstrated that incorporating subsurface tile drainage in the region led to reductions in surface runoff (-7% to -29%), groundwater recharge (-43% to -50%), evapotranspiration (-7% to 13%), and soil moisture (-2% to -3%), significantly improving model performance. While this capability was not utilized in the WRF-Hydro IWAA application, it is strongly recommended for future applications, as calibration alone has limited potential to address the model's shortcomings.

In the western U.S., the NWRFC exhibits unique behavior in terms of snow and streamflow biases. Snow biases in this region show a mix of positive and negative patterns: a slight positive snow bias at the start of the snow season, which shifts to a negative bias as the melt season begins. Interestingly, the streamflow bias is negatively correlated with snow biases, even when considering lagged time series correlations. However, both snow and streamflow biases are relatively small throughout most of the season, placing this region among the best-performing areas in the country. Note, the significant negative snow bias (~50%) observed in June, coinciding with the end of the melt season when snow water equivalent (SWE) values are typically low. The positive streamflow bias peaks at the end of the snow season and persists through the summer. This high streamflow bias can be attributed to the calibration adjustments made to account for exaggerated peak flows. These adjustments helped reduce the intensity of the peak flows, leading to a reduction in simulated streamflow biases and improving the KGE *values across the region. However, this improvement in peak flow representation came with trade-offs. The calibration introduced higher streamflow estimates during the recession limb of the hydrograph, leading to an overestimation of baseflow (Figure NS3). Additionally, the model exhibited high biases in soil moisture during this period. These issues likely stem from inadequate groundwater representation in the model, with the calibration attempting to compensate for this shortcoming by misplacing water in the system. Another contributing factor could be the improper partitioning of evapotranspiration, as indicated by the persistent negative ET bias throughout the season. This issue warrants further attention to improve model accuracy.*

The CBRFC and CNRFC exhibit similar bias patterns across different components. Both regions perform well at the start of the snow season in representing the snowpack, but they have lower peak SWE values and experience an earlier peak compared to SNODAS. A key area for improvement is the faster snowmelt rate observed in these regions compared to SNODAS, which could be addressed through better calibration. Currently, the MFSNO parameter—representing the melt factor in the snow depletion curve—is calibrated using streamflow observations to optimize streamflow performance. However, this approach may negatively impact the snowmelt rate. An ideal approach would be a stepwise calibration process, where snow-related parameters are first calibrated using snow-specific observations to maximize snow performance metrics. Although stepwise calibration was tested on a small subset of basins and showed superior accuracy for both snow and streamflow, time constraints prevented its full implementation for the IWAA WRF-Hydro application. In addition to MFSNO, other snow-sensitive parameters in the NoahMP scheme could be fine-tuned to improve snow representation that we recommend for future work. Both RFCs also suffer from an underestimation of ET for most of the year, except during the summer. The combined low snowpack and ET lead to significant overestimates of root-zone soil moisture across all seasons, as the model compensates for the shortcomings in snow and ET. In high-elevation areas of these RFCs, similar calibration artifacts as those observed in the NWRFC exist, where reducing the high streamflow peaks results in elevated baseflow values.

One of the deficiencies of the WRF-Hydro model in low-elevation semiarid regions of the Southwest is its lack of channel infiltration, which can be an important component of the water balance. Lahmers et al. (2019) introduced a conceptual channel infiltration function into the WRF-Hydro model architecture and found that accounting for channel losses not only improved streamflow performance but also reduced ET biases. However, high biases in soil moisture persisted in their simulations. Although this approach has shown promising results for the limited number of basins studied by Lahmers et al., it has yet to be tested on a regional or large-scale level. This capability may not need to be activated across the entire CONUS and currently, there is no study to determine where it should be implemented. It's also worth noting that in the implementation by Lahmers et al. (2019), the infiltrated water is lost from the system and does not contribute to soil moisture or groundwater recharge, meaning the water budget will not close if applied as-is. Given the time constraints of the current project, we have not *implemented the channel infiltration loss in the IWAA WRF-Hydro configuration. However, this approach may offer potential improvements for simulating water balance in the semiarid regions of the western U.S.*

The MBRFC, ABRFC, and WGRFC share several common features. All three regions exhibit spatially varied model performance, with poor simulations in the western areas and more reasonable performance along the eastern boundary. These regions are characterized by extensive agricultural land use, a large number of water diversions, active reservoirs (National Inventory of Dams, NID), and significant groundwater pumping (Scanlon et al., 2012). However, none of these factors are adequately represented in the current WRF-Hydro application. Previous studies using WRF-Hydro have shown similar challenges in model performance (Cosgrove et al., 2024), and difficulties in representing this area are not exclusive to WRF-Hydro. Other models also struggle with accurately simulating the region's behavior (Towler et al., 2022). Missing physical processes, such as water diversions and active reservoir management, as well as inadequate representation of groundwater, make it difficult to calibrate the model effectively. As shown in Figure S1, while calibration reduces high biases during the calibration period for the basins in these areas, these improvements do not persist during the validation period. Furthermore, after regionalization, the model still displays unsatisfactory performance, with high streamflow biases. Root zone soil moisture also shows a positive bias in these regions, suggesting that the model is incorrectly placing excess surface runoff into the soil. ET estimates are mostly unbiased, except in late spring and summer when significant biases are observed. Overall, the model struggles to partition water correctly within its current structure and requires modifications to better represent missing or poorly captured phenomena. As an example, the WRF-Hydro development team has been recently working toward adding diversion into the model code that could have a great potential, but it is still at the early stage of research. Another area of active research is the coupling of MODFLOW and WRF-Hydro which was mentioned earlier, and could enhance the quality of model simulations in this region to some degree.

LMRFC is among the RFCs with reasonable overall performance, for this region the ET biases are mostly the opposite sign of ET biases, suggesting the region could potentially be improved with a refined calibration process and improved partitioning of the available water. SERFC is a unique area also, with high streamflow biases before the calibration which was reduced with parameter estimation. However, the parameters did not transfer very well and southern Florida still suffers from high streamflow biases. The biases of the other water budget components, low soil moisture estimates along with low ET estimates, suggest this could be improved across the region with an improved water partitioning.

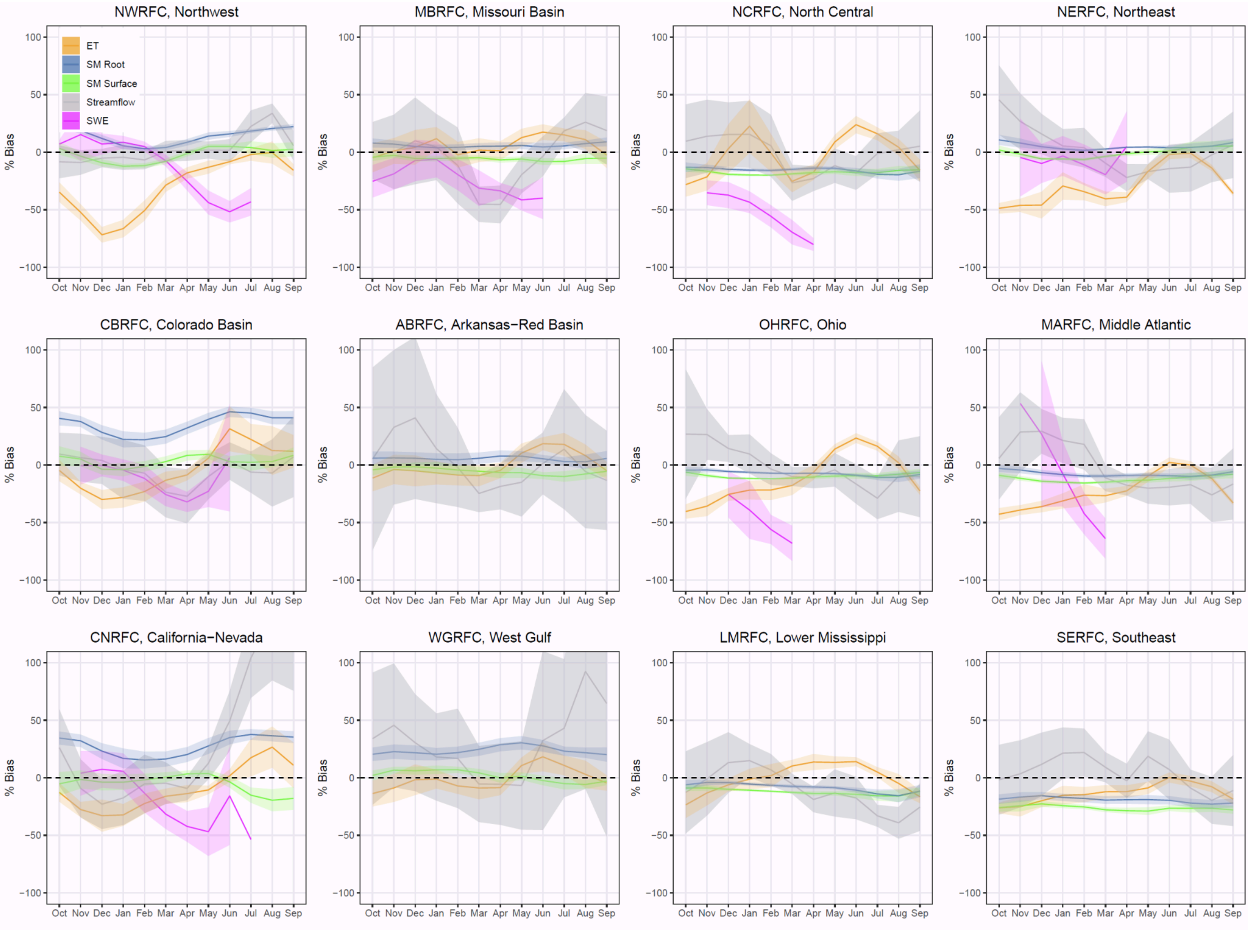


Figure N1. Time series of percent bias of monthly SWE, ET, surface SM, root-zone SM, and streamflow for each RFC region. The shaded area reflects the standard deviation of each variable across the years (2009-10 to 2021-10).

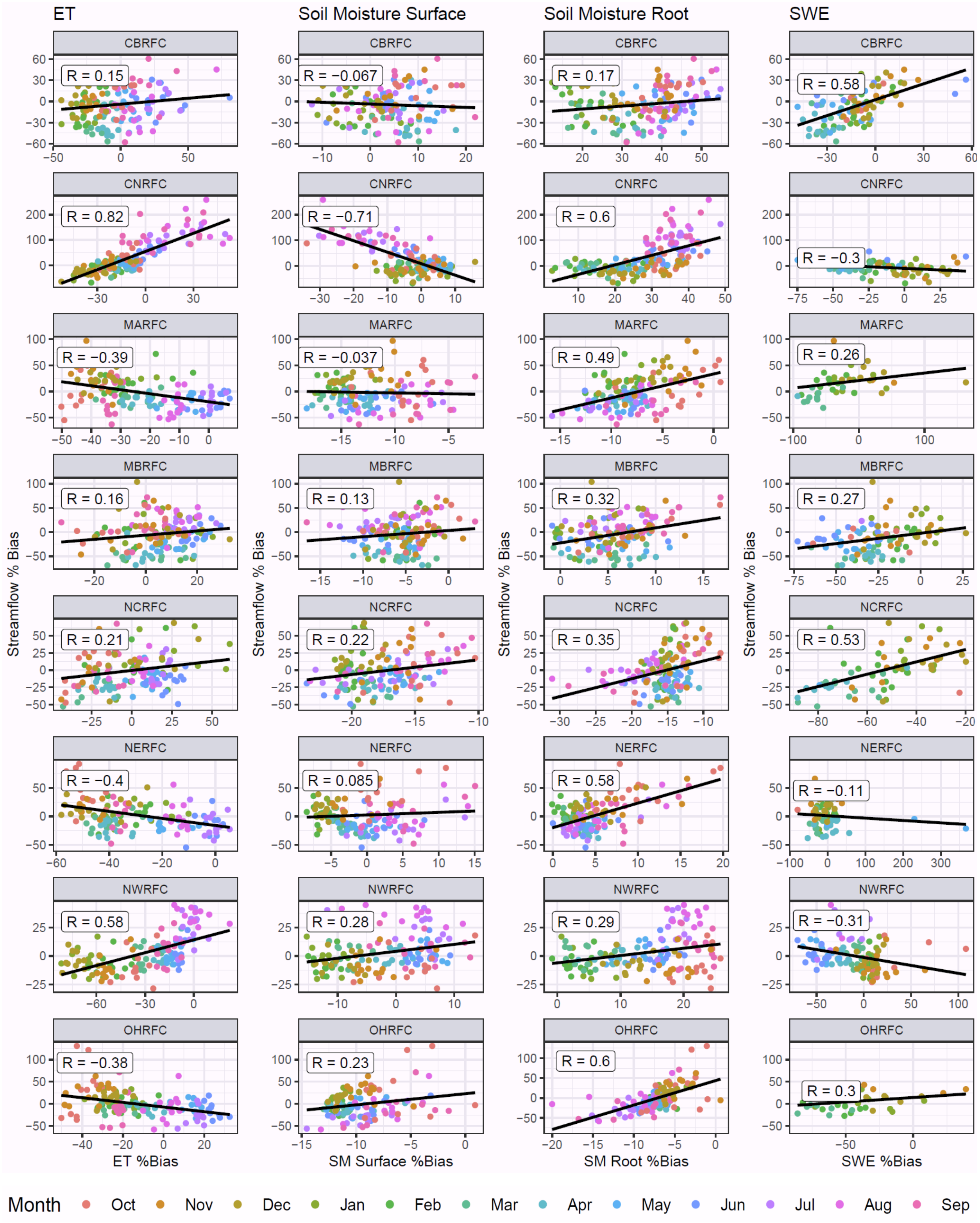

ET
Soil Moisture Surface
Soil Moisture Root
SWE
CBRFC
R = 0.15
R = −0.067
R = 0.17
R = 0.58
CNRFC
R = 0.82
R = −0.71
R = 0.6
R = −0.3
MARFC
R = −0.39
R = −0.037
R = 0.49
R = 0.26
MBRFC
R = 0.16
R = 0.13
R = 0.32
R = 0.27
NCRFC
R = 0.21
R = 0.22
R = 0.35
R = 0.53
NERFC
R = −0.4
R = 0.085
R = 0.58
R = −0.11
NWRFC
R = 0.58
R = 0.28
R = 0.29
R = −0.31
OHRFC
R = −0.38
R = 0.23
R = 0.6
R = 0.3
Streamflow % Bias
ET %Bias
SM Surface %Bias
SM Root %Bias
SWE %Bias
Month
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep

Figure NS1. Scatter plot of percent bias of monthly SWE, ET, surface SM, root-zone SM, against the streamflow bias (water years 2010-21) for each RFC region that receives large seasonal snow accumulation (> 5mm peak annual SWE). Color coding shows different months of the year.

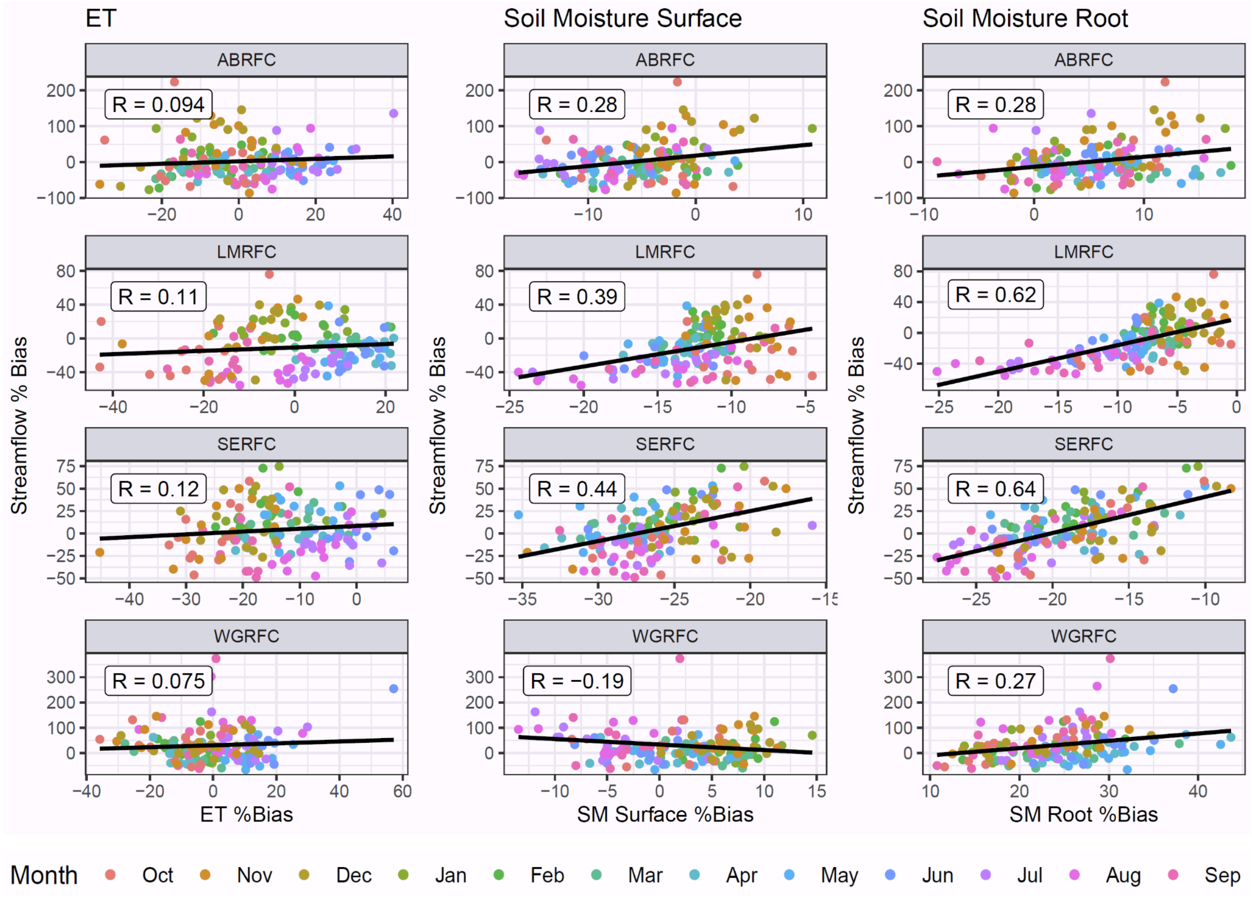
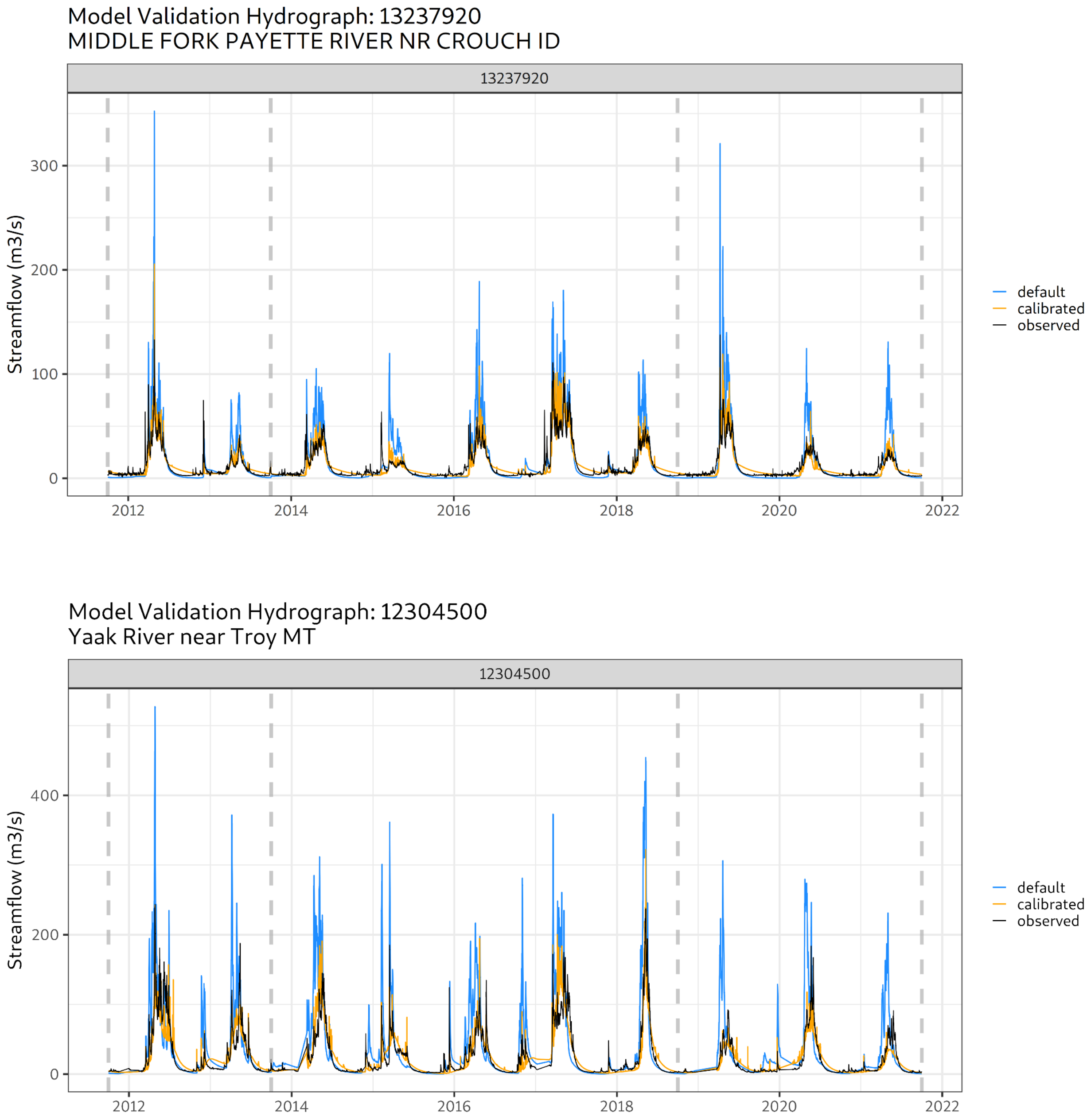


Figure NS2. Scatter plot of percent bias of monthly SWE, ET, surface SM, root-zone SM, against the streamflow bias (water years 2010-21) for RFC regions with insignificant seasonal snow accumulation (< 5mm peak annual SWE). Color coding shows different months of the year.



Model Validation Hydrograph: 13237920
MIDDLE FORK PAYETTE RIVER NR CROUCH ID

Model Validation Hydrograph: 12304500
Yaak River near Troy MT

Figure S3. Two sample hydrographs in the NWRFC with the default (in blue) and calibrated parameters (in orang) against the streamflow observations (in black) during the calibration period (2013-10 to 2018-10) and validation period (2011-10 to 2013-10 and 2018-10 to 2021-10).

Newly Added References:

Valayamkunnath, Prasanth, et al. "Modeling the hydrologic influence of subsurface tile drainage using the National Water Model." Water Resources Research 58.4 (2022): e2021WR031242.

Naple, P. W., 2021: Evaluating the performance of National Water Model snow simulations in the Northeastern United States using advanced mesonet observations. M.S. thesis, Dept. Atmospheric and Environmental Sciences, University at Albany, 98 pp.

Justin R. Minder , Theodore W. Letcher, Arezoo RafieeiNasab, Patrick W. Naple, Sierra Liotta, Junhong Wang. 2025. Evaluating and Improving Snow in the National Water Model, using Observations from the New York State Mesonet. Journal of Hydrometeorology, Volume 26, Issue 1, 69-90. https://doi.org/10.1175/JHM-D-24-0057.1

Rummler, Thomas, et al. "Lateral terrestrial water fluxes in the LSM of WRF‐Hydro: benefits of a 2D groundwater representation." Hydrological Processes 36.3 (2022): e14510.

Lahmers, Timothy M., et al. "Enhancing the structure of the WRF-hydro hydrologic model for semiarid environments." Journal of Hydrometeorology 20.4 (2019): 691-714.

Stets, E.G., Archer, A.A., Degnan, J.R., Erickson, M.L., Gorski, G., Medalie, L., and Scholl, M.A., 2025, The National integrated water availability assessment, water years 2010–20, chap. A of U.S. Geological Survey Integrated Water Availability Assessment—2010–20: U.S. Geological Survey Professional Paper 1894–A, 24 p., https://doi.org/10.3133/pp1894A.

Gorski, G., Stets, E.G., Scholl, M.A., Degnan, J.R., Mullaney, J.R., Galanter, A.E., Martinez, A.J., Padilla, J., LaFontaine, J.H., Corson-Dosch, H.R., and Shapiro, A., 2025, Water supply in the conterminous United States, Alaska, Hawaii, and Puerto Rico, water years 2010–20 (ver. 1.1, January 17, 2025), chap. B of U.S. Geological Survey Integrated Water Availability Assessment—2010–20: U.S. Geological Survey Professional Paper 1894–B, 60 p., https://doi.org/10.3133/pp1894B.

Scanlon, Bridget R., et al. "Groundwater depletion and sustainability of irrigation in the US High Plains and Central Valley." Proceedings of the national academy of sciences 109.24 (2012): 9320-9325.

Towler, E., Foks, S. S., Staub, L. E., Dickinson, J. E., Dugger, A. L., Essaid, H. I., Gochis, D., Hodson, T. O., Viger, R. J., and Zhang, Y.: Daily streamflow performance benchmark defined by the standard statistical suite (v1.0) for the National Water Model Retrospective (v2.1) at benchmark streamflow locations for the conterminous United States (ver 3.0, March 2023), US Geological Survey data release [data set], https://doi.org/10.5066/P9QT1KV7, 2023.

Felfelani, Farshid, et al. "Simulation of groundwater-flow dynamics in the US Northern High Plains driven by multi-model estimates of surficial aquifer recharge." Journal of Hydrology 630 (2024): 130703.

Farhsid Felfelani et al. 2024. Progress in development of WRF-Hydro/MODFLOW coupled hydrologic modeling system: Interface Structure and showcase demonstration. AGU Fall Meeting, Washington D.C., US. (https://agu.confex.com/agu/agu24/meetingapp.cgi/Paper/1639136)

Regan, R.S., Markstrom, S.L., Hay, L.E., Viger, R.J., Norton, P.A., Driscoll, J.M., and LaFontaine, J.H., 2018, Description of the national hydrologic model for use with the precipitation-runoff modeling system (prms): U.S. Geological Survey Techniques and Methods, book 6, chap. B9, 38 p., accessed September 21, 2023, at https://doi.org/ 10.3133/ tm6B9.

Response to Reviewer 2: review posted on Dec 20th.

Thank you so much for the detailed review, and providing helpful comments. We have made changes throughout the manuscript to hopefully address your concerns and also moved some of the text to Supplement in an effort to reduce the length of the main manuscript and keep the point of the study clear. We also added a new section called “Discussion of Water Budget Components” to address multiple raised concerns about the discussion section. The significant changes made to the manuscript are mentioned in the general response to the reviewers, and below we provide point-by-point responses to the questions and comments. The text from the manuscript is in *italic* and the newly added text is provided in ***bold***.

The manuscript titled "A WRF-Hydro based retrospective simulation of water resources for US integrated water availability assessment" by Rafieeinsasab et al. discusses a nationwide setup of the WRF-Hydro modeling system aimed at assessing water resources and availability in the United States. The paper is well written and clearly structured; however, it seems more like a report than an exploration of scientific questions. There are already numerous studies available that describe the calibration of WRF-Hydro. What's new here is the parameter regionalization approach but it seems to have a limited applicability for some of the regions. The discussion of the calibration results doesn't highlight potential starting points for improvements of the model (or does touch the sore spots).

We agree that there are multiple studies on the WRF-Hydro calibration. In an effort to address the concern of yours we moved most of the details of the calibration and regionalization to the supplement as well as the results and kept the essential part of it only in the text. One of the regionalization scenarios is similar to Cosgrove et al. 2024 and the other approach (CAMEL based) is a newly developed approach and not reported in any other manuscript and therefore we detailed the whole process in this paper. We have made several comments on the limitations of the calibration strategy in the newly added section called “Discussion of Water Budget Components” in response to your comments and in an attempt to explain the model performance. Please refer to the general comments to review the changes made and the raised points related to the calibration/regionalizations deficiencies. In the same section, we also added more details on the model itself, the current deficiencies, and areas for improvement.

On the other hand, the water budget study remains quite general with only long-term average analyses and no in depth discussion of the model's applicability to the different climatic conditions of the CONUS. Thus, the study's objective is ambiguous as it is neither a proper physically based analysis of the model itself nor an detailed examination of the model's capability to simulate water budgets under various climatic conditions. Therefore, I'd like to invite the authors to focus more on one of these aspects and I would recommend to do this in favor of a more nuanced analysis of the model's capability and shortcomings to simulate water budgets and resources.

This was a very helpful recommendation. The newly added section to the results entitled "Discussion of Water Budget Components" attempts to address the concern raised. Please refer to the general response to the reviewer to see the changes made. We also added the reference to the newly released USGS report based on the model simulations of this work as well as the PRMS model in the text for readers who would like to see more in-depth analysis of each water budget component and also comparison against the PRMS model.

Specific comments:

Here the model was used in a similar setup as it has been developed for the NWM which primarily addresses discharge prediction. However, since this study aims to examine water budgets – also on a longer term perspective – it is worth questioning whether the chosen NWM-based setup is appropriate for the task. Groundwater, for instance, is considered only in terms of a discharge contribution, approximated by a conceptual linear reservoir model. However, for water resources assessment Groundwater represents an important storage body. As shown in Rummler et al. (2022, https://doi.org/10.1002/hyp.14510) a more sophisticated description of groundwater processes WRF-Hydro can lead to significantly improved discharge estimates. Further improvements to the model were demonstrated in a study focused on semi-arid environments, in which even some of the co-authors were involved (Lahmers et al. 2019, https://doi.org/10.1175/JHM-D-18-0064.1). However, these findings were not considered here even though they could have enhanced simulations for the southwestern regions.

Thank you for raising the concern. While the NWM was developed to represent processes beyond just discharge, we agree with the statement made by the reviewer that groundwater plays a critical role in water supply and the current configuration of the model is mostly suitable for surface water analysis. In the recently released USGS report, which the IWAA simulations was performed for, GW analysis is being performed using observational data from groundwater wells instead (https://doi.org/10.3133/pp1894B). In the same report, incorporation of groundwater flow processes and groundwater surface water interactions is identified as an area of potential improvement for future water supply assessments. That being acknowledged, the WRF-Hydro group is currently working on coupling the MODFLOW with WRF-Hydro, which could address the issues with GW simulations to some extent in the future studies. The offline coupling results have been published in Felfelani et al. 2024 (https://www.sciencedirect.com/science/article/abs/pii/S0022169424000970), and the full coupling using BMI was presented by Felfelani at AGU 2024 (https://agu.confex.com/agu/agu24/meetingapp.cgi/Paper/1639136) and a manuscript is in preparation. To address the raised concerns, we added the following to the new section before diving into verification of other components.

"We recognize that the current configuration of the IWAA may not be fully suitable for all water budget components, particularly the groundwater component. The existing setup is more

appropriate for surface water analysis due to its simplified representation of groundwater and baseflow. Rummler et al. (2022) and Felfelani et al. (2024a) also emphasize the need for a more accurate representation of groundwater in the WRF-Hydro model. Ongoing research is exploring the integration of the U.S. Geological Survey's modular finite-difference flow model (MODFLOW) with WRF-Hydro, a development that could lead to significant improvements in model performance (Felfelani et al., 2024b). Given the limitation of the current WRF-Hydro model in presenting groundwater, we do not evaluate this water budget component here. Gorski et al., 2025 also performed the groundwater analysis based on well observational data rather than model simulations, and highlighted the groundwater modeling as an area for improvement in future IWAA studies.*"*

Regarding the comment made about Lahmers et al. 2019, we agree with the reviewer that this work has shown benefit in semiarid regions. However, the work has been tested on only a few basins and we have not done testing of the model performance on regional to CONUS-scale simulations. Also, in the current configuration implemented by Lahmers, infiltrated water through the channel is lost, and does not feed back into the soil or groundwater, and therefore the water budget will not close. This enhancement/capability needs a careful strategy where to implement it to be consistent with local conditions and not mask other important processes, and this analysis was beyond the scope of the IWAA project. For the IWAA project, we had to operate under a very tight timeline for the production of the model simulations, which didn't allow for that broader assessment of where channel losses would be active. To address the concern of the reviewer the following paragraph in the newly added section (discussion) of the manuscript.

"One of the deficiencies of the WRF-Hydro model in low-elevation semiarid regions of the Southwest is its lack of channel infiltration, which can be an important component of the water balance. Lahmers et al. (2019) introduced a conceptual channel infiltration function into the WRF-Hydro model architecture and found that accounting for channel losses not only improved streamflow performance but also reduced ET biases. However, high biases in soil moisture persisted in their simulations. Although this approach has shown promising results for the limited number of basins studied by Lahmers et al., it has yet to be tested on a regional or large-scale level. This capability may not need to be activated across the entire CONUS and currently, there is no study to determine where it should be implemented. It's also worth noting that in the implementation by Lahmers et al. (2019), the infiltrated water is lost from the system and does not contribute to soil moisture or groundwater recharge, meaning the water budget will not close if applied as-is. Given the time constraints of the current project, we have not implemented the channel infiltration loss in the IWAA WRF-Hydro configuration. However, this approach may offer potential improvements for simulating water balance in the semiarid regions of the western U.S."

For the bias correction, what is the reasoning for using a "day-of-the-year" approach? Is it that you want to do a kind of climatological bias correction? You stated that Daymet is only available until 2017. Looking to the data portal, one can find data until 2023. So data including 2021 should have been available at the time of your analysis. Nevertheless, using a long term climatological correction does not account for inter-annual variability (e.g., enduring extraordinary drought periods) and may further disregard fundamental changes in climate between the overall averaging period (1980-2017) and the study period (2010-2021). With the available data you could have pursued a more time-variant approach.

The reviewer is correct that Daymet data are available for more recent dates in the portal, and our use of Daymet data up to 2017 at the time of the processing was based on quite some analyses through trial and error. First, we obtained the original PRISM data and examined CONUS404 bias structures in reference to PRISM and we noticed large biases especially in the snow dominated regions. We consulted the literature and also compared PRISM with SNOTEL data and we realized that there might be some inherent biases in PRISM (e.g., related to snow undercatch in some of the observational datasets as well as cold biases at the SNOTEL sites). Then we were informed about the availability of a better quality PRISM dataset with some of the biases corrected and we purchased the better quality PRISM up to 2017 and we indeed noticed some improvements in the better quality PRISM compared to the original version. Subsequently, we obtained and examined Daymet and compared CONUS404 with both Daymet and better quality PRISM and we got quite similar bias structures between Daymet and PRISM. Since Daymet has a high spatial resolution of 1-km compared to PRISM's 4-km, we decided to use Daymet (up to 2017 similar to PRISM) as the reference dataset for doing the bias corrections. We have since obtained Daymet up to the present time and also the better quality PRISM up to 2022, however, the data used at the time of the study was limited to 2017. We now have added additional wording to the text to read:

"This time frame and Daymet were used based on extensive analyses of the CONUS404 bias structures in reference to a number of observational datasets including Daymet and PRISM and also the time frame covered the concurrent dates from the Daymet, PRISM and CONUS404 datasets we had available in our local repository."

As the reviewer suspected, we chose to do a day-of-year correction to adjust for how the mean bias changes climatologically throughout the year. We now better describe this with a new sentence (in bold here):

"The day-of-year biases were then calculated at every pixel by averaging the biases for each day of the year from the 38-year data set, applying a 31-day smoothing to both precipitation and temperature biases to remove anomalous day-to-day fluctuations in calculated biases. ***This was done to account for how the bias varies climatologically on an annual basis; interannual***

variability was not considered, since there was more uncertainty in its sign and magnitude. *Figure 3 shows the domain-averaged temperature and precipitation biases."*

We opted not to do an inter-annual variability correction, as there would have been greater uncertainty in how much of the residual bias is from inter-annual variability, and how much is from other factors. We also address this in the text (see new sentence above.)

Concerning the selection of relevant parameters (Table 1) it is mentioned that it also relies also on literature review; however, the extensive Noah MP parameter study by Cuntz et al. (2016, https://doi.org/10.1002/2016JD025097) is not considered or at least not mentioned. How has the relevance of the parameters been assessed in general? In terms of calibration results it would be interesting to learn about the ranges, patterns, and relationships of the optimized values but this relates to my initial comment about the focus of this study –whether it should concentrate more on the model itself or on the budget analysis.

Thank you for the comment. The focus of the paper is introduction of the application in hand, and evaluating the model performance against the widely used products, to highlight the potential of the models for water availability study. Therefore, we avoid shifting the discussion toward in-depth analysis of the model parameters and how the calibration is impacting the parameter range. Following the concerns of both reviewers, we have reduced the calibration/regionalization section, cited Cosgrove et al. 2024 and Rafieeinasab et al. 2025 (https://doi.org/10.1029/2024WR038048) where possible, and moved content to the Supplement. Also we agree with the reviewer that Cuntz et al. 2016 has been a leading paper in sensitivity analysis of the NoahMP parameters and we added the missing citation to the manuscript.

Why do you consider hourly measures for the evaluation of the calibration basins and daily measures for the regionalized stream flow?

Due to time constraints in the start of the project we performed hourly calibration, which is the common time resolution most WRF-Hydro applications use, and did not have enough time to experiment with other temporal resolutions or other objective functions. However, as we proceeded into the project, we decided to put less emphasis on the sub-daily error metrics given the goal of the project is water availability assessment, and performed regionalization and also verifications of streamflow at the daily time step. One could argue that the calibration also should have been done at a daily timescale and that is fair; therefore, we added the following to the text to add context for the reader.

The choice of the hourly streamflow calibration and also use of KGE as the objective function is based on previous WRF-Hydro applications (Cosgrove et al. 2024, RafieeiNasab et al. 2025). Due to time limitations of the project, we did not experiment with other temporal scales (daily or coarser) or different objective functions.

That being acknowledged, even though the first use of the model simulation is at a coarse spatial (HUC12) and temporal (monthly) scale by USGS for water availability assessment (Stets et al. 2024: https://doi.org/10.3133/pp1894A), we foresee other applications could benefit from the sub-daily calibration and sub-daily outputs.

It is anticipated that for the basins with regionalized parameter sets performance will be reduced. Have you considered employing also other regionalization approaches, such as proposed by Schweppe et al. (2022, https://doi.org/10.5194/gmd-15-859-2022)?

Yes, we have considered this work in the past. MPR provides smoother parameter fields, however, according to Mizukami et al. 2017, the performance at the calibration sites using individual basin calibration slightly outperforms the model performance using MPR approach (Figure 5 and 6, in particular the NSE error metric). Also, there is not enough guidance on transfer function forms and geophysical predictors to be used for each parameter and we usually calibrate a relatively large number of parameters (impacting soil, vegetarian, snow and runoff). So we decided to calibrate all basins with reliable streamflow data (and limited human interference) and then transfer the calibrated parameters from these basins to uncalibrated areas based on catchment similarity.

In section 5.2 it would be very interesting to read more about the physical reasons for the mismatch of the simulations. Perhaps this could also be discussed in a final overarching interpretation of the results that includes all the different water budgets analyzed in this study.

Thank you for the comment. This is now discussed more in depth along with other water budget components in the newly added section. Please refer to the general response to the reviewers.

The rest of the analysis for snow, evaporation, soil moisture only shows that the model can reproduce a 10-year climatological mean. This may also be achievable from an reanalysis. Here I would expect a more in-depth analysis of the water budgets, regional features, the shortcomings of the model and potential for improvement, as well as the benefits of simulating the water budgets at such high resolution and over this large spatial extent. Additionally, the general suitability of the model in its current configuration for longer-term trend analysis and future water availability projections as is intended by the IWAA initiative should be discussed.

Thank you so much for the useful comment. In order to address multiple concerns raised by both reviewers, we have added a discussion section to explain further the impact of different processes on the streamflow error metrics.

Finally, my last point is about long term storage, i.e. groundwater. There is no evaluation of the changes in terrestrial water storage which I assume is important for water budget analysis and water balance closure. GRACE derived deviations should be well suited for such an analysis given the large spatial extent of the study region.

Thank you for the comment. To address your comments earlier, we mentioned in the discussion section that the GW simulation in this study is based on a simple GW conceptualization, and so the USGS assessment report is using well observations for the groundwater analysis instead. That being acknowledged, another colleague is currently working on verification of the total water storage of IWAA (as well as other models such as NWM) against GRACE, which he plans to publish in a different manuscript and, therefore, we refrain from bringing that into this manuscript.

Minor comments:

- L263: There are several studies that point out the ineptitude of NSE as a goodness of fit measure for daily to subdaily hydrographs.
  - NSE is a widely used metric and generally we do not see any issue in using it as a verification error metric along other metrics such as bias, correlation and KGE. We agree that there are studies questioning the suitability of using NSE as an error metric and that is one of the reasons for using KGE as the objective function in the calibration process. We did not use NSE in the optimization or decision-making anywhere in the process, and only presented it as part of a suite of supplementary error metrics for model evaluation. However, during the revision effort, to shorten the paper and figures, we dropped all the instances where we report NSE values from the paper. Therefore, NSE error metrics have been removed from Figure 8 (now Figure S3) and Figure 10 (now Figure 7) and text has been edited accordingly.
- L263-275: There's no information about the temporal resolution used for model validation.
  - The following text has been added to the manuscript to address the raised concern. ***"Since calibration of model parameters is performed using hourly streamflow, the verification of the results presented in the Supplement is also presented at the hourly time step. However, given the main model application is focused on water availability assessment, the regionalized streamflow verification is performed at the daily time step. The verification of other variables such as evapotranspiration, soil moisture and snowpack is performed at the monthly time scale."***
- Figs. 7 & 8: Captions doe not include information about the time period used for the analysis.
  - Figure 7 and 8 are now in the Supplement and labeled as Figure S2 and Figure S3. As suggested, the time period has been added to the captions.
- Figs. 14–19: Captions doe not include information about the time period used for the analysis.
  - Time period of the verification is added to the caption of all figures.