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.

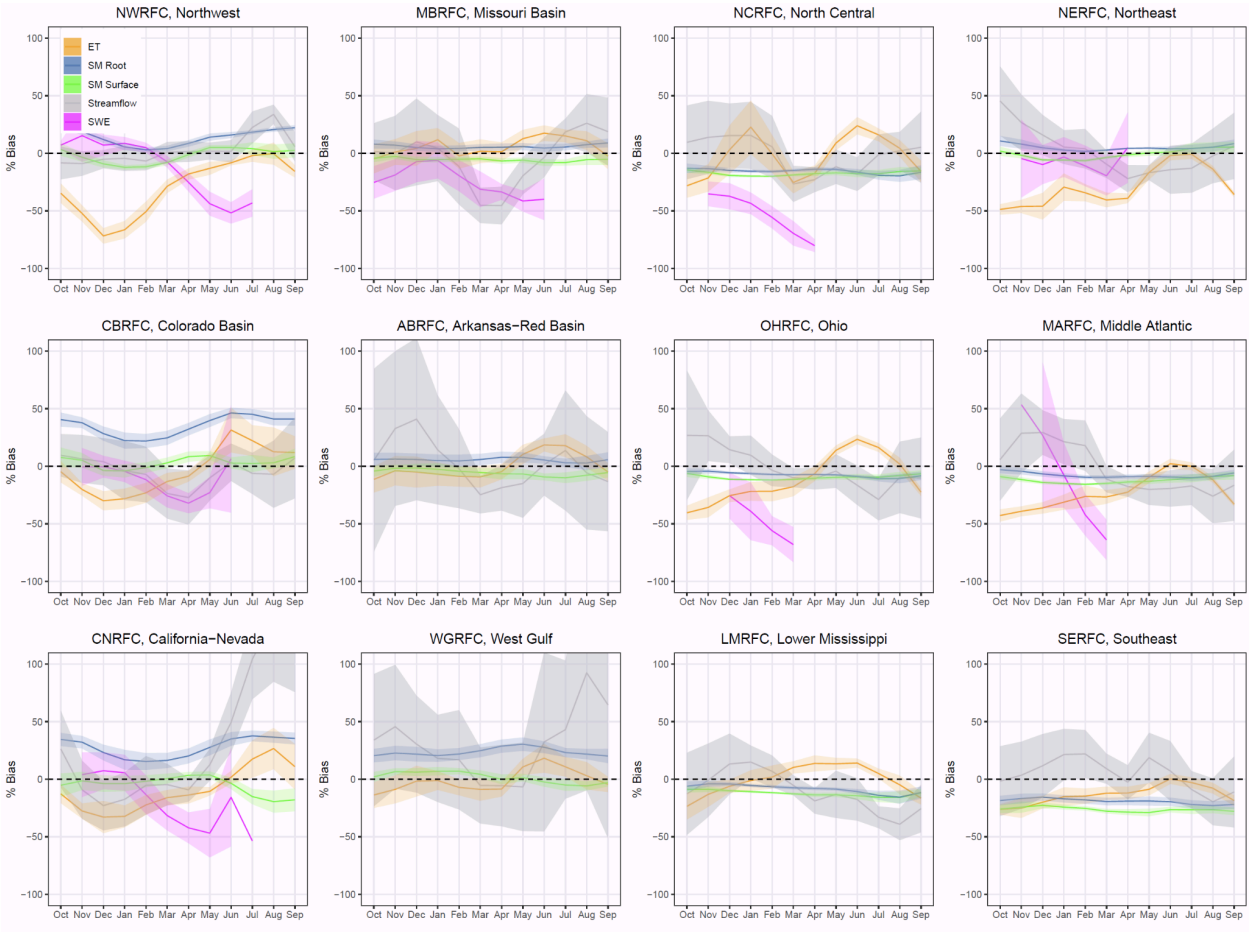


Figure N1. Time series of percent bias of monthly SWE, ET, surface SM, root-zone SM, and streamflow for each RFC region. The shaded area reflects the standard deviation of each variable across the years (2009-10 to 2021-10).

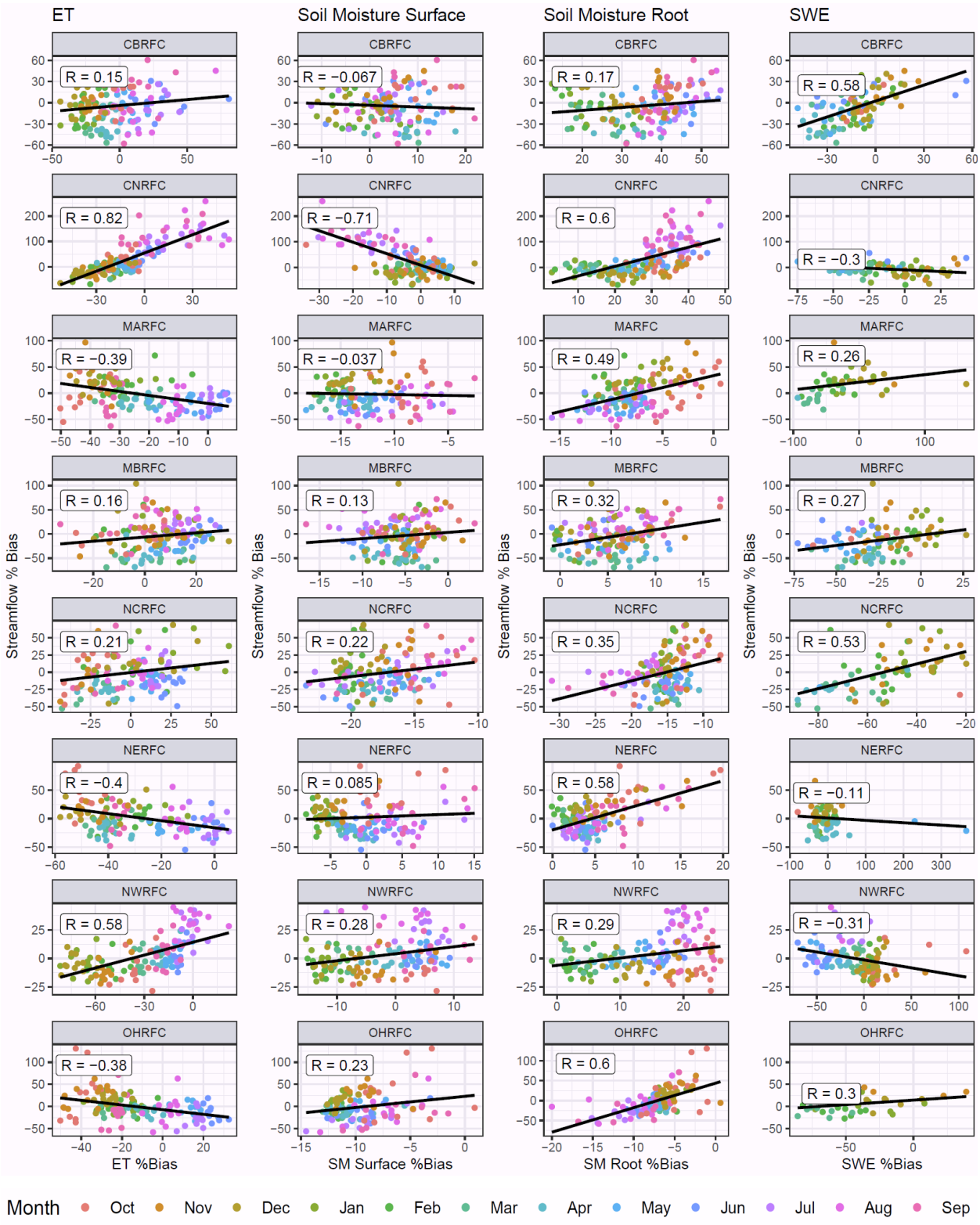

ET
Soil Moisture Surface
Soil Moisture Root
SWE
CBRFC
R = 0.15
R = −0.067
R = 0.17
R = 0.58
CNRFC
R = 0.82
R = −0.71
R = 0.6
R = −0.3
MARFC
R = −0.39
R = −0.037
R = 0.49
R = 0.26
MBRFC
R = 0.16
R = 0.13
R = 0.32
R = 0.27
NCRFC
R = 0.21
R = 0.22
R = 0.35
R = 0.53
NERFC
R = −0.4
R = 0.085
R = 0.58
R = −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.

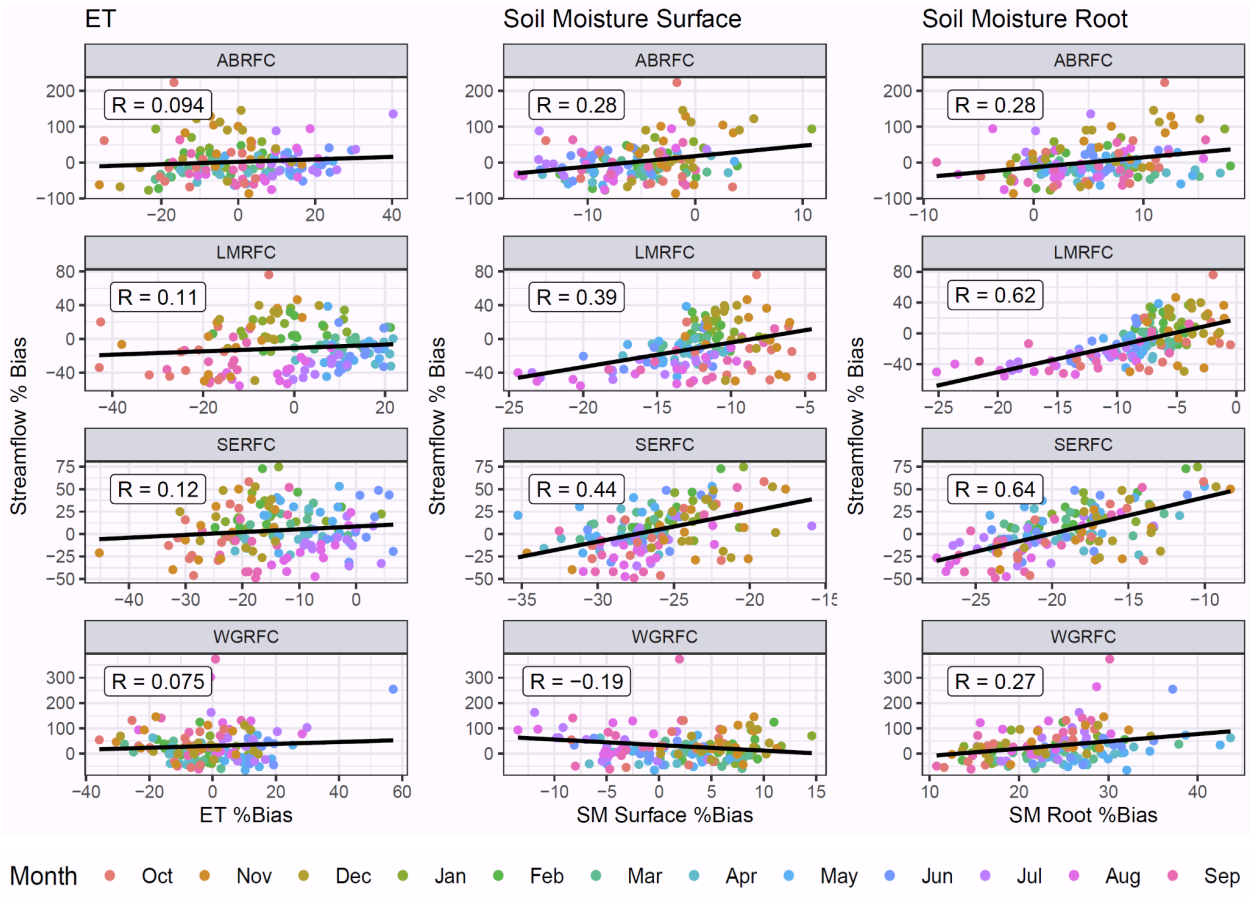


Figure NS2. Scatter plot of percent bias of monthly SWE, ET, surface SM, root-zone SM, against the streamflow bias (water years 2010-21) for RFC regions with insignificant seasonal snow accumulation (< 5mm peak annual SWE). Color coding shows different months of the year.

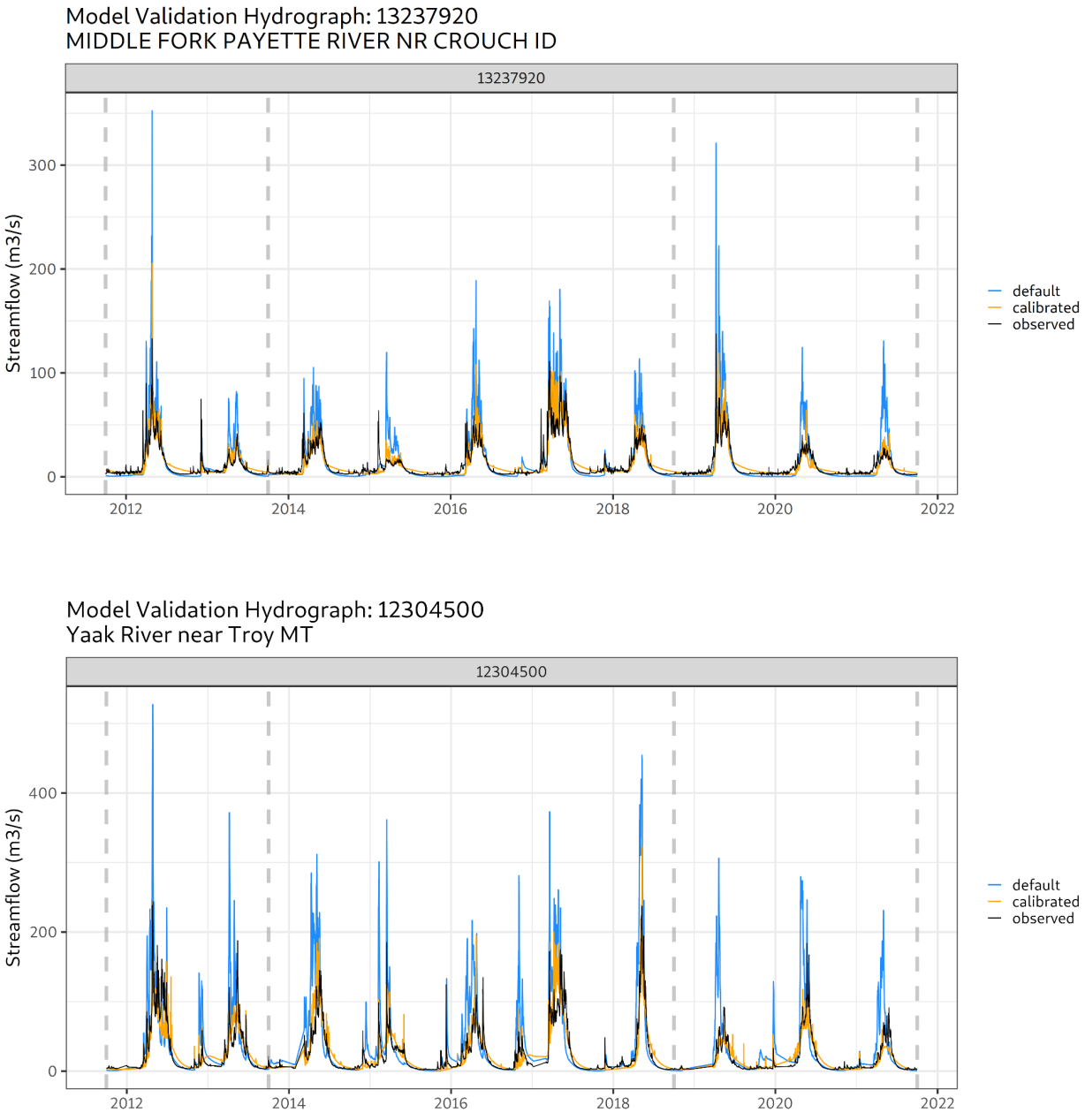


Figure S3. Two sample hydrographs in the NWRFC with the default (in blue) and calibrated parameters (in orang) against the streamflow observations (in black) during the calibration period (2013-10 to 2018-10) and validation period (2011-10 to 2013-10 and 2018-10 to 2021-10).

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 1, review posted on Dec 10th.

General Comments:

This paper documents the WRF-Hydro IWAA modeling application in terms of its configuration and performance. It is generally well-written, and provides a detailed and comprehensive documentation of the WRF-Hydro IWAA. However, it has several main issues that need to be addressed in relation to the three HESS Review Criteria (i. Scientific significance, ii. Scientific quality, iii. Presentation quality). I recommend "Major Revision".

Thank you so much for your valuable comments, and thorough revision. 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 a point-by-point response to the questions and comments. The text from the manuscript is in italic and the newly added text is provided in bold.

· **Scientific Significance is "Fair"** *(Does the manuscript represent a substantial contribution to scientific progress within the scope of HESS (substantial new concepts, ideas, methods, or data)?"* The paper is within the scope of HESS, but the scientific contribution needs to be much better contextualized. The paper states: "*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 and evaluating its performance"*. As stated, the paper does comprehensively document the single experiment, the WRF-Hydro IWAAs. In that respect, it covers the "what" and the "how". However, as is, it doesn't provide enough context to help the reader understand if this is a substantial scientific advance – i.e., it's missing the "why". Three suggestions to help with this contextualization:

1. The paper needs to more clearly state the "why", this includes the research gap (the need) and how what you are putting forth fills that (the advance). Why does the IWAA need hydro model simulations (rather than just observations?). Why did you need a new configuration of WRF-Hydro for the IWAAs? (i.e., why didn't you just use an existing application, like the operational NWM? Why use a national-scale model over catchment-based models?). Why did the IWAA need to be forced with a new met product like a bias-adjusted CONUS404? (i.e., why not just use the AORC?). Once the need is more clearly identified (see specific comments for Introduction), you need to show how this new application (i.e., CONUS404BA + IWAAs WRF-Hydro) is better (or an advance) in terms of design and/or performance, or at least better suited to your IWAAs application than other options might have been (see next point):

Thank you for raising the concern. We added some text to the introduction and other sections which hopefully helps to address some of the questions raised here. Below is the response to each question asked.

Why does the IWAA need hydro model simulations (rather than just observations?).

The decision was driven by the need to have continuous analysis even in sparsely observed regions. The following text is being added to the introduction section to address this question.

"To enable continuous, nationwide analysis—even in regions with sparse observational data—two national-scale hydrological models were utilized in the IWAA framework."

Why did you need a new configuration of WRF-Hydro for the IWAAs? (i.e., why didn't you just use an existing application, like the operational NWM?

The initial intent was to tailor the model setup to the IWAA application however, time constraints did not allow us to do a lot of modifications from the NWM application. The main difference between the two in terms of physics option is the exclusion of reservoir routing from the IWAA configuration to provide natural flows. Also the new forcing dataset warrants a new calibration and regionalization to tune the model to the new dataset. The following text is being added to the introduction and the model description to make this point.

"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."

Why use a national-scale model over catchment-based models?).

This is being addressed by Stets et al. 2025, and we added the following to the start of the introduction to make the point clear.

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 initiatives (Stets et al. 2025).

Why did the IWAA need to be forced with a new met product like a bias-adjusted CONUS404? (i.e., why not just use the AORC?).

A goal of IWAA is to provide water availability assessment under historical, current, and future climate. Observation-based products are inconsistent in space and time since they are limited by the spatial coverage of observation networks and are subject to observations coming online and offline over time. Observation-based products are also generally limited to temperature and precipitation, while hydrologic response can be influenced by other meteorological factors such as radiation, wind, and humidity. The CONUS404 dataset provides a continuous (in space and time) distributed estimate of a wide range of meteorological conditions, allowing a more physically-based hydrologic response to climate. Also, although the AORC or other observation-based products could be suited for analyzing the historical and current periods, they do not provide a future scenario. All above are the reasons for creation of the CONUS404 dataset. We have the following text in the introduction section:

"*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.*"

We also had the following in the result section conveying the similar point.

"*Additionally, CONUS404 boasts capability to generate future climate scenarios, a feat not attainable with observation-based atmospheric forcing. Hence, while acknowledging the trade-offs, the utilization of CONUS404 offers several advantages over other available products.*"

We added the following to the introduction section, to make this point clear upfront. Thank you for the recommendation!

"***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 IWAA studies.*"**

Once the need is more clearly identified (see specific comments for Introduction), you need to show how this new application (i.e., CONUS404BA + IWAAs WRF-Hydro) is better (or an advance) in terms of design and/or performance, or at least better suited to your IWAAs application than other options might have been (see next point):

Please see the next point below for a response.

2. The results need to be contextualized with another benchmark or model baseline. Currently, the only benchmark is the default vs calibrated, calibration vs validation periods, which follow the expected relative performance. There is one mention of the streamflow performance being lower than in Cosgrove et al. (2024); is there a way to obtain this data for comparison? Or to

compare with NWMv2.1 retro performance (Towler et al. 2023)? Or a climatological benchmark like in Knoben et al 2020? If this can't be done quantitatively, then more qualitative comparison and justification for why for the performance is (or is not) acceptable for the purpose of the IWAA is needed (this discussion could go in the Conclusions).

IWAAs WRF-Hydro is one of two ensemble members used in compilation of the USGS report. Gorski et al. 2025 provide comparison of WRF-Hydro with PRMS, the other hydrological model used for the IWAA, and therefore we refrain from comparing it against a different model. Instead we focus on providing comparison with observations or widely used model estimates. The current application is close to the NWM in terms of model physics; however, the application use is very different and comparison of the IWAA against NWM is beyond the scope of the current study. Another colleague is currently preparing a manuscript dedicated to diving into the differences of the two configurations.

3. The methods need to be contextualized with another benchmark or model baseline. Throughout the manuscript, there's a lot of details of the WRF-Hydro model configuration, and often they are in comparison to other NWMs (v2.1, v3.0). I suggest that the methodological parts in the manuscript should focus more on what's different or novel about this application – and why these changes are well-suited to the IWAAs. I suggest maybe a table or two to highlight these differences against some baseline, like the operational NWM, and perhaps some of the in-depth details can go in a Supplemental.

We agree with your points raised here. We reduced the content of the calibration/regionalization section significantly and moved the text and discussion mostly to a Supplement. We edited the WRF-Hydro description slightly to reduce the text; however, we refrained from moving it to the Supplement completely to provide a good overview of what the WRF-Hydro model does for the reader before diving into the model performance verification. From the physics perspective, the only difference is IWAA does not perform lake routing. Here is the description of that in the text:

"*WRF-Hydro also includes options to represent lakes and reservoirs (i.e., waterbodies). However, 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.*"

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., and Woods, R. A.: A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments, Water Resourc. Res., 56, e2019WR025975, https://doi.org/10.1029/2019WR025975, 2020.

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.

· **Scientific quality is "good": "***Are the scientific approach and applied methods valid? Are the results discussed in an appropriate and balanced way (consideration of related work, including appropriate references)?*" The approach and methods are valid, but the results should be discussed more in terms of related work/references (see previous three points about contextualization), and what conclusions can be drawn from the results. Related to this, the Conclusions is more of a summary, and could benefit from revision and additional discussion and what conclusions can be drawn (see Scientific Significance point 2 and specific comments).

Thanks for the comments. We have added a new section before the conclusion providing more discussion of the results. We also reformatted the conclusion, reducing the text and keeping only the key points. The edited conclusion is provided below where we provided responses to comments specific to the conclusion section.

· **Presentation quality is "fair":** "*Are the scientific results and conclusions presented in a clear, concise, and well-structured way (number and quality of figures/tables, appropriate use of English language)?*" The paper is generally well-written, but the content and overall structure need to be tightened and made more cohesive. The majority of the paper is focused on streamflow, and the other water budget component results aren't introduced or well-integrated in the paper (see Specific Comments on these sections). As previously mentioned, there are a lot of dense, methodological details included; these might be better summarized in a table and details in a Supplemental (see Scientific Significance point 3). Similarly, the performance metric results could be tightened up, and I suggest where some figures could be combined or go in a Supplemental. In addition, if possible, a new figure or comparison with a benchmark or model baseline would be a good addition (see Scientific Significance point 2).

Thank you for the comments. We have addressed some of these concerns in more detail in the specific comments. To summarize, we moved the technical discussion to the supplement, moved the calibration results to the supplement, and moved a few figures to the supplement. We also added new figures and discussion points. However, we do not provide comparison results as that has been done by Gorski et al. 2025.

Specific comments:

Abstract: lines 1-5. This is missing the research gap or the "why" of this article. For instance, it is not clear why hydrological model simulations (in this case from WRF-hydro) are need to characterize water availability for the IWAAs (why not just use observations?). Making this case to readers up front is going to be critical to showing the value of your paper.

We reworded the initial part of the abstract to address your concern.

*"**The Weather Research and Forecasting model hydrological modeling extension package (WRF-Hydro) is one of the selected hydrologic models used in the IWAAs to generate a spatially and temporally continuous estimate of hydrological fluxes and storage across the conterminous United States (CONUS), even in regions with sparse observations.**"*

Line 12. "*specific emphasis on temporal accuracy issues"* is not clear.

To address the concern below we moved the streamflow results to the end and removed this text. It was referring to the low correlation coefficients in the IWAA simulation compared to NWM results. Since it is not really essential for water availability studies, the text is removed from the abstract.

Abstract – this is written almost like two abstracts, where the first paragraph is about streamflow, and the second paragraph is about snow, ET, and soil moisture. These should be integrated. Would help to introduce earlier what is done in the paper (like around line 5, say in this paper we evaluate simulations of streamflow, snow, et, soil moisture, etc). Final line of the abstract is very vague, can these results be tied to the streamflow results more definitively? What are the conclusions drawn from the study?

Thank you for the comment. We added the introduction of what is being verified and edited the text slightly to shorten it. Below is the edited text.

A systematic and periodic evaluation of water supply across the United States is critical for gaining comprehensive insights into the present state of the nation's water resources and strategically planning for the future. The U.S. Geological Survey (USGS) Integrated Water Availability Assessments (IWAAs) is a national initiative designed to characterize past, present, and future water availability in the United States. The Weather Research and Forecasting model hydrological modeling extension package (WRF-Hydro) is one of the selected hydrologic models used in the IWAAs to generate a spatially and temporally continuous estimate of hydrological fluxes and storage across the conterminous United States (CONUS), even in regions with sparse observations. The state-of-the-art CONUS404 dataset, a regional hydroclimate dataset over the CONUS, is bias-adjusted for precipitation and temperature fields and utilized to calibrate and drive the hydrologic applications in IWAAs. The WRF-Hydro model simulation provides estimates of water budget components, from which we verified the snow water equivalent (SWE), evapotranspiration (ET), surface (top 10 cm) and root zone soil moisture (SM) estimates and streamflow in this study.

Throughout the CONUS, WRF-Hydro IWAAs simulations of snow water equivalent closely align with the Snow Data Assimilation System (SNODAS) during the snow accumulation season, but show low biases during the snow ablation season. The lower SWE peak estimates, combined with a faster melt rate, results in low biases in streamflow in most snow-dominant basins. WRF–Hydro IWAAs actual evapotranspiration (ET) simulations generally exhibit close

agreement with Global Land Evaporation Amsterdam Model (GLEAM) ET estimates. Despite this overall agreement, simulated WRF-Hydro IWAAs ET is higher in parts of the central U.S. and lower in parts of the northeast, southeast, and northwest regions of the U.S., and in urban areas when compared to GLEAM. Focusing on seasonal patterns, in many regions the biases in evapotranspiration (ET) and streamflow show opposite signs, highlighting areas where enhanced parameter estimation could lead to improvements. There is also a strong agreement between WRF-Hydro IWAAs simulation and GLEAM surface soil moisture (top 10 cm) values, with the WRF-Hydro IWAAs model simulating some lower estimates particularly over the eastern U.S. Similarly, simulated WRF-Hydro IWAAs root-zone soil moisture is underestimated in the southeast U.S. while there are positive biases observed in the western U.S., relative to the GLEAM simulations. Streamflow performance is reasonable at USGS gages, particularly in the eastern and western regions. However, certain challenges arise in the central U.S., Arizona, and southern Florida, where the model exhibits poor performance. The observed shortcomings in these regions can be attributed to missing or poor presentation of physical processes in the model.

Introduction: Similar to comments on abstract, the research gap is not clear, and the reader is missing the "why" we need hydrologic simulations for the IWAAs, or what the main contribution of this work. Why do we need a new IWAA application, rather than use the operational NWM for instance? Some of the pieces of this are in the Introduction, but I suggest reworking the Introduction to make the value clear to the reader.

We improved the introduction based on the recommendations here for "Introduction" and also the "Atmospheric Forcing" section. More details are spread throughout other responses. Now the introduction section starts with a brief introduction to the IWAA program, refers the reader to the Stets et 2025 study for details, provides background information on the CONUS404 and why we performed bias adjustment, then introduces the model used in this study and discusses what is different in this study and why we need to have an IWAA configuration. We finish the introduction with a short explanation of what components we evaluate and why.

Line 27: After this sentence: "*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."* Prior to this usgs would do trend assessment on observations, right? It seems after that sentence you should help the reader understand why hydro modeling is needed for this (ie. observations are sparse/many ungauged basins). Depending on how you reorganize the paper, this might be where you introduce the unified national-scale framework of the NWM, which provides the continuous spatial/temp coverage needed for this type of assessment.

Thank you for the suggestion. With the USGS report being released, we edited the introduction section. In particular, the following points discuss why national-scale modeling is used:

"*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).*"

"*To enable continuous, nationwide analysis—even in regions with sparse observational data—two national-scale hydrological models were utilized in the IWAA framework. …*"

Line 34: Need to define acronym and describe to the reader what CONUS404 is, and why it is relevant to IWAA application.

Although this is not an acronym that could be described like others, we added the following to the text to clarify the reasoning behind calling the downscaled product as CONUS404.

"***CONUS404 provides 40+ years of data at a spatial resolution of 4-km across CONUS and hence called CONUS404.***"

Line 36. Would it help to give more background on the IWAA (what are the other modeling applications contributing?).

Thank you for the recommendation. As stated in the earlier comment, we added a little more background on the IWAA and mentioned the other model application (*Precipitation Runoff Modeling System)* used alongside WRF-Hydro; however, the reader is referred to the USGS report for more detailed information.

Line 39: "*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 and evaluating its performance."* I agree that this is what the paper focuses on, and while this is worth documenting for the IWAA community, it is harder to identify the new contribution within this for HESS' international readership.

We have expanded the manuscript to include a comprehensive discussion on model evaluation and interaction of water budget components. We hope that the revisions made throughout the manuscript adequately address your concerns.

Line 57: Increasing the forecast points seems like something that should be highlighted sooner (or at least the need for this).

Although the IWAA WRF-Hydro application benefits from fine spatial resolution similar to NWM and provides simulation on the NHDPlus streams and catchments, the wording of "increase of the forecast points" is a referring to NWM application which its main purpose is for

forecasting applications and hence we made the comment related to the increase of forecast points. To avoid confusion for readers we dropped the sentence pointing to the increase of forecast points.

Line 60: You mention it aligns with the hydrography of NWM3.0, are there other differences with the operational model? For the model, what (if anything) makes the IWAAs different than the operational model – why did you make those decisions? How do those decisions better support the purpose of the IWAAs estimates?

The only physics difference is the exclusion of lake/reservoir routing from the model compared to NWM, which we mention now in the introduction. The use of a different forcing dataset, however, motivates the need for a new model application with different parameters and responses. We could not simply take the operational NWM model, which is forced and calibrated using the AORC dataset, and apply it for the IWAA application. The following is added to the introduction to convey the information.

"*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 to represent meteorological conditions. Therefore, it is necessary to calibrate the model to the new atmospheric forcing dataset and adjust the parameters accordingly.*"

Line 61: This paragraph doesn't seem provide information relevant to the background or motivation for the study. Suggest moving to the model description unless this is relevant to the research gap or why of the paper.

As suggested, this paragraph is removed from the introduction.

Line 70: This is abrupt. Seems like it would be useful to have a preceding paragraph, especially on the need to evaluate national-scale hydrologic fluxes. Many studies look at streamflow, but do many others look at snow, soil moisture, and ET? What additional insight does evaluating those provide? Can you tie it back to the why of the paper (like the IWAAs?).

Thank you for the suggestion. We added the following to the prior paragraph to address the "why" of going beyond streamflow analysis. Addition:

"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 Description:

This is well written, but it is very dense, and it might help if you can better differentiate between what is unique about the IWAAs application versus another benchmark application (say the operational NWM) and why you made those decisions (see previous comment on Introduction). If there are quite a few differences, it might be easier to see them in a table(?). As part of this table, you could also include comparisons of how the model is Calibrated/Regionalized with a benchmark, from that section later in the paper. Then you could focus the main text on describing the key differences and why you made them for the IWAA, and put detailed descriptions in a Supplemental.

We provided a general overview of the WRF-Hydro model for readers to quickly know about the flow of the processes and some short descriptions of each. We believe providing such high-level information helps the reader to know the model and its strengths and limitations better. However, following this recommendation, we made minor modifications to reduce the text.

Figure 1. Maybe bold or highlight somehow the variables that you evaluate in this paper in this figure. You could say that at line 78 when you introduce it.

We have highlighted the variables that are evaluated in this study in Figure 1, made modification to the figure caption, and added it to the text as suggested (**in bold**).

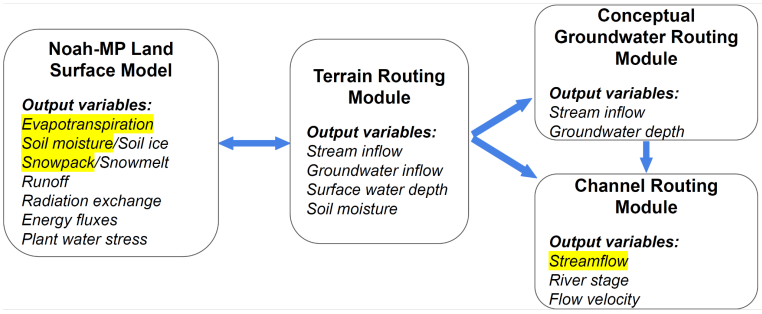


Figure 1. Existing model physics in WRF-Hydro used in IWAAs study. The arrows show the direction of information passing between the modules. ***The highlighted variables are evaluated in this study.***

Forcing:

Line 113-119. I wonder if some of this background information might be better in the Introduction, perhaps after the need for a unified continental scale hydro model is introduced, you could move to the forcings.

Thank you for the comment. This text is moved to the introduction section as suggested.

Line 127-129. Maybe some of this justification should go in the Intro as well, and it would be good if there was a reference to qualify how the CONUS404 performs.

Thank you for the comment. This text is moved to the introduction section as suggested. Rasmussen et al. 2023 (https://doi.org/10.1175/BAMS-D-21-0326.1) provides a general overview of the product as well as evaluation of different fields such as precipitation, temperature and downward solar radiation. Therefore we did not provide assessment of different fields and only provided spatial bias maps of precipitation and temperature, which we intend to bias adjust in this manuscript.

Line 130 - This is a very detailed description of the bias-correction. I wasn't clear if this bias-correction (CONUS404BA) is also publicly available – suggest explicitly mentioning that this is available as a data release reference here to clarify for the reader. Also, when I first read this, based on what was stated as the purpose of the paper in the Intro, I thought this was not necessary and should go in a Supplemental, but then after reading the entire paper I wasn't sure if the CONUS404BA was one of the more important parts. This goes back to more clearly describing the scientific advance, and setting up the reader in the Intro to know if this is one of the main contributions of the paper.

Thank you for stating the concern. CONUS404 forcing dataset is a key component of this study, and here we attempt to introduce and document the bias-adjusted version of the CONUS404 dataset. It is also a main reason why we had to go through the calibration of the model even though the model configuration is very similar to the NWMv30. We now explicitly mention in the manuscript that the dataset is publicly available in the last sentence of the Forcing section:

"The version of the CONUS404 dataset that includes bias adjustments to air temperature and precipitation is referred to as CONUS404BA (Zhang et al., 2024). This dataset ***is publicly available and covers the time period from October 1979 to October 2022****, although only the water years of 2010–2021 are used in the IWAAs configuration of WRF-Hydro in this study (Rafieeinasab et al., 2024)".*

Zhang et al. 2024 is the reference to the publicly available CONUS404BA dataset and Rafieeinasab et al. 2024 is reference to the publicly available WRF-Hydro model outputs generated by this study.

Figure 4 is not well explained in the text (either provide the explanation in the text or caption and/or maybe put in Supplemental).

We now provide a more detailed description of the figure in both the text and the caption (shown in bold). The figure caption now reads:

"Figure 4. Temperature bias by elevation for four different products over the central Rockies. ***Different colored dots in the scatterplot provide the mean annual bias for each pixel of CONUS 404 vs. each product (Daymet, PRISM, ASOS, and SNOTEL). Description of how the bias correction is applied at different elevations is provided on the right side of the plot.****"*

The text has been altered as follows:

*"****Figure 4 shows how temperature bias varies with elevation when compared against four different products: ASOS, SNOTEL, Daymet, and PRISM.*** *Locations above 2,500 m elevation were not corrected because the Daymet (and PRISM) values were influenced by biased SNOTEL stations (Oyler et al. 2015), and pixels between 2,000– 2,500 m elevation had their biases linearly corrected between no correction (2,500 m) to full correction (2,000 m), depending on their altitude."*

Line 183. Figure 5 shows the calibration basins in terms of the RFCs – it would be helpful to introduce/define RFCs here, and why you use them in your analysis. How do the RFCs fit into the IWAAs? Also, suggest editing Figure 5 caption (maybe remove: *as it will be used frequently in the results and discussions)*.

Given this figure was not too busy compared to other U.S. maps in the paper, we added the RFC boundaries here in order to introduce the full names of the RFCs as well as the abbreviated names. Later in the manuscript, we aggregated the results spatially to be able to offer summary statistics and regional conclusions. We looked into different ways of aggregating the results and chose the RFC boundaries. These regions have been used in other studies and generally follow watershed boundaries, so would allow analysis of basin-integrated variables (e.g., streamflow) and comparison of water budget partitioning.

The figure caption has been edited as suggested. Also, we have now added a section called "Verification" which outlines how we proceed with model evaluation, and in this section we add the following narrative to explain that the aggregation is at the RFC level.

"We also provide aggregated statistics across River Forecast Center (RFC) regions, with boundaries shown in Figure 5. **These regions generally follow watershed boundaries, so would allow analysis of basin-integrated variables (e.g., streamflow) and comparison of water budget partitioning."**

Table 1. Is WRF-Hydro always calibrated with these 17 model parameters? Is this the same/similar table as in Cosgrove 2024? If so, maybe it can go in the Supplemental (or just pull out parameters in a table highlighting what the differences are between IWAA WRF-Hydro and NWM). Is DDS always used? Again, it would be good to have a WRF-Hydro benchmark/comparison for context, and a table or tables could be a good way to show this.

Thank you for the comments and the questions. The latest version of NWM discussed in Cosgrove et al. 2024 is NWMv2.1. While most of the parameters are the same, some new parameters were introduced in NWMv3.0 due to change of physics options, so we could not refer directly to Cosgrove et al. 2024. The list is also not exactly the same as NWMv3.0 (RafieeiNasab et al. 2025), as there are two less parameters in IWAA configuration compared to NWMv30. Following the recommendations of both reviewers, in an effort to shorten paper and not to focus too much on calibration/regionalization, the calibration table is moved to the supplement. We have also referenced RafieeiNasab et al. 2025, which has the list of the calibrated parameters in NWMv3.0 and also provides a good overview of the previous calibration activities with WRF-Hydro.

DDS has been used in all the NWM configurations; however, many non-NWM studies have used other types of calibrations (listed in RafieeiNasab et al. 2025). As mentioned earlier, there are multiple studies that are underway which do in-depth comparisons against NWM. Here, we would like to focus on the introduction of the modeling framework and comparison against either observational datasets (such as streamflow observations from USGS) or widely used model estimates (such as SNODAS and GLEAM). Therefore, we avoid comparing the IWAA configurations against NWM.

Figure 6 & Regionalization. I wonder if it's more important to document the differences between the regionalization approach used in IWAA versus in NWMv2.1 and NWMv3.0 (like in a table), than to provide the workflow in Figure 6 (which could maybe go in Supplemental?).

We agree with the reviewer. We moved most of the regionalization content to the supplement and left only a general description of the procedure in the main text. The comparison against NWM versions is provided in the Supplement. Please see the next comment to see the general response to the reviewers document on how the new main text is being modified and reduced. The regionalization methodology is mostly the same between NWMv3.0 and IWAA; however, the NWMv3.0 regionalization methodology has not been published yet, so we are using the current study to share the details.

Section 4. Model Calibration: This is a very detailed and dense section. To repeat a previous Main comment, I suggest that the parts of this that go in the manuscript should focus more on what's different or novel about this application – and why this is well-suited to the IWAAs. I suggest maybe a table or two to highlight these differences against some baseline, like the

operational NWM, and perhaps some of the details that are already published can go in a Supplemental.

Thank you for the comments. We have provided essential information about calibration and regionalization and moved most of the content to the supplement. The reduced text is provided in the general response to the reviewers. Please refer to that.

5. Results and Discussion

5.1. Add “Streamflow” to Evaluation of Calibration Basins

Added “Streamflow” to the title of the section. This section is now moved to Supplement as suggested in the next comment.

Figure 8. Can you put the calibration and validation on the same plot but have one of the lines (say validation) as dashed? Otherwise, it is hard to see the differences between the model runs for the different periods. This is expected relative performance, so perhaps some of the metrics could go in a Supplemental.

Great suggestion. This figure is now moved to Supplemental material (Figure S3). Following the suggestion, the calibration period is shown as solid line, and the validation period is shown as dashed line. Time information is added to the figure caption following suggestions of the reviewers. Given NSE related discussion did not add new information to the results section, in an effort to reduce the content we dropped all the figures related to NSE. Figure caption is provided below.

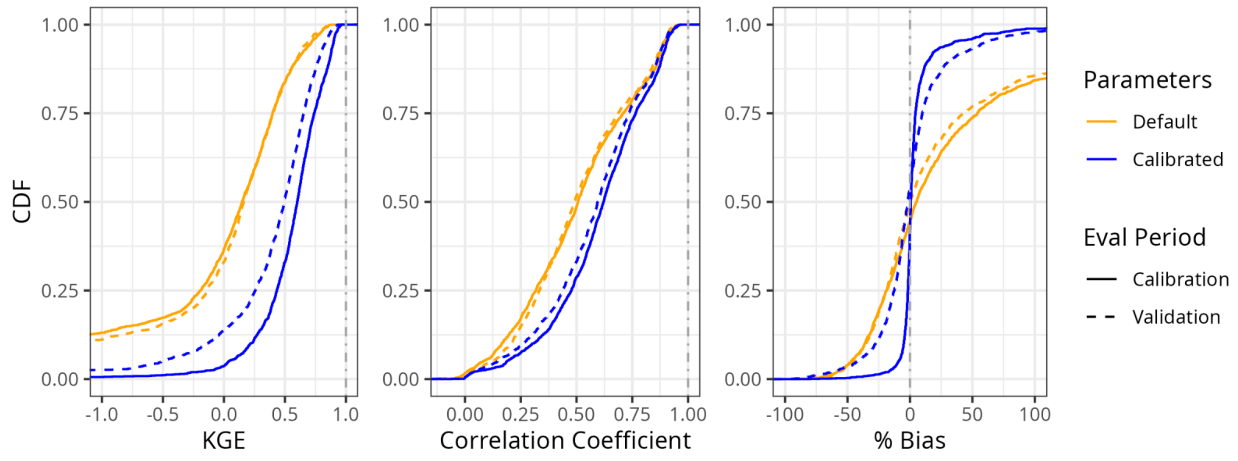


Figure S3: Cumulative density function of hourly streamflow metrics (KGE, correlation coefficient and %Bias) across the 1,522 calibration gages with the default parameters (orange) and calibrated parameters (blue). The solid lines show the CDF of error metrics over the calibration period (2013-10 to 2018-10) while the dashed lines show the CDF error metrics over

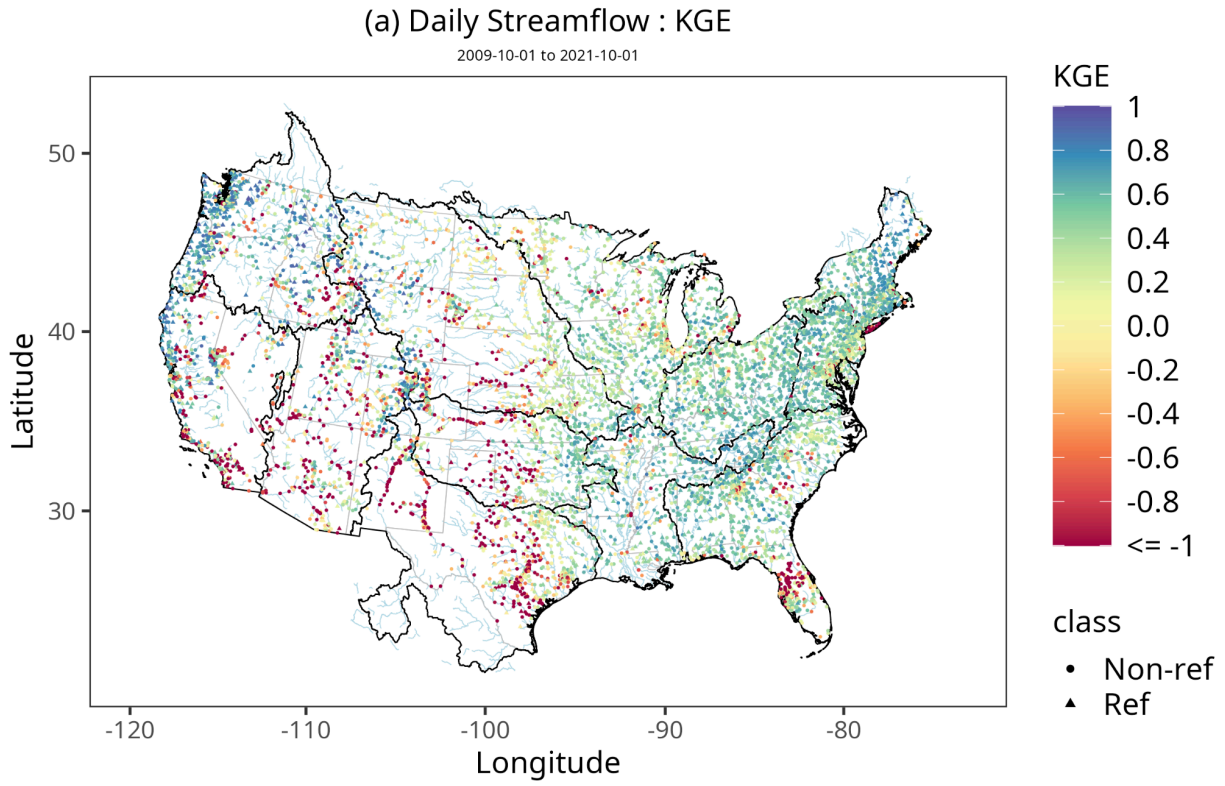
the validation period (2011-10 to 2013-10 and 2018-10 to 2021-10). The dot-dashed line in gray shows the optimal value of each error metric.

5.2. To be consistent with the 5.1, maybe make both “Evaluation” or both “Verification”

Section 5.1 is moved to Supplemental. The title is now using “Evaluation” in both places.

Figure 9. It is hard to see all the points… can you make all the points the same size to try to improve the visibility? %Bias and correlation coefficient are also quite small (since they are in the same row). Maybe put one in the supplemental so the other can be better seen like KGE?

KGE and %Bias are kept in the main text as suggested and the correlation coefficient spatial map is moved to the Supplemental material. Below are the new remade figures for KGE and %Bias with same size for both shapes to improve visibility as suggested.



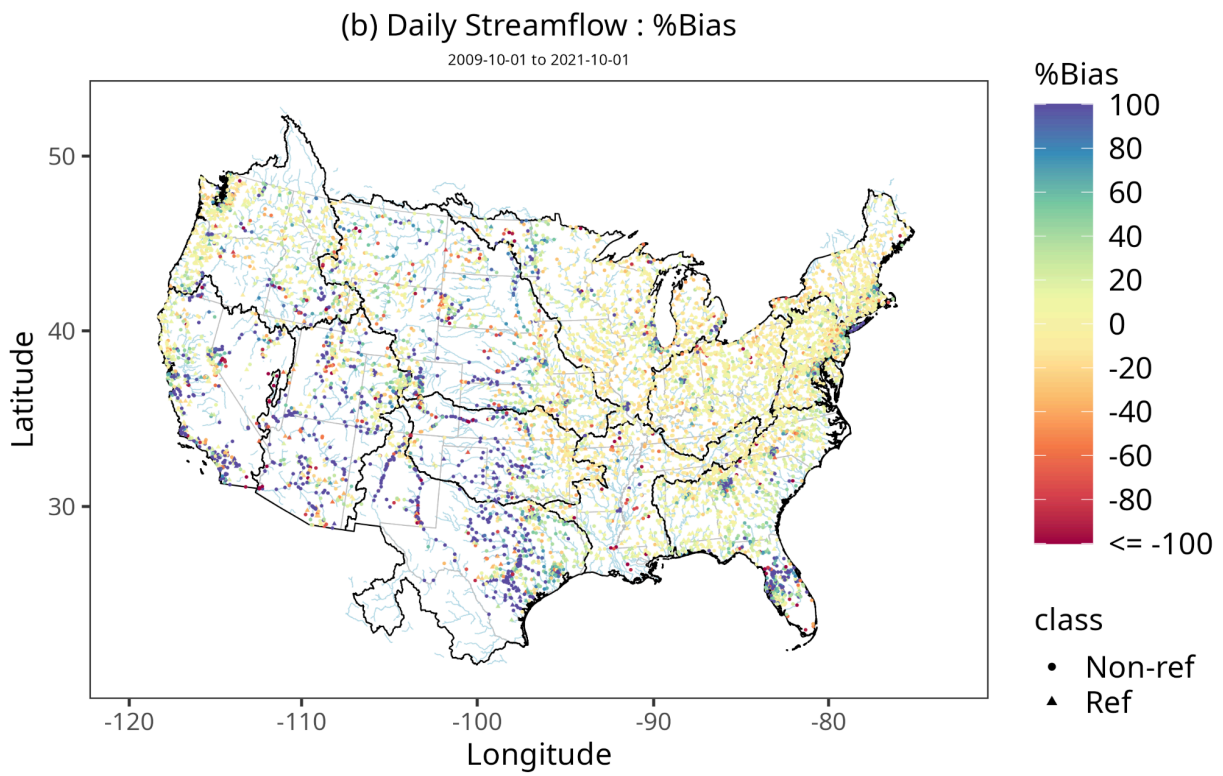


Figure 6. Spatial maps of (a) KGE and (b) percent bias (%Bias) of daily streamflow over water year (WY) 2010 through WY2021 at USGS reference (Ref) and non-reference gages (Non-ref).

Line 325 – this paragraph is abrupt/out of place as is, but I think you can tie it in if you put this in reference to Figure 9, which has Ref and Non-Ref gages. Do you see better/worse performance depending on if it is Ref or Non-Ref? Hard to tell now since there are so many points in Figure 9 (see previous point).

Thanks for pointing out the discontinuity. The first part of this section is now moved to the Supplement, and we moved this text prior to the mention of Figure 9. Now the text is better connected and more coherent. The boxplots also break down the results by Ref and Non-Ref to show the clear different performance level, and this difference is discussed in the text. Below is the reorganized text.

"*It is important to acknowledge that the model application does not account for human interventions to unimpeded channel flow. Consequently, suboptimal model performance is anticipated in regions with extensive stream regulation, such as large rivers where flows are heavily managed for water supply or hydropower. Despite this limitation we opted to include regulated gages for comprehensive reporting and analysis. Figure 6 present spatial maps of daily streamflow KGE and percent bias derived from the conclusive WRF-Hydro IWAAs model run across all USGS gages nationwide including GAGES-II reference basins (with minimal human impacts) and non-reference basins (with more significant human impacts). The spatial map of*

daily streamflow correlation coefficient is provided in the supplement (Figure S4). Figure 7 encapsulates comparable information through the use of boxplots, delineating key performance metrics for each RFC. The metrics, including KGE, percent bias, and correlation coefficient, are categorized into GAGES-II reference basins and non-reference basins. Notably, across all metrics and for every RFC, the model consistently demonstrates superior performance in reference basins, aligning with anticipated outcomes. In the majority of the non-reference basins, the model is likely missing a critical process, such as water diversion for irrigation or hydropower regulation. This deficiency highlights the challenges associated with accurately representing complex hydrological processes in non-reference basins, impacting overall model performance in these areas."

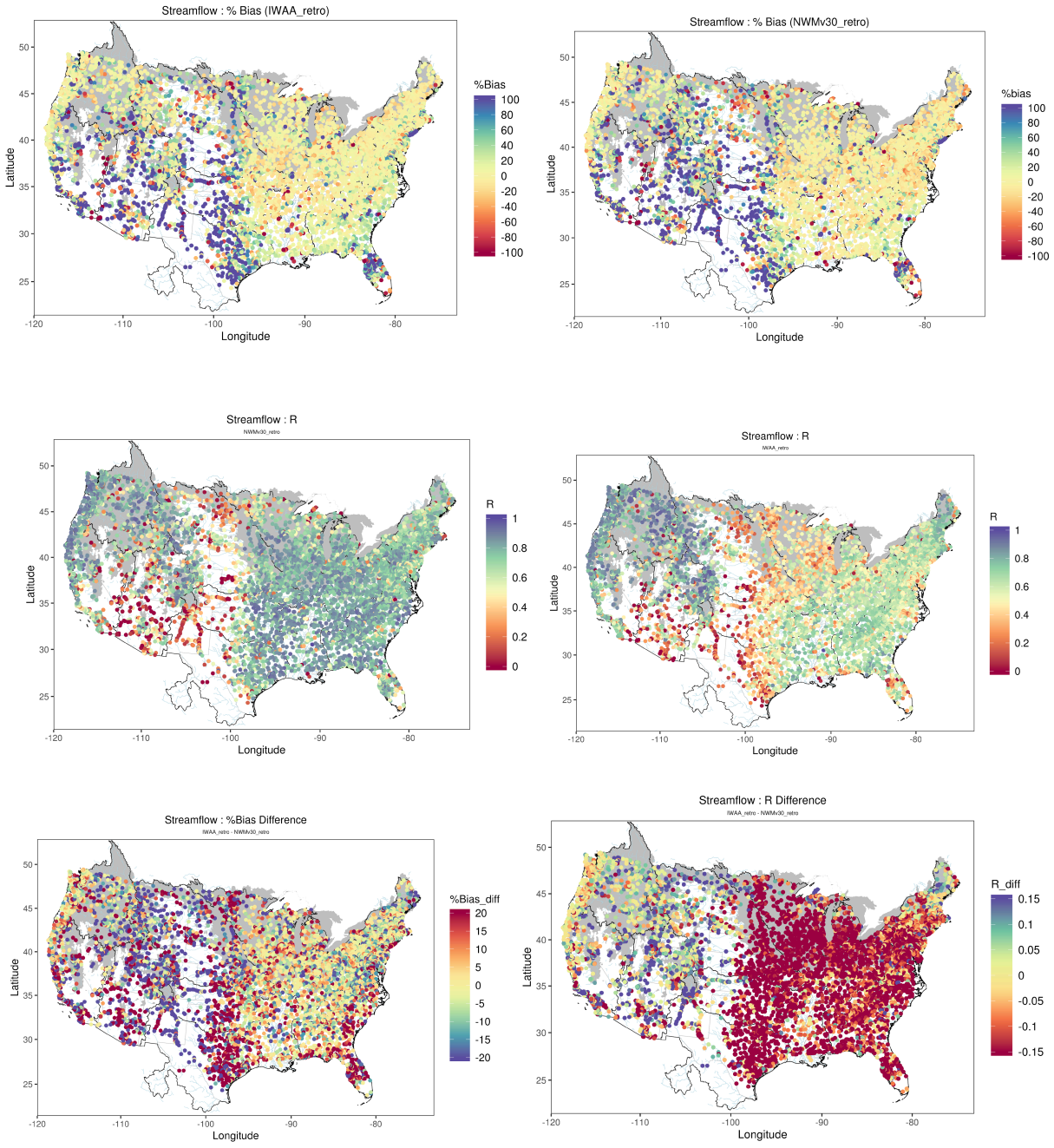
Line 344. It would be useful to contextualize the performance difference if you can quantify the comparison with Cosgrove 2024, which would be a great baseline. Can you obtain this data? See General Comments for other suggestions on benchmarking. Either way, this is a useful discussion point even if it stays as qualitative, but seems buried here (this might fit better in a discussion and conclusions section).

Thank you for the comments, as mentioned in the general comments we think it is best if we keep this manuscript focused on the introduction of the model, outputs, and verification against observation or widely used model estimates. Another colleague of ours is specifically working toward extensive verification of NWM and IWAA against each other, and therefore I avoid bringing it directly to the manuscript. Gorski et al. 2025 also provided comparison of the WRF-Hydro simulations against PRMS and therefore we avoid the comparative study.

Below are a few figures that show the difference between the streamflow results of the two products. As shown, the percent bias is very similar between the two products and CONUS404 forced simulations have superior performance in the rocky mountain range in many locations. The obvious difference between the two products is in the correlation coefficient. Correlation coefficients drop in the IWAA model compared to the NWM. This had been acknowledged in the text and we added another sentence to comment on the reasoning (in bold):

The relatively low correlation coefficients of the WRF-Hydro IWAAs simulation compared to the NWM (Cosgrove et al., 2024) is likely due to the use of CONUS404 which is a model-based dataset. NWM retrospective analysis uses Analysis of Record for Calibration (AORC) dataset (Fall et al., 2023), an observation-based precipitation dataset. ***CONUS404 provides a physically coherent depiction of mesoscale events and this could come at the expense of local accuracy in certain regions when comparing CONUS404 against observational based dataset such as AORC.*** *Employing CONUS404, however, offers additional advantages. Notably, modeled precipitation often outperforms observation-based products in mountainous regions (Lundquist et al., 2019), thereby enhancing hydrological model accuracy in these areas. Additionally, CONUS404 boasts capability to generate future climate scenarios, a feat not attainable with*

observation-based atmospheric forcing. Hence, while acknowledging the trade-offs, the utilization of CONUS404 offers several advantages over other available products.



Line 339: I think you mean “Figure 10” here, not “9”. Also, suggest only showing 1 or 2 of these metrics unless the story is different or relevant to IWAAs, and putting the others in the Supplemental.

The figure numbering is correct, it was pointing to the spatial maps of correlation coefficients in Figure 9. As suggested, we moved the correlation coefficient spatial map to the supplement, and therefore we revised the text accordingly.

Figure 11. Caption is wrong. Interesting figure, but need to indicate why it was included (what is it telling us, how can it help us to interpret the other results?).

Since we added a few figures to explain the connection between the errors in SWE, SM, ET and streamflow, we opted to drop this figure to keep the length of the manuscript at a reasonable level.

5.3. Snow Analysis

As stated previously, this is a bit abrupt, since there is no background/motivation in the Introduction for why this is useful for IWAAs, or diagnosing streamflow errors, etc. Or will IWAA be doing trend analysis on these variables too? It almost feels like we are starting a new paper at this point in the manuscript.

10 year simulations are not enough to perform such studies. There is a longer term model simulation completed for current (water years 1980 to 2022) which enables such study, we are also in the process of creating the future scenario that allows looking into the trends under current and future climate.

Line 401. Here you mention potential reasons for the snow biases, but if the paper is focused on streamflow (which the majority of it was), can you say anything about how these biases relate to the streamflow results we've seen?

Please refer to the general response to the reviewers, the new section called "Discussion of Water Budget Components" attempts to answer this comment.

Is there a baseline to compare the snow results to, even if it is qualitative? Seems like you could reference Cosgrove et al. 2024 here, but maybe others too.

Cosgrove et al. 2024 provide an overview of SWE analysis relative to SNODAS, which we have now referenced in the Snow Analysis section (6.2) . We reference an additional study (Garousi-Nejad and Tarboton, 2022), which evaluated NWM snow against SNOTEL stations. We did not add a SNOTEL analysis to this paper, however, given the scale-mismatch between in situ SNOTEL stations and RFC-averaged snow states. The text in bold has been added to the the text:

"*Snow model performance is also commonly evaluated against SNOTEL (Snow Telemetry) in situ measurements (e.g. Garousi-Nejad and Tarboton 2022 and Cosgrove et al., 2024). We*

chose not to evaluate IWAAs SWE against SNOTEL, given the substantial spatial scale mismatch between SNOTEL SWE (<50 m^2) and RFC-aggregated IWAAs SWE (1000s of km^2).

Broadly, relative to SNODAS, WRF-Hydro IWAAs SWE exhibits low biases which develop over the course of the snow accumulation season (~December-February), peak near the time of peak SWE (~February-April, depending on the region), and persist through the snow ablation season (~March-May) (Figures 8 and 9). ***This pattern is similar to that observed in the NWM relative to SNODAS, which is expected given the similarity in snow model configuration (Cosgrove et al. 2024)****."*

5.4 ET Analysis

Again, it would be great if this could be better integrated into the paper. As with snow: Are there other efforts/studies that have evaluated ET using WRF-Hydro (to compare baseline or benchmark)? Or how do these analyses help to explain the streamflow results we've seen?

Please refer to the general response to the reviewers, the new section called "Discussion of Water Budget Components" attempts to answer this comment. The relationship between ET and streamflow are discussed and addressed in the new section. Colleague of ours has an ET verification manuscript under prep, however, the work cannot be cited at this time.

5.5. Soil Moisture Analysis

Similar to snow and ET: Are there other efforts/studies that have evaluated soil moisture using WRF-Hydro (to compare baseline or benchmark)? Or how do these analyses help to explain the streamflow results we've seen?

Please refer to the general response to the reviewers, the new section called "Discussion of Water Budget Components" attempts to answer this comment. There is a high correlation between streamflow and root zone soil moisture almost in all locations. We have made some connections between streamflow biases against the soil moisture biases in the newly added text.

Conclusions (suggest removing potential model enhancements from section name)

Thank you for the suggestion, it is been renamed to "Conclusions"

Line 492. No need to redefine, just say WRF-Hydro IWAA.

Corrected as suggested!

The "Conclusions…" section is mostly a summary and needs to be revisited once the paper is revised. Suggest adding more discussion and conclusions based on results (see General

Comments on this). A few thoughts (you don't need to answer all of these, they are just to get you thinking):

· How will places where there is poor model performance be handled in the IWAA? Will trend analyses still be done?

· Can you discuss how do these results compare to other related work and studies? If this can't be done quantitatively, then more qualitative comparison and justification for why for the performance is acceptable for the purpose of the IWAA is needed.

· What parts of this study (methods or results) might you recommend that others in the hydro community adopt or compare with? What are the "lessons learned" for using CONUS404 and a national-scale hydro model for the IWAA? What caveats would you provide in terms of using these results for water availability and trend assessment?

· Can you use some of the other water budget results to better understand the streamflow results? This could help to flesh out more of the potential model enhancements.

Thank you for the comment, to address your comment, here is the modified text to the conclusion.

In this paper, we describe the WRF-Hydro modeling effort under the USGS IWAAs, a nationwide water supply study across the CONUS. The atmospheric forcing used is the publicly available CONUS404 dataset, a mesoscale hydroclimate dataset available over the CONUS for the most recent 43 years. The CONUS404 precipitation and temperature are bias-adjusted relative to the Daymet data. WRF-Hydro IWAAs calibration is performed across 1,522 basins in the US resulting in substantial improvements in streamflow model simulations in the majority of basins. The model parameters are then extrapolated from high quality calibration basins to all other uncalibrated locations based on similarity between the calibration basins and the regionalization units (HUC10 scale). Then, we conduct model simulations spanning the period from October 2009 to October 2021 and evaluate model performance for snow water equivalent, soil moisture, evapotranspiration, and streamflow to paint a more complete picture of the model behavior.

Snow performance is evaluated using SNODAS SWE. We reiterate that, as a model-based product, SNODAS SWE is considered a benchmark rather than ground truth. Evaluating IWAAs snow state variables against in situ and remotely sensed snow observations could support the full characterization of IWAAs snow performance. Results show a reasonable agreement between SNODAS and WRF-Hydro IWAAs SWE across CONUS during the snow accumulation season; however, a lower peak SWE and a broad low bias develops during the ablation season. The timing of peak SWE in WRF-Hydro IWAAs coincides with that of SNODAS in most RFCs. However, WRF-Hydro IWAAs SWE peaks early in certain RFCs, notably the California-Nevada, Colorado Basin, and Northeast regions. This SWE behavior relative to SNODAS is similar to

that observed in the NWM (Cosgrove et al., 2024). The low snow biases along with the early melt of snow results in high streamflow biases ealy in snow season followed by low biases in spring and summer season particularly in the northeast US. Calibrating snow processes, either stepwise or in conjunction with streamflow (Parajka and Blöschl, 2008; Duethmann et al., 2014), has proven effective in improving snowpack simulation accuracy while preserving reasonable streamflow performance. This approach is strongly recommended for future IWAAs, given the focus on water availability and the large number of tunable snow parameters in the NoahMP model.

WRF-Hydro IWAAs ET and soil moisture simulations are evaluated against the Global Land Evaporation Amsterdam Model (GLEAM) dataset. As with SNODAS, the model-based GLEAM serves as a benchmark rather than ground truth. ET comparison reveals a reasonable agreement between the two models when comparing the cumulative distribution functions across CONUS, except over urban areas where the WRF-Hydro IWAAs implementation underestimates ET severely. In addition, WRF-Hydro IWAAs has slightly higher estimates of ET over the central US and lower ET estimates over the Northeast (NERFC), Southeast (SERFC), and Pacific Northwest regions (NWRFC), as opposed to GLEAM. Monthly ET biases show opposite signs to the streamflow bias in many regions across CONUS suggesting the need for a better refinement of the water budget components; this could potentially be improved with a more granular calibration strategy instead of defining a univariate objective function over the full calibration period.

Surface and root-zone soil moisture analyses suggest a strong agreement between WRF-Hydro IWAAs simulations and GLEAM. Generally, over eastern RFCs, IWAAs SM estimates are slightly lower than GLEAM, and over western RFCs IWAAs SM estimates are slightly higher than GLEAM. Overall, the WRF-Hydro IWAAs streamflow performance is superior at Northwest (NWRFC) and California-Nevada (CNRFC) RFCs and has a reasonable performance in the Northeast (NERFC), Middle Atlantic (MARFC), Ohio (OHRFC), Southeast (SERFC) and Lower Mississippi (LMRFC) and Colorado Basin (CBRFC) RFCs. As discussed, the snow, ET, and soil moisture biases results in a sub-optimal streamflow performance in these regions could be further improved with a more refined water budget partitioning through a different calibration strategy which prioritizes all water budget components, rather than focusing solely on streamflow. Improved physics representation is another area where the model performance could be improved, for example, including channel infiltration loss scheme (Lahmers et al. 2019) could improve the model performance in semiarid regions.

The model has poor performance in the Missouri Basin (MBRFC), North Central (NCRFC), West Gulf (WGRFC) and Arkansas-Red Basin (ABRFC) RFCs where the median KGE of daily streamflow values are below 0.5. These suboptimal model behaviors could be rooted in deficiencies in model process representation such as poor presentation of agricultural activities, missing water diversions and active reservoir management, as well as inadequate representation

of groundwater or atmospheric forcing errors. Previous studies such as Valayamkunnath et al. 2022 has shown value including the subsurface tile drainage scheme in regions with heavy agricultural presence and is recommended for future IWAAs simulations. This capability was not in the current study, due to the tight project schedules. Finally, a better presentation of the groundwater processes is highly recommended and an active area of research. Due to limitations of the current scheme, the groundwater model simulations were not used in the IWAA final report (Gorski et al. 25), and the analysis was based on the well dataset.

NCAR has extended model simulations from the existing 12-year span to 43 years, leveraging the comprehensive CONUS404 dataset for long-term analysis. This expansion facilitates various studies, including trend analysis that is under the second phase of IWAAs. Additionally, a CONUS404 future scenario has been completed, which can be used as atmospheric forcing for the WRF-Hydro model to support investigation of climate-change effects on water budget components.