

Intercomparison of the U.S. National Water Model with OpenET over the Bear River Basin, U.S.

Ayman Nassar^{1,2}, David Tarboton¹, Martha Anderson³, Yun Yang⁴, Joshua B. Fisher⁵, Adam J. Purdy⁶, Furqan Baig⁷, Cenlin He⁸, David Gochis⁸, Forrest Melton⁹, John Volk¹⁰

¹Utah Water Research Laboratory, Department of Civil and Environmental Engineering, Utah State University, 8200 Old Main Hill, Logan, Utah 84322, USA

²Department of Civil and Environmental Engineering, University of Utah, Salt Lake City, UT, 84112, USA

³Hydrology and Remote Sensing Laboratory, United States Department of Agriculture, Agricultural Research Service, 10300 Baltimore Avenue, Building 003, BARC-West, Beltsville, Maryland, USA

⁴Department of Forestry, Mississippi State University, Mississippi 39762, USA

⁵Joint Institute for Regional Earth System Science and Engineering, University of California Los Angeles, California, USA

⁶Department of Applied Environmental Sciences, California State University Monterey Bay, Seaside, California, USA

⁷Department of Geography & Geographic Information Science, University of Illinois at Urbana-Champaign, Illinois, USA

⁸National Center for Atmospheric Research, Boulder, Colorado, USA

⁹NASA Ames Research Center Moffett Field, California, USA

¹⁰Hydrologic Sciences, Desert Research Institute, Reno, Nevada, USA

This is the accepted version of the article published in final form at:

Nassar, A., D. Tarboton, M. Anderson, Y. Yang, J. B. Fisher, A. J. Purdy, F. Baig, C. He, D. Gochis, F. Melton and J. Volk, (2025), "Intercomparison of the U.S. National water model with OpenET over the Bear River Basin, U.S.," *Journal of Hydrology*, 656: 132826, <https://doi.org/10.1016/j.jhydrol.2025.132826>.

This version is available under the terms of the Creative Commons CC BY-NC-ND 4.0 Attribution-NonCommercial-NoDerivatives 4.0 International License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

Abstract:

This study compared evapotranspiration (ET) data from the diagnostic, satellite-driven OpenET modeling platform with ET from the prognostic U.S. National Water Model (NWM), in the Bear River Basin, U.S. ET estimates from each national-scale modeling system were compared, and evaluated against water balance ET, derived from gridded precipitation and streamflow measurements. This analysis provides an example of how prognostic-diagnostic modeling systems can be used synergistically, at basin scale, to evaluate the spatial and temporal biases and errors in both systems. Monthly ET simulations from the NWM version 2.1 retrospective analysis over the Bear River Basin were compared with OpenET data from 2017 to 2020 at monthly and seasonal timescales, aggregated to match the 1-km NWM grid. OpenET provides estimates of ET calculated using six different diagnostic remote sensing models, as well as an ensemble average estimate. Results suggest agreement between the NWM and OpenET assessments at the 1-km scale, but with notable discrepancies for some land cover types, such as agriculture and riparian areas. The NWM showed less spatial variability and tended to predict lower ET fluxes compared to OpenET, particularly from June to August. In comparison with water balance estimates of ET in four natural sub-watersheds within the Bear River Basin, OpenET model estimates were generally biased high in two watersheds dominated by evergreen forest. Results from this study provide useful information for both NWM and OpenET developers and researchers, demonstrating the power of comparing prognostic and diagnostic modeling systems. This study serves as a prototype for broader assessment of both NWM and OpenET via intercomparison in other regions, as well as an approach for quantifying uncertainty in both prognostic and diagnostic models where observational data are limited.

Highlights:

- Compared with OpenET, the U.S. National Water Model tends to underpredict evapotranspiration fluxes in all seasons.
- OpenET overpredicts evapotranspiration compared to water balance estimates from streamflow and precipitation in two forested sub-watersheds.
- Discrepancies between NWM evapotranspiration and OpenET were observed in irrigated lands, riparian areas, and one mis-calibrated watershed.

Keywords: Evapotranspiration, U.S. National Water Model, WRF-Hydro, water balance, Noah-MP, OpenET, remote sensing

1. Introduction

Various hydrologic forecasting services are currently employed across the U.S., with ongoing efforts to enhance their accuracy. The U.S. National Water Model (NWM) is one of these services, which implements the community Weather Research and Forecasting Model Hydrological modeling system (WRF-Hydro) used by the National Weather Service (NWS) of the National Oceanic and Atmospheric Administration (NOAA) for operational hydrologic forecasting (Gochis, et al 2020; <https://water.noaa.gov/about/nwm>). The system provides hourly streamflow forecasts for approximately 2.7 million river reaches across the U.S. and generates spatially continuous estimates of key hydrologic variables, such as evapotranspiration (ET), soil moisture, infiltration variables, snowpack characteristics, and shallow groundwater depth. This is part of increasing effort to operationally model hydrology at national and continental scales, providing timely, place-specific warnings aimed at saving lives and property. Operational model outputs are freely accessible on data servers (<https://registry.opendata.aws/nwm-archive/>). Multiple studies have evaluated the NWM in a wide range of research and applications. These include streamflow (Seo et al. 2021; Hansen et al. 2019), water management operations (Viterbo et al. 2020) and

snowpack simulation (Garousi-Nejad and Tarboton 2022). In a recent study conducted by Abdelkader et al. (2023) to assess the NWM's streamflow retrospective version 2.1 dataset for the entire CONUS, favorable agreement was found between the NWM and observed streamflow in catchments with natural flow. However, in examining snow water equivalent (SWE), Garousi-Nejad and Tarboton (2022) found that NWM version 2.0 retrospective (NWM-R2.0) analysis data tend to underestimate SWE as measured by the SNOwpack TELemetry Network (SNOTEL) early in the season. Later in the season, this underestimation bias further increases due to errors in input data, particularly precipitation and air temperature. While the NWM has been investigated in different applications, limited studies have been conducted to evaluate its performance in estimating ET.

Comparisons with spatially distributed ET data from remote sensing methods can be an effective means to evaluate NWM produced ET estimates. Many studies have been conducted to assess the estimates of ET from interpretive or diagnostic remote sensing models, such as those that comprise OpenET, as well as predictive, or prognostic models, such as the NWM. Diagnostic models typically combine measurements and energy balance principles to interpret what a quantity (in this case ET) is in a given situation, while prognostic models use both energy and water balance equations, and input or forcing variables to predict the evolution of the quantities involved (here temperature, soil moisture and evapotranspiration). For example, Hain et al. (2015) found that ET from the Noah Land Surface Model (LSM) (Chen and Dudhia 2001; Chen et al. 1996; Ek et al. 2003) had positive and negative biases across the contiguous United States (CONUS) when compared with the Atmosphere Land Exchange Inverse (ALEXI) remote sensing model (Anderson et al. 1997; Mecikalski et al. 1999) due to neglect of soil water sources and consideration of the impact of soil water sinks. Another study by Yilmaz et al. (2014) compared three different

approaches, including ALEXI, Noah LSM, and the Moderate Resolution Imaging Spectroradiometer (MODIS) (Mu et al. 2007) to compute ET fluxes. Their results showed that ALEXI performed better in areas where ET was not directly linked to local rainfall, such as irrigated lands or regions affected by shallow groundwater. Furthermore, Lin et al. (2018) conducted an evaluation of WRF-Hydro simulated ET compared to MODIS and FLUXNET ET data (Baldocchi et al. 2001; Pastorello et al. 2017) and found that ET predictions were more accurate in wet years compared to dry years due to bias in the baseflow. Abolafia-Rosenzweig et al. (2023) found that WRF (Skamarock et al. 2008) coupled with Noah LSM with multiparameterization options (Noah-MP) (Niu et al. 2011) simulations generally overestimated the ET compared with MODIS over the western U.S. from 2001-2020. The comparison between prognostic and diagnostic model estimates of ET has proven beneficial, identifying missing physical processes and persistent biases in the prognostic system. However, many of these studies have been limited in scope by using individual diagnostic models or data sources for comparison.

In this research we compare the NWM ET (which is built on the Noah-MP LSM) with OpenET, a satellite-driven diagnostic ET modeling and data access framework that provides high resolution ET data from multiple approaches, primarily using Landsat remotely sensed inputs. OpenET computes ET using six physically-based ET models at 30-m spatial resolution and at daily, monthly and annual time steps (Melton et al. 2022). OpenET also provides an ensemble value that averages non-outlier estimates as described below. There are multiple benefits of using diagnostic OpenET data in this analysis of NWM. First, the ensemble ET value has been demonstrated to have generally higher accuracy than any individual model in the ensemble, although this is not always the case over all locations or land cover types (Volk et al., 2024a; Melton et al. 2022). The inter-model agreement across the satellite-driven model ensemble provides additional insights into

the reliability of the diagnostic estimate. Finally, the relatively high resolution of the OpenET data enables us to investigate spatial variability in moisture fluxes at the sub-NWM-pixel scale, enabling improved interpretation with respect to physical features on the ground. While OpenET benefits from an ensemble approach, the NWM does not currently include ensembles. This may change in the future, as the development of the Next Generation Water Resources Modeling Framework (NextGEN) is expected to facilitate the incorporation of multiple models (<https://noaa-owp.github.io/ngen/>). However, for this study, we are limited to publicly available retrospective results from the NWM team.

The goal of this research was to gain a deeper understanding of how the NWM model ET behaves across different land surfaces, and to identify opportunities for improvement. In turn, the comparison provides useful evaluation of the OpenET models (many based on the energy balance approach) provided by the water balance constraints inherent in NWM, and by water balance estimates constrained by measured streamflow. While focus here is on the Bear River Basin, this study serves as an example of prognostic-diagnostic comparison can be used to model performance in other significant regions and for addressing uncertainties in alternative modeling systems, especially in situations where observational data are scarce.

Section 2 of this paper provides a description of the study domain. Following that, Section 3 presents the models and data utilized. The results and discussion in Section 4 and 5, respectively focus on the temporal and spatial comparison between NWM ET and various OpenET approaches. Additionally, we assess geographic variables associated with model differences and evaluate differences using the water balance approach. The last section presents conclusions derived from this study.

2. Study Domain

The model comparison was conducted over the Bear River Basin (BRB) between 2017 and 2020. BRB is located on the border of three U.S. states, Utah, Idaho and Wyoming, with an area of 19,425 km² (Figure 1). The basin is characterized by a complex network of streams and rivers that flow through a variety of landscapes, including mountains, plateaus, and valleys. The Bear River is the largest river in the watershed, originating in Utah then flowing 500 miles through parts of Idaho, Wyoming, and back into Utah where it enters the Great Salt Lake (GSL). BRB was chosen for this study because of interest in better understanding the water balance and water use in the GSL basin as part of efforts to inform management to reduce declines in the level of the GSL. It is a watershed that is representative of other watersheds that drain to the GSL in terms of its spatial variability in elevation and landscape composition, with both agricultural and natural land covers, but is more manageable in scale, in comparison to the entire GSL (Wurtsbaugh & Sima, 2022; Utah Division of Water Resources, 2004). Beyond the GSL basin, the BRB is also representative of many other Western U.S. mountain watersheds, with snowmelt-driven streamflow being used for irrigated agriculture. Shrubland is the most prominent land cover type, covering approximately 50% of the total area, followed by evergreen needleleaf (~14%), irrigated cropland and pasture (~11%), dryland cropland and pasture (~8%) and deciduous broadleaf forest (~8%). Other land cover types account for less than 10% of the total area of the watershed. Evergreen needleleaf dominates the higher elevation, while grasses and irrigated crops/pasture dominate the lower elevations. Urban areas, mainly located in valleys, occupy less than 1% of the watershed.

The climate of the basin is dry and cold with elevation varying between 1280 m and 3870 m. Precipitation falls mainly as snow during the winter months. The average annual precipitation varies spatially, ranging from as low as 250 mm in the lower valleys to approximately 1650 mm

in the high elevations (Utah Division of Water Resources, 2004). On average, the basin receives about 940 mm of precipitation annually. The average amount of water the basin loses to the atmosphere due to ET is approximately 850 mm annually (Utah Division of Water Resources, 2004). During the study period, the basin experienced varying levels of drought severity according to a time-series of drought index values obtained from the U.S. Drought Monitor (USDM) (Figure 2). According to the USDM, the BRB experienced abnormally dry (D0), moderate drought (D1), and severe drought (D2) conditions during the study period (2017 – 2020).

As shown in Figure 1c, we identified four sub-watersheds within BRB dominated by natural vegetation to evaluate the ET obtained from NWM and OpenET with the water balance ET (ET_{wb}) estimated from precipitation and streamflow at the water year time scale where storage changes are small ($ET_{wb}=P-Q$). These natural sub-watersheds have been selected based on several attributes, including: (1) land use/land cover primarily consist of forests, grasslands, shrubs, or other forms of natural vegetation; (2) the basin is not impacted by significant irrigation water withdrawals, and (3) the availability of USGS streamflow data that covers the study period (2017-2020).

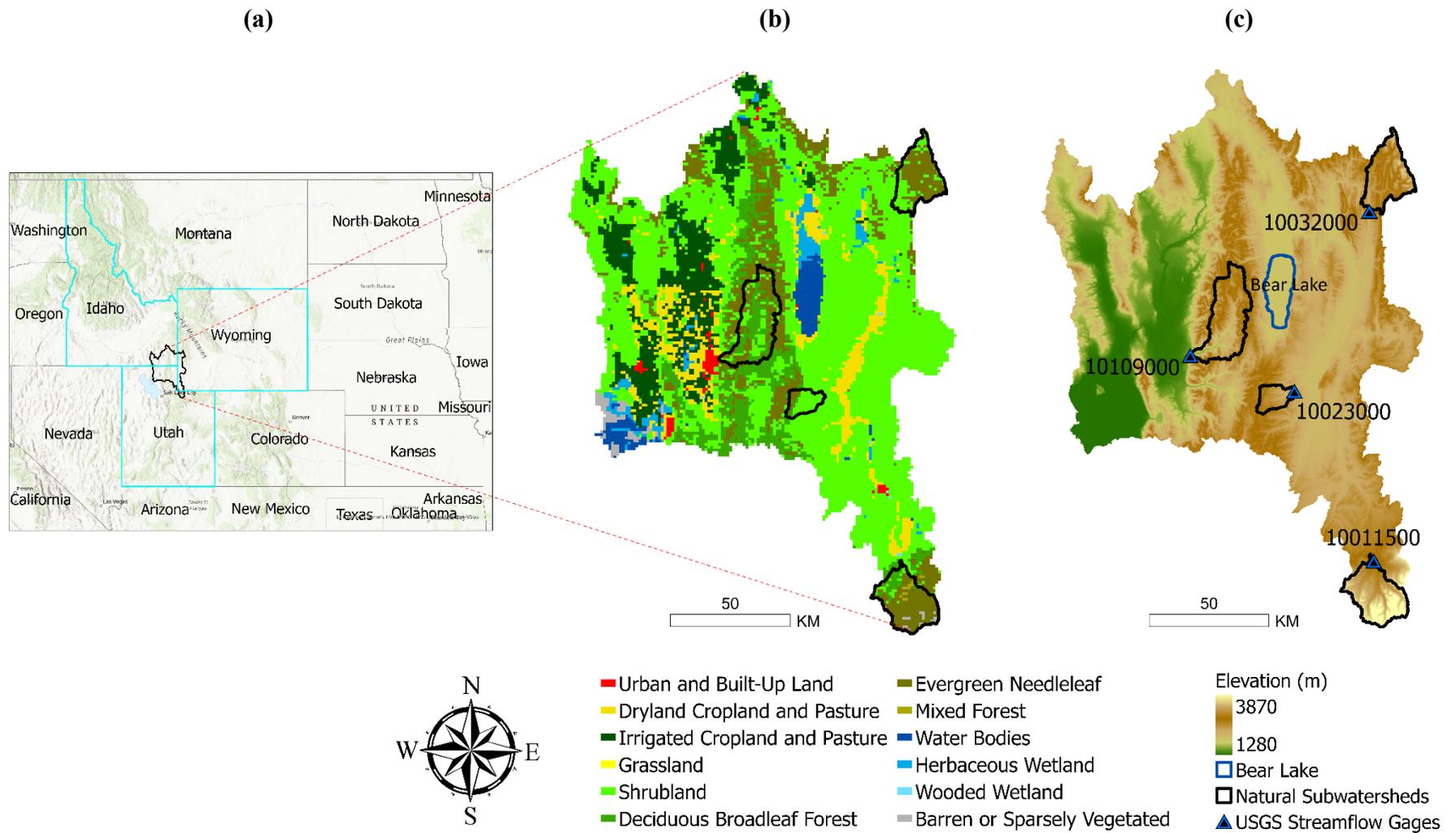


Figure 1. (a) Study area location within the U.S. at the junction of Utah, Wyoming, and Idaho. (b) Land cover from the NWM domain dataset and (c) Elevation. Also noted on b and c are sub-watersheds within the basin (black outline) and USGS streamflow gages (blue triangles) used in the water balance assessment.

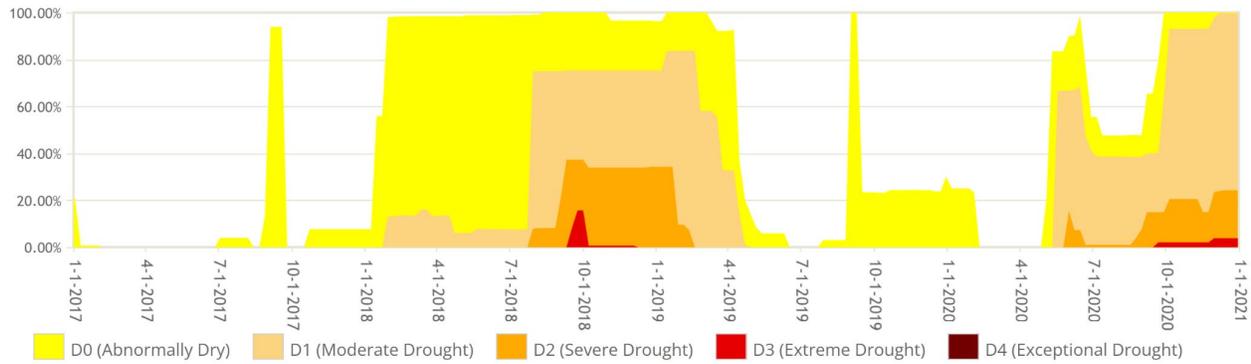


Figure 2. Time-series of U.S. Drought Monitor (USDM) drought class areal coverage over the Bear River Basin from 2017 to 2020.

Table 1. Natural sub-watersheds within the BRB used to calculate the water balance ET for comparison with NWM ET and OpenET estimates.

Sub-watershed Name	USGS Gage ID.	Area (km²)	Elevation Range (m)	Mean Annual Precipitation* (mm/year)
Smith Fork Near Border, WY	10032000	424	2046 - 3268	785
Logan River Above State Dam, Near Logan, UT	10109000	555	1425 - 3040	919
Bear River Near UT-WY State Line	10011500	455	2393 - 3870	833
Big Creek Near Randolph, UT	10023000	131	1961 - 2709	617

*GridMET 30-year mean (1990-2020).

3. Model Descriptions and Experimental Design

This study relies on data obtained from two different platforms that calculate ET; namely, NWM V2.1 and OpenET. The methods of ET calculation and the primary input datasets used in these platforms are different, yet complementary. The NWM V2.1 uses the prognostic Noah-MP LSM forced by meteorological data including precipitation rates. All processes represented by the model are explicitly modeled and contained within the equation set, and need to be identified accurately, both spatially and temporally across the landscape. This can be a challenge in some cases, requiring a priori knowledge of management practices (e.g., irrigation, tile drainage) and accurate representation of sub-surface water storage.

The OpenET platform provides ET obtained from six fully automated satellite-based models, where evaporative fluxes are calculated using remote sensing inputs of land-surface temperature and vegetation cover, as well as gridded meteorological data and land surface datasets (Melton et al., 2022). OpenET calculates a single ensemble ET value for each pixel and timestep as the mean of all models after flagging and removing up to two outliers from the ensemble using the median absolute deviation approach (Volk et al., 2024a; Melton et al., 2022). Individual OpenET modeling approaches are described in Section 3.2. An advantage of this approach is that the remote sensing inputs may diagnostically capture patterns of water management and ancillary moisture sources that are not known a priori, either through the impact to the land-surface temperature or to vegetation indices via locally enhanced biomass production. OpenET is also at a higher spatial resolution (30 m) than the NWM (1 km).

3.1 NWM Background

In August 2016, the NWM was made operational (<https://water.noaa.gov/about/nwm>) by providing real-time spatially distributed hydrologic forecasts over the entire CONUS. The NWM uses the community WRF-Hydro model framework (Viterbo et al. 2020) for simulating different complex hydro-climatic processes such as ET, snowmelt, infiltration, runoff, and others that vary significantly due to changes in elevation, soils, and vegetation types as well as meteorological forcing conditions. The WRF-Hydro model includes the Noah-MP LSM (Yang et al. 2011; He et al., 2023) at 1-km spatial resolution as well as an overland routing scheme at 250 m. The use of Noah-MP in WRF-Hydro allows users to select among multiple physics options. Further details about Noah-MP can be found in the technical description (He et al., 2023). In this study we used the retrospective simulation from NWM V2.1 obtained from the Amazon Web Services (AWS) portal (<https://noaa-nwm-retrospective-2-1-pds.s3.amazonaws.com/index.html>). More details about the NWM V2.1 general configurations and its retrospective run are given in Appendix A. The NWM has tunable parameters that were used for calibration and can be categorized into two types, as listed in Appendix A: (1) constants, which are held fixed across the calibration region (as indicated by ‘type constant’ in Appendix A), or (2) multiplier adjusted, which are adjusted from spatially variable a-priori values using a scalar multiplier that serves as a calibration parameter. A-priori values were obtained from soil and other physical properties as described by Lahmers et al., (2021) and Gochis et al. (2020). The use of multipliers and constants avoids the challenges of high dimensionality in the calibration of distributed models, while still taking advantage of spatially distributed information from datasets such as STATSGO (<https://sdmdataaccess.sc.egov.usda.gov>). Parameters over watersheds upstream of stream gages used in calibration are adjusted separately to match the streamflow at that gage, a process that can

result in watershed scale spatial differences in parameter patterns that may manifest in watershed scale process differences, an effect we observed in some of our results. Here we did not attempt to change any of the NWM parameters or calibration procedures; however, we mention this because the parameters calibrated by the NWM team that underpin the retrospective results used in this study do impact our comparisons.

3.2 OpenET

The OpenET project is a broad collaborative effort to provide spatially continuous ET data for the western U.S. (Melton et al., 2022). The project provides daily, monthly and annual ET at 30-m spatial resolution. The ET models included in OpenET are summarized in Table 2. The platform primarily utilizes Landsat satellite data, along with grid-based weather data, including solar radiation, air temperature, humidity, and wind speed. Some models in the OpenET framework also integrate data from GOES, Suomi NPP, and Terra and Aqua satellites.

Four of the models (ALEXI/DisALEXI, eeMETRIC, SSEBop, and geeSEBAL) are based on principles of surface energy balance and use the Landsat surface temperature product as a key remote sensing input, along with vegetation indices and surface albedo. PT-JPL is based on a Priestley-Taylor formulation for ET and is most sensitive to optical vegetation index remote sensing inputs, but also integrates land surface temperature (LST) in constraining net radiation (Fisher et al., 2008). The SIMS model uses a reflectance-based approach, principally driven by Landsat NDVI along with a crop coefficient computed from vegetation density and condition, reference ET data, and soil evaporation coefficients computed from a gridded soil water balance model (Pereira et al., 2020; Melton et al., 2012). Since SIMS applies primarily to agricultural

areas that are a small part of the basin, it was excluded from our model-specific comparisons, although it does factor into the ensemble ET value for agricultural areas.

Table 2. OpenET models used and their inputs

Model acronym	Full name	Satellite and ancillary inputs	Meteorological inputs
ALEXI/DisALEXI (Anderson et al., 2018; Anderson et al., 1997)	Atmosphere-Land Exchange Inverse/ALEXI disaggregation (ver. 0.0.27)	<i>Primary:</i> Thermal data from GOES (ALEXI) and Landsat (DisALEXI); surface reflectance from MODIS and Landsat TM/ETM+/OLI <i>Secondary:</i> NLCD land cover data	Insolation, near-surface wind, air temperature, vapor pressure and atmospheric pressure from the Climate Forecast System Reanalysis (CFSR); ALEXI additionally uses CFSR atmospheric temperature profile data
eeMETRIC (Allen et al., 2011; Allen et al., 2005)	Mapping Evapotranspiration at High Resolution with	<i>Primary:</i> Surface reflectance and thermal radiation	Insolation, near-surface wind speed, air temperature, and

	Internalized Calibration (ver. 0.20.15)	from Landsat TM/ETM+/OLI <i>Secondary:</i> NLCD land cover data (for USA) and GlobCover for the globe, SRTM DEM, SSURGO (USA) and FAO Harmonized World Soil Database v 1.2 (globe)	vapor pressure from CIMIS and North American Land Data Assimilation System (NLDAS) for the USA, and from Climate Forecast System Ver. 2 (CFSV2) for the globe; Precipitation from gridMET
geeSEBAL (Bastiaanssen et al., 1998; Laipelt et al., 2021)	Surface Energy Balance Algorithm for Land using Google Earth Engine (ver. 0.2.1)	<i>Primary:</i> Surface reflectance and thermal radiation from Landsat TM/ETM+/OLI <i>Secondary:</i> Elevation from SRTM; Cropland data layers from USDA NASS	Daily shortwave incident radiation from GRIDMET; Hourly near-surface wind speed, air temperature, specific humidity and atmospheric pressure from NLDAS

<p>PT-JPL (Fisher et al., 2008)</p>	<p>Priestley-Taylor Jet Propulsion Laboratory (ver. 0.2.1)</p>	<p><i>Primary:</i> Surface reflectance and thermal radiation from Landsat TM/ETM+/OLI <i>Secondary:</i> MODIS maximum fraction of absorbed photosynthetically active radiation (fAPAR)</p>	<p>Insolation, near- surface wind speed, air temperature, and vapor pressure from CIMIS and North American Land Data Assimilation System (NLDAS)</p>
<p>SIMS (Melton et al., 2012; Pereira et al., 2020)</p>	<p>Satellite Irrigation Management Support (ver. 0.0.20)</p>	<p><i>Primary:</i> Surface reflectance from Landsat TM/ETM+/OLI and Sentinel-2A/2B <i>Secondary:</i> USDA Cropland Data Layer and state crop mapping data products; Surface</p>	<p>ET_o* data from Spatial CIMIS (in California); gridMET Eto and precipitation data for other states</p>

		reflectance from Terra/Aqua MODIS and Suomi NPP VIIRS can be used for gap-filling	
SSEBop (Senay, 2018; Senay et al., 2013)	Operational Simplified Surface Energy Balance (ver 0.1.5)	<i>Primary:</i> Thermal radiation from Landsat <i>Secondary:</i> NDVI from Landsat and SRTM DEM	ET _o data from Spatial CIMIS (in California) and gridMET; Daymet Daily Maximum Air Temperature (long- term average)

* ET_o is the grass reference evapotranspiration used as a primary scaling flux in multiple OpenET models determined from radiation and other weather variables (Melton et al., 2022).

3.3 Model comparison strategy

Our analysis is based on monthly data from 2017 to 2020, which was the time interval of ET information available from OpenET at the time of data download (<https://openetdata.org/>, accessed on October 11, 2022). To facilitate comparison, we aggregated the NWM V2.1 3-hour simulated data to monthly intervals. Additionally, for seasonal ET comparisons, both NWM and OpenET data were aggregated to 3-month intervals: March-May (MAM), June-August (JJA),

September-November (SON), and December-February (DJF). OpenET datasets (individual models and ensemble average) were spatially aggregated (through simple averaging) from 30-m resolution to match the 1-km model grid of NWM V2.1. This difference in spatial scale is a potential source of uncertainty and is addressed in the interpretation of the results.

We also evaluated the NWM ET and OpenET approaches against water balance estimates, where, at an annual scale, storage changes may be taken to be small, and the difference between precipitation (P) and streamflow (Q) approximates evapotranspiration (ET), ($ET_{wb} = P - Q$). In general, the water balance equation is:

$$\Delta S = P - Q - ET \quad (1)$$

where ΔS is change in storage. This can be expressed as:

$$ET + \Delta S = P - Q \quad (2)$$

which provides the basis for comparing cumulative P-Q with cumulative ET with an interpretation of storage changes. In the snowmelt-driven western U.S., the water year (October to September) ends at a time when snow has melted, runoff has occurred, and much of the seasonal soil moisture depleted, resulting in watershed water storage being close to its annual low point. While there is interannual storage, to a reasonable level of approximation, the water balance ET serves as a check on annual ET. This becomes an even better check over multiple years when the effect of storage is even less. However, water balance ET is still subject to uncertainties in precipitation and streamflow.

This analysis was conducted over four gaged sub-watersheds within the BRB (see Table 1 and Figure 1), using streamflow data obtained from the USGS stream gage network. We used precipitation data from two sources: the Analysis of Record for Calibration version 1.1 (AORC 1.1) (<https://hydrology.nws.noaa.gov/pub/AORC/V1.1/>), which serves as the forcing product for

NWM V2.1 simulations, and gridMET (Abatzoglou, 2013; <https://www.climatologylab.org/gridmet.html>), employed in some OpenET models. To assess the uncertainty in these two gridded precipitation datasets, we compared them with precipitation data from PRISM, DayMET, and NCLIM, obtained from the Climate Engine website (<https://www.climateengine.org/>). While the NWM calibrates hydrologic parameters to match streamflow observations and maintain water balance closure (as discussed in Section 3.1), it is important to note that the AORC precipitation input forcing is not adjusted by calibration.

3.4 Quantitative statistics

We used well-known quantitative statistics to compare the NWM ET and different OpenET models. Computed statistics included the SPAtial EFficiency metric (SPAEF), spatial mean, standard deviation (SDEV), and the standard error (SE).

SPAEF (Koch et al. 2018; Soltani et al. 2021) was used to evaluate the similarity of spatial patterns of ET from the NWM and the OpenET models. SPAEF ranges from $-\infty$ to 1, with 1 representing a perfect pattern match.

$$SPAEF = 1 - \sqrt{(\alpha - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2} \quad (3)$$

where α is the Pearson correlation coefficient between NWM ET and OpenET evaluated spatially across 1 km² grid cells; β is the fraction of coefficient of variations between NWM ET and OpenET, which quantifies the spatial variability; γ quantifies the fraction of the histogram intersection based on the z -scores of NWM ET and OpenET, with histogram bins determined based on the square root of the number of data values (i.e., number of grid cells in the domain). SPAEF was evaluated between NWM ET and each OpenET model for each season across the four

comparison years to assess changes in spatial similarity over time. Additionally, SPAEF was evaluated for the four-year averages of each season of NWM ET and the OpenET ensemble.

Spatial mean, \bar{x} , and standard deviation, SDEV, were calculated spatially across the entire watershed for each model at 1 km grid cell scale.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (4)$$

$$SDEV = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \quad (5)$$

where x_i denotes each pixel ET value, and n denotes the number of pixels for the entire ET map.

Standard error, SE, is a map computed based on the differences between the members of the OpenET ensemble. SE is derived from the standard deviation across the models, divided by the square root of the number of models (here 5) at each grid cell. This SE map serves as an indicator of OpenET inter-model uncertainty.

$$SE = \frac{\sigma}{\sqrt{N}} \quad (6)$$

where σ is the standard deviation across different OpenET models, N is the number of models (here 5)

4. Results

4.1 Temporal comparisons

Figure 3a shows the time-series of seasonal (3-month) ET from the NWM, the OpenET ensemble (calculated as the mean of the ensemble after filtering and removing outliers using the median absolute deviation approach), and individual OpenET models, each averaged over the BRB. Figure 3b illustrates the spatial variability in ET, as quantified by the standard deviation (SDEV) across the entire watershed for each model for each 3-month season evaluated across 1 km² grid cells with aggregated 3-month ET. Seasonal patterns in ET from both modeling systems

(NWM and OpenET) reflect seasonal insolation rates, vegetation leaf growth phenological stages, evaporative demand, and rainfall rates. ET values in the BRB peak during the warm season (JJA) and are at their lowest during the winter season (DJF) when solar radiation load and surface temperature are low. Comparing NWM ET with the suite of OpenET models reveals that NWM consistently estimated lower ET (by about 45 mm/season on average) throughout the study period, with more significant ET differences observed during the summer season (JJA). In Figure 3b, the variability in ET across the basin from NWM is often similar to the OpenET ensemble, except for 2018, where it was lower by 20-30 mm/season (Figure 3b). Since 2018 was a drought year (Figure 2), one possible explanation for this reduced variability is that NWM may not adequately account for ancillary sources of moisture (e.g., irrigation, shallow groundwater) that could sustain higher transpiration rates during drought years in some parts of the watershed.

The seasonal variability in spatial similarity between OpenET and NWM ET, using the SPAEF metric, reveals distinct spatial pattern differences among the various OpenET approaches (Figure 3c). The SPAEF values generally peak during JJA, or occasionally SON, indicating the highest spatial pattern alignment in summer, with the ensemble and SSEBop approaches generally higher. Conversely, the lowest SPAEF values are observed in DJF, suggesting a decrease in spatial pattern match during winter. Spatial pattern differences are compared in more detail using maps in Section 4.2 below.

To explore possible drivers of the ET differences between NWM and OpenET evident at the basin scale (Figure 3a), we examined seasonal biases in primary forcing variables from AORC (NWM) and gridMET (OpenET) data sources (Figure 4). Evaporative fluxes in NWM are strongly forced by precipitation, because it is a prognostic water balance model and ET is limited by model water availability, while measured precipitation has only a secondary influence through soil

evaporation in eeMETRIC and SIMS (excluded from basin-scale analyses here). The other OpenET models do not use precipitation as an input. Figure 4a shows that precipitation rates from AORC compared to gridMET are systematically lower by 13 mm/season on average. If all the extra precipitation was converted into ET in NWM, this would account for 29% of the total average difference between NWM and OpenET seasonal ET. OpenET models are most sensitive to forcings like insolation (primary), temperature, wind speed and vapor pressure deficit. Figure 4b indicates a strong agreement between the downwelling shortwave radiation (R_{si}) values obtained from gridMET and AORC datasets across different seasons.

ET_o is a primary scaling flux for eeMETRIC, SSEBop, and geeSEBAL, combining the impacts of insolation, wind, air temperature and vapor pressure on evaporative fluxes. Figure 4c compares bias-corrected gridMET ET_o that is used by OpenET with ET_o which we computed from AORC V1.1 data using the ET_o reference calculation described in Melton et al. (2022). OpenET bias-corrected gridMET ET_o uses ET_o data from over 800 weather stations in irrigated agriculture to correct long-term biases in ET_o calculated from gridMET (Volk et al., 2024b). In Figure 4c the bias-corrected gridMET ET_o is higher than AORC by about 7 mm/season on average, or about 16% of the NWM-OpenET difference.

Based on the above observations, biases in input forcing data do not appear to completely explain ET differences observed between NWM and OpenET at the basin scale. To gain a deeper understanding of these discrepancies, especially regarding where and how they manifest across the landscape, we examine the spatial patterns and spatial statistics of ET in greater detail in the following section (Section 4.2).

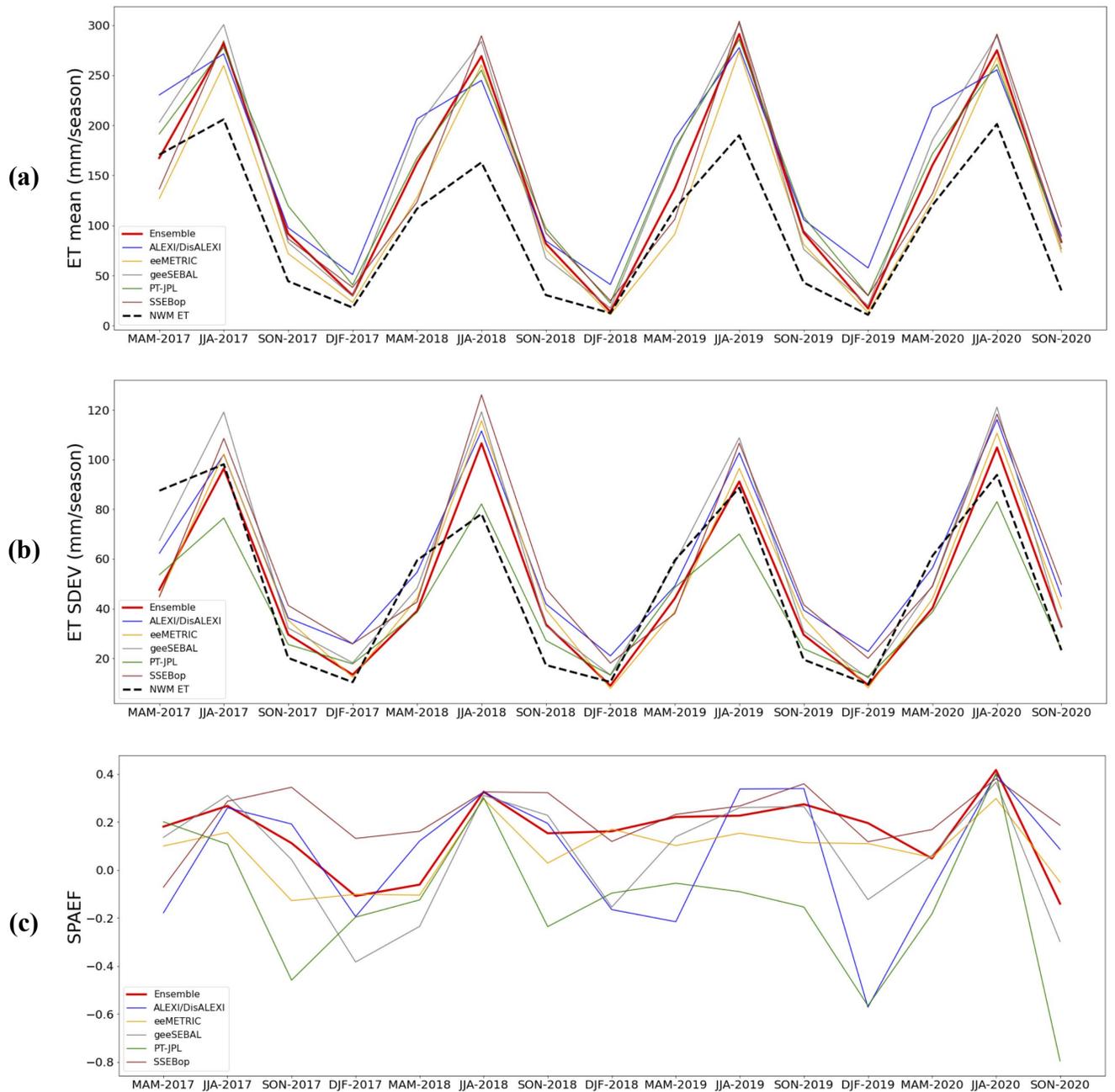


Figure 3. (a) Seasonal ET spatial mean, (b) Seasonal ET spatial standard deviation (SDEV) from NWM, OpenET ensemble, and individual OpenET models for the BRB and (c) Seasonal spatial pattern (SPAEF) between NWM ET and different OpenET members.

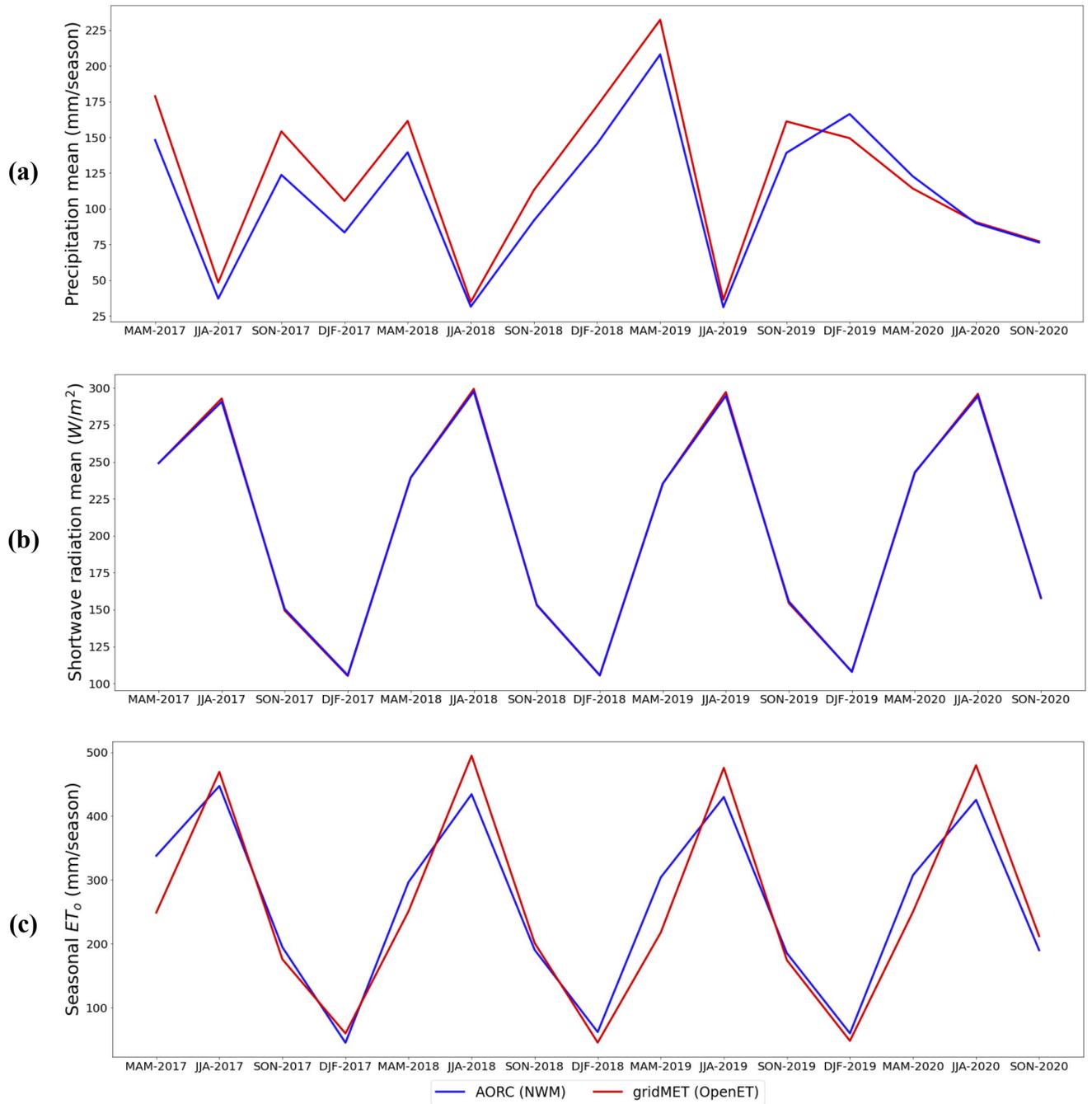


Figure 4. Time-series plots comparing (a) precipitation, (b) shortwave radiation and (c) ET_0 computed as a spatial watershed-average over seasons spanning (2017 - 2020) from gridMET and AORC.

4.2 Spatial comparisons

Spatial maps of seasonal ET, averaged over the 4 years, from both NWM and the OpenET ensemble help to explain where the differences apparent in Figure 3a arise within the basin (Figure 5). Comparisons with individual OpenET models are shown in Appendix B. In general, the spatiotemporal patterns from both modeling approaches are broadly similar, with lower ET typically observed in the eastern part of the domain, representing shrubland at higher elevations. However, notable differences are apparent in the spatial details of flux magnitude (Figure 6a). Also shown in Figure 6b are maps of time-averaged inter-model standard error (SE) by season, computed between the members of the OpenET ensemble. Small standard error indicates where the five basin-covering models agree well and, through convergence of evidence, we have high confidence in the ensemble estimate. Low standard error relative to the difference between NWM and OpenET estimates indicates disagreement between these different ET quantities.

NWM-OpenET differences are largest in JJA (Figure 5), but the general spatial structure/pattern of the differences is persistent between seasons (Figure 6a). The magnitude of NWM-OpenET differences tend to be larger than the standard error (Figure 6b), indicated by the difference scale ranging from -300 to 300 mm (Figure 6a) while the standard error scale tops out at 50 mm (Figure 6b). This suggests real differences between ET as estimated by OpenET and the NWM. Applying the SPAEF metric to assess spatial agreement between seasonal average ET maps in Figure 5, we obtain values of 0.08 (DJF), 0.11 (MAM), 0.32 (JJA), and 0.16 (SON), generally following the patterns in monthly assessments shown in Figure 3c ($-\infty$ being poor and 1 being good for SPAEF). The highest pattern match was observed in JJA, while the lowest SPAEF was obtained in DJF, likely due to the fact that many OpenET models are not specifically designed to handle snow-covered conditions.

NWM-OpenET differences during JJA when ET is highest are examined in greater detail in Figure 7, which also shows a Google Earth image of the basin to facilitate interpretation in relation to landcover and surface features. Some of the strongest persistent differences emerge in the riparian corridor along the Bear River in the eastern part of the basin. The land cover in this region is a mixture of riparian vegetation and irrigated agriculture. This is more apparent during summer as exhibited by the large contrast of very dry and very wet regions depicted in the ET ensemble. These riparian moisture sources and their impact on vegetation growth are captured by the remote sensing inputs to OpenET, primarily through higher vegetation indices (NDVI and/or LAI) and lower LST. These features do not appear to be depicted in the NWM JJA ET (Figure 5). Similarly, in the wetland areas in the southwest of the watershed near its outlet to the Great Salt Lake, OpenET is significantly higher than NWM. Patches of irrigated agriculture elsewhere in the basin are also associated with higher rates of ET from OpenET in comparison with NWM. These differences indicate limitations in the NWM representation of ET in riparian and irrigated agriculture areas, noting limitations in the representation of agricultural water management in the NWM.

One NWM area that stands out in JJA (and to a lesser extent in MAM) in Figures 5, 6a and 7a is associated with the Blacksmith Fork Watershed located in the southern central part of the basin (see identification Figure 7). The NWM ET in this sub-watershed of BRB is distinctly lower than its surroundings, with sharp boundaries that are not related to any specific physical features in Figure 7. We believe that this is an artifact of watershed-specific model calibration discussed further in Section 5.

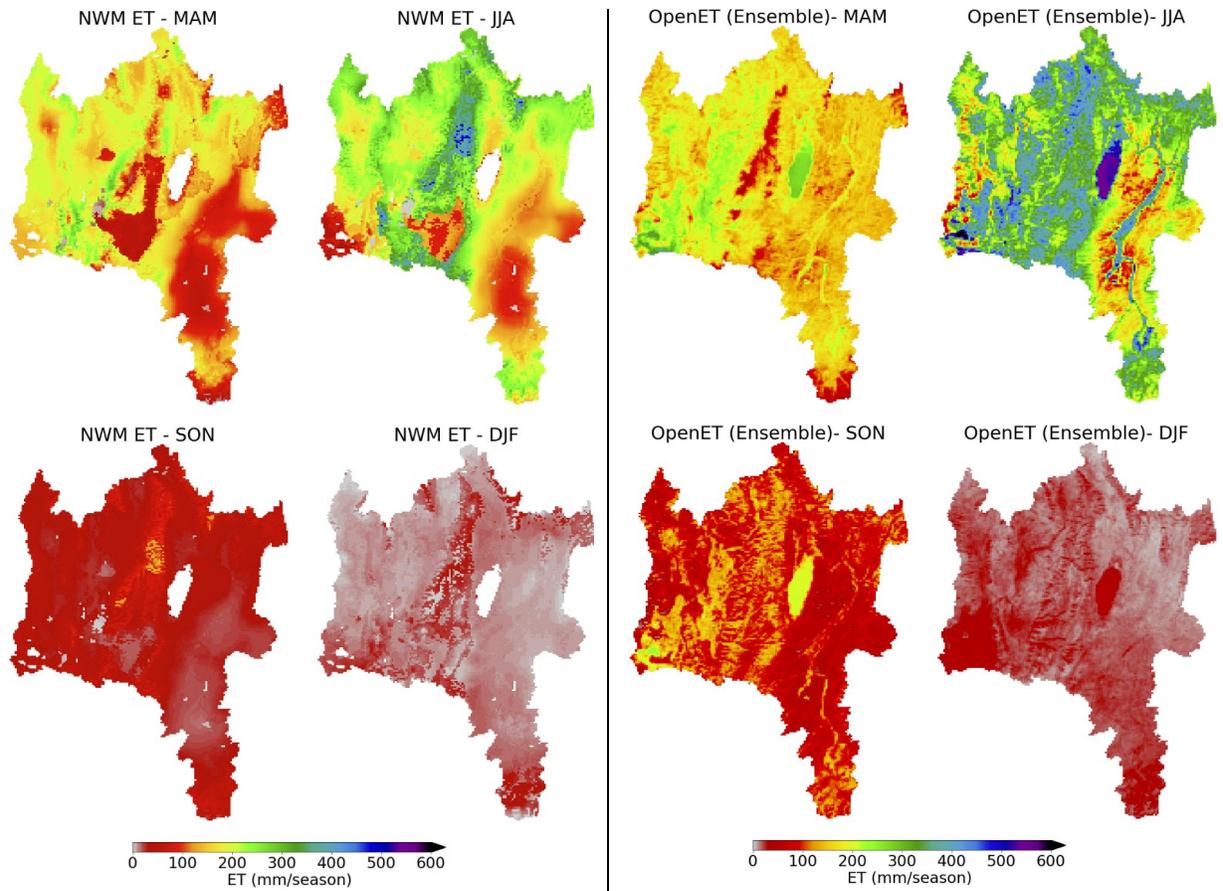


Figure 5. Comparison between maps of seasonal ET from both NWM and OpenET (ensemble).

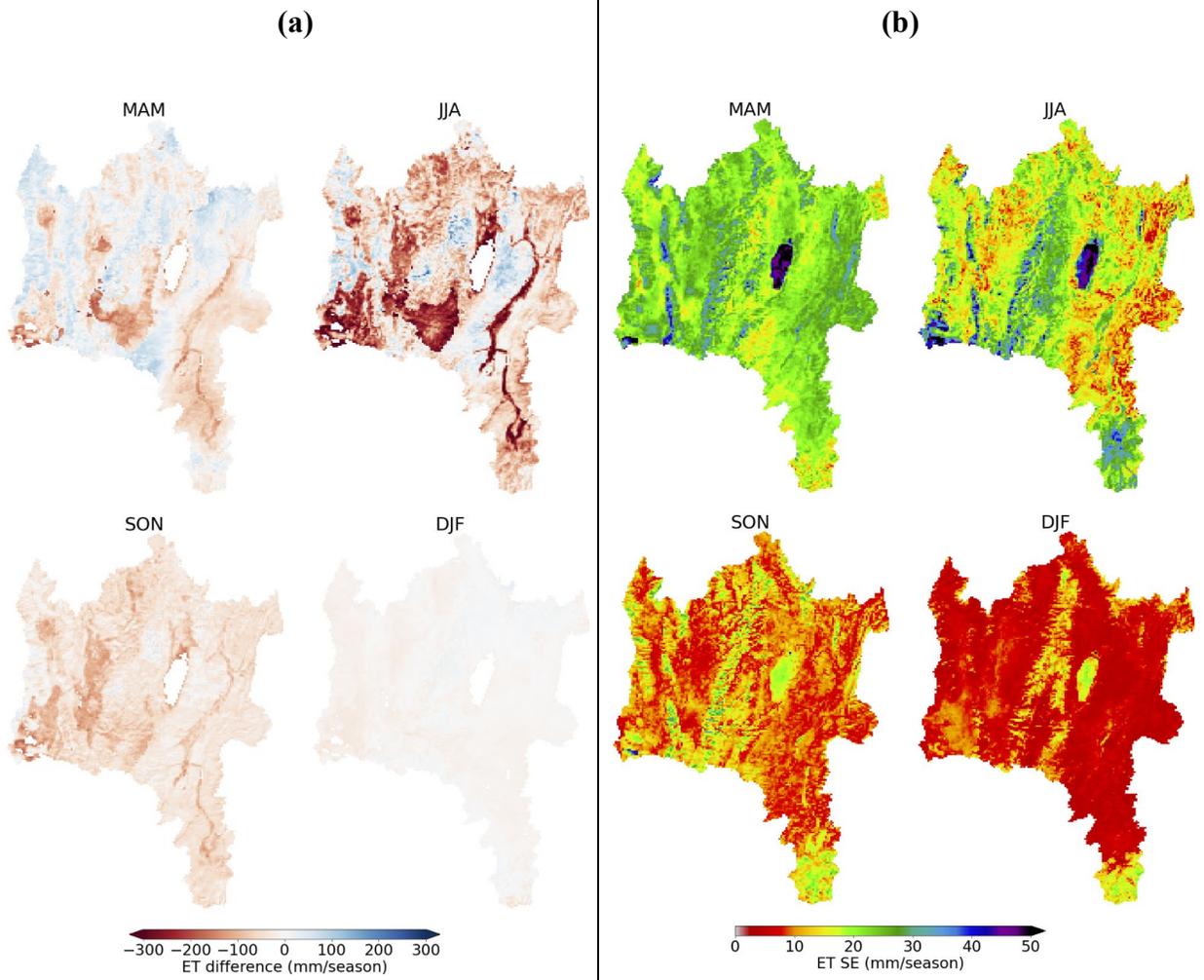


Figure 6. (a) Seasonal ET differences between NWM and ensemble OpenET; and (b) OpenET inter-model standard error (SE) maps among the five satellite-based methods used in this study.

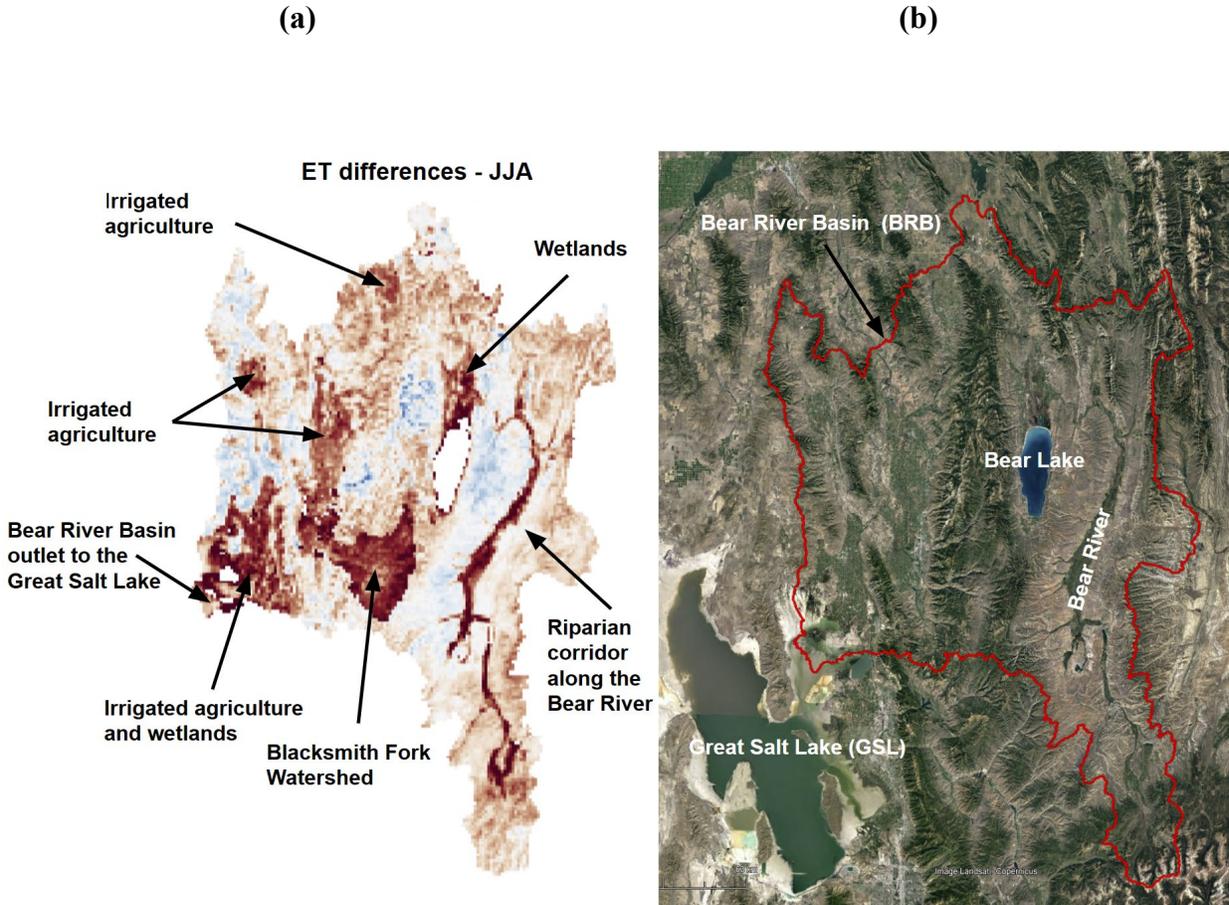


Figure 7. (a) ET differences in JJA between NWM ET and ensemble OpenET approach with details to facilitate the interpretation of results in relationship to landcover, and (b) a Google Earth image of the BRB.

4.3 Assessment of geographic variables related to model differences.

Given the relative spatial stability of NWM-OpenET difference patterns, we further examine relationships of these differences with geographic variables such as elevation, aspect and land use/land cover. Considering that the discrepancies are most significant during JJA when ET rates are highest, our primary focus will be on analyzing the factors influencing JJA variations.

Elevation and aspect, if not appropriately accounted for, can have a significant impact on remotely sensed LST and can add false variability to ET retrievals, particularly via the energy

balance. eeMETRIC explicitly accounts for topography using the Mountain Model package (Allen et al., 2013), and effectively flattens the LST field by an elevation based vertical lapse rate correction and by correcting the solar radiation flux inputs for slope and aspect. Noah-MP in the NWM accounts for the impact of topography on ET implicitly through the differences in atmospheric forcing (e.g., surface temperature, humidity, downward solar radiation) but with topography represented at a 1-km scale. While topographical features are important for surface energy balance, they seem to have a minimal influence on the discrepancies observed among the OpenET models and NWM ET (Figure 8a, b); however, we acknowledge that some uncertainty in the discrepancies maybe attributed to the upscaling of OpenET data to 1 km. Significant differences are observed in lower elevations ranging from 1200 to 1500 m and topographically flat regions that are dominated by wetlands, irrigated agricultural areas, and areas near the outlet of BRB to the GSL.

Based on Figure 8c, it appears that the magnitude of differences in JJA ET are similar for most land cover types. However, there are high differences obtained from herbaceous wetland and wooded wetland, with differences ranging approximately from -400 mm to 200 mm/season, but these are a small fraction of total area. However, in cropland and forest areas notable differences were also found. Since the NWM does not represent irrigation, ET from irrigated areas is limited by the NWM simulated soil moisture, which may be impacted by parameters adjusted during calibration.

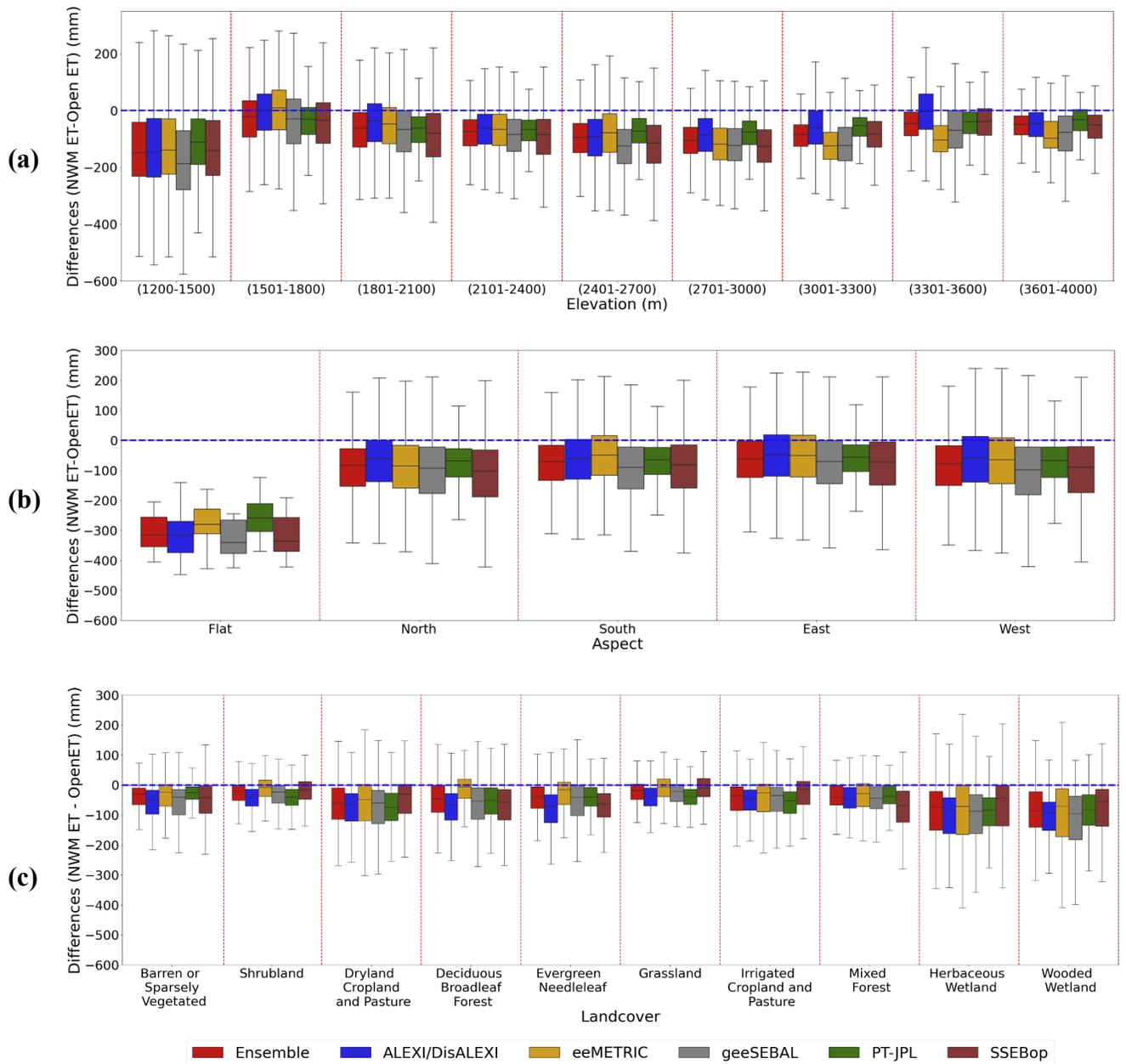


Figure 8. (a) Differences between JJA ET from NWM and different OpenET approaches based on elevation, (b) aspect, and (c) landuse/landcover classes. Note that in (b), flat includes grid cells characterized as flat by the ArcGIS evaluation of slope and aspect from NWM 1 km grid scale elevation.

4.4 Assessment of differences by water balance

The water balance calculation was done for four upstream sub-watersheds with minimal irrigation diversions or human impacts, that we designated as natural. These were: the Bear River Near UT-WY State Line; Logan River Above State Dam, Near Logan, UT; Smiths Forks Near Border, WY; and Big Creek Near Randolph, UT. First, we compared the NWM V2.1 streamflow estimates at daily timesteps with observations obtained from USGS gages (Figure 9). Overall, the results indicated a good agreement between the NWM V2.1 streamflow estimates and natural flow observations, although there is room for improvement in reducing the NWM overestimation of peak flow in one watershed.

We also compared sub-watershed precipitation (P) from AORC, which is used to force the NWM retrospective analysis dataset, with multiple gridded P datasets to help evaluate the uncertainty in the P data and to illustrate its potential effect on NWM ET. This analysis involved estimating ET by subtracting simulated streamflow (Q) from the different P datasets.

Figure 10 illustrates the monthly accumulated P and ET derived from different data sources. In Figure 10a, the black line represents the cumulative P from the AORC dataset, while in Figure 10b, we used a blue-shaded range constructed from gridMET, AORC, DayMET, PRISM and NCLIM precipitation data sources. The variability across these data sources reflects uncertainty in precipitation. Accumulated ET values were derived from NWM, OpenET, and the water balance calculation of precipitation (P) minus streamflow (Q) for the period of 2017 – 2020 for these sub-watersheds. Note that while accumulated ET is strictly increasing, cumulative P-Q includes the effects of snow and terrestrial water storage and thus increases more steeply from October (the beginning of the water year) until April/May and then declines as spring snowmelt results in high streamflow and reductions in snow storage.

AORC P was the input to the NWM, so in general, apart from storage effects, the red (cumulative NWM ET) and brown (cumulative P (AORC) – Q (NWM)) lines are consistent across different sub-watersheds as expected (Figure 10a). Accumulated storage within the watershed occurs when the brown line is above the red line, while there is accumulated deficit when the brown line is below the red line. The fact that the brown and red lines end each water year (and the full 3-year period) very close to each other reflects that, in the NWM, storage and deficit essentially balance out over the 3 years. This outcome is expected and is a result of the NWM's design/construction as a water balance prognostic model. The green line in Figure 10a used Q (observed) in the water balance ET calculation instead of Q (NWM). The differences between P (AORC) – Q (observed) (Green line) and NWM ET (Red line) reflect errors between model and observed NWM Q. This is particularly evident in the Bear and Smith rivers, where the NWM underestimates cumulative Q, suggesting that the NWM ET is relatively higher than ET evaluated from water balance, P (AORC) – Q (observed). However, even though relatively higher, NWM ET does not get as high as most OpenET estimates (Figure 10a). In the case of Logan River, Q (NWM) has better agreement with Q (observed) but is slightly higher, which suggests that NWM ET is relatively low, and thus being less than OpenET is not inconsistent.

In Figure 10b, we assessed the uncertainty associated with precipitation using different datasets. When comparing the four watersheds, we observed a higher amount of precipitation received in Logan River watershed than Smith, Bear and Big Creek watersheds. The annual precipitation rate in the Logan River watershed is approximately 900 mm, whereas in the Smith, Bear and Big Creek watersheds annual precipitation decreases to around 700 mm. Considering the different precipitation datasets, the variation between these sources was about 150 mm/year. This variance may contribute to differences between water balance ET, NWM ET and OpenET approaches. The

orange-shaded area represents the calculated range of water balance ET, which is derived from minimum and maximum values obtained when calculated as the difference between P (obtained from different datasets) and Q (observed). This range should be consistent and balanced by cumulative ET as watersheds do not generally accumulate or lose storage over the long term (multiple years). Comparing the four watersheds, we see that OpenET, and the ensemble ET value in particular, closely aligns with the water balance ET in the Logan River and Big Creek watersheds, falling within the uncertainty range denoted by the orange shading in Figure 10b. However, in the Bear and Smith watersheds, OpenET tends to overestimate cumulative ET by around 60-100% across OpenET models over the 4-year period when compared to the water balance ET derived from multiple precipitation sources.

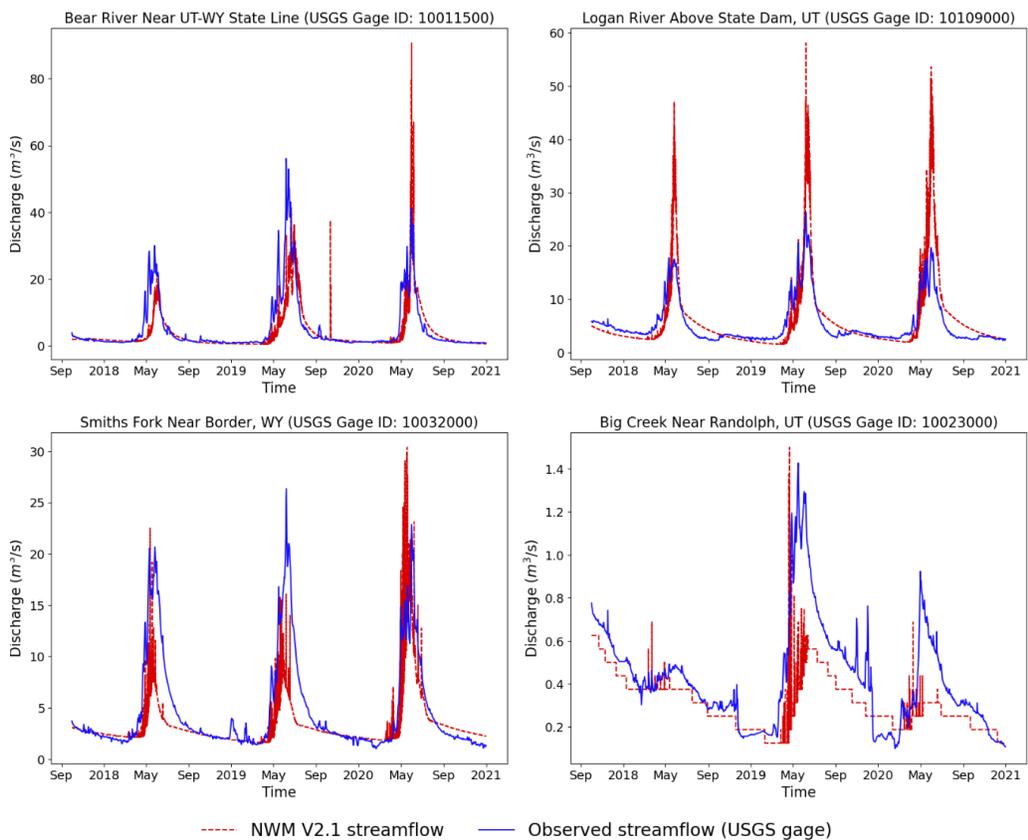


Figure 9. Hydrograph comparison between simulated NWM V2.1 streamflow and observed USGS streamflow gage.

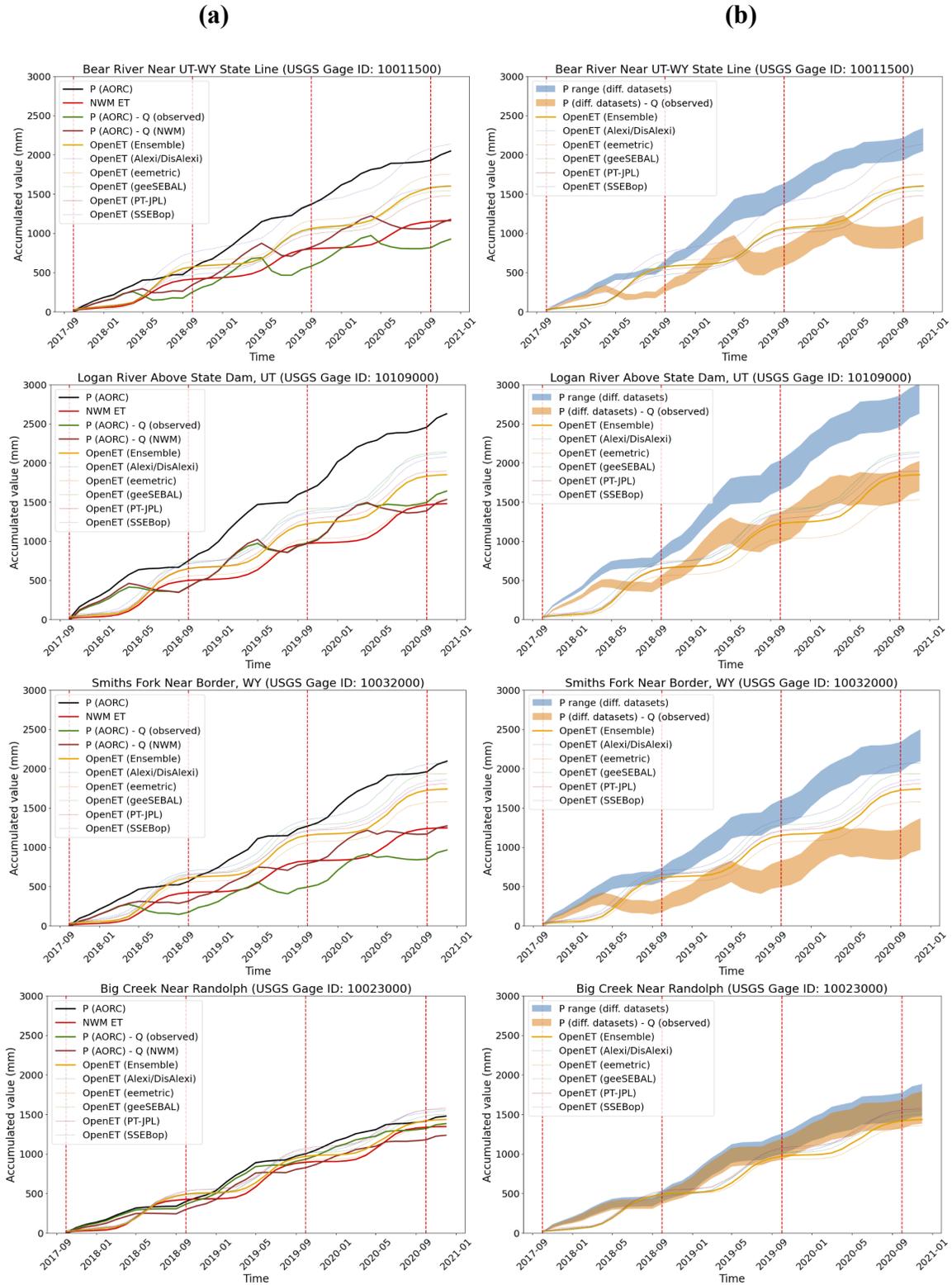


Figure 10. Accumulated P and ET curves based on NWM, OpenET models evaluated, and water balance P-Q which is comparable to ET when accounting for storage. (a) NWM water balance

components compared with OpenET models; (b) Variability due to precipitation uncertainty in water balance compared with OpenET models.

5. Discussion

The results above show that relative to OpenET the NWM estimates of ET are consistently lower in the study domain (Figure 3). In upstream sub-watersheds, the NWM estimated similar or slightly higher ET than water balance ET estimates. Discrepancies between OpenET and the NWM ET estimates are more apparent when viewing their spatial distribution and seasonal variations as opposed to long-term basin-wide totals, particularly in regions of irrigated agriculture, riparian corridors, and wetlands during the summer. This leads to questions as to the sources of these differences. Broadly there may be errors in the NWM, errors in OpenET, and errors in the inputs to each. We found that uncertainty in the forcing data to the NWM, namely precipitation, shortwave radiation and reference ET, are not likely sources of major uncertainties in NWM ET estimates. Precipitation is the largest source of uncertainty. We found (Figure 10) about a 150 mm/year typical difference across different precipitation data sources. The AORC precipitation input used in the NWM was at the low end of the gridded precipitation datasets used and a higher precipitation input to the NWM would result in higher NWM ET. While there was no direct adjustment of precipitation in the model calibration, it is possible that calibration adjusted other parameters to compensate for precipitation errors. However, the precipitation uncertainty, up to around 150 mm/year or 40 mm per season does not by itself appear big enough to explain ET differences. We base this on the observation in Figure 10b that even with the upper bound of P, the water balance estimate of ET is substantially less than OpenET for two of the four natural watersheds examined. This leads us to infer that a significant part of the difference between OpenET and NWM ET is overestimation by OpenET.

Comparing OpenET and NWM ET estimates by water balance estimation of ET as precipitation minus streamflow allows us to further investigate model discrepancies and causes. However, analysis was limited to four upstream natural watersheds not significantly impacted by diversions or riparian areas where river source water could supply ET. They are also geographically situated in locations unlikely to be impacted by regional groundwater inflows, noting that other researchers have resorted to consideration of regional groundwater in efforts to close watershed water balances (Soltani et al. 2021). For two of the four watersheds the cumulative observed P-Q range, accounting for gridded P data source differences, is below the range of cumulative OpenET values, from all methods, and from the ensemble. These two sub-watersheds, Bear (10011500) and Smith (10032000) are located in areas dominated by evergreen forest, a small component of the land cover characteristic of the full basin (Figure 1). A large-scale model intercomparison and evaluation study conducted by Volk et al. (2024a) using close to 150 Ameriflux towers identified evergreen forest as the land cover class with highest systematic bias in OpenET, with a mean bias error of approximately 25% for the ensemble ET value at water year scales. In addition, the mountainous terrain in these sub-watersheds presents an additional challenge to remote sensing models based on surface energy balance. By contrast, for the Logan River and Big Creek, OpenET and the water balance ET compare relatively well. These are drier sub-watersheds with less evergreen forest cover, and smaller runoff ratios. In the case of Big Creek, the runoff ratio is very small with essentially all precipitation being translated into ET, a process represented well by OpenET and the NWM. These findings suggest caution regarding the use of OpenET for analyses related to the water balance of evergreen forested mountain watersheds, which are where much streamflow originates in the intermountain western U.S. Alternatively, due to its strong agreement with water-balance ET estimates and streamflow, the

NWM showed strong results in partitioning discharge in forested upstream basins, suggesting that it may be a useful tool in assessing OpenET or other remote sensing ET products in similar basins. A broader evaluation of water balance sampling across other land cover types and topography present in the basin will be required to draw inferences regarding basin-scale biases between NWM and OpenET.

Looking at the differences between NWM ET and OpenET at the scale of the entire watershed, not limited to the four natural sub-watersheds, we found other important sources of differences and areas for further investigation and improvement in the NWM. During the peak ET months of the summer, we found two important patterns. First, where there is irrigation along river corridors, riparian areas and wetlands, OpenET systematically predicts higher ET than the NWM. This suggests additional water sources in these areas may be available for ET but are not partitioned as such in the NWM. Similarly, in low elevation, flat, and wetland areas NWM ET is systematically lower than OpenET. Further analyses of the NWM should evaluate these regions and the local processes related to ET partitioning for opportunities for improvement. Secondly, the outline of the Blacksmith Fork watershed stands out in the OpenET NWM ET difference map. This can be traced to a discontinuity in NWM ET at this watershed boundary. There is no physical reason for a discontinuity in ET associated with a watershed boundary, and we believe this is due to a difference in NWM parameters associated with calibration for specific basins. Calibrated parameters relate to soil properties, and thus affect the modeling of ET. This also provides an opportunity for considering independent ET datasets, such as from OpenET in the regional calibration of NWM parameters to move it towards better physical parameterization of the processes involved with ET that are important for water balance partitioning.

Another consideration is spatial variability. Notwithstanding the bias in mean between NWM ET and OpenET, the variability across the BRB represented by OpenET resampled to the NWM 1 km grid cell resolution (Figure 3b) is reasonably well captured by the NWM. The NWM spatial standard deviation is, for most seasons, within the range of OpenET spatial standard deviation, and close to the OpenET ensemble spatial standard deviation. This speaks to the utility of an ensemble quantity where multiple models are available. It also shows that even with omission of some physical ET processes from NWM, the spatial variability of NWM ET is consistent with OpenET, which incorporates satellite observations of land surface conditions to capture spatial patterns in ET. For one year, 2018, the most extreme drought year in our study period, the spatial standard deviation of NWM ET is less than that of OpenET. We surmise that possibly fewer wetter grid cells modeling higher ET values in the drought year, perhaps those with irrigation, and that acted to suppress the spatial standard deviation. Here the 30-m resolution OpenET data were aggregated to the 1-km resolution of the NWM, and part of these differences may be due to this aggregation. Another potential source of systematic bias in NWM simulated ET in the BRB may be caused by misrepresentation of hydrologic processes that can result in increased soil and shallow groundwater available for ET such as hillslope scale lateral flow, regional flow, and groundwater-vadose interactions. Simulation of lateral redistribution of water at hillslope scales (~1 m – 1 km) in land surface models can result in more accurate water balance estimates from land surface models and can result in increased soil moisture and ET and reduced dry bias (Fan et al., 2019; Yang et al., 2021; Ji et al., 2017). WRF-Hydro implements lateral connectivity between 250-m grid cells; it uses a shallow diffusive wave solution for overland lateral flow and a Boussinesq approximation for shallow subsurface lateral flow (within the 2-m soil column). The lateral fluxes are aggregated and disaggregated back to the 1 km Noah-MP grid. This scale

mismatch may result in spatial bias in the estimated lateral water subsidies that occur at hillslope scales. Also, in the NWM soil drainage is routed directly to a stream network and does not have the chance to resurface or be available for valley floor ET. In the study area, mountain block recharge is the major source of groundwater (Meixner et al., 2016) that is discharged through phreatophyte shrubland ET in valleys (Meyers et al., 2021; Beamer et al., 2013; Nichols 1993) and direct evaporation in the GSL. The model assumption of groundwater discharge to be only contributing to streamflow may be reasonable in some upstream sub-watersheds, however, at the basin-wide scale, a lack of representation of these larger scale processes may be responsible for some of the low seasonal ET biases in NWM results (relative to OpenET) that we see in the lowland areas of the BRB and near the GSL.

This work focused on one major subbasin draining into the Great Salt Lake and has highlighted the challenges associated with the NWM and its overall representation of the water balance and ET. It has also shown some of the uncertainties associated with independently estimated ET computed from methods such as the ensemble of satellite-driven models used by OpenET. While these specific findings are limited to this specific area, this watershed is typical of others in the intermountain western U.S. region, and we suspect that the findings will apply in other areas too. By comparing with the diagnostic ET estimates of OpenET we identified areas for further targeted studies that may lead to improvements in the NWM's partitioning of ET.

Conclusions

For our study area, the Bear River Basin centered in northern Utah and southeastern Idaho, we found that NWM estimates of ET are consistently lower with respect to OpenET, with biases attributed to both modeling systems and likely related to errors in model inputs and calibration,

missing processes in the represented in the NWM framework, and biases in OpenET related to land cover/topography.

We found that some NWM uncertainty is due to uncertainty in the input precipitation datasets, with the AORC precipitation that was used as input to the NWM being at the low end of the range of other gridded precipitation products used in the water balance evaluation. This suggests that improving precipitation input to the NWM offers an opportunity to improve NWM ET and water balance outputs.

We found that NWM ET was underestimated in lowland, riparian, wetland, and irrigated agriculture areas suggesting that it may not accurately model ancillary water supplies or represent processes that lead to increased partitioning of water to ET. There are thus opportunities to improve the NWM through better representation of these processes.

We found that NWM ET has discontinuities along watershed boundaries. These stand out when looking at differences between OpenET and NWM ET and appear to be caused by the NWM calibration that adjusts parameters or parameter multipliers across watersheds, and some of these parameters have an impact on soil moisture which then plays a role in the modeling of ET. There is in general no physical reason for these discontinuities and research to improve the calibration parameter adjustment approach that avoids these discontinuities would improve the model by advancing it closer to a more physical representation of the processes involved.

A water balance assessment in two evergreen forest dominated natural sub-watersheds within the Bear River Basin is consistent with biases reported in an OpenET evaluation study using eddy covariance instrumentation, where the OpenET ensemble ET value overpredicted flux tower observations in evergreen forest. These results point to paths towards improvement in OpenET in forested systems.

More broadly, we found that the comparisons between a diagnostic dataset such as those from satellite driven remote sensing ET from the OpenET platform with a prognostic model like the NWM is a valuable approach for evaluating the spatial and temporal biases and error in both systems. This comparative analysis approach may be useful in addressing model performance in other regions and in addressing uncertainties in other modeling systems particularly where observational data are limited.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Data availability

This paper used NOAA National Water Model CONUS Retrospective Data accessed on October 11, 2022 from <https://registry.opendata.aws/nwm-archive> (NOAA, 2023), and OpenET data retrieved from <https://openetdata.org/> (OpenET, 2023). Data and Jupyter Notebooks documenting the methods used to produce these results are available in HydroShare (Nassar and Tarboton, 2025).

Acknowledgements

This work was supported by the National Science Foundation (NSF) under collaborative grant 1835569 for Computational and Data Innovation Implementing a National Community Hydrologic Modeling Framework for Scientific Discovery, 1928369 for the integration of Reproducibility methods into HydroShare, and 2118329 Harnessing the Data Revolution Institute for Geospatial Understanding through an Integrative Discovery Environment. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. This work benefited from discussions with Gabriel Senay, the OpenET and WRF-Hydro teams, WRF-Hydro being the framework used to implement the U.S. National Water Model. We also appreciate the comments of anonymous Journal of Hydrology reviewers that improved the paper.

Appendix A: This includes both the NWM v2.1 general configuration as well as the calibrated parameters used for the retrospective analysis.

Table A1. NWM v2.1 General Configurations

Item	Description
Model resolution	1-km land surface grid; 250-m terrain routing grid; NHDPlusV2 vector channel routing network and conceptual groundwater basins.
Spin up period	Warm started with final states from a 10-year simulation, then acclimated by running

	February 1979 through December 1979 twice
Driving data	Analysis of Record for Calibration (AORC)
Output frequency	<p>CHROUT: Every hour, channel network output</p> <p>LAKEOUT: Every hour, reservoir (lake) output</p> <p>GWOUT: Every hour, conceptual groundwater output.</p> <p>LDASOUT: Every 3 hours, land model output</p> <p>RTOUT: Every 3 hours, high resolution terrain routing output</p>
Accumulation periods	<p>For the accumulation variables (3 hourly UGDNOFF, ACCET, ACSNOM), the accumulation takes place between restart dates:</p> <ol style="list-style-type: none"> 1. 00Z January 1 - 21Z March 31 2. 00Z April 1 - 21Z June 30 3. 00Z July 1 - 21Z September 30 4. 00Z October 1 - 21Z December 31
Model time step	<p>Forcing data: 3600 seconds</p> <p>Land surface model: 3600 seconds</p> <p>Channel routing: 300 seconds</p> <p>Terrain routing: 10 seconds</p>

Table A2. Calibrated parameters of NWM based on (Gochis et al. 2020), including their calibration range (Max. and Min.) and type.

Parameter	Description	Min	Max.	Type	Unit
<i>Soil parameters</i>					
BEXP	Pore size distribution index.	0.4	1.9	Multiplier	Dimensionless
SMCMAX	Porosity, saturated value of soil moisture (volumetric).	0.8	1.2	Multiplier	Volumetric fraction
DKSAT	Saturated hydraulic conductivity.	0.2	10	Multiplier	m^{-1}
RSURFEXP	Soil evaporation resistance	1	6	Constant	Dimensionless
Vegetation					
CWPVT	Empirical canopy wind parameter.	0.5	2	Multiplier	m^{-1}
VCMX25	Maximum rate of carboxylation at 25°C	0.6	1.4	Multiplier	$\mu mol m^{-2} s^{-1}$
MP	Slope of conductance-to-photosynthesis relationship	0.6	1.4	Multiplier	Unitless
<i>Snow parameters</i>					
MFSNO	Melt factor for snow depletion curve; larger value yields a smaller snow cover fraction	0.25	2	Multiplier	Dimensionless

<i>Runoff parameters</i>					
REFKDT	Reference soil infiltration parameter (used in runoff formulation). It significantly impacts surface infiltration and hence the partitioning of total runoff into surface and subsurface runoff. Increasing REFKDT decreases surface runoff.	0.1	4	Constant	Unitless
SLOPE	Slope index for soil drainage.	0	1	Constant	0–1
RETDEPRTFAC	Surface retention depth.	0.1	20000	Constant	Unitless
LKSATFAC	Multiplier on lateral hydraulic conductivity (controls anisotropy between vertical and lateral conductivity).	10	10000	Constant	Unitless
<i>Groundwater parameters</i>					
ZMAX	Maximum groundwater bucket depth.	10	250	Constant	mm
EXPON	Exponent controlling rate of bucket drainage as a function of depth.	1	3	Constant	Dimensionless

Appendix B: Comparison between maps of seasonal ET from both NWM and OpenET approaches

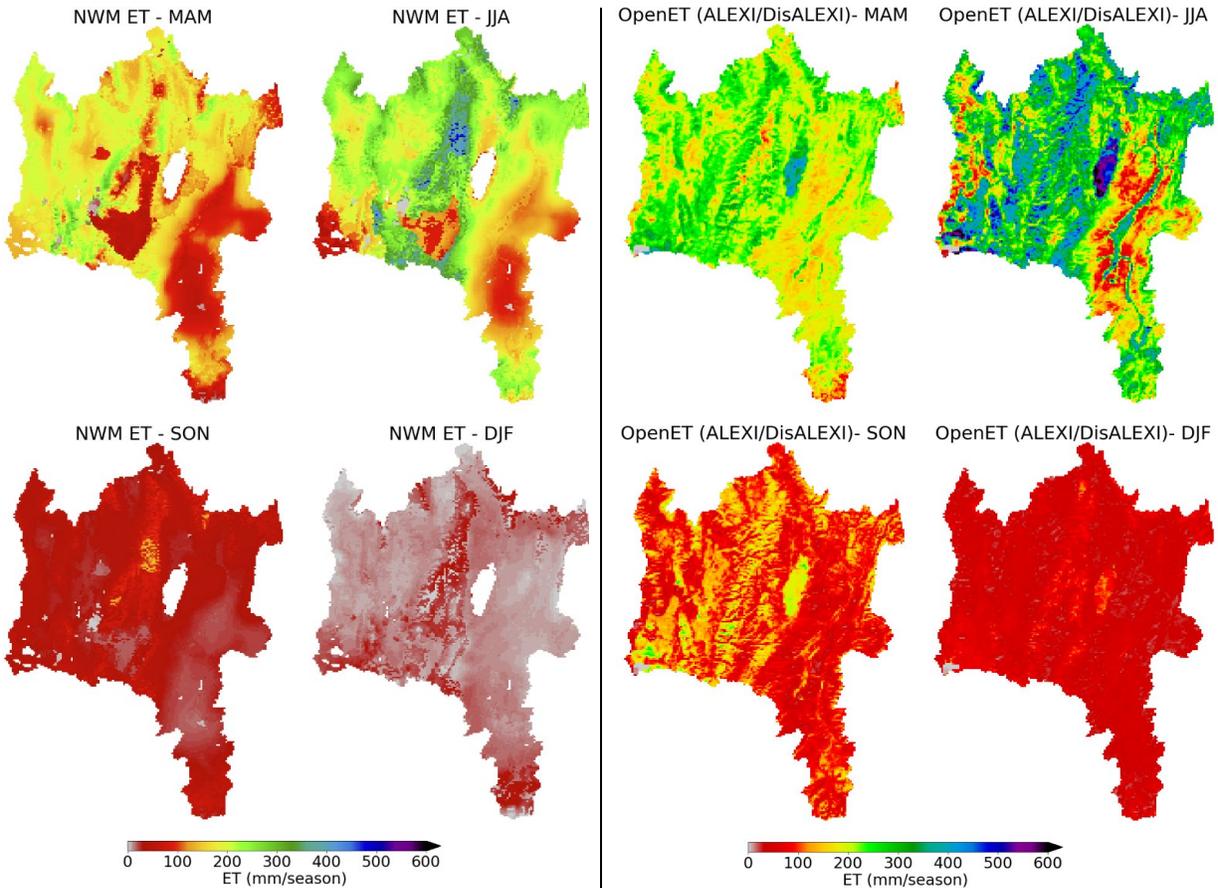


Figure B1. Comparison between maps of seasonal ET from both NWM and OpenET (ALEXI/DisALEXI).

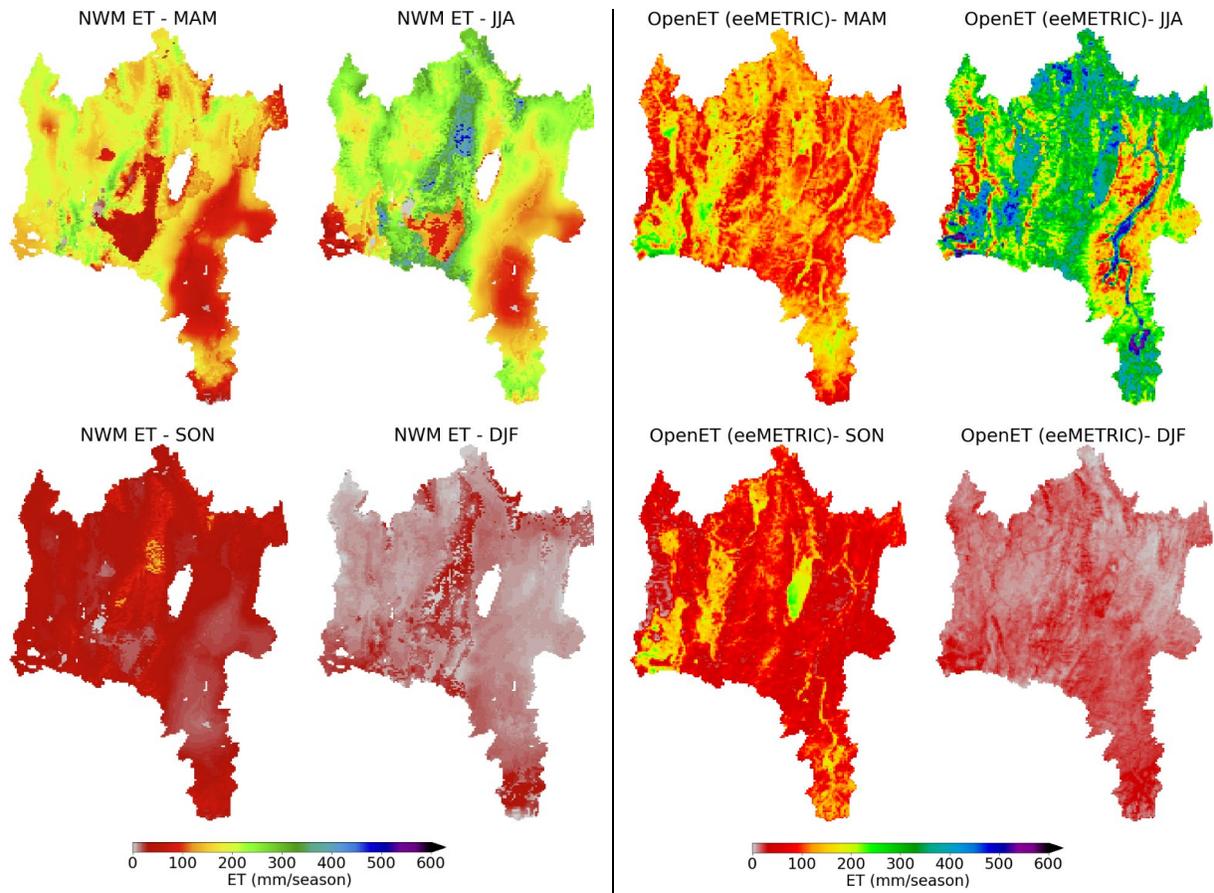


Figure B2. Comparison between maps of seasonal ET from both NWM and OpenET (eeMETRIC).

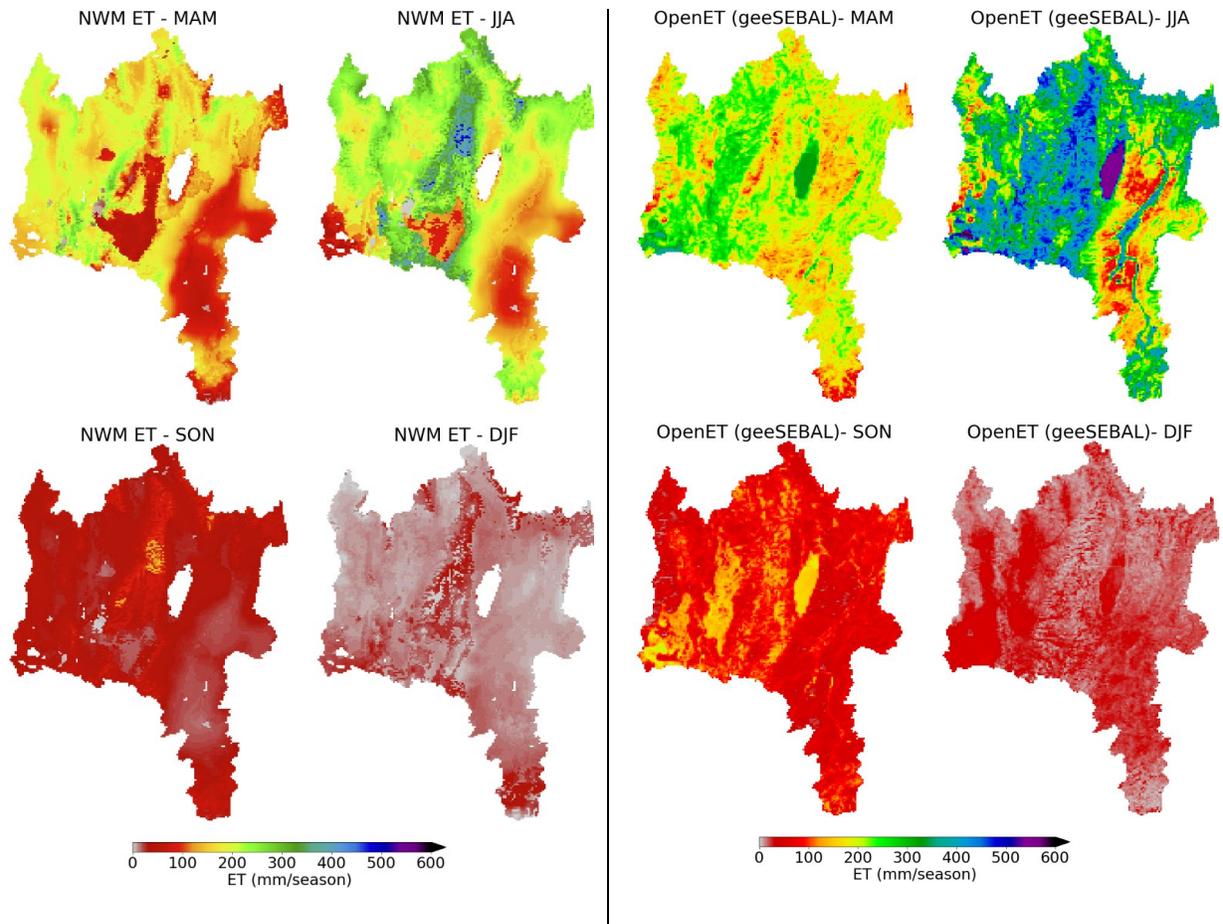


Figure B3. Comparison between maps of seasonal ET from both NWM and OpenET (geeSEBAL).

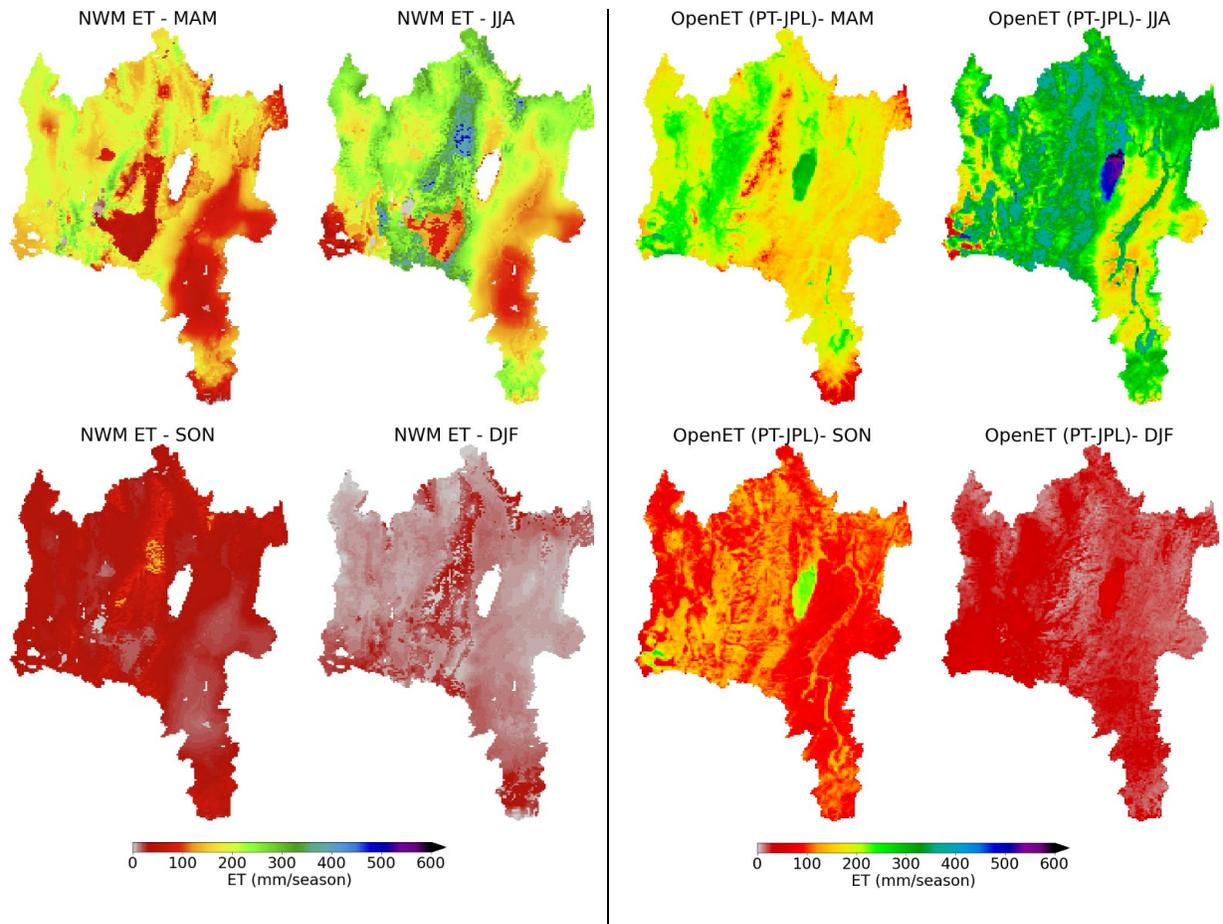


Figure B4. Comparison between maps of seasonal ET from both NWM and OpenET (PT-JPL).

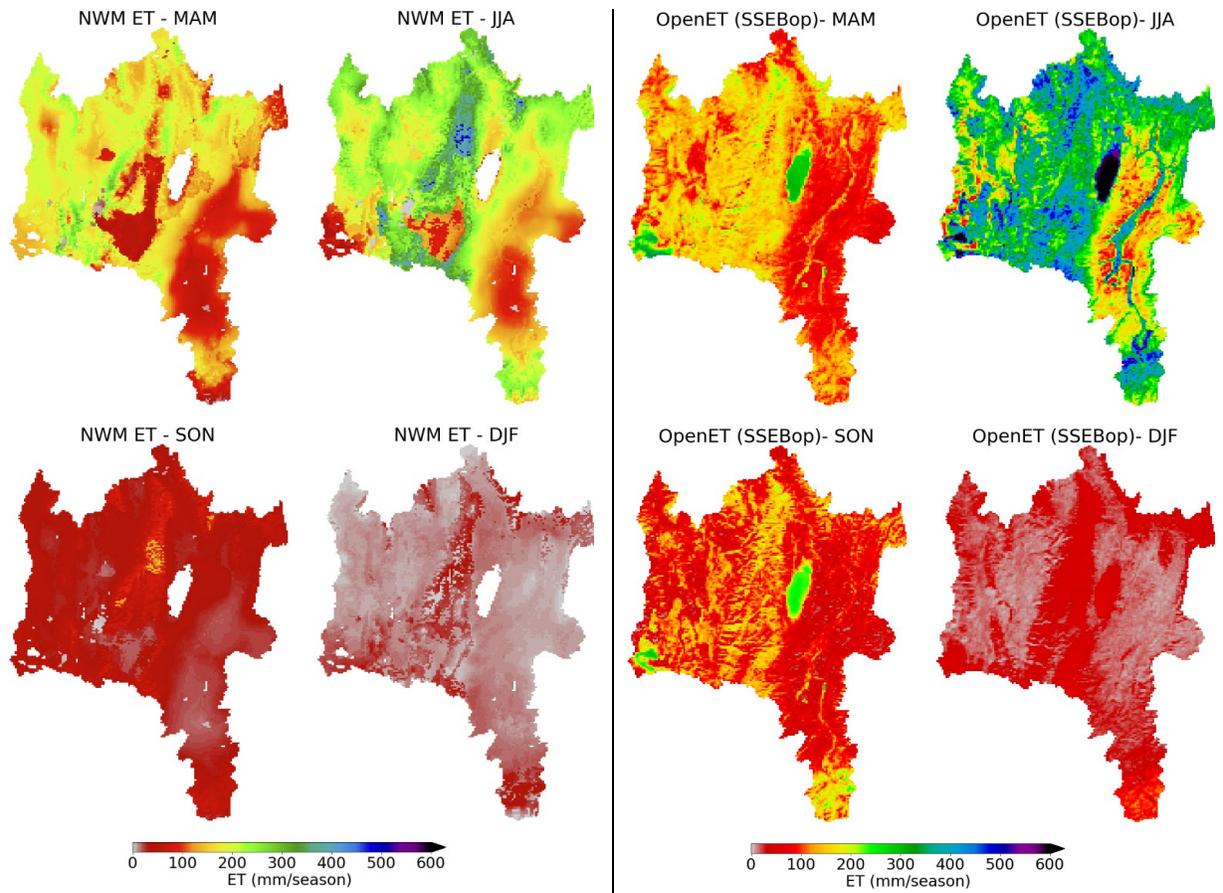


Figure B5. Comparison between maps of seasonal ET from both NWM and OpenET (SSEBop).

References

1. Abatzoglou, J.T., 2013. Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, 33(1), 121-131. <https://doi.org/10.1002/joc.3413>.
2. Abdelkader, M., Temimi, M., Ouarda, T.B., 2023. Assessing the National Water Model's Streamflow Estimates Using a Multi-Decade Retrospective Dataset across the Contiguous United States. *Water*, 15(13), 2319. <https://doi.org/10.3390/w15132319>.
3. Abolafia-Rosenzweig, R., He, C., Chen, F., Ikeda, K., Schneider, T., Rasmussen, R., 2023. High resolution forecasting of summer drought in the western United States. *Water Resources Research*, 59, e2022WR033734. <https://doi.org/10.1029/2022WR033734>.
4. Allen, R., Irmak, A., Trezza, R., Hendrickx, J.M.H., Bastiaanssen, W., Kjaersgaard, J., 2011. Satellite-based ET estimation in agriculture using SEBAL and METRIC. *Hydrological Processes*, 25(26), 4011–4027. <https://doi.org/10.1002/hyp.8408>.
5. Allen, R.G., Tasumi, M., Morse, A., Trezza, R., 2005. A Landsat-based energy balance and evapotranspiration model in Western US water rights regulation and planning. *Irrigation and Drainage Systems*, 19(3-4), 251–268. <https://doi.org/10.1007/s10795-005-5187-z>.
6. Allen, R.G., Burnett, B., Kramber, W., Huntington, J., Kjaersgaard, J., Kilic, A., et al., 2013. Automated calibration of the METRIC-Landsat evapotranspiration process. *Journal of the American Water Resources Association*, 49(3), 563–576. <https://doi.org/10.1111/jawr.12056>.
7. Anderson, M., Gao, F., Knipper, K., Hain, C., Dulaney, W., Baldocchi, D., et al., 2018. Field-scale assessment of land and water use change over the California Delta using remote sensing. *Remote Sensing*, 10(6), 889. <https://doi.org/10.3390/rs10060889>.
8. Anderson, M.C., Norman, J.M., Diak, G.R., Kustas, W.P., Mecikalski, J.R., 1997. A two-source time-integrated model for estimating surface fluxes using thermal infrared remote sensing. *Remote Sensing of Environment*, 60(2), 195–216. [https://doi.org/10.1016/S0034-4257\(96\)00215-5](https://doi.org/10.1016/S0034-4257(96)00215-5).
9. Amazon Web Services (AWS). National Water Model Archive on AWS. Available at: <https://registry.opendata.aws/nwm-archive/>.

10. Amazon Web Services (AWS). U.S. National Water Model Retrospective Dataset Version 2.1. Available at: <https://noaa-nwm-retrospective-2-1-pds.s3.amazonaws.com/index.html>.
11. Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., et al., 2001. FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society*. [https://doi.org/10.1175/1520-0477\(2001\)082<2415:FANTTS>2.3.CO;2](https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2)
12. Bastiaanssen, W.G.M., Menenti, M., Feddes, R.A., Holtslag, A.A.M., 1998. A remote sensing surface energy balance algorithm for land (SEBAL). 1. Formulation. *Journal of Hydrology*, 212-213, 198–212. [https://doi.org/10.1016/S0022-1694\(98\)00253-4](https://doi.org/10.1016/S0022-1694(98)00253-4).
13. Beamer, J.P., Huntington, J.L., Morton, C.G., Pohll, G.M., 2013. Estimating annual groundwater evapotranspiration from phreatophytes in the Great Basin using Landsat and flux tower measurements. *Journal of the American Water Resources Association*, 49(3), 518–533. <https://doi.org/10.1111/jawr.12058>.
14. Chen, F., Dudhia, J., 2001. Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Monthly Weather Review*, 129(4), 569–585. [https://doi.org/10.1175/1520-0493\(2001\)129<0569:CAALSH>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2)
15. Chen, F., Mitchell, K., Schaake, J., Xue, Y., Pan, H.-L., Koren, V., et al., 1996. Modeling of land surface evaporation by four schemes and comparison with FIFE observations. *Journal of Geophysical Research, D: Atmospheres*, 101(D3), 7251–7268. <https://doi.org/10.1029/95JD02165>.
16. Climate Engine. Desert Research Institute and University of California, Merced. Available at: <http://climateengine.org>.
17. Climatology Lab. gridMET Dataset Platform. Available at: <https://www.climatologylab.org/gridmet.html>.
18. Ek, M.B., Mitchell, K.E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., et al., 2003. Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *Journal of Geophysical Research, D: Atmospheres*, 108(D22). <https://doi.org/10.1029/2002JD003296>.

19. Fan, Y., Clark, M., Lawrence, D.M., Swenson, S., Band, L.E., Brantley, S.L., et al., 2019. Hillslope hydrology in global change research and earth system modeling. *Water Resources Research*, 55(2), 1737–1772. <https://doi.org/10.1029/2018WR023903>.
20. Fisher, J.B., Tu, K.P., Baldocchi, D.D., 2008. Global estimates of the land-atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites. *Remote Sensing of Environment*, 112(3), 901-919. <https://doi.org/10.1016/j.rse.2007.06.025>.
21. Garousi-Nejad, I., Tarboton, D.G., 2022. A comparison of National Water Model retrospective analysis snow outputs at snow telemetry sites across the western United States. *Hydrological Processes*. <https://doi.org/10.1002/hyp.14469>.
22. Gochis, D.J., and Coauthors, 2020. The NCAR WRF-Hydro Modeling System Technical Description. NCAR Technical Note. https://ral.ucar.edu/projects/wrf_hydro/documentation.
23. Koch, J., Demirel, M., Stisen, S., 2018. The SPAtial Efficiency metric (SPAEF): multiple-component evaluation of spatial patterns for optimization of hydrological models. *Geoscientific Model Development*, 11(5), 1873-1886. <https://doi.org/10.5194/gmd-11-1873-2018>.
24. Hain, C.R., Crow, W.T., Anderson, M.C., Tugrul Yilmaz, M., 2015. Diagnosing neglected soil moisture source–sink processes via a thermal infrared–based two-source energy balance model. *Journal of Hydrometeorology*, 16(3), 1070–1086. <https://doi.org/10.1175/JHM-D-14-0017.1>.
25. Hansen, C., Shafiei Shiva, J., McDonald, S., Nabors, A., 2019. Assessing retrospective national water model streamflow with respect to droughts and low flows in the Colorado River Basin. *Journal of the American Water Resources Association*, 55(4), 964–975. <https://doi.org/10.1111/1752-1688.12784>.
26. He, C., and Coauthors, 2023. The NCAR Noah-MP Technical Documentation. <https://doi.org/10.5065/ew8g-yr95>.
27. Ji, P., Yuan, X., Liang, X.-Z., 2017. Do lateral flows matter for the hyper-resolution land surface modeling? *Journal of Geophysical Research*, 122(22), 12,077–12,092. <https://doi.org/10.1002/2017jd027366>.

28. Lahmers, T.M., Hazenberg, P., Gupta, H., Castro, C., Gochis, D., Dugger, A., et al., 2021. Evaluation of NOAA National Water Model parameter calibration in semiarid environments prone to channel infiltration. *Journal of Hydrometeorology*, 22(11), 2923–2942. <https://doi.org/10.1175/JHM-D-21-0032.1>.
29. Laipelt, L., Henrique Bloedow Kayser, R., Santos Fleischmann, A., Ruhoff, A., Bastiaanssen, W., Erickson, T. A., & Melton, F., 2021. Long-term monitoring of evapotranspiration using the SEBAL algorithm and Google Earth Engine cloud computing. *ISPRS Journal of Photogrammetry and Remote Sensing: Official Publication of the International Society for Photogrammetry and Remote Sensing*, 178, 81–96. <https://doi.org/10.1016/j.isprsjprs.2021.05.018>
30. Lin, P., Rajib, M. A., Yang, Z.-L., Somos-Valenzuela, M., Merwade, V., Maidment, D. R., et al., 2018. Spatiotemporal evaluation of simulated evapotranspiration and streamflow over Texas using the WRF-Hydro-RAPID modeling framework. *JAWRA Journal of the American Water Resources Association*, 54(1), 40–54. <https://doi.org/10.1111/1752-1688.12585>
31. Mecikalski, J. R., Diak, G. R., Anderson, M. C., & Norman, J. M., 1999. Estimating fluxes on continental scales using remotely sensed data in an atmospheric–land exchange model. *Journal of Applied Meteorology and Climatology*, 38(9), 1352–1369. [https://doi.org/10.1175/1520-0450\(1999\)038<1352:EFOCSU>2.0.CO;2](https://doi.org/10.1175/1520-0450(1999)038<1352:EFOCSU>2.0.CO;2)
32. Meixner, T., Manning, A. H., Stonestrom, D. A., Allen, D. M., Ajami, H., Blasch, K. W., et al., 2016. Implications of projected climate change for groundwater recharge in the western United States. *Journal of Hydrology*, 534, 124–138. <https://doi.org/10.1016/j.jhydrol.2015.12.027>
33. Melton, F. S., Johnson, L. F., Lund, C. P., Pierce, L. L., Michaelis, A. R., Hiatt, S. H., et al., 2012. Satellite irrigation management support with the terrestrial observation and prediction system: A framework for integration of satellite and surface observations to support improvements in agricultural water resource management. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(6), 1709–1721. <https://doi.org/10.1109/JSTARS.2012.2214474>
34. Melton, F. S., Huntington, J., Grimm, R., Herring, J., Hall, M., Rollison, D., Erickson, T., Allen, R., Anderson, M., Fisher, J., Kilic, A., Senay, G., Volk, J., Hain, C., Johnson, L.,

- Ruhoff, A., Blanenau, P., Bromley, M., Carrara, W., Daudert, B., Doherty, C., Dunkerly, C., Friedrichs, M., Guzman, A., Halverson, G., Hansen, J., Harding, J., Kang, Y., Ketchum, D., Minor, B., Morton, C., Ortega-Salazar, S., Ott, T., Ozdogan, M., Revelle, P., Schull, M., Wang, T., Yang, Y., & Anderson, R., 2022. OpenET: Filling a critical data gap in water management for the western United States. *JAWRA Journal of the American Water Resources Association*, 58(6), 971–994. <https://doi.org/10.1111/1752-1688.12956>
35. Meyers, Z. P., Frisbee, M. D., Rademacher, L. K., & Stewart-Maddox, N. S., 2021. Old groundwater buffers the effects of a major drought in groundwater-dependent ecosystems of the eastern Sierra Nevada (CA). *Environmental Research Letters*, 16(4), 044044. <https://doi.org/10.1088/1748-9326/abde5f>
36. Mu, Q., Heinsch, F. A., Zhao, M., & Running, S. W., 2007. Development of a global evapotranspiration algorithm based on MODIS and global meteorology data. *Remote Sensing of Environment*, 111(4), 519–536. <https://doi.org/10.1016/j.rse.2007.04.015>
37. Nassar, A., & Tarboton, D., 2025. Comparison of evapotranspiration from the National Water Model retrospective analysis with remotely sensed estimates from OpenET. *HydroShare*. <https://doi.org/10.4211/hs.a2f48c5949704ec5822202cbcf631287>
38. Nichols, W. D., 1993. Estimating discharge of shallow groundwater by transpiration from greasewood in the northern Great Basin. *Water Resources Research*, 29(8), 2771–2778. <https://doi.org/10.1029/93WR00930>
39. NOAA Office of Water Prediction (OWP). Analysis of record for calibration: Version 1.1. Available at: <https://hydrology.nws.noaa.gov/pub/AORC/V1.1/>
40. NOAA Office of Water Prediction (OWP). Next generation water resources modeling framework. Available at: <https://noaa-owp.github.io/ngen/>
41. NOAA Office of Water Prediction (OWP). U.S. National Water Model (NWM). Available at: <https://water.noaa.gov/about/nwm>
42. OpenET Data Platform. Available at: <https://openetdata.org/>
43. Pastorello, G., Papale, D., Chu, H., Trotta, C., Agarwal, D., Canfora, E., et al., 2017. A new data set to keep a sharper eye on land-air exchanges. *Eos*. <https://doi.org/10.1029/2017eo071597>

44. Pereira, L. S., Paredes, P., Melton, F. S., Johnson, L. F., López-Urrea, R., Cancela, J., & Allen, R. G., 2020. Prediction of basal crop coefficients from fraction of ground cover and height. *Agricultural Water Management*, 241, 106197.
<https://doi.org/10.1016/j.agwat.2020.106197>
45. Senay, G. B., 2018. Satellite psychrometric formulation of the operational simplified surface energy balance (SSEBop) model for quantifying and mapping evapotranspiration. *Applied Engineering in Agriculture*, 34(3), 555–566. <https://doi.org/10.13031/aea.12614>
46. Senay, G. B., Bohms, S., Singh, R. K., Gowda, P. H., Velpuri, N. M., Alemu, H., & Verdin, J. P., 2013. Operational evapotranspiration mapping using remote sensing and weather datasets: A new parameterization for the SSEB approach. *Journal of the American Water Resources Association*, 49(3), 577–591.
<https://doi.org/10.1111/jawr.12057>
47. Seo, B.-C., Krajewski, W. F., & Quintero, F., 2021. Multi-scale hydrologic evaluation of the national water model streamflow data assimilation. *Journal of the American Water Resources Association*, 57(6), 875–884. <https://doi.org/10.1111/1752-1688.12955>
48. Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., et al., 2008. A description of the advanced research WRF version 3. NCAR Technical Note, 475, 113. Retrieved from
<https://www.academia.edu/download/38038853/WRFV3technote.pdf>
49. Soltani, M., Bjerre, E., Koch, J., & Stisen, S., 2021. Integrating remote sensing data in optimization of a national water resources model to improve the spatial pattern performance of evapotranspiration. *Journal of Hydrology*, 603, 127026.
<https://doi.org/10.1016/j.jhydrol.2021.127026>
50. Soltani, M., Koch, J., & Stisen, S., 2021. Using a groundwater adjusted water balance approach and copulas to evaluate spatial patterns and dependence structures in remote sensing derived evapotranspiration products. *Remote Sensing*, 13(5), 853.
<https://doi.org/10.3390/rs13050853>
51. United States Department of Agriculture (USDA). Soil data access. Available at:
<https://sdmdataaccess.sc.egov.usda.gov>

52. Utah Division of Water Resources, 2004. Bear River Basin: Planning for the future, Utah state water plan. Available at: <https://water.utah.gov/wp-content/uploads/2020/06/Bear-River-2004.pdf>
53. Viterbo, F., Read, L., Nowak, K., Wood, A. W., Gochis, D., Cifelli, R., & Hughes, M. , 2020. General assessment of the operational utility of National Water Model reservoir inflows for the Bureau of Reclamation facilities. *Water*, 12(10), 2897. <https://doi.org/10.3390/w12102897>
54. Volk, J. M., Huntington, J. L., Melton, F. S., et al., 2024a. Assessing the accuracy of OpenET satellite-based evapotranspiration data to support water resource and land management applications. *Nature Water*. <https://doi.org/10.1038/s44221-023-00181-7>
55. Volk, J. M., Dunkerly, C., Kilic, A., Pearson, C., Huntington, J. L., Melton, F., Allen, R., & Minor, B., 2024b. Mapping and correcting the bias in gridded reference evapotranspiration over irrigated agricultural lands across the contiguous United States. *Agricultural Water Management*. Manuscript in preparation.
56. Wurtsbaugh, W. A., & Sima, S. (2022). Contrasting management and fates of two sister lakes: Great Salt Lake (USA) and Lake Urmia (Iran). *Water*, 14(19), 3005. <https://doi.org/10.3390/w14193005>
57. Yang, Z.-L., Niu, G.-Y., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al., 2011. The Community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins. *Journal of Geophysical Research*. <https://doi.org/10.1029/2010jd015140>
58. Yilmaz, M. T., Anderson, M. C., & Zaitchik, B., 2014. Comparison of prognostic and diagnostic surface flux modeling approaches over the Nile River basin. *Water Resources*. <https://doi.org/10.1002/2013WR014194>