JGR Atmospheres

RESEARCH ARTICLE
10.1029/2023JD038522

Key Points:

- Retrospective NOAA National Water Model soil moisture and streamflow were evaluated for drought-monitoring applications
- The National Water Model was comparable in skill to land model guidance currently used to inform the US drought monitor
- The National Water Model's operational forcing strategy currently limits its application for real-time drought monitoring

Supporting Information:

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

Correspondence to:

M. Hughes,
mimi.hughes@noaa.gov

Citation:

Hughes, M., Jackson, D. L., Unruh, D., Wang, H., Hobbins, M., Ogden, F. L., et al. (2024). Evaluation of retrospective National Water Model Soil moisture and streamflow for drought-monitoring applications. *Journal of Geophysical Research: Atmospheres*, *129*, e2023JD038522. https://doi.org/10.1029/2023JD038522

Received 11 JAN 2023
Accepted 23 FEB 2024

Evaluation of Retrospective National Water Model Soil Moisture and Streamflow for Drought-Monitoring Applications

M. Hughes[1], D. L. Jackson[1,2], D. Unruh[3], H. Wang[4], M. Hobbins[1,2], F. L. Ogden[3], R. Cifelli[1], B. Cosgrove[3], D. DeWitt[4], A. Dugger[5], T. W. Ford[6], B. Fuchs[7], M. Glaudemans[3,8], D. Gochis[5], S. M. Quiring[9], A. RafieeiNasab[5], R. S. Webb[1], Y. Xia[10], and L. Xu[4,11]

[1]NOAA Physical Sciences Laboratory, Boulder, CO, USA, [2]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA, [3]NOAA Office of Water Prediction, Silver Spring, MD, USA, [4]NOAA Climate Prediction Center, University Research Court, College Park, MD, USA, [5]National Center for Atmospheric Research, Boulder, CO, USA, [6]Illinois State Water Survey, Prairie Research Institute, University of Illinois, Urbana, Champaign, IL, USA, [7]National Drought Mitigation Center, University of Nebraska, Lincoln, NE, USA, [8]Now at NOAA Analyze, Forecast, Support Office, Silver Spring, MD, USA, [9]Ohio State University, Columbus, OH, USA, [10]SAIC at NOAA Environmental Modeling Center, National Centers for Environmental Prediction, College Park, MD, USA, [11]ERT, Inc, Laurel, MD, USA

Abstract The National Oceanic and Atmospheric Administration (NOAA)'s National Water Model (NWM) provides analyses and predictions of hydrologic variables relevant to drought monitoring and forecasts at fine time and space scales (hourly, 0.25–1 km). We present results exploring the potential for NWM soil moisture and streamflow analyses to inform operational drought monitoring. Both agricultural and hydrologic drought monitoring rely either explicitly or implicitly on an accurate representation of anomalous soil moisture values. Much of our analysis focuses on comparisons of soil moisture anomalies in the NWM to those from in-situ observations. To establish benchmarks for NWM soil moisture skill, we also include other gridded data sets currently used to inform the US Drought Monitor, specifically those from the North American Land Data Assimilation System phase 2 (NLDAS-2) land surface models. We then compare NWM streamflow low flows with ~500 stream gauges from the United States Geological Survey (USGS) Hydro-Climatic Data Network of undisturbed basins. The NWM soil moisture simulation's skill parallels that from NLDAS-2. The accuracy of drought condition identification from NWM streamflow exceeds that based on soil moisture as determined by Critical Success Index scores for extreme dry percentiles. Different meteorological forcings are used in the operational NWM cycles than those used in this retrospective analysis. This forcing disconnect, together with concerns about current-generation land surface model soil moisture-transport schemes, inhibit its current operational use for drought monitoring.

Plain Language Summary The National Oceanic and Atmospheric Administration's National Water Model offers output relevant for drought monitoring. This paper evaluates the National Water Model soil moisture and streamflow with drought applications in mind, and compares those evaluations to other modeling tools currently used to inform drought monitoring. We find the model's ability to estimate anomalous low flows exceeds its ability to estimate anomalously dry soils, and discuss its current potential to inform operational drought monitoring.

1. Introduction

Droughts rank among the most destructive and expensive natural hazards. The National Center for Environmental Information (NCEI) counts 28 separate billion-dollar drought events since 1980, together costing $261.5 billion and nearly 4,000 lives. The most expensive of these was the 1988 drought and heat wave in the Midwestern U.S. that directly cost over $45 billion and 454 lives (NCEI website). Due to these extreme socioeconomic costs, Congress established the National Integrated Drought Information System (NIDIS) in 2006, reauthorizing it in 2014 and 2019, to coordinate drought monitoring, forecasting, and early warning capabilities across the country (https://www.drought.gov/about/). In addition, the National Oceanic Atmospheric Administration's (NOAA) National Weather Service (NWS) works toward minimizing the effects of these disasters, targeting improved extreme weather monitoring and forecasts to "drive a better response to high-impact weather," including drought.

Since spring 1999, the US Drought Monitor (USDM) has undertaken operational drought monitoring across the U.S. (Svoboda et al., 2002), including the contiguous U.S. (CONUS), Alaska, Hawaii, and Puerto Rico. The USDM is a joint effort of the National Drought Mitigation Center (NDMC, which preceded NIDIS and explicitly targets drought mitigation and monitoring with its mission), NOAA, and the U.S. Department of Agriculture (USDA). The USDM derives a weekly drought-assessment map from a blend of objective drought metrics and expert opinion gathered from hundreds of regional contributors. This map is used by decision-makers across the U.S. and across sectors (Noel et al., 2020), ranging from federal agencies, such as the USDA and the Internal Revenue Service, who use it to trigger disaster declarations and determine eligibility of those impacted for low-interest loans and tax deferral on drought-forced sales of livestock, respectively, down through regional, state, local, tribal, and basin scales to support decision makers' drought responses.

The USDM takes a “convergence of evidence” approach, integrating data from a wide array of drought-relevant metrics that monitor all aspects of the hydrologic cycle, examining drought both through its supply and demand fluxes and through surface-moisture states, as well as using evidence from other indicators (e.g., reservoir levels). The moisture supply side (i.e., precipitation) is estimated by standardized drought metrics such as the Standardized Precipitation Index (SPI; McKee et al., 1993), and the demand side often by the Evaporative Stress Index (ESI; Anderson et al., 2011), or the Evaporative Demand Drought Index (EDDI; Hobbins et al., 2016; McEvoy et al., 2016). The state of surface-moisture availability is estimated through various metrics, both as a function of observed and modeled soil moisture (SM), and from streamflow indications given by USGS streamflow gauges.

The current observational SM monitoring infrastructure in the United States is inadequate for drought-monitoring purposes: in-situ SM measurements are limited spatially to a few thousand locations across the CONUS and temporally to a few decades (Quiring et al., 2016). Most current satellites only estimate SM in the top ∼5 cm of the soil (e.g., Champagne et al., 2016; Entekhabi et al., 2010; Kerr et al., 2012); satellite gravimetry offers estimates of changes to water in deeper soils, but at coarse spatial resolution and limited accuracy (Chen et al., 2022). Instead, current drought-monitoring efforts rely on spatially distributed modeling of SM (Ford & Quiring, 2019). One such model-based dataset is the multi-agency North American Land Data Assimilation System, phase 2 (NLDAS-2) suite of land surface models (LSMs; Xia et al., 2012a, 2012b) that estimate SM at a ∼12-km resolution with a roughly 4-day operational latency (NLDAS, 2022).

In August 2016, the National Water Model (NWM) version 1.0 became operational at NOAA’s National Centers for Environmental Prediction (NCEP). The NWM simulates and forecasts SM, streamflow, and other hydrologic quantities over CONUS at 1-km (LSM component) to 250-m (hydrologic routing) spatial resolutions with lead times ranging from hours to weeks (Cohen et al., 2018; Viterbo et al., 2020). The NWM’s high-resolution, distributed numerical hydrological model guidance provides streamflow for 2.7 million river reaches, supplementing the authoritative hydrological forecasts provided by NOAA’s River Forecast Centers at their ∼3,600 forecast points across CONUS.

The NWM has some advantages over existing products for drought monitoring. First, the NWM’s hourly cycling offers decreased latency (∼45 min) compared to other existing drought monitor-relevant numerical guidance (e.g., ∼4 days for NLDAS-2). This low latency could offer essential information for real-time drought monitoring, particularly for detecting the effects of rapidly developing synoptic events (e.g., flooding, hurricanes) on changes in soil moisture and streamflow, and in particular could offer evidence for quickly evolving “flash droughts” (Otkin et al., 2018). The NWM provides outputs in a spatially consistent manner for the entire forecast domain, and benefits from the significant effort NOAA invests in developing accurate meteorological forcing fields (e.g., precipitation, Martinaitis et al., 2021, 2020; NOAA, 2020). A crucial component of the NWM’s usage in operational decision-making is its quantification of skill in estimating SM and streamflow in the context of drought monitoring, which allows decision-makers to discern whether it meets their usability requirements.

Since its operationalization, the NWM has undergone several rapid version upgrades, with version 2.0 operational from June 2019–April 2021. The current (November 2022) operational version is v2.1. These version upgrades have targeted previously known and operationally identified limitations in the NWM’s performance and formulation. While this paper evaluates a recent NWM version with a long, publicly available retrospective analysis, several upgrades are on the horizon for the NWM, including improvements targeting the physics of moisture transport through the soil (La Follette et al., 2023). These planned improvements in the physical representation of soil moisture transport in the NWM will increase its accuracy.

This paper presents drought-targeted evaluations of SM and streamflow from a 26-year (1993–2018) retrospective simulation of the NWM v2.0. We present skill scores for SM and streamflow compared against in-situ SM and gauged streamflow, using NLDAS-2 as a benchmark (e.g., Best et al., 2015; Newman et al., 2017). We then present case studies of the NWM's simulation of one historical drought identified in Wood et al. (2015) as one of four relevant tests for new drought-monitoring products. We conclude with a discussion of the limitations of the current NWM land surface model formulations and complications introduced by its operational configuration.

2. Data and Methods

The model and observational data used are summarized in Table S1 in Supporting Information S1.

2.1. Land Surface Models

This paper focuses on retrospective NWM v2.0 simulations. The modeling framework that underlies the NWM is the Weather Research and Forecasting (WRF)-Hydro v5.1.0 (Gochis et al., 2020), configured as shown in Table S2 in Supporting Information S1.

In NWM v2.0, land-surface processes were modeled using the Noah-MP land surface scheme (Niu et al., 2011), with namelist options also shown in Table S2 in Supporting Information S1. The Noah-MP code was optimized to perform partitioning of latent and sensible heat fluxes from the total radiation budget and provide lower boundary conditions for the WRF mesoscale meteorological model. The NWM v2.0 was calibrated to streamflow using its retrospective simulations with atmospheric forcings from NLDAS-2 (Viterbo et al., 2020). In addition to soil column physics representation from the Noah-MP LSM, soil hydrology in the WRF-Hydro/NWM system is greatly influenced by overland and saturated subsurface routing processes and the allowance for ponded water which, through their redistribution of water and subsequent feedback on the land surface states, influence spatial and temporal patterns of SM across complex terrain landscapes (e.g., Arnault et al., 2016; Lahmers et al., 2022). The NWM's advanced calibration also directly influences these SM patterns directly through its impact on infiltration (Sofokleous et al., 2022).

SM from four NLDAS-2 LSMs–Noah, SAC, VIC, and Mosaic–was used to benchmark the NWM evaluations. Note that although SAC, unlike the other three NLDAS-2 models and Noah-MP, was never conceptualized as a lower bound of an atmospheric model to describe energy and water fluxes, for brevity hereafter we refer to all four NLDAS-2 models and the Noah-MP component of the NWM as "land surface models". The Noah (Ek et al., 2003; Wei et al., 2013) and Mosaic (Koster & Suarez, 1992, 1994) models were developed for the purpose of providing surface energy and water fluxes to atmospheric models, whereas SAC (Koren et al., 2004) and VIC (Wood et al., 1997) were developed as uncoupled hydrology models for the purpose of predicting hydrologic fluxes such as runoff and streamflow. Vegetation classes in the NLDAS-2 LSMs are derived from the satellite-based University of Maryland vegetation classification scheme (Hansen et al., 2000), with a monthly seasonal cycle of vegetation in VIC, Noah, and Mosaic (Mitchell et al., 2004; Xia et al., 2012b). A more detailed description of the formulation of these models is found in Xia et al. (2015b).

The NWM v2.0 retrospective simulation and the four NLDAS-2 LSMs used NLDAS-2 meteorological data as common upper boundary forcing. The forcing data were used at their native 1-hr temporal resolution for all models and at 1/8° grid spacing for the NLDAS-2 LSMs. For the NWM simulation, the NLDAS-2 forcings were first interpolated to the 1-km resolution NWM Noah-MP grid: conservative remapping was used for precipitation and bilinear interpolation was used for all other variables. Then, several variables were also downscaled: pressure, relative humidity, and temperature were adjusted based on the 1-km grid point elevation with a North American Regional Reanalysis (Mesinger et al., 2006) lapse-rate-based adjustment; and shortwave radiation was downscaled using a topographic effect adjustment from the WRF model code (Dudhia, 1989).

The five LSMs vary in their SM depths and thus our analysis (on 0–100 cm average SM for the primary results) required some LSM-dependent processing. Given their common heritage, the NWM's Noah-MP and NLDAS-2 Noah estimated SM at four computational soil-depth discretizations: 0–10 cm, 10–40 cm, 40–100 cm, and 100–200 cm. We calculated two further integrated levels–0–100 cm and 0–200 cm–using weighted averaging of the individual levels for the NWM. Mosaic output does not directly provide SM for the 40–100 cm and 100–200 cm layers, so we computed these from weighted averaging of other layers. The SAC is a storage-type model without

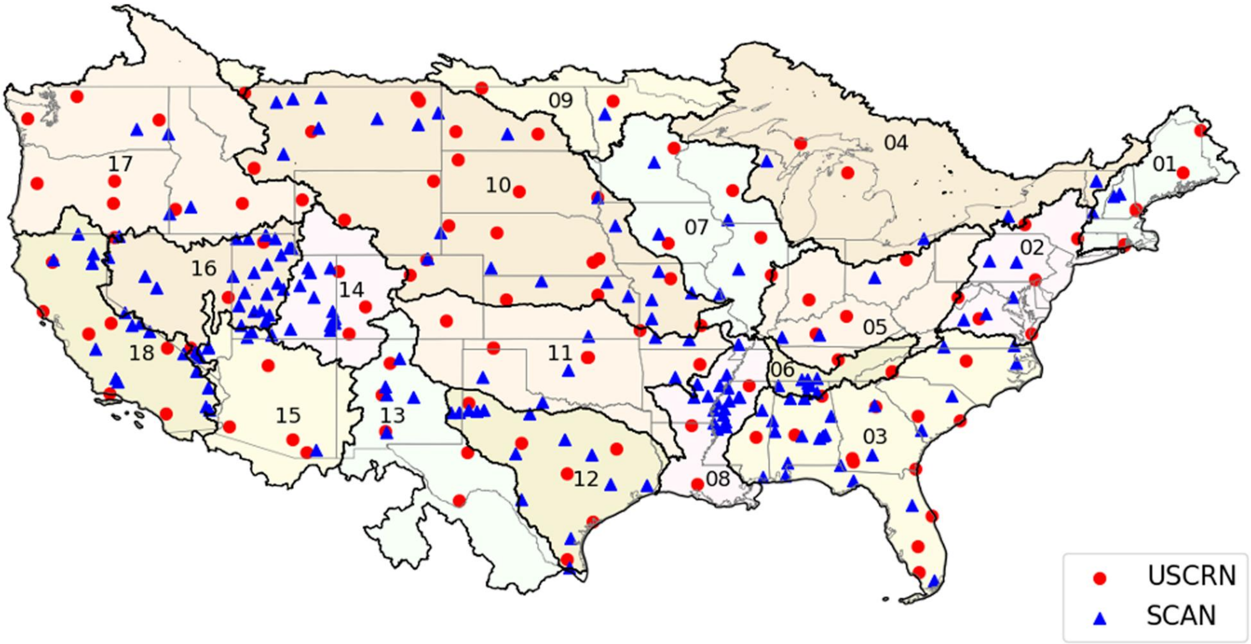

Figure 1. Location of soil moisture stations for the SCAN (blue) and USCRN (red) and code values for the 18 HUC2 regions. Names of HUC2 regions are given in Table 1.

any specific soil-layer scheme, so a linear interpolation technique was used to map the output onto the four Noah model levels (Xia et al., 2015b) and weighted averaging was applied using those four levels to generate soil moisture values for the 0–100 cm and 0–200 cm layers. VIC soil moisture was provided at three levels: a top layer defined at 0–10 cm, and two further layers whose depths vary spatially depending on vegetation type and root distribution, as well as on the 0–100 cm level which was used for most of the paper's analysis.

2.2. Soil Moisture Observations

SM observations from over 200 stations from the Soil Moisture Climate Analysis Network (SCAN; Schaefer et al., 2007) and the U.S. Climate Reference Network (USCRN, Bell et al., 2013), whose locations are shown in Figure 1, were acquired through the National Soil Moisture Network website (NSMN, www.nationalsoilmoisture.com, Quiring et al., 2016). Both networks are national with stations observing SM and atmospheric climate parameters (air temperature, relative humidity, precipitation, etc.). Both networks derive SM measurements from the Stevens Water Hydra probe dielectric reflectometers at depths of 5, 10, 20, 50 and 100 cm. The earliest data from SCAN used in this study are from 1998, while the USCRN SM probes were established starting in 2008.

All SM network data were averaged to daily mean values, with additional quality controls applied uniformly to both networks. Quiring et al. (2016) describe four tests, including a range test that limit SM from 0.0 to 0.6 cm^3 cm^{-3}, a streak test that removes invariant data occurring over periods exceeding 10 days, and two variability tests that remove data exceeding 3 standard deviations from the climatology and sudden magnitude changes. These controls primarily removed erroneously extreme values and artificial jumps (Quiring et al., 2016; Xia et al., 2015b). A weighted average of the 0–100 cm was computed assuming each depth was the center of a layer with boundaries equidistant between adjacent levels.

The in situ data sets have some values missing, and this is more of an issue in the deeper soil layers. For the USCRN network, the highest number of valid observations occurs at 10 cm and the percentage of observations relative to 10 cm at the other levels is 99.0% (5 cm), 81.0% (20 cm), 77.0% (50 cm), and 72.2% (100 cm). For the SCAN network, the level with the highest number of valid observations occurs at 20 cm. The percentage of observations relative to 20 cm was 99.7% (5 cm), 99.5% (10 cm), 97.9% (50 cm), and 89.1% (100 cm). 0–100 cm layer values are only computed when valid data existed at all 5 levels for a given day.

Table 1
HUC2 Codes and Associated Region Names and Abbreviations

HUC2 number	HUC2 region name	HUC2 name abbreviation
01	New England	N_ENG
02	Mid-Atlantic	M_ATL
03	South Atlantic-Gulf	S_ATL
04	Great Lakes	G_Lakes
05	Ohio	OH
06	Tennessee	TN
07	Upper Mississippi	U_MS
08	Lower Mississippi	L_MS
09	Souris-Red_Rainy	SOU
10	Missouri	MO
11	Arkansas-White-Red	ARK
12	Texas-Gulf	TX
13	Rio Grande	RIO
14	Upper Colorado	U_CO
15	Lower Colorado	L_CO
16	Great Basin	G_Basin
17	Pacific Northwest	P_NW
18	California	CA

2.3. Streamflow Gauges

The Hydro-Climatic Data Network (HCDN-2009), a subset of the Geospatial Attributes of Gages for Evaluating Streamflow, Version II (GAGES II) is a USGS-maintained network of 743 stream gauges (Lins, 2012). Located on streams that represent "natural" flow (i.e., flow that is unimpaired by human activity), these gauges have long, stable periods of record suitable for hydro-climatic studies. The sites must fulfill the criteria outlined in Lins (2012).

For the streamflow evaluation, only HCDN-2009 stations in the contiguous US that had a corresponding stream reach in the NWM were analyzed. Because some skill metrics used can be large for minor errors for locations with small streamflow, we also eliminated gauges with 28-day average, 30th percentile flow below 1 cf s^{-1} (0.028 m^3 s^{-1}). This reduces the number of streamflow stations to 507. Only the years from 1993 to 2018 were used from these stations to overlap with the NWM 2.0 retrospective time coverage.

2.4. U.S. Drought Monitor (USDM)

Rasterized grids of USDM drought categories were used as an additional dataset for SM comparison. The USDM categories were determined based on a review of the hazards literature at the time of their development (i.e., 1998–1999). These categories use a ranking percentiles methodology by USDM authors based on the balance of available evidence (e.g., numerical model guidance, in situ data and local on-the-ground reports) using the "convergence of evidence" approach (Svoboda et al., 2002). The weekly maps combine both objective data (which comprises the majority of the analysis) and subjective feedback into the final product. The USDM drought categories are: D0 (abnormally dry), D1 (moderate drought), D2 (severe drought), D3 (extreme drought), and D4 (exceptional drought), which correspond to below 30th, 20th, 10th, 5th, and 2nd percentile SM, respectively.

2.5. Calculation of Percentiles

2.5.1. Soil Moisture

Two climatologies were used to determine SM percentiles from model and observational data. The first climatology spanned the entire 26-year NWM retrospective period (1993–2018) and was used for model-to-model comparisons. For these analyses, an empirically derived probability density function (ePDF) was estimated from the model output for each grid point, soil level, and each day of the year using a 15-day centered aggregate of hourly volumetric SM values sampled every three hours. Data from February 29s were not included in the ePDF statistics. At each grid point we calculated the 2nd, 5th, 10th, 20th, and 30th percentile values, which correspond to the percentiles defining drought categories in the USDM objective blend.

For comparison of modeled SM to observations, a second climatology of model data was derived for the length of station data record (which varies by station). Since NSMN data were daily averages, both observations and model ePDFs were constructed using daily means. Station data were collocated with the nearest model grid cell. Although the comparison of model grid cells with in-situ observations at points is known to suffer from representativeness issues (Vergopolan et al., 2020; Xia et al., 2015a), most results and salient conclusions were aggregated to larger areas (i.e., level-2 Hydrological Unit Codes, or HUC2s) to alleviate this concern (Figure 1). Model data were masked to remove any periods with missing observations. Frozen soil conditions were eliminated from both model and station data when NLDAS-2-Noah daily mean soil temperatures at 0–10 cm were <0°C (NWM soil temperatures are not presently archived). This impacted only SCAN stations since USCRN data were already quality controlled to address frozen soils. Of the 186 SCAN stations used in this study, 131 stations had less than 20% of daily data with frozen soil conditions, and 84 had no frozen soils. HUC2 regions 04, 07, 09, 10, 14, and 16 of the northern and intermountain western regions were the most affected. The most data removal occurred in the mountains of Utah (where 43.2% of the data were removed), and in stations near the Canadian border east of the Rocky Mountains. The frozen soil methods used here are conservative in the sense that they err

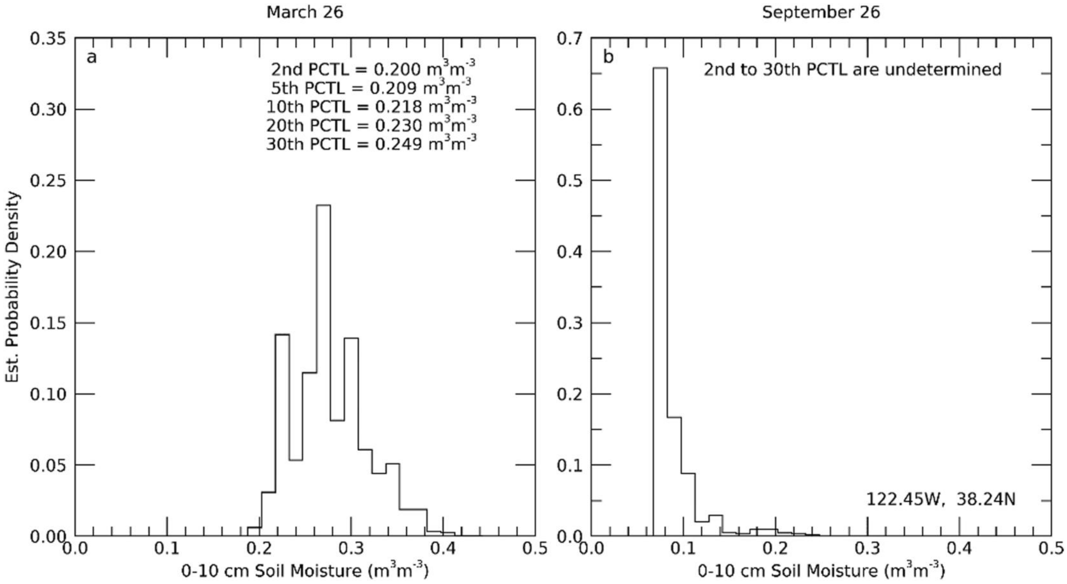

Figure 2. Examples of probability densities for a grid cell near Sonoma, California for March 26th (a) and for September 26th (b). This location has a Mediterranean climate with a winter wet season and a summer dry season. Percentile values for this grid cell are shown for (a), but cannot be determined for (b) since probability density of left-most soil moisture bin is 0.657. Note difference in *y*-axis scale between (a) and (b).

on the side of removing more data than less, because the Noah soil temperatures were generally colder than the SCAN soil temperatures at most locations.

The lowest percentile values (i.e., 2nd and 5th percentiles) for some grid cells could not be determined due to the skewed shape of their ePDFs and limited record length. Figure 2 gives examples of two ePDFs for a grid cell on selected days during the wet and dry seasons near Sonoma, California. Panel (a) shows a typical ePDF with the computed percentiles for the five percentile levels. Panel (b) provides an example of how the skewed ePDF prevents determination of these percentiles, since 66% of the data on this day and location occupied the left-most (driest) bin of the distribution. These grid cells were flagged and assigned the lowest possible percentile that could be determined. Some of these irregular ePDFs occurred in desert regions where the climatological SM varies across a small range of values during the dry season (when rainfall is virtually nonexistent). This issue occurred in all the models and at all levels in the models for specific locations, and was most prevalent in the near-surface levels. More details about this issue, including percent of grids impacted and seasonality, are provided in the SI.

2.5.2. Streamflow

Daily mean streamflow data available at 2.7 million stream reaches from the raw hourly NWM v2.0 26-year (1993–2018) retrospective simulation were used to develop a cumulative distribution function for each calendar day from the daily means. The 2nd, 5th, 10th, 20th, and 30th low-flow percentile classes were derived from this climatology for 28-day durations. The USGS National Water Information System (NWIS) method of diagnosing hydrologic drought was used to compare each day's average streamflow value to the pre-defined percentile classes for each duration at each location (Helsel et al., 2020).

2.6. Evaluation Statistics

Several statistics were used to evaluate model performance of SM (denoted Θ in Equations 2–4), using station data as ground truth, and NLDAS-2 LSMs as benchmarks of skill. Let p50 represent the 50th percentile for a given ePDF, and *i* represent day of year (note February 29th is ignored). Bias was defined from the ePDF data as:

$$\mathrm{BIAS} = \frac{1}{365}\sum_{i=1}^{365} p50_i^M - p50_i^O \tag{1}$$

where M represents model and O represents observations. Bias has the volumetric units of the SM values from which it is calculated.

Anomaly correlation was determined by first removing the SM annual cycle. Removing the annual cycle rather than the climatological annual mean eliminated correlation arising from the seasonal cycle and thereby better assessed correlation of higher frequency variability between model and observations. The climatology of the seasonal cycle (denoted by overbar) was described by:

$$\overline{\theta_i} = \frac{1}{N}\sum_{j=1}^{N}\theta_{ij}, \tag{2}$$

where i represents day of the year, j represents each year of data in the SM time series and N is the total number of years. Using this climatology, the SM anomaly was written as:

$$\theta_{ij}^{A} = \theta_{ij} - \overline{\theta_{i.}} \tag{3}$$

Anomalies were computed for both model $\theta_{ij}^{A,M}$ and station $\theta_{ij}^{A,O}$ data. The anomaly correlation coefficient (ACC) was computed using a standard Pearson correlation coefficient as:

$$\text{ACC} = \frac{\sum_{j=1}^{N}\sum\limits_{i=1}^{365}\theta_{ij}^{A,M}\theta_{ij}^{A,O}}{\sqrt{\sum_{j=1}^{N}\sum\limits_{i=1}^{365}\left(\theta_{ij}^{A,M}\right)^2}\sqrt{\sum_{j=1}^{N}\sum\limits_{i=1}^{365}\left(\theta_{ij}^{A,O}\right)^2}}. \tag{4}$$

The possible range for ACC is −1 to 1, with 1 being perfect correlation between the two data sets and −1 being perfectly anti-correlated. ACC has been used in past evaluations of SM data sets for drought applications (e.g., Xia et al., 2015b).

The contingency table-based evaluation metric critical success index (CSI) was used to assess model skill in detecting drought conditions; contingency table-based metrics, similar to CSI, have also been used previously for evaluation of SM data sets for drought applications (e.g., Probability of Detection, Ford & Quiring, 2019). Table S3 in Supporting Information S1 shows a contingency table defining model and observational drought conditions when SM is below the 20th percentile. The 20th percentile was used in our analyses rather than more severe/extreme percentiles due to the difficulty of deriving reliable ePDFs with data records shorter than 26 years. CSI was defined as:

$$\text{CSI} = \text{Hit}/(\text{Hit} + \text{Miss} + \text{False Positive}). \tag{5}$$

The possible range for CSI is 0–1 with perfect prediction of drought by the model resulting in CSI = 1. CSI is a commonly used evaluation statistic for weather forecasting, albeit with some known weaknesses, in particular that it is automatically lower for rare events (Schaefer, 1990). Because CSI decreases as events become more rare (e.g., for lower percentiles), we focused on the CSI of the NWM relative to those of the NLDAS-2 models.

The streamflow percentiles at NWIS gauge locations were evaluated by calculating the normalized root mean square error (nRMSE) and the Nash-Sutcliffe efficiency (NSE) for the same USDM-relevant (i.e., 2nd, 5th, 10th, 20th, and 30th) percentile classes:

$$\text{nRMSEp} = \frac{\sqrt{\frac{\sum_1^n(Qm,p-Qo,p)^2}{n}}}{\overline{Qm,p}} \tag{6}$$

$$\text{NSEp} = 1 - \frac{\sum_1^n(Qo,p - Qm,p)^2}{\sum_1^n\left(Qo,p - \overline{Qo,p}\right)^2}, \tag{7}$$

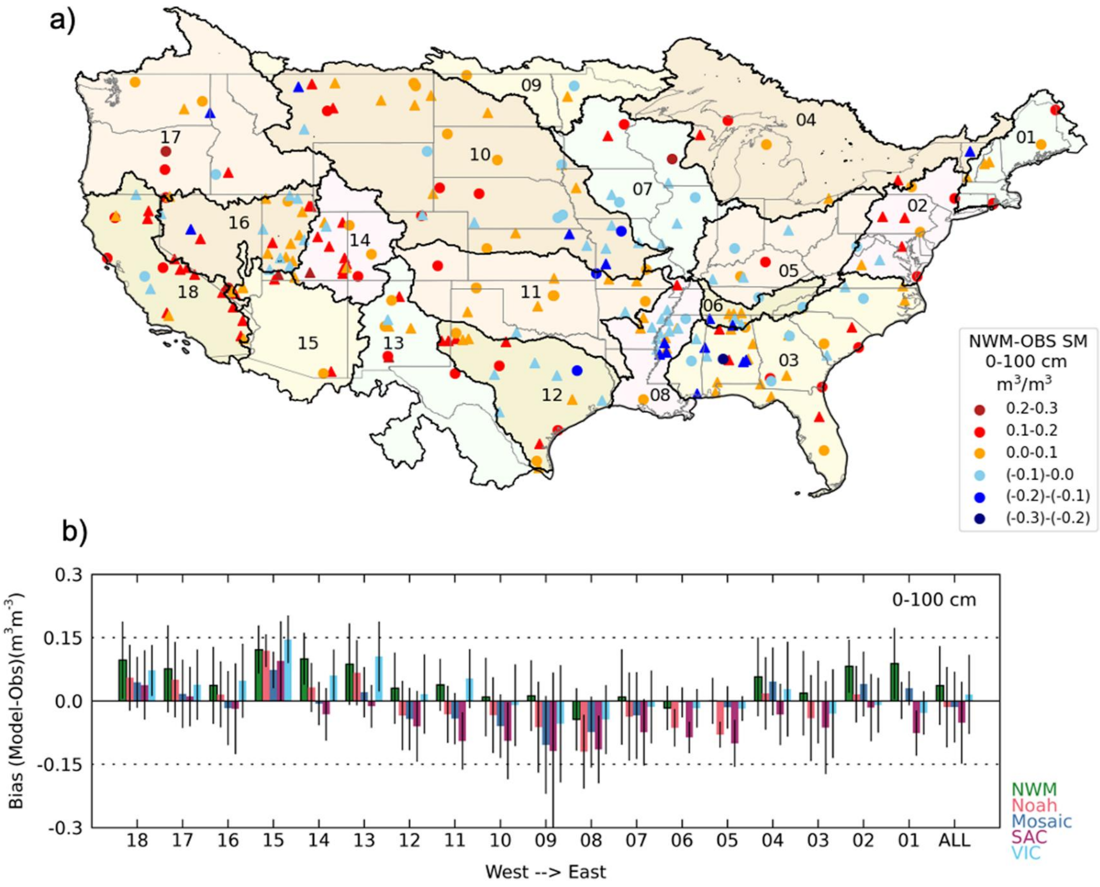

Figure 3. Soil moisture bias map (a) and bar plots (b) for NWM and four NLDAS-2 models compared with USCRN and SCAN observation networks. Map gives 0–100 cm SM bias between the NWM and observations. Bar plots show mean bias for each of the 18 HUC2 regions, with error bars to indicate the standard deviation across the stations. *X*-axis labels denote each HUC2 region shown in Figure 1b. ALL indicates mean bias for the CONUS. Triangles are SCAN and circles are USCRN. Background colors on map denote HUC2 regions.

where Qm,p and Qo,p are model and observed streamflow for each percentile class. As a normalized, guaranteed-positive metric, nRMSE ranges from 0 to infinity (although we only show values up to 2.25), with 0.5 indicating a 50% RMSE, 1 a 100% RMSE, etc. NSE ranges from -infinity to 1, with one indicating best performance and scores greater than 0 indicating meaningful/useful representation of variability (in this case the annual variability of low flows). In addition, the CSI of daily low-flow events (calculated from Equation 5) was used as a third metric for streamflow evaluation, providing a common metric for comparison to the SM results.

3. Results

3.1. Summary Statistics

3.1.1. Soil Moisture

We begin our evaluation with SM summary statistics from the NWM compared with NLDAS-2 LSMs as benchmarks for NWM performance. Here we show results for the 0–100 cm depth-average SM, due to the relevance of deeper soils for drought monitoring; results for individual levels are shown in the supplemental information (SI). Aggregating across CONUS, NWM 0–100 cm volumetric SM is positively biased (i.e., wetter) compared to the in-situ observations (Figure 3b, rightmost cluster of bars labeled “ALL”). The standard deviation across sites is larger in magnitude than the average bias, which indicates large site-to-site variation in bias, including many locations with negative bias. This wide range in bias is apparent on the map of NWM bias

(Figure 3a), where several regions contain adjacent in-situ sites with biases of opposite sign. Comparison of NWM's SM bias with its root mean squared error (RMSE; Figure S2 in Supporting Information S1) reveals that for most locations the bias and RMSE are comparable, suggesting the bias is systematic. The NWM's positive bias is larger in near-surface SM than in deeper soil levels (Figure S1 in Supporting Information S1). The positive bias is a result of the calibration strategy of the NWM, which limits drainage from the bottom of the soil column to reduce high biases in streamflow.

Compared to most NLDAS-2 LSMs, the NWM is more positively biased (i.e., wetter)—the exception being NLDAS-2-VIC (VIC 0–100 cm SM is wetter than the NWM in a few HUC2s, and VIC 0–10 cm SM is wetter than the NWM in most HUC2s, ure S1 in Supporting Information S1). NWM's RMSE is larger than NLDAS-2 LSM RMSEs in several HUC-2s, although only slightly larger compared to the site-to-site variations (Figure S2 in Supporting Information S1). NLDAS-2-Noah and NLDAS-2-Mosaic have the smallest biases but still rather large site-to-site standard deviations. All models become drier compared to observations in deeper levels (SI): in the NWM this results in smaller positive biases, whereas in the NLDAS-2 models this results in biases that either shift from positive to negative, or become more negative, in deeper levels.

The NWM and NLDAS-2 models follow a similar pattern in the geographical distribution of biases when aggregated across each of the HUC2s: relative to observations, model biases are (comparably) more positive in the northeastern (HUC2 01–02 and 04) and western (HUC2 13–18) HUC2s, and more negative in the southern and central HUC2s (HUC2 03 and 05–12). Note that, to some degree, the variation across HUC2s could be influenced by both the density of SM observations in the individual HUCs and by the quality of the precipitation forcing in NLDAS-2, which is higher in the east than the west (Mo et al., 2012). In the NWM, this tendency results in larger positive biases along the coastal HUC2s and near-zero or small negative biases in the central US HUC2s. As was the case for the CONUS aggregate, across the HUC2s, the standard deviation of biases is large, often larger than the mean bias.

Despite these biases, SM information from the NWM may still be applicable for drought monitoring if it reasonably represents drier-than-usual conditions (i.e., anomalously dry SM values). We evaluate this ability to represent dry anomalies using the ACC and CSI.

Similar to SM biases, ACCs vary widely across CONUS (Figure 4), with many examples of adjacent locations exhibiting markedly different performance (Figure 4a). ACCs decrease with depth, with CONUS-wide NWM ACC of approximately 0.6 in the 0–10 cm depth, and ~0.45 in the 40–100 cm depth (SI). Unlike bias, where a decrease with depth sometimes improves performance, reductions in ACC unilaterally indicate reduced skill. The higher ACCs in the upper soil depths could stem from the increased variability of near-surface SM and stronger relationship to precipitation forcing. ACCs vary widely by HUC2, although the lowest ACCs tend to be in the western HUC2s. CONUS-aggregate ACCs (Figure 4b, rightmost bars) in the NWM are generally comparable to those of the NLDAS-2 models; in some HUC2s the NWM has slightly higher ACCs than the four NLDAS-2 models (e.g., 02 – Mid Atlantic), but the differences between the LSMs are generally rather small compared to the site-to-site variation of ACC.

NWM CSIs also vary widely across CONUS (Figure 5a). The western US has extremely low 0–100 cm NWM CSIs (typically below 20%), whereas in the central and eastern US CSIs typically range from 20% to 50%. Similar to the results for ACC, the NWM's CSIs are comparable to the CSIs of the NLDAS-2 models (Figure 5b), and decrease in deeper layers (SI). Notably, for all three of these metrics (bias, ACC, and CSI), the NWM's scores are similar to NLDAS-2 LSM scores in most HUC2s.

As a way to summarize the results from the three different performance metrics for each HUC2, Table 2 ranks each HUC2 for each metric, with the best-scoring HUC2 in position 1 and worst in position 18, and then presents an overall ranking based on the sum of the ranks for the three metrics (the lowest scoring is the "best"). HUC2s in the north-central US (05 – Ohio, 07 – Upper Mississippi, and 06 – Tennessee) rank the highest using this method and combination of skill scores, whereas the southwestern US HUC2s (including 14 and 15 – the upper and lower Colorado, 13 – Rio Grande, and 18 – California) rank the lowest. Tables S4–S7 in Supporting Information S1 present the same information for NLDAS-2-Noah, NLDAS-2-Mosaic, NLDAS-2-SAC, and NLDAS-2-VIC, and Table S8 in Supporting Information S1 compares the rank-performance of the four models. The HUC2 regions' rankings are somewhat consistent across the four LSMs, although there are some exceptions (most notably NLDAS-2-SAC's rankings).

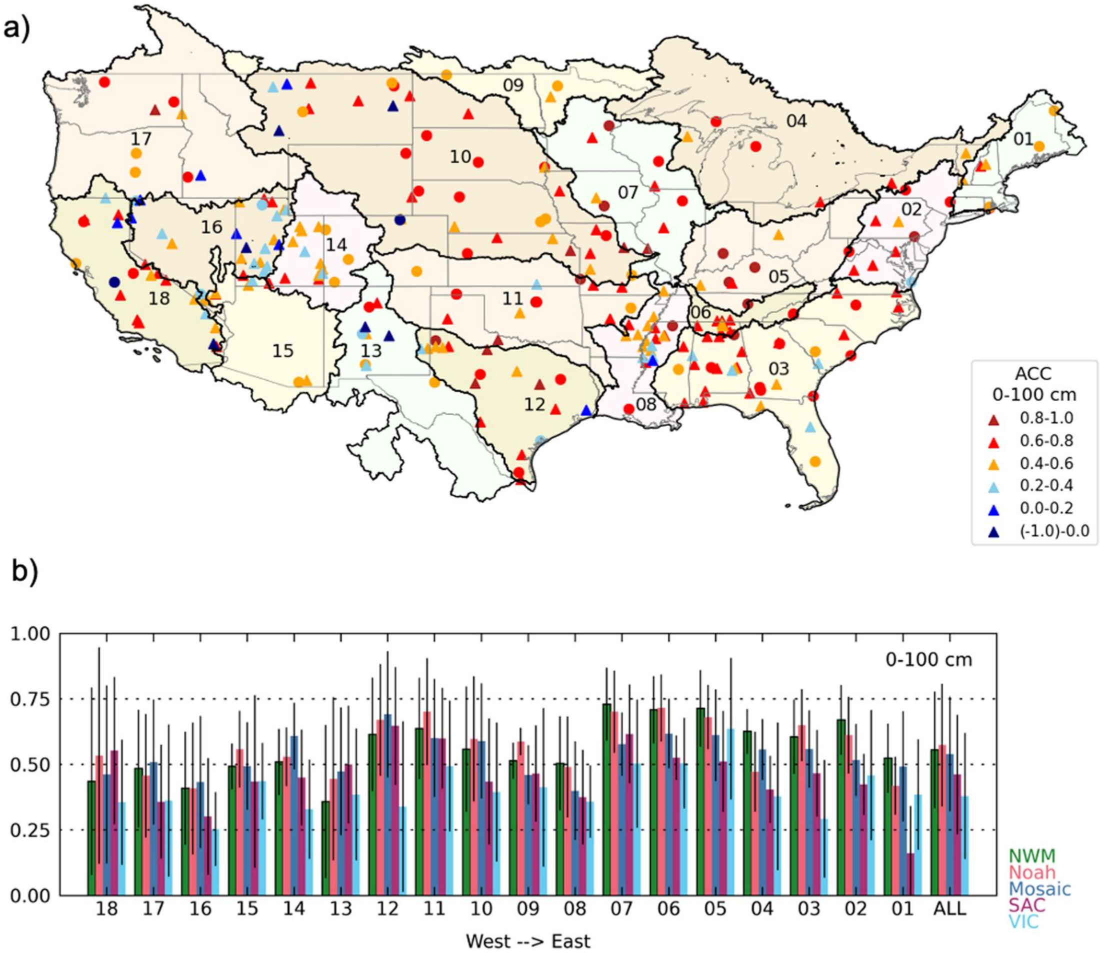

Figure 4. Same as Figure 3 but for ACC. Displayed range is 0–1 since LSM SM is never anticorrelated with observed SM.

3.1.2. Streamflow

Currently, USGS weekly observation-based WaterWatch streamflow percentiles are used to inform the USDM. However, the NWM's network of 2.7 million stream reaches offers unprecedented additional insight into hydrologic drought conditions across CONUS. Thus, this section evaluates NWM retrospective streamflow conditions during low-flow conditions.

Normalized RMSEs (nRMSE) and Nash Sutcliffe Efficiency scores (NSEs) of the NWM 20th-percentile flows (Figures 6 and 7) show how well the retrospective NWM represents the volume of these low flows in a climatological sense. nRMSE indicates whether the volume of flow is percentage-wise close to that observed. NSE indicates whether or not year-to-year variations in the low-flow volumes are well captured. The spatial distribution of both skill metrics is similar: the NWM streamflow is best represented in the Pacific Northwest and along the Eastern Seaboard, with nRMSE generally below 0.5 and NSEs generally above 0.5 in these regions. nRMSEs are slightly higher (i.e., worse) in the north central and southwestern US; in particular the Souris-Red-Rainy HUC2 (09) has extremely high nRMSEs. NSEs are reasonably high (with across-station median NSE > 0.5) for all HUC2s except the Lower Colorado (15), and the four central US HUC2s (12, 11, 10, 09).

CSIs of NWM low-flow streamflows (Figure 8) similarly are highest along the Eastern Seaboard and Pacific Northwest, with values above 45% at most locations in these regions. Streamflow CSIs are lowest in the western Missouri River, Souris-Red-Rainy, and Great Basin HUC2s (10, 09, and 16, respectively). CSIs are lower for the more extreme percentiles (not shown), perhaps to some extent due to the formulation of the CSI score (Schaefer, 1990).

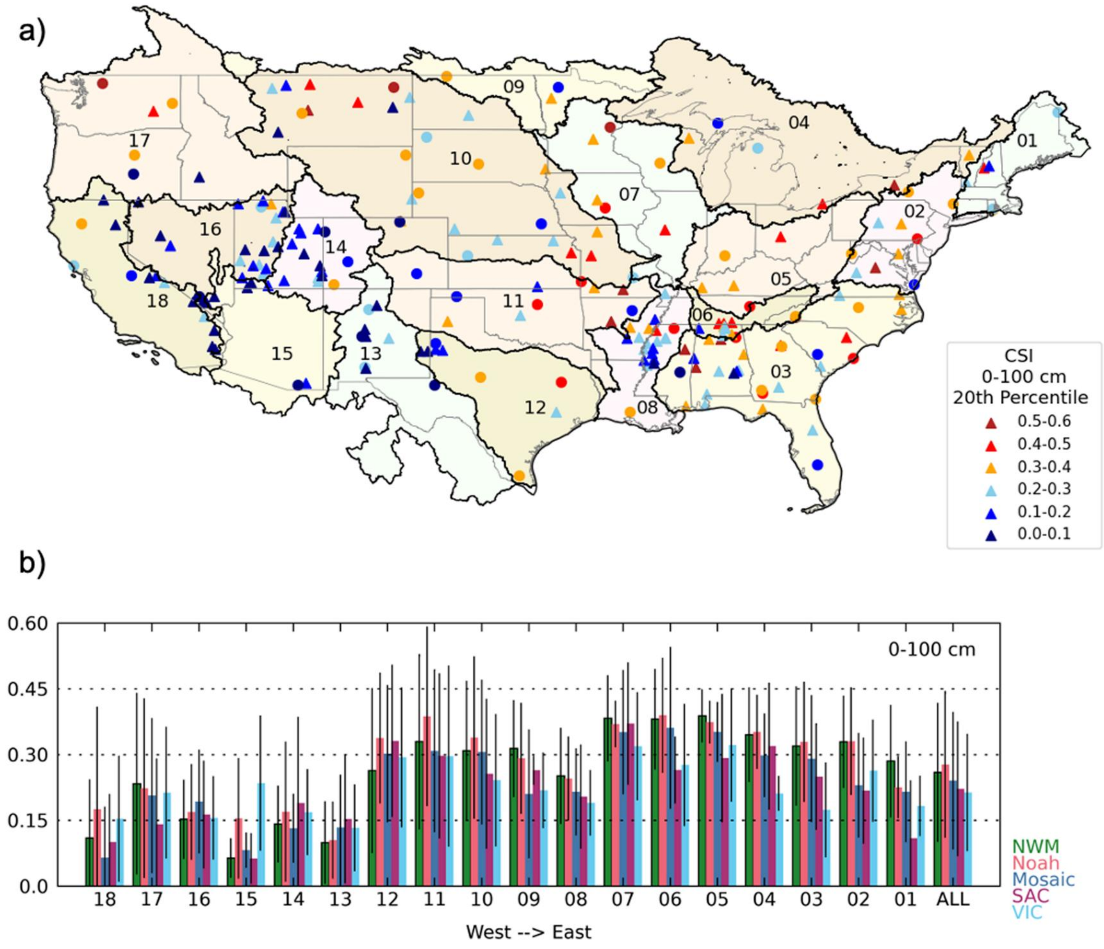

Figure 5. Same as Figure 3 only for critical success index (CSI, %) for SM below the 20th percentile.

3.2. Comparison of NWM SM to the USDM

The USDM represents the most widely accepted estimates of US drought conditions. For this reason, we provide a brief comparison of NWM SM percentiles to the USDM categories. This comparison offers an additional, stakeholder-relevant perspective on the ability of NWM SM anomalies to capture drought conditions beyond that obtained from the comparison to the in-situ and NLDAS-2 SM.

We start by noting some of the limitations of using the USDM as a comparison dataset to our SM- and streamflow-based drought monitoring products. First, the USDM is not completely independent from other data sets presented here. For example, since ~2009 (when NLDAS Phase 1 was made available to USDM authors), its authors have considered NLDAS (first Phase 1, and starting in ~2015, version 2) LSM SM as a piece of evidence in assigning USDM drought categories. Second, the USDM represents more than agricultural drought (i.e., anomalously low SM). Third, the NWM was only available for a short period of record relative to the reference period for some of the data sets considered by USDM authors (e.g., NOAA's Climate Prediction Center SM data uses a ~100-year reference period for the derivation of its percentiles). This mismatch in period of record would disproportionately impact regions for which the shorter, more-recent record was anomalous climatologically (e.g., the southwestern US was in drought conditions for much of the 26-year NWM retrospective period; Williams et al., 2020). These points are discussed in Wang et al. (2022). For these reasons, we restrict our USDM analysis to qualitative comparisons.

Figure 9 illustrates the percent of CONUS under each USDM drought or SM category from 2000 to 2018 in the USDM (top) and NWM (bottom). The USDM generally shows more land area in each of the four drought

Table 2
NWM 0–100 cm Bias, ACC, and CSI Values and Relative Rank for Each HUC2 Region

HUC-2 number	HUC-2 name	Bias m^3/m^3	Bias rank	Acorr	Acorr rank	CSI	CSI rank	Overall rank
05	OH	0.000	1	0.714	2	0.388	1	1
07	U_MS	0.010	3	0.730	1	0.383	2	2
06	TN	−0.016	5	0.709	3	0.381	3	3
11	ARK	0.038	9	0.637	5	0.330	5	4
10	MO	0.009	2	0.558	9	0.309	9	5
04	G_Lakes	0.057	11	0.626	6	0.345	4	6
03	S_ATL	0.018	6	0.605	8	0.319	7	7
09	SOU	0.013	4	0.514	11	0.314	8	8
02	M_ATL	0.082	13	0.670	4	0.329	6	9
12	TX	0.030	7	0.615	7	0.264	11	10
08	L_MS	−0.044	10	0.504	13	0.251	12	11
01	N_ENG	0.089	15	0.524	10	0.285	10	12
16	G_Basin	0.037	8	0.409	17	0.152	14	13
17	P_NW	0.076	12	0.484	15	0.234	13	14
14	U_CO	0.100	17	0.510	12	0.141	15	15
18	CA	0.097	16	0.436	16	0.110	16	16
13	RIO	0.087	14	0.358	18	0.990	17	17
15	L_CO	0.122	18	0.493	14	0.640	18	18

Note. Overall rank determined using ranks from all three parameters. Ranks ordered with lowest values having best score.

categories than the corresponding NWM SM percentile (e.g., the percent of grid cells rated as D0 ranges from ∼20% to 80%, whereas grid cells with NWM 30th percentile SM values range from ∼10% to 50%). The USDM also varies more smoothly than the NWM. Despite these differences, broad similarities are clear: when there are large shifts in the area covered by each category, both data sets tend to shift in the same direction, indicating that the NWM appears to capture the large-scale trends in drought revealed in the USDM time series.

3.3. 2012 Great Plains Drought Case Study

Here, we document NWM retrospective SM conditions during the 2012 Great Plains drought. The NOAA Drought Task Force formalized a protocol for evaluating new products for drought monitoring and prediction (Wood et al., 2015). This protocol suggested investigation of four historic case studies of extreme drought (between black lines in Figure 9), both due to their historic impact and to facilitate comparison to previous research. For brevity, we provide detailed results and discussion in the main text for only the 2012 Great Plains drought, however the SI includes figures and limited discussion regarding NWM SM conditions during the other three historical droughts.

This drought started in May of 2012 (Figures 10a and 10c and Figure 11) and intensified rapidly from mid-June through July. The intensification was due to a failure of summertime rainfall and heat waves in the Central Great Plains (Hoerling et al., 2014). This drought also proved difficult to predict – the 17 May 2012 outlook for June-August 2012 predicted near-normal precipitation (see Figure 9 in Hoerling et al., 2014) – and resulted in large losses to corn yields. Near its peak extent, most of Oklahoma, Kansas, and Nebraska were classified by the USDM as in either extreme (D3) or exceptional (D4) drought (Figure 10b); nearly half the US was in severe drought by the end of summer 2012 (Hoerling et al., 2014).

On 1 May 2012 the NWM SM percentiles (Figure 10c) indicated normal conditions through central Nebraska and across much of Kansas, Iowa, and Missouri, and drier-than-normal conditions across much of Colorado, eastern Nebraska, and southeastern Missouri. These SM conditions largely agree with the USDM map for that week (Figure 10a), albeit with more heterogeneous conditions in NWM SM and normal NWM SM in some locations

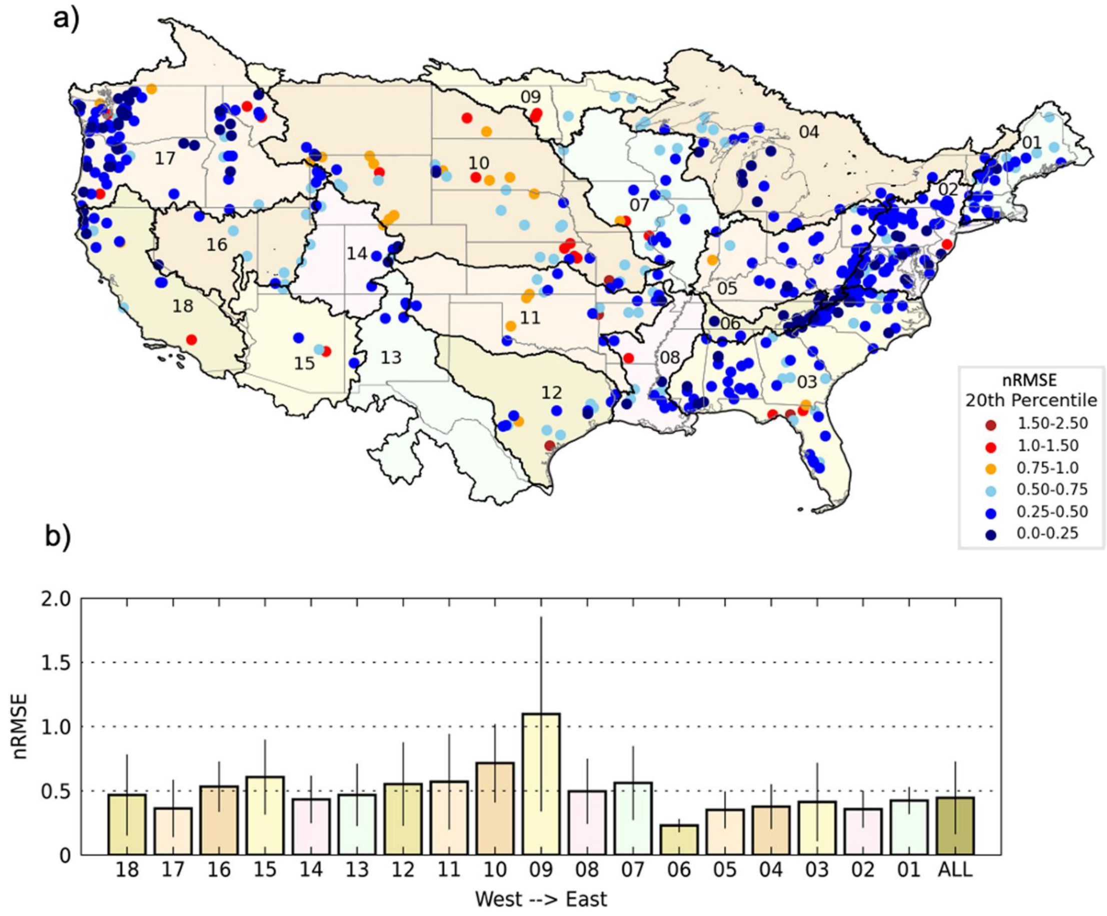

Figure 6. (a) Normalized RMSE (nRMSE) for NWM 20th percentile streamflow. (b) Corresponding nRMSE mean and standard deviation for each HUC2 region. Colors on bar plot are associated with the HUC region on the map.

(e.g., northwestern Iowa and the Oklahoma panhandle), where the USDM indicated drought conditions. The increased heterogeneity in the NWM SM compared to the USDM is expected given the unfiltered NWM high-resolution model grid and USDM's multi-faceted, human-developed maps. By August, NWM SMs were below 5th percentile for most of the region east of the Rocky Mountains, and below the 2nd percentile for much of Kansas and Missouri. Through most of the region these SM percentiles agree well with that week's USDM drought conditions for the region (Figure 10b). Some differences are visible, though. The NWM is more heterogeneous than the USDM; however, we recall that the USDM is a composite product and reflects the data sets considered each week (i.e., if the NWM had been considered the map might look different). In addition, western Kansas and Nebraska have localized regions that the USDM classifies as D3 and D4 but where the NWM SM is less extreme. Satellite data assimilation has been used in other studies to mitigate model deficiencies (e.g., Chen et al., 2021), and could plausibly improve NWM representation of extreme dry conditions. Finally, NWM SM percentiles in the intermountain west are more heterogeneous than the USDM drought categories and also generally less extreme. The less-extreme indications in the NWM SM are possibly partially due to its limited period of record (Wang et al., 2022).

This pattern of reduced areal extent and less-extreme dry conditions in the NWM SM relative to the USDM is also visible in a time series of areal extent for each drought category and SM percentile for the region (Figure 11). In addition, the time series reveal that the areal extent of low NWM SM tends to lead the increase in the corresponding USDM category by a few to several weeks, highlighting its potential for drought early warning (e.g., Ford et al., 2015). This advanced lead time is more pronounced in the 0–10 cm SM than the 0–100 cm SM.

To understand the factors contributing to these differences, we next compare the percent of days during May-Dec 2012 with NWM 0–100 cm SMs below the 10th percentile with the same metric from the NLDAS-2 LSMs

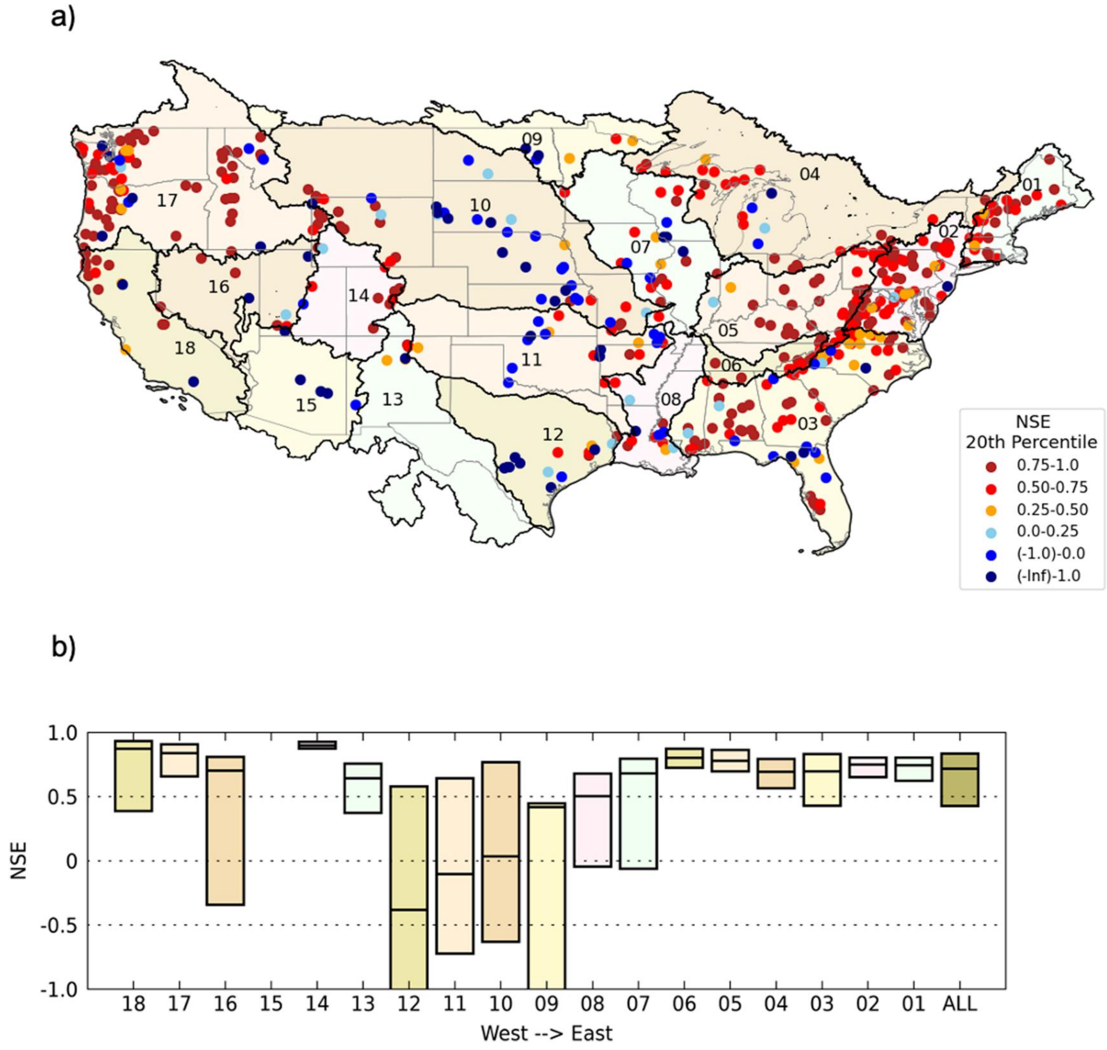

Figure 7. (a) Nash Sutcliffe Efficiency (NSE) for the NWM 20th percentile streamflow climatology. (b) Corresponding NSE median, 25th, and 75th percentiles for each HUC2 region. Note that all values for region 15 are <−1.0. Colors on bar plot are associated with the HUC region on the map.

(Figure 12). We focus this comparison on two regions with somewhat different behaviors in Figures 10b and 10d: southwestern Kansas, where the NWM has SM percentiles above those of the corresponding USDM drought category, and Missouri, where the August NWM SM percentiles are comparable the USDM drought category.

All LSMs have swaths across southwest Kansas with ~20% of days below the 10th percentile SM, likely indicating a response to a rain event, which is confirmed upon inspection of NLDAS-2 precipitation for the time period (not shown). This swath is not apparent in the USDM D2 percentage map (Figure 12f)—in fact, the USDM has nearly 100% weeks of the May-December period in at least D1 in SW Kansas. This longer persistence in the USDM may be explained by the long-term nature of this drought and the different reference periods for data sets considered by the USDM. The difference for this region may partially explain the differences in temporal areal extent visible in Figure 11.

In contrast to this LSM-consistent signal, the pattern in Missouri varies across the LSMs. The NWM SM has persistently low SM in much of Missouri, with a broad swath across the state with ~80% of May-December days having SM below the 10th percentile. NLDAS-2-SAC exhibits a similar pattern. In contrast, NLDAS-2-Noah, NLDAS-2-VIC, and NLDAS-2-Mosaic have far fewer days with below-10th percentile SM, which is more consistent with the USDM; we note here that although the USDM now considers NLDAS-2 SMs in its balance of evidence, in 2012 it used NLDAS phase 1 information. These LSMs use the same meteorological forcing data (although these forcings are downscaled and regridded to drive the 1-km NWM); thus, this difference in drought

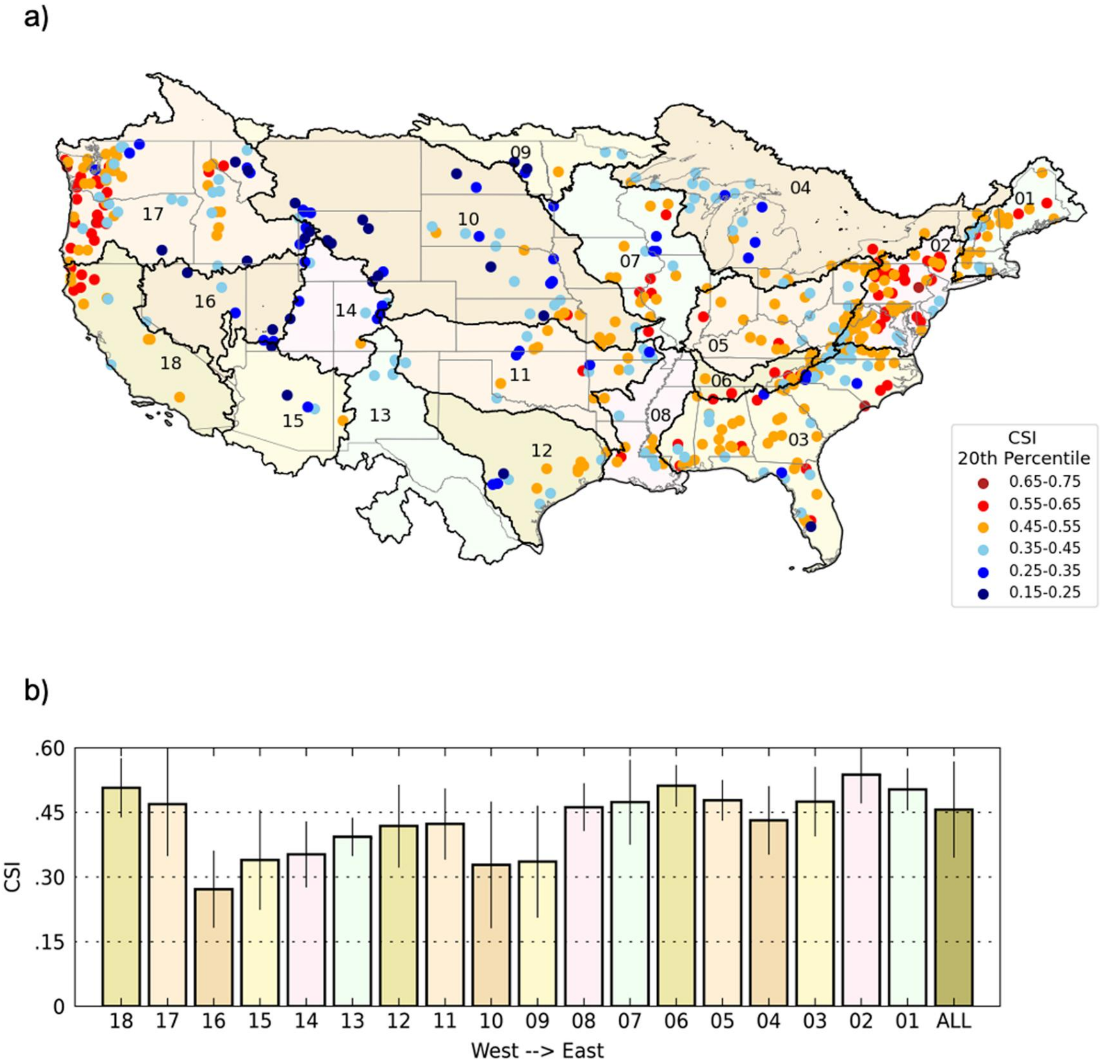

Figure 8. (a) Critical Success Index (CSI) for the daily streamflow below the 20th percentile. (b) Bar plot of mean and standard deviation CSI for each HUC2 region. Colors on bar plot are associated with the HUC region on the map.

duration in the four models is not a direct response to forcing and instead is likely due to differences in model structure or in parameter choices (e.g., groundwater scheme or vegetation parameterization, Wu et al., 2021).

4. Discussion

This section contextualizes and interprets some results from Section 3, briefly discusses the soil-moisture physics of the Noah-MP (in its WRF-Hydro context) and NLDAS-2 LSMs that potentially impact their performance (particularly in arid regions), and then discusses the operational utility and stakeholder relevance given this discussion and noted limitations.

4.1. Connecting the Dots

4.1.1. Interpreting Summary Statistics

Here we examine the summary statistics for NWM SM and streamflow collectively, and discuss their implications. As noted in Section 3, the NWM SM biases are in general larger (i.e., wetter) than the biases of NLDAS-2-Noah, NLDAS-2-Mosaic, and NLDAS-2-SAC. Since the NLDAS-2 LSMs and the NWM use essentially the same meteorological forcings (Section 2.1), this implies the NWM has proportionately less evapotranspiration or runoff than these three NLDAS-2 LSMs. We hypothesize that these positive NWM SM biases may be a result of

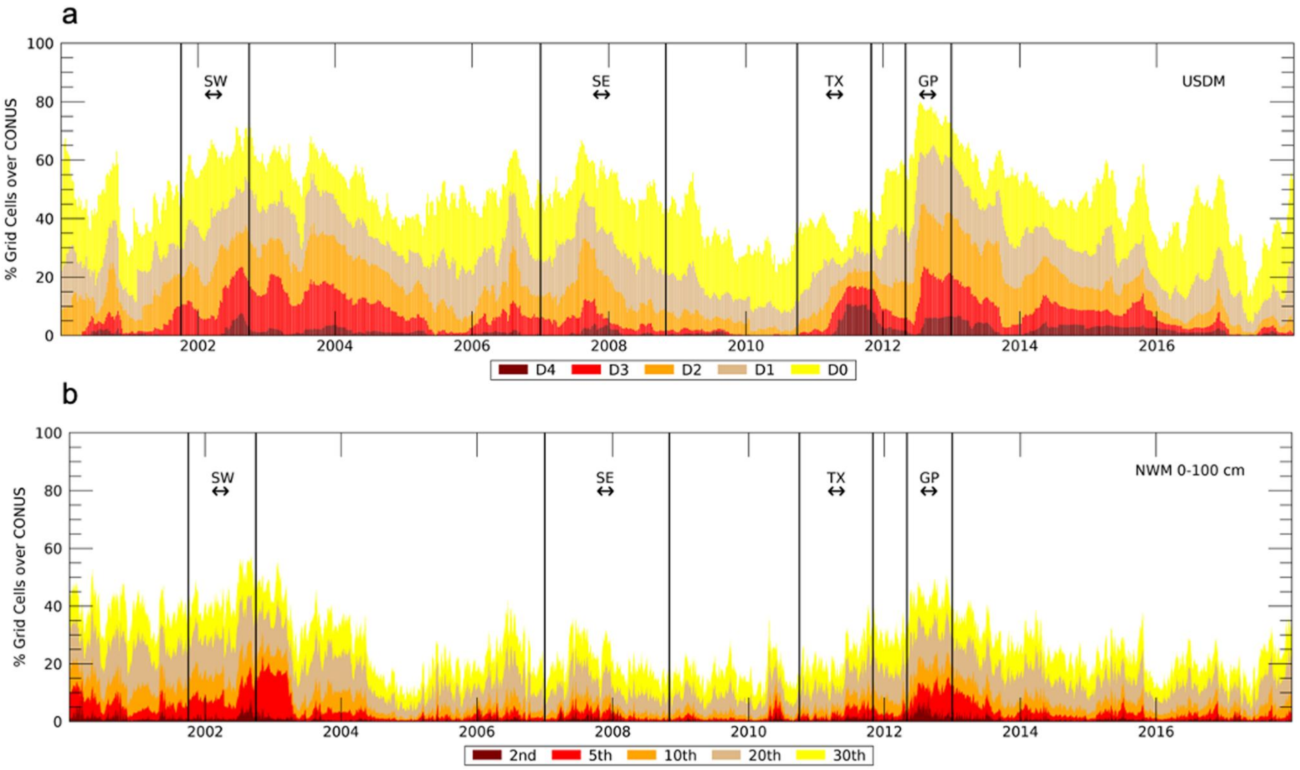

Figure 9. Time series of % grid cells over the CONUS with (a) USDM drought D0–D4 conditions and (b) NWM 0–100 cm soil moisture below 2nd–30th percentile conditions. Time periods surrounding the four major US droughts noted as test cases in Wood et al. (2015) are shown by the vertical black lines: Southwest (SW), Southeast (SE), Texas (TX), and Great Plains (GP). Undefined percentile locations for the NWM are not considered in percentage calculation.

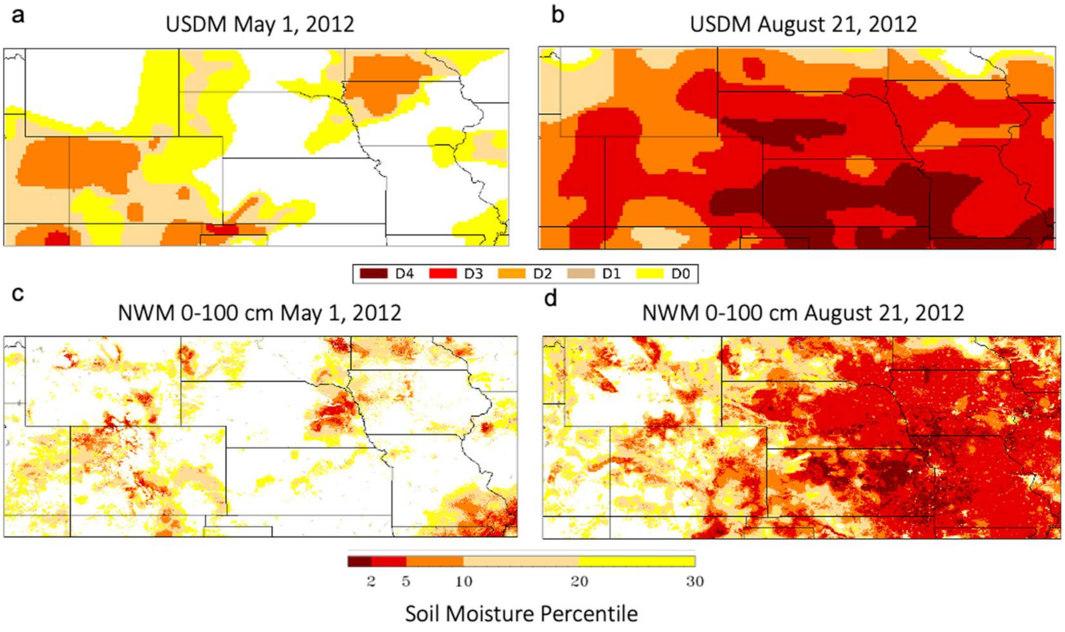

Figure 10. USDM drought analysis and NWM 0–100 cm soil moisture percentiles maps during 2012 Great Plains drought. USDM (a) and NWM soil moisture percentiles (c) on 1 May 2012 give conditions just before rapid onset and intensification, and USDM (b) and NWM soil moisture percentiles (d) on 21 August 2012 are during extreme drought conditions.

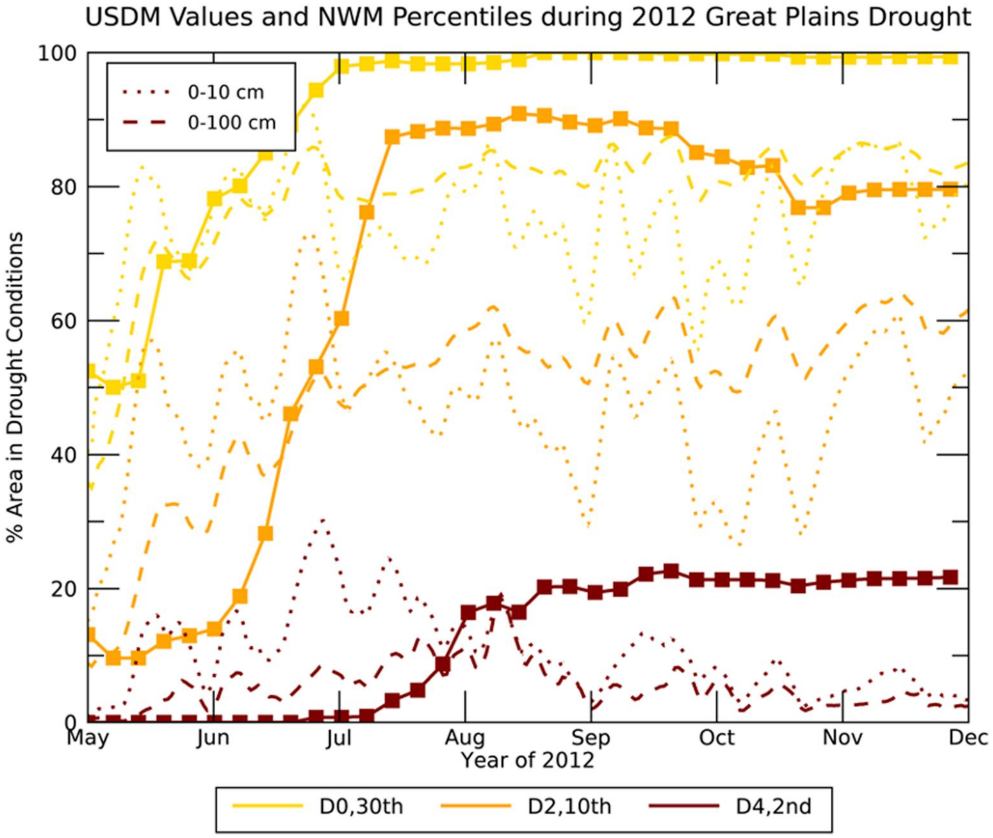

Figure 11. Time series of % area for Great Plains region (112°W–80°W, 36°N–45°N) undergoing drought conditions defined by USDM values in solid curves and NWM 0–10 cm (dotted) and 0–100 cm (dashed) SM percentiles. Colored squares indicate weekly time steps with USDM product.

the NWM calibration methods, which calibrate to match streamflow observed in unimpaired basins across CONUS. This hypothesis is consistent with the generally higher NWM streamflow skill scores relative to its SM. Nevertheless, the NWM SM wet bias does not seem to impact its performance relative to NLDAS-2 for ACC and CSI, both of which are arguably more relevant for drought monitoring than bias (e.g., Ford & Quiring, 2019). Finally, both NWM SM and streamflow have their highest CSIs in the Midwestern and Eastern U.S., suggesting that NWM outputs might have the most value for drought monitoring in those regions. We note that, to some degree, these summary statistics could be impacted by limitations of the in situ data: in addition to heterogeneity in the network density noted in Section 3.1.1, random missing data and sensor measurement error both could impact the in situ data accuracy.

4.1.2. NWM SM and the USDM

Comparisons of NWM SM to the USDM are somewhat difficult to interpret (see Section 3.2). Compared to the USDM (Figures 9 and 11), the NWM SM anomalies capture the overall trends represented in the USDM but display generally less extensive drought conditions. For the 2012 Great Plains drought, the NWM SM tends to precede the USDM into deepening drought; this is also true for some drought categories in the 2011 Texas and 2008 Georgia droughts, but not for the 2002 Western US drought (Figures S8–S10 in Supporting Information S1). The NWM SM tends to recover from dry conditions more quickly than the USDM reduces drought categories. This tendency is not ubiquitous, however, as Figure 12 provides an example of the NWM persisting drought too long in Iowa and Missouri. These differences between the USDM and the NWM SM arise from several factors, many of which are discussed in Wang et al. (2022). In addition, since the USDM is a convergence of evidence from several dozen inputs, any single indicator may have a different drought value (e.g., the various indices may have range from D1 to D4 but the USDM will "converge" at D3). Drought lessening/termination in the USDM also arises from a "convergence" approach, whereas a single precipitation event can shift LSM SM to a wetter percentile.

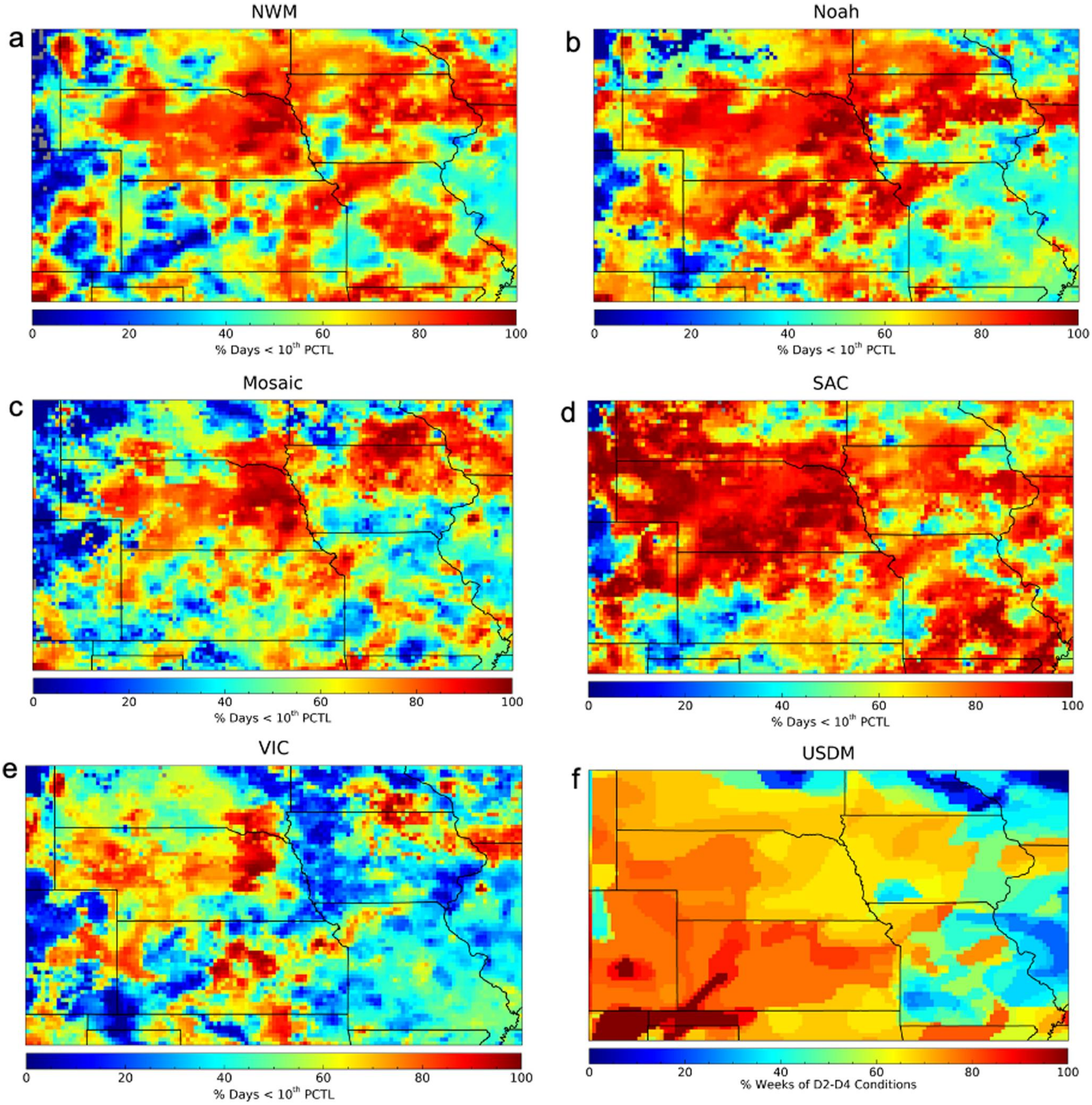

Figure 12. % days from May 1 to 31 December 2012 with 0–100 cm soil moisture percentiles below the 10th percentile for (a) NWM, (b) Noah, (c) Mosaic, (d) SAC, and (e) VIC. Grayed points show gridpoints for which the 10th percentile SM is not defined. (f) % weeks with D2–D4 conditions for USDM data.

4.2. Limitations of LSM Formulations

4.2.1. Soil Moisture Transport Scheme

The NWM's Noah-MP LSM employs a numerical solution of the moisture-content form of the Richards equation (He et al., 2023; Ek et al., 2003), which is valid only for homogeneous, non-layered, and unsaturated soils (Caviedes-Voullieme et al., 2013). These simplifications are inaccurate under certain conditions (e.g., near saturation and in layered soils), and can also degrade the partitioning of precipitation, snowmelt, and throughfall inputs into surface runoff and soil moisture (Lee & Abriola, 1999). As such, infiltration is not a strength of Noah-

MP's formulation (e.g., Crow et al., 2018). The solution scheme employs two features to reduce the occurrence of saturation. First, a scaled (between 0 and 1) free-drainage boundary condition is applied at the bottom of a 2-m thick soil column; the values of this scaling parameter change across the country with model calibration, with values near 0 typically resulting in wetter soils. Second, the soil is discretized into thick layers (i.e., 0–10 cm, 10–40 cm, 40–100 cm, and 100–200 cm) that help to promote solution smoothness, reduce run times, and increase solution robustness. However, these thick discretizations greatly dampen simulated SM dynamics and limit solution accuracy (Downer & Ogden, 2004).

When saturation occurs in the applied Richards solver, an empirical approach is used to estimate surface runoff from a given input of precipitation/melt/throughfall. This empirical approach (Moore, 1985) is materially the same as the Soil Conservation Service Curve Number (SCS-CN) method (Schaake et al., 1996): it is a statistically based method that represents a non-linear regression between storm total runoff as a function of storm-total precipitation, snowmelt, or throughfall (Crow et al., 2018). The SCS-CN regression method is mostly unbiased in certain watersheds. However, some watersheds exhibit behavior that is non-monotonic, while others exhibit threshold behaviors that the SCS-CN method cannot describe (Hawkins et al., 2009). Because the SCS-CN approach applies best to agricultural watersheds but poorly in forested catchments and those with deep, well-drained soils (Hawkins et al., 2009), application of the SCS-CN-like method is inappropriate in some of the NWM's hydro-physiographic regions.

While the limitations arising from the physics of the soil moisture transport scheme are problematic, they are not limited to NWM's Noah-MP. All four of the NLDAS-2 land surface models have similar or more severe limitations: Noah and Noah-MP share similar physics, as the primary improvements to Noah-MP from Noah were in the snow and canopy layers (Niu et al., 2011). Similarly, NLDAS-2-VIC uses a very similar Richards formulation (Liang et al., 1996) that is based on Mahrt and Pan (1984). NLDAS-2-Mosaic and NLDAS-2-SAC differ from these three models in that they are conceptual. NLDAS-2-Mosaic uses a three-layer isothermal model for soil moisture that is driven by hydraulic diffusion and gravitational drainage (Koster & Suarez, 1992); although similar to the other models, water moves between the layers with a 1D version of the Richards equation (Sellers et al., 1986). NLDAS-2-SAC differs the most, incorporating a conceptual two-layer soil column with tension and free water in each layer (Koren et al., 2004).

Finally, we note that, unlike standalone Noah-MP, routing processes and ponding in the NWM impact the surface soil moisture. Routing and ponding processes tend to increase SM, particularly in areas with low soil conductivity (e.g., Fersch et al., 2020; Lahmers et al., 2020). This tendency for routing and ponding processes to increase SM could also help explain the NWM's tendency to be slightly wetter than the other LSMs.

4.2.2. Calibration Strategy

NWM calibration (minimizing errors in streamflow across calibration basins; Viterbo et al., 2020) focuses on long-term water balances. Without human influence, changes in storage should be minimal over long periods of time. However, long-term human manipulation of water resources in some regions (e.g., large-scale water table lowering, pumping, surface detention ponds, etc.) impact the model's water budget, even in areas with more obvious, larger-scale water impoundments such as reservoirs and canal diversions. The NWM calibration and regionalization strategy does not explicitly address biases and inconsistencies in hydrologic processes driven by anthropogenic influences (e.g., diversions, impoundments, pumping, irrigation, etc.). Thus, the model's calibration scheme will adjust model parameters, including but not limited to the scaled free-drainage boundary condition, to compensate for some of these missing processes. Further, this calibration strategy emphasizes long-term climatic inputs of water minus long-term average streamflows and significantly understates the role of short-term variability in groundwater and SM storage. Event-scale errors in runoff prediction will translate directly into errors in input to the soil and affect SM states within the model. As the NWM cycles, sequential errors will lead to biases in SM state and adversely affect the ability of the model to predict SM.

4.2.3. Groundwater Module

The NWM employs a nonlinear conceptual reservoir to simulate groundwater discharge to streams. This approach is widely used in place of numerical modeling of deep groundwater, particularly in cases where aquifer properties are highly uncertain. The application of this approach in the NWM is one-way coupling, which only allows flow from the soil to the reservoir. This is a major limitation in parts of CONUS where the groundwater table is

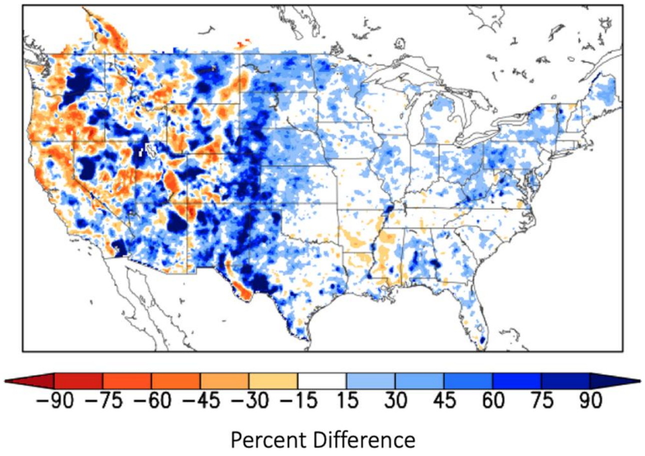

Figure 13. Percent difference between NLDAS-2 precipitation and precipitation from the NWM analysis and assimilation forcing files (i.e (NWM – NLDAS-2)/NLDAS-2*100) over the period 20 June 2019 through 19 June 2020. We use the “tm02” (i.e., “time minus 2”) analysis and assimilation files.

shallow, as tends to be the case in humid areas and semi-humid areas during extensive wet periods, although this limitation is somewhat mitigated by the impact of terrain routing which tends to increase soil moisture in valley bottoms. This is particularly true during the non-growing season as reduced evapotranspiration tends to limit the plant uptake of soil moisture. Fortunately, most SM sensors used to evaluate NWM SM in this study were likely installed in places where shallow groundwater does not play a role. Across the wetter parts of CONUS, however, this structural limitation may contribute to biases in SM. Whether available SM observations accurately represent CONUS-wide performance is an open question.

4.2.4. Impact of Formulation on Performance

These formulation limitations (along with the forcing errors that all five LSMs share) likely contribute to deficiencies in the LSMs representing SM and streamflow. Since all five LSMs use somewhat similar formulations for SM infiltration, this formulation is probably not the reason for differences in scores across models (e.g., SM bias). However, it might explain why all the LSMs struggle in certain regions (e.g., the arid western US, consistent with Lin et al., 2018 who found an earlier version of WRF Hydro struggled more in arid than wet regions of TX), although that could also be explained by deficiencies in the common meteorological forcing in that region (Henn et al., 2018; Zheng et al., 2020). We speculate that, similar to the NLDAS-2-based results of Xia et al. (2015b), the high SM biases in the NWM result from low-biased evapotranspiration (similar to NLDAS-2-Noah); the NWM is structurally similar to NLDAS-2-Noah except for its inclusion of routing and ponding processes, but its improved canopy processes likely impact its evapotranspiration-SM balance. Finally, the NWM’s streamflow-focused calibration strategy likely results in its generally higher streamflow CSIs than SM CSIs.

4.3. Potential Utility for Drought Monitoring

Our results suggest that, despite the limitations noted in Section 4.2, the retrospective NWM simulations are comparable to the NLDAS-2 LSM simulations at representing anomalously dry SM conditions, and furthermore offer a representation of low-flow streamflow conditions at ungauged reaches across the country. These results suggest that the NWM could potentially be used to inform drought monitoring, since the NLDAS-2 LSMs are used as one source of information for the USDM (https://droughtmonitor.unl.edu/). This potential application depends critically on the historical context the long-term retrospective simulations provide: the large, spatially heterogeneous biases in SM preclude usage of the raw SM model output for drought monitoring and it is only after the fields are converted to anomalies (i.e., with reference to the long-term SM and streamflow percentiles at each grid point/channel segment) that the information becomes useful and useable for drought monitoring.

That said, this long-term retrospective NWM simulation differs in an important way from the operational NWM analyses, and this difference limits their current application to drought monitoring. The version 2.0 NWM operational analysis and assimilation (A&A) cycle used meteorological forcings deemed the best available in the low-latency, near-real-time operational computing environment. These A&A forcings differ from the NLDAS-2 forcings used (after downscaling and regridding for the NWM 1-km grid) for the retrospective simulations: for example, the A&A precipitation is derived from a blend of Multi-Radar Multi Sensor (MRMS, Zhang et al., 2016) gauge-corrected precipitation estimates and—in places where radar is blocked/not available—short-lead-time precipitation forecasts from the High-Resolution Rapid Refresh (HRRR, Benjamin et al., 2016) model. A&A precipitation is significantly different from NLDAS-2 precipitation in many locations across CONUS, especially the western US, even when averaged across an entire year (Figure 13). These differences in precipitation likely cause substantial differences in the land-surface and hydrological states in the NWM (note that there was no overlap in NWM simulations forced by these two meteorological data sets, so the impact on streamflow and SM could not be formally evaluated). However, no long-term archive of A&A forcings exists; it would in any case be limited in duration by the availability of radar data required for its precipitation estimates and would require significant computational resources to generate. The NWM change to version 2.1 in April 2021 partially mitigated this issue: v2.1 uses retrospective forcings based on the Analysis of Record for Calibration (AoRC; Fall et al., 2023). For dates after 2002, AoRC precipitation is based on StageIV precipitation, and StageIV precipitation is used in the 28-hr extended range A&A cycle, thus providing overlap between the retrospective and operational NWMv2.1. Although an evaluation of NWMv2.1 is beyond the scope of this paper, Cosgrove et al., 2024 shows that for several evaluation metrics NWMv2.0 and v2.1 are comparable.

4.4. Stakeholder Relevance

Despite the limitations described above, the NWM could be a valuable tool for drought applications, particularly for consideration by USDM authors. The USDM process is always looking to incorporate new data and tools. Data, such as the NWM, with very high spatial resolution and covering the entire CONUS with little to no latency would be especially helpful in gauging the response to soil moisture in a more real-time analysis (currently only possible with limited in situ data). This would improve how soil moisture and hydrological data are incorporated, especially if they were in a GIS format and available via data services. The NWM retrospective performs comparably to the NLDAS-2 LSMs, and the operational NWM is run hourly with very low latency. In addition, the physical consistency (constrained by conservation of mass and energy) between its different land surface and hydrologic variables – most of which are potentially relevant for drought monitoring – offers a comprehensive view of water availability across CONUS at high spatial and temporal resolution. Finally, the NWM is a forecast model run at various lead times (currently out to 30 days); with evaluation to understand its reliability for drought outlooks, and retrospective forecasts to contextualize its output in a climatological sense, it also has potential to add value as a forecasting tool.

5. Summary

We evaluated soil moisture (SM) and streamflow outputs from a 26-year retrospective simulation of NOAA's National Water Model (NWM), version 2.0, from a drought-monitoring perspective. NWM SM was compared with two national networks of in-situ SM, and to NLDAS-2 LSM-derived SM and the U.S. Drought Monitor. NWM streamflow was compared to the USGS HCDN streamflow gauges.

NWM SM had the highest aggregate skill scores in the north-central US, and the lowest in the southwestern US. NWM's SM was comparable in skill to the NLDAS-2 models in terms of bias, anomaly correlation coefficient, and critical success index (CSI). NWM streamflow scores were highest across the Pacific Northwest and eastern US; notably the 20th-percentile streamflow CSIs were above 50% at most gauges in these regions.

Evaluation of NWM SM during the 2012 Great Plains drought revealed rapidly drying soils from May through August 2012. NWM's SM percentiles during this drought were comparable to the drought categories of the USDM and SM percentiles from the other NLDAS-2 LSMs, with the following caveats: (a) NWM SM percentiles decreased somewhat more rapidly than the drought condition area in the USDM; (b) the peak drought area of anomalously dry SM conditions was smaller and more heterogeneous than the comparable USDM drought categories; and (c) in areas of Missouri and Iowa the NWM low-SM percentiles persisted longer than SM from NLDAS-2-Noah, NLDAS-2-VIC, and NLDAS-2-Mosaic, and the comparable USDM drought categories. Caveats (a) and (b) above also applied somewhat for the other drought case studies shown in the SI.

These evaluations, including comparisons to current operationally used NLDAS-2 LSMs as benchmarks, indicate potential for NWM application in operational drought monitoring. Advantages of the NWM include its high temporal and spatial resolution and low latency. However, a major limitation in the current NWM formulation prevents this application: the operational meteorological forcing differs from that of the retrospective simulation, a situation that is somewhat mitigated in NWM v2.1. In addition, NWM soil physics formulations – like those from the NLDAS-2 LSMs – should be improved for certain land surface types of CONUS. These two issues confound the NWM's ability to provide "actionable information" for the drought-monitoring community. Efforts are underway to improve the next operational version of the NWM with an eye toward a flexible, regionally tailored model structure.

Data Availability Statement

- The NLDAS-2 LSM data that were used for a benchmark for the NWM are available in two locations. NLDAS-2-Noah, -Mosaic, and -VIC data are available at the NASA DISC repository (Xia et al., 2012c, 2012d, 2012e). The NLDAS-2-SAC model output is available on Zenodo (NLDAS-2-SAC, 2023).
- The NWMv2.0 retrospective analysis is available on the Registry of Open Data on AWS (NWMv2.0, 2019).
- Streamflow observations used to evaluate the NWM streamflow data is available from the USGS National Water Information System (US Geological Survey 2016).
- In situ soil moisture data used to evaluate the NWM streamflow data is available from the National Soil Moisture database maintained at Ohio State University (Quiring, 2022).
- The US drought monitor weekly conditions used in comparison to the NWM were downloaded as shapefiles from the National Drought Mitigation Center at the University of Nebraska, Lincoln (US Drought Monitor, 2023).
- IDL code to process soil moisture data is provided on Zenodo (Jackson, 2023).

Acknowledgments

The participation of MHu, DJ, DU, MHo, HW, and MG was supported by a NOAA Ocean and Atmospheric Research-National Weather Service service level agreement supported by NIDIS and the NOAA CPO. The work of DJ and MHo was also supported in part by the NOAA cooperative agreement NA17OAR4320101. The participation of SQ and TF was supported by NIDIS and the NOAA CPO.

References

Anderson, M. C., Hain, C., Wardlow, B., Pimstein, A., Mecikalski, J. R., & Kustas, W. P. (2011). Evaluation of drought indices based on Thermal remote sensing of evapotranspiration over the continental United States. *Journal of Climate*, *24*(8), 2025–2044. https://doi.org/10.1175/2010jcli3812.1

Arnault, J., Wagner, S., Rummler, T., Fersch, B., Bliefernicht, J., Andresen, S., & Kunstmann, H. (2016). Role of runoff–infiltration partitioning and resolved overland flow on land–atmosphere feedbacks: A case study with the WRF-hydro coupled modeling system for West Africa. *Journal of Hydrometeorology*, *17*(5), 1489–1516. https://doi.org/10.1175/JHM-D-15-0089.1

Bell, J. E., Palecki, M. A., Baker, C. B., Collins, W. G., Lawrimore, J. H., Leeper, R. D., et al. (2013). U.S. Climate reference network soil moisture and temperature observations. *Journal of Hydrometeorology*, *14*(3), 977–988. https://doi.org/10.1175/jhm-d-12-0146.1

Benjamin, S. G., Weygandt, S. S., Brown, J. M., Hu, M., Alexander, C. R., Smirnova, T. G., et al. (2016). a North American hourly assimilation and model forecast cycle: The rapid Refresh. *Monthly Weather Review*, *144*(4), 1669–1694. https://doi.org/10.1175/mwr-d-15-0242.1

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., et al. (2015). The plumbing of land surface models: Benchmarking model performance. *Journal of Hydrometeorology*, *16*(3), 1425–1442. https://doi.org/10.1175/jhm-d-14-0158.1

Caviedes-Voullieme, D., Garcia-Navarro, P., & Murillo, J. (2013). Verification, conservation, stability and efficiency of a finite volume method for the 1D Richards equation. *Journal of Hydrology*, *480*, 69–84. https://doi.org/10.1016/j.jhydrol.2012.12.008

Champagne, C., Rowlandson, T., Berg, A., Burns, T., L'Heureux, J., Tetlock, E., et al. (2016). Satellite surface soil moisture from SMOS and Aquarius: Assessment for applications in agricultural landscapes. *International Journal of Applied Earth Observation and Geoinformation*, *45*, 143–154. https://doi.org/10.1016/j.jag.2015.09.004

Chen, J. L., Cazenave, A., Dahle, C., Llovel, W., Panet, I., Pfeffer, J., & Moreira, L. (2022). Applications and challenges of GRACE and GRACE follow-on satellite gravimetry. *Surveys in Geophysics*, *43*(1), 305–345. https://doi.org/10.1007/s10712-021-09685-x

Chen, W.-J., Huang, C.-L., & Yang, Z.-L. (2021). More severe drought detected by the assimilation of brightness temperature and terrestrial water storage anomalies in Texas during 2010–2013. *Journal of Hydrology*, *603*, 126802. https://doi.org/10.1016/j.jhydrol.2021.126802

Cohen, S., Praskievicz, S., & Maidment, D. R. (2018). Featured collection introduction: National water model. *Journal of the American Water Resources Association*, *54*(4), 767–769. https://doi.org/10.1111/1752-1688.12664

Cosgrove, B., Gochis, D., Flowers, T., Dugger, A., Ogden, F., Graziano, T., et al. (2024). NOAA's national water model: Advancing operational hydrology through continental-scale modeling. *JAWRA Journal of the American Water Resources Association*, *00*(0), 1–26. https://doi.org/10.1111/1752-1688.13184

Crow, W. T., Chen, F., Reichle, R. H., Xia, Y., & Liu, Q. (2018). Exploiting soil moisture, precipitation, and streamflow observations to evaluate soil moisture/runoff coupling in land surface models. *Geophysical Research Letters*, *45*(10), 4869–4878. https://doi.org/10.1029/2018gl077193

Downer, C. W., & Ogden, F. L. (2004). Appropriate vertical discretization of Richards' equation for two-dimensional watershed-scale modelling. *Hydrological Processes*, *18*, 1–22. https://doi.org/10.1002/hyp.1306

Dudhia, J. (1989). Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *Journal of the Atmospheric Sciences*, *46*(20), 3077–3107. https://doi.org/10.1175/1520-0469(1989)046<3077:nsocod>2.0.co;2

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-Atmospheres*, *108*(D22). https://doi.org/10.1029/2002jd003296

Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., et al. (2010). The soil moisture active passive (SMAP) mission. *Proceedings of the IEEE*, *98*(5), 704–716. https://doi.org/10.1109/jproc.2010.2043918

Fall, G., Kitzmiller, D., Pavlovic, S., Zhang, Z., Patrick, N., St. Laurent, M., et al. (2023). The Office of water prediction's analysis of record for calibration, version 1.1: Dataset description and precipitation evaluation. *JAWRA Journal of the American Water Resources Association*, *00*(0), 1–27. https://doi.org/10.1111/1752-1688.13143

Fersch, B., Senatore, A., Adler, B., Arnault, J., Mauder, M., Schneider, K., et al. (2020). High-resolution fully coupled atmospheric–hydrological modeling: A cross-compartment regional water and energy cycle evaluation. *Hydrology and Earth System Sciences*, *24*(5), 2457–2481. https://doi.org/10.5194/hess-24-2457-2020

Ford, T. W., McRoberts, D. B., Quiring, S. M., & Hall, R. E. (2015). On the utility of in situ soil moisture observations for flash drought early warning in Oklahoma, USA. *Geophys. Res. Lett.*, *42*(22), 9790–9798. https://doi.org/10.1002/2015GL066600

Ford, T. W., & Quiring, S. M. (2019). Comparison of contemporary in situ, model, and satellite remote sensing soil moisture with a focus on drought monitoring. *Water Resources Research*, *55*(2), 1565–1582. https://doi.org/10.1029/2018wr024039

Gochis, D. J., et al. (2020). The WRF-Hydro® modeling system technical description, (Version 5.1.1) (p. 107).

Hansen, M. C., Defries, R. S., Townshend, J. R. G., & Sohlberg, R. (2000). Global land cover classification at 1 km spatial resolution using a classification tree approach. *International Journal of Remote Sensing*, *21*(6–7), 1331–1364. https://doi.org/10.1080/014311600210209

Hawkins, R. H., Ward, T. J., Woodward, D. E., & Van Mullem, J. A. (2009). *Curve number hydrology: State of the practice*. American Society of Civil Engineers.

He, C., Valayamkunnath, P., Barlage, M., Chen, F., Gochis, D., Ryan, C., et al. (2023). Modernizing the open-source community Noah with multi-parameterization options (Noah-MP) land surface model (version 5.0) with enhanced modularity, interoperability, and applicability. *Geoscientific Model Development*, *16*(17), 5131–5151. https://doi.org/10.5194/gmd-16-5131-2023

Helsel, D. R., Hirsch, R. M., Ryberg, K. R., Archfield, S. A., & Gilroy, E. J. (2020). *Statistical methods in water resources: US Geological Survey techniques and methods*. US Geological Survey.

Henn, B., Newman, A. J., Livneh, B., Daly, C., & Jessica, D. L. (2018). An assessment of differences in gridded precipitation datasets in complex terrain. *Journal of Hydrology*, *556*, 1205–1219. https://doi.org/10.1016/j.jhydrol.2017.03.008

Hobbins, M. T., Wood, A., McEvoy, D. J., Huntington, J. L., Morton, C., Anderson, M., & Hain, C. (2016). The evaporative demand drought index. Part I: Linking drought evolution to variations in evaporative demand. *Journal of Hydrometeorology*, *17*(6), 1745–1761. https://doi.org/10.1175/jhm-d-15-0121.1

Hoerling, M., Eischeid, J., Kumar, A., Leung, R., Mariotti, A., Mo, K., et al. (2014). Causes and predictability of the 2012 great plains drought. *Bulletin of the American Meteorological Society*, *95*(2), 269–282. https://doi.org/10.1175/bams-d-13-00055.1

Jackson, D. J. (2023). IDL code for soil moisture analysis [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.10055467

Kerr, Y. H., Waldteufel, P., Richaume, P., Wigneron, J. P., Ferrazzoli, P., Mahmoodi, A., et al. (2012). The SMOS soil moisture retrieval algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, *50*(5), 1384–1403. https://doi.org/10.1109/tgrs.2012.2184548

Koren, V., Reed, S., Smith, M., Zhang, Z., & Seo, D. J. (2004). Hydrology Laboratory Research Modeling System (HL-RMS) of the US National Weather Service. *Journal of Hydrology*, *291*(3–4), 297–318. https://doi.org/10.1016/j.jhydrol.2003.12.039

Koster, R. D., & Suarez, M. J. (1992). Modeling the land surface boundary in climate models as a composite of independent vegetation stands. *Journal of Geophysical Research*, *97*(D3), 2697–2715. https://doi.org/10.1029/91jd01696

Koster, R. D., & Suarez, M. J., (1994). The components of a SVAT scheme and their effects on a GCMS hydrological cycle. *Advances in Water Resources*, *17*, 61–78.

La Follette, P., Ogden, F. L., & Jan, A. (2023). Layered Green and Ampt infiltration with redistribution. *Water Resources Research*, *59*(7), e2022WR033742. https://doi.org/10.1029/2022WR033742

Lahmers, T. M., Castro, C. L., & Hazenberg, P. (2020). Effects of lateral flow on the convective environment in a coupled hydrometeorological modeling system in a semiarid environment. *Journal of Hydrometeorology*, *21*(4), 615–642. https://doi.org/10.1175/jhm-d-19-0100.1

Lahmers, T. M., Kumar, S. V., Rosen, D., Dugger, A., Gochis, D. J., Santanello, J. A., et al. (2022). Assimilation of NASA's Airborne snow observatory snow measurements for improved hydrological modeling: A case study enabled by the coupled LIS/WRF-hydro system. *Water Resources Research*, *58*(3), e2021WR029867. https://doi.org/10.1029/2021WR029867

Lee, D. H., & Abriola, L. M. (1999). Use of the Richards equation in land surface parameterizations. *Journal of Geophysical Research-Atmospheres*, *104*, 27519–27526.

Liang, X., Wood, E. F., & Lettenmaier, D. P. (1996). Surface soil moisture parameterization of the VIC-2L model: Evaluation and modification. *Global and Planetary Change*, *13*(1–4), 195–206. https://doi.org/10.1016/0921-8181(95)00046-1

Lin, P. R., 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. *Journal of the American Water Resources Association*, *54*(1), 40–54. https://doi.org/10.1111/1752-1688.12585

Lins, H. F. (2012). *USGS hydro-climatic data network 2009 (HCDN-2009)* (p. 3047). US Geological Survey Fact Sheet.

Mahrt, L., & Pan, H. (1984). A 2-layer model of soil hydrology. *Boundary-Layer Meteorology*, *29*, 1–20. https://doi.org/10.1007/bf00119116

Martinaitis, S. M., Cocks, S. B., Simpson, M. J., Osborne, A. P., Harkema, S. S., Grams, H. M., et al. (2021). Advancements and characteristics of gauge ingest and quality control within the Multi-Radar Multi-Sensor System. *Journal of Hydrometeorology*, *22*, 2455–2474. https://doi.org/10.1175/JHM-D-20-0234.1

Martinaitis, S. M., Osborne, A. P., Simpson, M. J., Zhang, J., Howard, K. W., Cocks, S. B., et al. (2020). A physically based multisensor quantitative precipitation estimation approach for gap-filling radar coverage. *Journal of Hydrometeorology*, *21*(7), 1485–1511. https://doi.org/10.1175/JHM-D-19-0264.1

McEvoy, D. J., Huntington, J. L., Hobbins, M. T., Wood, A., Morton, C., Anderson, M., & Hain, C. (2016). The evaporative demand drought index. Part II: CONUS-wide assessment against common drought indicators. *Journal of Hydrometeorology*, *17*(6), 1763–1779. https://doi.org/10.1175/jhm-d-15-0122.1

McKee, T. B., Doesken, N. J., & Kleist, J. (1993). The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology* (Vol. *17*, pp. 179–183). No. 22.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., et al. (2006). North American regional reanalysis. *Bulletin of the American Meteorological Society*, *87*(3), 343–360. 343-+. https://doi.org/10.1175/bams-87-3-343

Mitchell, K. E., Lohmann, D., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., et al. (2004). The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research: Atmospheres*, *109*(D7). https://doi.org/10.1029/2003jd003823

Mo, K. C., Chen, L. C., Shukla, S., Bohn, T. J., & Lettenmaier, D. P. (2012). Uncertainties in North American Land Data Assimilation Systems over the contiguous United States. *Journal of Hydrometeorology*, *13*(3), 996–1009. https://doi.org/10.1175/jhm-d-11-0132.1

Moore, R. J. (1985). The probability-distributed principle and runoff production at point and basin scales. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, *30*(2), 273–297. https://doi.org/10.1080/02626668509490989

Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., & Nearing, G. (2017). Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology*, *18*(8), 2215–2225. https://doi.org/10.1175/jhm-d-16-0284.1

Niu, G. Y., Yang, Z. L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (2011). The community Noah land surface model with multi-parameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research-Atmospheres*, *116*(D12), D12109. https://doi.org/10.1029/2010jd015139

NLDAS. (2022). North-American Land Data Assimilation System (NLDAS). Retrieved from https://www.drought.gov/data-maps-tools/north-american-land-data-assimilation-system-nldas.]

NLDAS-2-SAC. (2023). NLDAS-2-SAC output [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.10052520

NOAA. (2020). Precipitation prediction grand challenge strategy. In *NOAA Rep.* (p. 44). Retrieved from www.noaa.gov/sites/default/files/2022-01/PPGC-Strategy_FINAL_2020-1030.pdf

Noel, M., Bathke, D., Fuchs, B., Gutzmer, D., Haigh, T., Hayes, M., et al. (2020). Linking drought impacts to drought severity at the state level. *Bulletin of the American Meteorological Society*, *101*(8), E1312–E1321. https://doi.org/10.1175/bams-d-19-0067.1

NWMv2.0 (2019). NOAA national water model 2.0 CONUS retrospective, from the Registry of Open Data on AWS. [Dataset]. Retrieved from https://registry.opendata.aws/nwm-archive

Otkin, J. A., Svoboda, M., Hunt, E. D., Ford, T. W., Anderson, M. C., Hain, C., & Basara, J. B. (2018). Flash droughts: A review and assessment of the challenges imposed by rapid-onset droughts in the United States. *Bulletin of the American Meteorological Society*, *99*(5), 911–919. https://doi.org/10.1175/bams-d-17-0149.1

Quiring (2022). Soil moisture data from two national networks from the Ohio State University. [Dataset]. http://nationalsoilmoisture.com/

Quiring, S. M., Ford, T. W., Wang, J. K., Khong, A., Harris, E., Lindgren, T., et al. (2016). The North American soil moisture database development and applications. *Bulletin of the American Meteorological Society*, *97*(8), 1441–1459. 1441-+. https://doi.org/10.1175/bams-d-13-00263.1

Schaake, J. C., Koren, V. I., Duan, Q. Y., Mitchell, K., & Chen, F. (1996). Simple water balance model for estimating runoff at different spatial and temporal scales. *Journal of Geophysical Research-Atmosphere*, *101*(D3), 7461–7475. https://doi.org/10.1029/95jd02892

Schaefer, G. L., Cosh, M. H., & Jackson, T. J. (2007). The USDA Natural Resources Conservation Service Soil Climate Analysis Network (SCAN). *Journal of Atmospheric and Oceanic Technology*, *24*(12), 2073–2077. https://doi.org/10.1175/2007jtecha930.1

Schaefer, J. T. (1990). The critical success index as an indicator of warning skill. *Weather and Forecasting*, *5*(4), 570–575. https://doi.org/10.1175/1520-0434(1990)005<0570:tcsiaa>2.0.co;2

Sellers, P. J., Mintz, Y., Sud, Y. C., & Dalcher, A. (1986). A Simple Biosphere Model (SIB) for use within general-circulation models. *Journal of the Atmospheric Sciences*, *43*(6), 505–531. https://doi.org/10.1175/1520-0469(1986)043<0505:asbmfu>2.0.co;2

Sofokleous, I., Bruggeman, A., Camera, C., & Eliades, M. (2022). Grid-based calibration of the WRF-Hydro with Noah-MP model with improved groundwater and transpiration process equations. *Journal of Hydrology*, *617*(Part A), 128991. 2023, 128991, ISSN 0022-1694. https://doi.org/10.1016/j.jhydrol.2022.128991

Svoboda, M., LeComte, D., Hayes, M., Heim, R., Gleason, K., Angel, J., et al. (2002). The drought monitor. *Bulletin of the American Meteorological Society*, *83*(8), 1181–1190. https://doi.org/10.1175/1520-0477-83.8.1181

U.S. Drought Monitor. (2023). US Drought Monitor weekly shapefiles from the University of Nebraska, Lincoln. [Dataset]. Retrieved from https://droughtmonitor.unl.edu/DmData/GISData.aspx

U.S. Geological Survey. (2016). National Water Information System (USGS Water Data for the Nation) daily values service. [Dataset]. Retrieved from https://waterservices.usgs.gov

Vergopolan, N., Chaney, N. W., Beck, H. E., Pan, M., Sheffield, J., Chan, S., & Wood, E. F. (2020). *Combining hyper-resolution land surface modeling with SMAP brightness temperatures to obtain 30-m soil moisture estimates* (p. 242). Remote Sensing of Environment.

Viterbo, F., Mahoney, K., Read, L., Salas, F., Bates, B., Elliott, J., et al. (2020). A multiscale, hydrometeorological forecast evaluation of national water model forecasts of the May 2018 Ellicott City, Maryland, Flood. *Journal of Hydrometeorology*, *21*(3), 475–499. https://doi.org/10.1175/jhm-d-19-0125.1

Wang, H., Xu, L., Hughes, M., Chelliah, M., DeWitt, D. G., Fuchs, B. A., & Jackson, D. L. (2022). Potential caveats in land surface model evaluations using the US drought monitor: Roles of base periods and drought indicators. *Environmental Research Letters*, *17*(1), 014011. https://doi.org/10.1088/1748-9326/ac3f63

Wei, H. L., Xia, Y., Mitchell, K. E., & Ek, M. B. (2013). Improvement of the Noah land surface model for warm season processes: Evaluation of water and energy flux simulation. *Hydrological Processes*, *27*(2), 297–303. https://doi.org/10.1002/hyp.9214

Williams, A. P., Cook, E. R., Smerdon, J. E., Cook, B. I., Abatzoglou, J. T., Bolles, K., et al. (2020). Large contribution from anthropogenic warming to an emerging North American megadrought. *Science*, *368*(6488), 314–318. 314-+. https://doi.org/10.1126/science.aaz9600

Wood, E. F., Lettenmaier, D., Liang, X., Nijssen, B., & Wetzel, S. W. (1997). Hydrological modeling of continental-scale basins. *Annual Review of Earth and Planetary Sciences*, *25*(1), 279–300. 279-+. https://doi.org/10.1146/annurev.earth.25.1.279

Wood, E. F., Schubert, S. D., Wood, A. W., Peters-Lidard, C. D., Mo, K. C., Mariotti, A., & Pulwarty, R. S. (2015). Prospects for advancing drought understanding, monitoring, and prediction. *Journal of Hydrometeorology*, *16*(4), 1636–1657. https://doi.org/10.1175/jhm-d-14-0164.1

Wu, W. Y., Yang, Z. L., & Barlage, M. (2021). The impact of Noah-MP physical parameterizations on modeling water availability during droughts in the Texas-Gulf Region. *Journal of Hydrometeorology*, *22*, 1221–1233. https://doi.org/10.1175/jhm-d-20-0189.1

Xia, Y., Ek, M. B., Wu, Y. H., Ford, T., & Quiring, S. M. (2015a). Comparison of NLDAS-2 simulated and NASMD observed daily soil moisture. Part II: Impact of soil texture classification and vegetation type mismatches. *Journal of Hydrometeorology*, *16*(5), 1981–2000. https://doi.org/10.1175/jhm-d-14-0097.1

Xia, Y., Ek, M. B., Wu, Y. H., Ford, T., & Quiring, S. M. (2015b). Comparison of NLDAS-2 simulated and NASMD observed daily soil moisture. Part I: Comparison and analysis. *Journal of Hydrometeorology*, *16*(5), 1962–1980. https://doi.org/10.1175/jhm-d-14-0096.1

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., et al. (2012a). Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. *Journal of Geophysical Research-Atmospheres*, *117*(D3). https://doi.org/10.1029/2011jd016051

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., et al. (2012b). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research-Atmospheres*, *117*(D3). https://doi.org/10.1029/2011jd016048

Xia, Y., Mitchell, K., Ek, M. B., Sheffield, J., Cosgrove, B., Wood, E., et al. (2012c). NCEP/EMC, NLDAS Noah Land Surface Model L4 hourly 0.125 x 0.125 degree V002, Edited by David Mocko, NASA/GSFC/HSL, Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). [Dataset]. https://doi.org/10.5067/47Z13FNQODKV

Xia, Y., Mitchell, K., Ek, M. B., Sheffield, J., Cosgrove, B., Wood, E., et al. (2012d). NCEP/EMC (2014), NLDAS VIC Land Surface Model L4 hourly 0.125 x 0.125 degree V002, Edited by David Mocko, NASA/GSFC/HSL, Greenbelt, Maryland, USA, GoddardEarth Sciences Data and Information Services Center (GES DISC). [Dataset]. https://doi.org/10.5067/ELBDAPAKNGJ9

Xia, Y., Mitchell, K., Ek, M. B., Sheffield, J., Cosgrove, B., Wood, E., et al. (2012e). NCEP/EMC (2009), NLDAS Mosaic Land Surface Model L4 hourly 0.125 x 0.125 degree V002, Edited by David Mocko, NASA/GSFC/HSL, Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). [Dataset]. https://doi.org/10.5067/EN4MBWTCENE5

Zhang, J., Howard, K., Langston, C., Kaney, B., Qi, Y., Tang, L., et al. (2016). Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation initial operating capabilities. *Bulletin of the American Meteorological Society*, *97*(4), 621–637. https://doi.org/10.1175/bams-d-14-00174.1

Zheng, H., Yang, Z.-L., Lin, P., Wu, W.-Y., Li, L., Xu, Z., et al. (2020). Falsification-oriented signature-based evaluation for guiding the development of land surface models and the enhancement of observations. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2020MS002132. https://doi.org/10.1029/2020MS002132

References From the Supporting Information

Ball, J. T., Woodrow, I. E., & Berry, J. A. (1987). *A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions*. In *Progress in Photosynthesis Research* (pp. 221–224). Springer.

Chen, F., Mitchell, K., Schaake, J., Xue, Y., Pan, H., Koren, V., et al. (1996). Modeling of land surface evaporation by four schemes and comparison with FIFE observations. *Journal of Geophysical Research-Atmospheres*, *101*(D3), 7251–7268. https://doi.org/10.1029/95jd02165

Jordan, R. E. (1991). A one-dimensional temperature model for a snow cover: Technical documentation for SNTHERM (p. 89).

Niu, G. Y., & Yang, Z. L. (2006). Effects of frozen soil on snowmelt runoff and soil water storage at a continental scale. *Journal of Hydrometeorology*, *7*(5), 937–952. https://doi.org/10.1175/jhm538.1

Sakaguchi, K., & Zeng, X. B. (2009). Effects of soil wetness, plant litter, and under-canopy atmospheric stability on ground evaporation in the Community Land Model (CLM3.5). *Journal of Geophysical Research-Atmospheres*, *114*(D1). https://doi.org/10.1029/2008jd010834

Slack, J. R., & Landwehr, J. M. (1992). Hydro-climatic data network (HCDN); a U.S. Geological Survey streamflow data set for the United States for the study of climate variations. Report 92-129 (pp. 1874–1988).

Verseghy, D. L. (1991). Class-A Canadian land surface scheme for GCMS 1. Soil model. *International Journal of Climatology*, *11*(2), 111–133. https://doi.org/10.1002/joc.3370110202