scientific data

OPEN

DATA DESCRIPTOR

Homogenized gridded dataset for drought and hydrometeorological modeling for the continental United States

Robert Erhardt[1] ✉, Courtney A. Di Vittorio[2], Staci A. Hepler[1], Lauren E. L. Lowman[2] & Wendy Wei[1]

We present a novel data set for drought in the continental US (CONUS) built to enable computationally efficient spatio-temporal statistical and probabilistic models of drought. We converted drought data obtained from the widely-used US Drought Monitor (USDM) from its native geo-referenced polygon format to a 0.5 degree regular grid. We merged known environmental drivers of drought, including those obtained from the North American Land Data Assimilation System (NLDAS-2), US Geological Survey (USGS) streamflow data, and National Oceanic and Atmospheric Administration (NOAA) teleconnections data. The resulting data set permits statistical and probabilistic modeling of drought with explicit spatial and/or temporal dependence. Such models could be used to forecast drought at short-range, seasonal to sub-seasonal, and inter-annual timescales with uncertainty, extending the reach and value of the current US Drought Outlook from the National Weather Service Climate Prediction Center. This novel data product provides the first common gridded dataset that includes critical variables used to inform hydrological and meteorological drought.

Background & Summary

The last two decades have seen strong development of gridded drought indices. The global standardized precipitation evapotranspiration index database (SPEIBASE) was first released in 2010[1,2]. It utilizes gridded public climate data available monthly at 0.5 degrees and constructs a drought index based on both precipitation as well as evapotranspiration. The Global Precipitation Climatology Centre Drought Index (GPCC-DI) extended this to produce an index which combined a standardized precipitation index (SPI) with the SPEI to achieve a new global, monthly gridded drought index, but at a lower 1 degree spatial resolution[3]. Several researchers have achieved higher spatial resolutions in regional data products. Two examples include a combined SPI/SPEI index achieved daily at 0.1 degrees in China[4], and a suite of drought indices at the high 12 km resolution produced over a small region covering three states in the United States[5].

Enhancements described above include new data sources, superior spatio-temporal resolutions, and region-specific data sets more suited to studying local drought. Here we present a drought database for the continental United States that has all three enhancements. That is, this database: includes an enhanced drought index which draws on both high quality environmental data sources as well as expert local regional judgments; includes a suite of hydrometeorological variables which permits scientific study of regional drought processes across a range of scales; and achieves a superior spatio-temporal resolution than what is currently available from global products which cover the contintental United States. We achieve these goals by constructing a gridded drought database from the U.S. Drought Monitor (USDM), and merging in a suite of hydrometeorological variables at the same spatio-temporal resolution.

The U.S. Drought Monitor (USDM)[6] provides a weekly snapshot of the drought status in the United States. The U.S. Drought Monitor is jointly produced by the National Drought Mitigation (NDMC) at the University of Nebraska-Lincoln, the United States Department of Agriculture (USDA), and the National Oceanic and Atmospheric Administration (NOAA). This data product was first released on January 4, 2000, and has been

[1]Wake Forest University, Department of Statistical Sciences, Winston-Salem, NC, USA. [2]Wake Forest University, Department of Engineering, Winston-Salem, NC, USA. ✉e-mail: erhardrj@wfu.edu

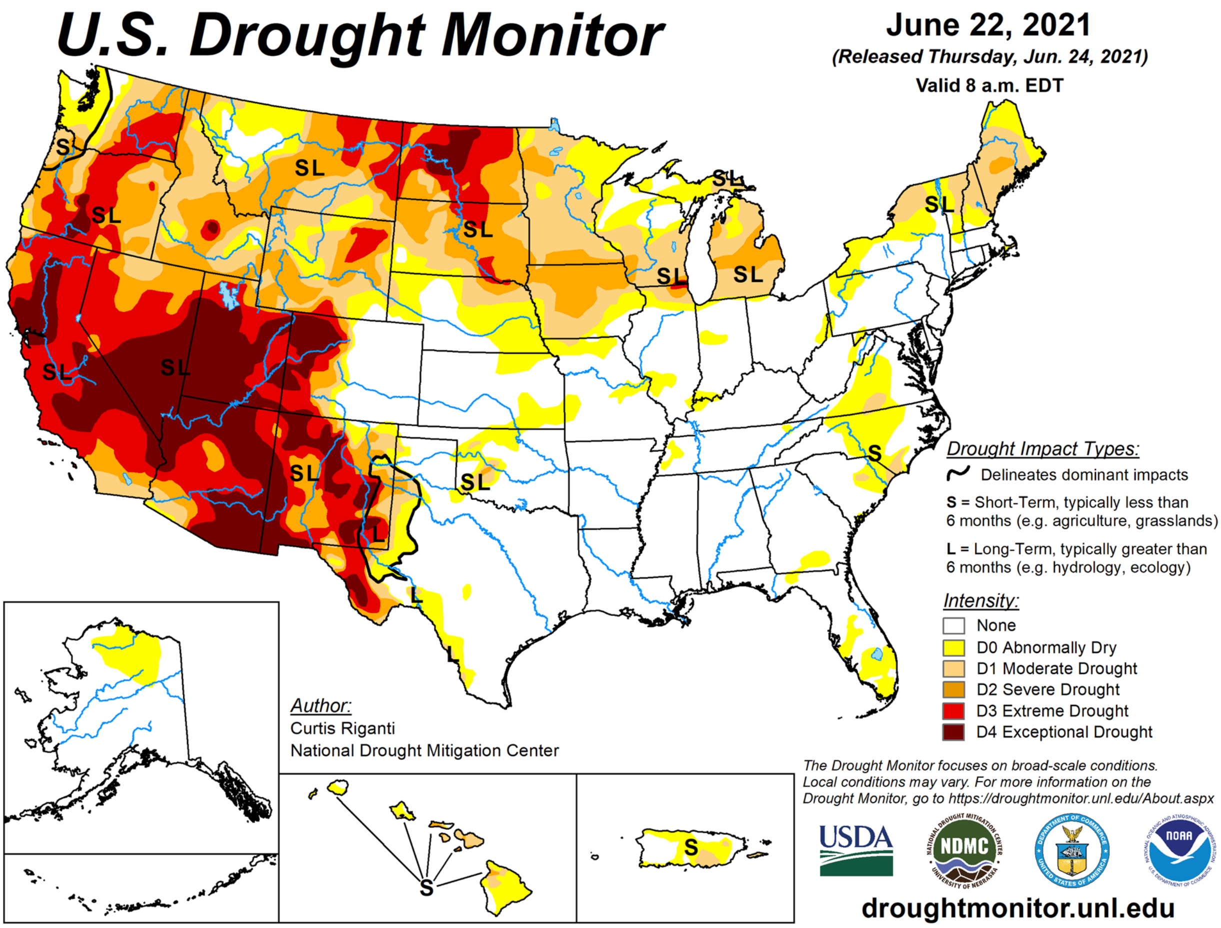
U.S. Drought Monitor
June 22, 2021
(Released Thursday, Jun. 24, 2021)
Valid 8 a.m. EDT
S
SL
SL
SL
SL
SL
SL
SL
SL
SL
SL
S
L
L
L
S
S
S
Drought Impact Types:
Delineates dominant impacts
S = Short-Term, typically less than 6 months (e.g. agriculture, grasslands)
L = Long-Term, typically greater than 6 months (e.g. hydrology, ecology)
Intensity:
None
D0 Abnormally Dry
D1 Moderate Drought
D2 Severe Drought
D3 Extreme Drought
D4 Exceptional Drought
Author:
Curtis Riganti
National Drought Mitigation Center
The Drought Monitor focuses on broad-scale conditions. Local conditions may vary. For more information on the Drought Monitor, go to https://droughtmonitor.unl.edu/About.aspx
USDA
NDMC
droughtmonitor.unl.edu

Fig. 1 An example of the raw data for the U.S. Drought Monitor for June 22, 2021. The U.S. Drought Monitor is jointly produced by the National Drought Mitigation Center at the University of Nebraska-Lincoln, the United States Department of Agriculture, and the National Oceanic and Atmospheric Administration. Map courtesy of NDMC.

updated weekly ever since. It characterizes drought severity as of 8AM EST/EDT each Tuesday for the entire U.S., though the images are released two days later on the following Thursday. The monitor classifies drought into one of six ordered categories: 0, no drought; D0, abnormally dry or "pre-drought"; and levels D1 through D4, which represent increasing levels of drought severity. Figure 1 shows the USDM as of 8AM EST on June 22, 2021. The USDM is defined everywhere in space over the contiguous United States, and also Alaska, Hawaii, Puerto Rico, and some outlying territories.

The classification of drought into the ordinal measure relies on inputs such as the Palmer Drought Severity Index[7], soil moisture[8], United States Geological Survey weekly streamflow data (https://waterwatch.usgs.gov/), and other environmental inputs. However, the USDM is *not* a deterministic calculation based on these inputs. Rather, the NDMC website describes the USDM as a blending of the "best available data, local observations, and experts' best judgment that makes the U.S. Drought Monitor more versatile than other drought indicators. (https://droughtmonitor.unl.edu/About/AbouttheData/DroughtClassification.aspx)".

USDM images are updated each week, and this is used to define the weekly time scale for our data product described in this paper. The USDM images are freely available as shapefiles (https://droughtmonitor.unl.edu/DmData/GISData.aspx). While this data structure is ideal for visualizing drought status continuously in space, it presents challenges for building statistical models to answer scientific questions. Examples of such questions include:

- How much variability in drought level can be explained by uncertainty in meteorological and land-surface conditions? In contrast, how much variability is explained by "local observations and experts' best judgment"?
- What is the relative importance of each individual USDM input to the overall estimated drought level? How much value would a proposed new scientific input add? How does the relative importance of individual inputs vary across space and time?
- How can we best make projections of drought status into the future? The current drought forecast tool is the US Drought Outlook, published each month and shows expectations of future droughts one month and one season ahead (Fig. 2). While informative for many purposes, the US Drought Outlook is limited in that it does not capture or convey forecast uncertainty. How could such forecasts be enhanced to quantify uncertainty across space and time?

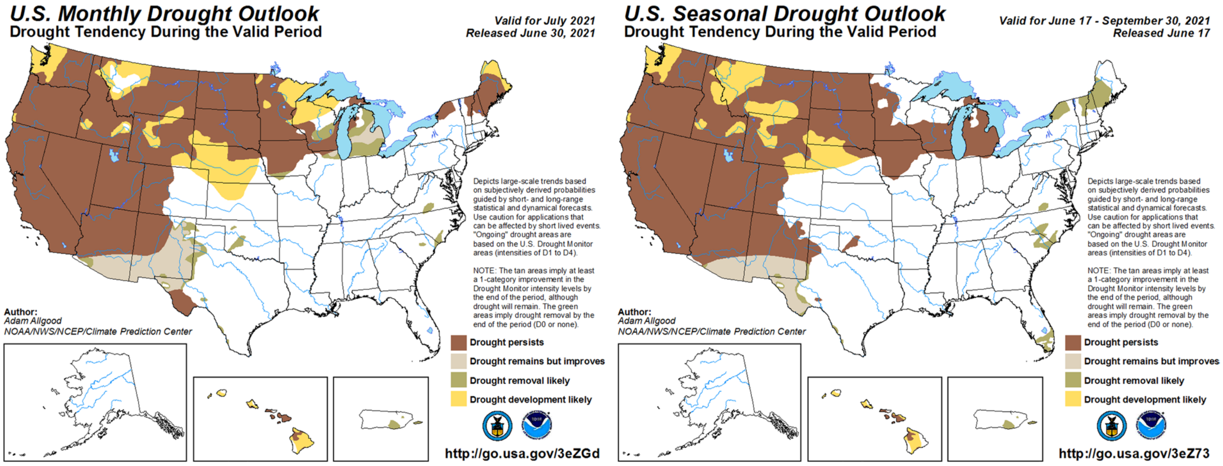
U.S. Monthly Drought Outlook
Drought Tendency During the Valid Period
Valid for July 2021
Released June 30, 2021
Depicts large-scale trends based on subjectively derived probabilities guided by short- and long-range statistical and dynamical forecasts. Use caution for applications that can be affected by short lived events. "Ongoing" drought areas are based on the U.S. Drought Monitor areas (intensities of D1 to D4).
NOTE: The tan areas imply at least a 1-category improvement in the Drought Monitor intensity levels by the end of the period, although drought will remain. The green areas imply drought removal by the end of the period (D0 or none).
Author:
Adam Allgood
NOAA/NWS/NCEP/Climate Prediction Center
Drought persists
Drought remains but improves
Drought removal likely
Drought development likely
http://go.usa.gov/3eZGd
U.S. Seasonal Drought Outlook
Drought Tendency During the Valid Period
Valid for June 17 - September 30, 2021
Released June 17
Depicts large-scale trends based on subjectively derived probabilities guided by short- and long-range statistical and dynamical forecasts. Use caution for applications that can be affected by short lived events. "Ongoing" drought areas are based on the U.S. Drought Monitor areas (intensities of D1 to D4).
NOTE: The tan areas imply at least a 1-category improvement in the Drought Monitor intensity levels by the end of the period, although drought will remain. The green areas imply drought removal by the end of the period (D0 or none).
Author:
Adam Allgood
NOAA/NWS/NCEP/Climate Prediction Center
Drought persists
Drought remains but improves
Drought removal likely
Drought development likely
http://go.usa.gov/3eZ73

Fig. 2 Examples of a monthly US Drought Outlook (left) and seasonal US Drought Outlook (right) released in June 2021, around the same time as the USDM shown in Fig. 1 was released. No quantification of forecast uncertainty is captured or conveyed.

To answer any of the previous questions, one would need a statistical model relating the ordinal response of drought level to other covariates. This immediately raises a question–what should the spatio-temporal support be for such a model? The response variable of drought level is defined everywhere in space, stored as geo-referenced polygons, and therefore not easily used as a response variable in its native form. On the other hand, the environmental covariates are available at different spatial and temporal resolutions. Some–such as USGS streamflow–are point-referenced and obtained at specific locations sampled unevenly and irregularly across the US. Others–such as meteorological variables–are provided as satellite or reanalysis products stored on a regular grid indexed by latitude and longitude. Modeling the USDM drought severity in a statistical framework therefore necessarily blends vectorized images, point-referenced data with incomplete coverage, and gridded aerial observations, all sampled at different time scales. Our solution presented here is to process all available data to a common spatio-temporal support.

With an eye towards scientific questions involving forecasting, we consider additional covariates beyond those used to define the USDM. One study used predictors obtained from the North American Multi-Model Ensemble (NMME)[9] climate forecasts to produce 1- and 3-month forecasts of the USDM[10]. They noted that more environmental variables beyond precipitation and temperature are used to construct the USDM, and therefore the NMME dataset should be supplemented with other climate information to produce the best forecasts. In our data product, we added accumulated precipitation, total evapotranspiration, potential evapotranspiration, potential latent heat flux, soil moisture, surface runoff, soil temperature, leaf area index, snow depth, snow melt, snow cover fraction, and water equivalent of accumulated snow depth. Numerous studies have also established the importance of using the El Niño Southern Oscillation (ENSO) index for drought modeling[10–18]. Other studies have considered climate teleconnections such as the North Atlantic Oscillation (NAO)[19]. Accordingly, we added four teleconnections to our data product. These teleconnections are notably different from the other hydrometeorological and hydrological drought-forcing variables in a number of ways. First, they are (at most) statistically related to drought, rather than directly and locally-acting as the other variables. Second, users should likely consider the teleconnections as statistically-linked to drought with spatio-temporally-varying coefficients, which would allow the statistical relationship between the teleconnection to vary in space and/or time. Users could also include subsets of teleconnections in a hierarchical model, using them as covariates for hydrometeorological and hydrological variables in a middle layer of model, leaving the hydrometeorological and hydrological variables alone to directly act on drought. Regardless of the particular use, the point is that teleconnections differ from the other locally-varying and direct variables in these data.

It is clear that there is a scientific need for a database to support statistical research using the USDM as well as seasonal to sub-seasonal (S2S) forecasting of drought, where predictive skill is relatively low[20]. This new database should include known scientific inputs for the USDM along with other potential predictors of drought (Fig. 3). All data should be recorded at the same spatial and temporal support to avoid the need for change of support models which could otherwise increase the computational cost of fitting models. In this paper, we describe the construction of a comprehensive database which discretizes the USDM to a 0.5 degree spatial resolution, and merges in summaries of NLDAS-2 data and USGS streamflow data along with common climate teleconnections, all at the same spatial resolution and weekly time scale (see Tables 1 and 2).

Methods

We process and combine four sources of raw data: USDM drought data, NLDAS-2 hydrometeorological data, USGS streamflow data, and teleconnections data. Figure 4 shows this process, each component of which is described below.

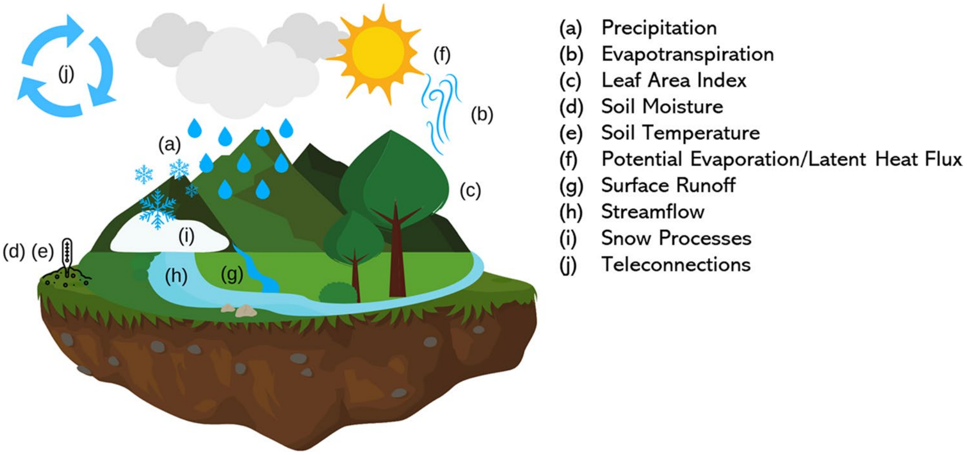

Fig. 3 Summary of climate, atmospheric, and land surface variables included in the homogenized dataset. All variables have been demonstrated in the literature to be informative of drought and/or its impacts.

US Drought Monitor. US Drought Monitor raw data are available as geo-referenced polygons stored as shapefiles, released each week. One example is shown in Fig. 1. The primary processing step consisted of discretizing these weekly geo-referenced polygons to a common gridded support covering the contiguous United States. We defined a regular grid whose centroids are a set of points evenly spaced every 0.5 degrees by both latitude and longitude, and this grid serves as the common spatial support for all variables in the data set. Numerous studies have used gridded drought data at various scales, including 1 degree[3,21], 0.5 degree[22], and 0.3 degrees[23,24]. Studies analyzing drought in a statistical or probabilistic framework have used spatial resolutions of 0.125 degrees[25,26], 0.25 degrees[27], 0.5 degrees[11,28,29], or 1 degree[10], depending on the spatial resolutions of the particular predictor variables.

We began with a bounding box which ranged from −124.75 degrees longitude to −65.25 degrees longitude, and 25.25 degrees latitude to 49.75 degrees latitude. Grid cells were labeled with the letter referring to each of the 50 rows (latitude) from letters ordered as A,…, Z, AA,…, XX beginning at 49.75 degrees north, and columns with numbers from 1 to 120 beginning at −124.75 degrees west. All lat/lon coordinates are defined at the grid centroid. Thus, the bounding box is indexed by A1 (−124.75, 49.75) to XX120 (−65.25, 25.25). However, only 3259 of these 6000 grid cell locations fall within the contiguous United States over land where the USDM is defined. We will later see that one grid cell over Great Salt Lake is primarily water with undefined land variables, and so $i = 1, \ldots, I = 3258$ will index spatial location in our data. Accordingly, the top-left grid that appears in the data product is C6 at (−122.25, 48.75), and the bottom-right is XX89 at (−80.75, 25.25). The particular value of the USDM is taken from the centroid of each grid cell, for each week–i.e. there is no averaging within a grid cell. This choice results in a gridded summary of drought that has the same support as the USDM itself (0, D0, D1,…, D4) and is therefore interpretable and familiar to all end users of the USDM (unlike, for example, averaging levels within a grid cell, which would produce values between the familiar support of the USDM). For grid cells with two or more levels of drought falling within a single cell, we explored taking the mode (by area within each cell), but found it made very little practical difference compared to the centroid, as the two values are identical in nearly every case. Figure 5 shows one example of the 0.5 degree grid overlaying USDM data in one region of the United States to demonstrate visually the consistency between the centroid and mode. Figure 6 shows the result of this discretization process for the week of June 22, 2021. The same process is applied to each week of the available USDM data, beginning 01/04/2000 ($t = 1$) and ending on 06/30/2022 ($t = 1174$). This results in the database having exactly 3,824,892 rows (3258 × 1174), each of which has columns indicating the grid name, latitude, longitude, time, and measurement of drought.

NLDAS-2. The North America Land Data Assimilation System Phase 2 (NLDAS-2) is an integrated observation and model reanalysis data set designed to drive offline land surface models[30,31]. NLDAS-2 land surface forcing fields are derived from the North American Regional Reanalysis (NARR) fields and are available at 0.125° × 0.125° spatial resolution, and at hourly or monthly temporal resolutions. Specific details on the spatial interpolation and temporal disaggregation methods adopted in NLDAS-2 are described in Cosgrove *et al.*[32]. Accumulated precipitation [kg/m^2] and potential evapotranspiration [kg/m^2] were obtained as covariates for the statistical model from the NLDAS-2 Forcing File A dataset. These variables constrain water availability for a given time period. Outputs from the NLDAS Noah land-surface model forced with the forcing fields were also obtained as covariates. The Noah land-surface model performs a water and energy balance for the land-surface, which is discretized into four soil layers, and parameterizes warm and cold season processes[33,34]. The specific outputs used as covariates are total evapotranspiration [kg/m^2], potential latent heat flux [W/m^2], soil moisture from the top soil layer with a thickness of 10 cm [kg/m^2], surface runoff [kg/m^2], temperature of the top soil layer [K], leaf area index [-], snow depth [m], snow melt [kg/m^2], snow cover fraction [-], and water equivalent of accumulated snow depth [kg/m^2]. All datasets are freely available to download from NASA's Goddard Earth Sciences Data and Information Services Center (GES DISC)(https://disc.gsfc.nasa.gov/).

Data type	Description	Time	Units	Spatial resolution	Temporal resolution
Precipitation[30,66]	Hourly total accumulated precipitation calculated using gauge-only daily precipitation analyzed by NCEP's Climate Prediction Center. NLDAS-2 Forcing file A	01/01/2000-06/30/2022	kg m^{-2}	0.125°	1-hour
Potential Evapotranspiration[67,68]	Evaporation occurring over a free water surface assuming unlimited water supply and accounting for aerodynamic resistance and the surface energy budget. NLDAS-2 Forcing file A	01/01/2000-06/30/2022	kg m^{-2}	0.125°	1-hour
Total Evapotranspiration[31,69]	Total land evapotranspiration as the sum of evaporation from bare soil, canopy interception, and transpiration from plants. NLDAS-2 Noah LSM	01/01/2000-06/30/2022	kg m^{-2}	0.125°	1-hour
Potential Latent Heat Flux[30,66]	Potential evapotranspiration in energy units calculated as the sum of evaporation from bare soil and canopy interception, transpiration from the canopy, and sublimation from snowpack. NLDAS-2 Noah LSM	01/01/2000-06/30/2022	W m^{-2}	0.125°	1-hour
Soil Moisture Content[30,31,66,70]	Amount of water held in 4 distinct soil layers (0–10 cm, 10–40 cm, 40–100, and 100–200 cm) estimated from the Noah land-surface model using coupled energy and water balances. NLDAS-2 Noah LSM	01/01/2000-06/30/2022	kg m^{-2}	0.125°	1-hour
Surface Runoff[30,31,66,70]	Accumulated liquid water on the land surface from precipitation that is not infiltrated into soils. NLDAS-2 Noah LSM	01/01/2000-06/30/2022	kg m^{-2}	0.125°	1-hour
Soil Temperature[30,31,66,71]	Temperature of 4 distinct soil layers (0–10 cm, 10–40 cm, 40–100 cm, and 100–200 cm) estimated from the Noah land-surface model. NLDAS-2 Noah LSM	01/01/2000-06/30/2022	K	0.125°	1-hour
Leaf Area Index[30,31,66]	Unitless parameter that quantifies vegetation canopy density as the ratio the one-sided area of leaf material in the canopy per unit ground area. NLDAS-2 Noah LSM	01/01/2000-06/30/2022	unitless	0.125°	1-hour
Snow Depth[33,72]	Instantaneous depth of snow accumulated on the land surface estimated from the Noah land-surface model. NLDAS-2 Noah LSM	01/01/2000-06/30/2022	m	0.125°	1-hour
Snow Melt[33,72]	Instantaneous amount of liquid water produced by snow melt on the land surface estimated from the Noah land-surface model. NLDAS-2 Noah LSM	01/01/2000-06/30/2022	kg m^{-2}	0.125°	1-hour
Snow Cover[33,72]	Fraction of land pixel covered with snow computed as a non-linear function of snow water equivalent (SWE) utilizing a generalized snow depletion curve. NLDAS-2 Noah LSM	01/01/2000-06/30/2022	unitless	0.125°	1-hour
Water Equivalent of Accumulated Snow Depth[33,72]	Equivalent amount of liquid water stored in snowpack estimated from the Noah land-surface model. NLDAS-2 Noah LSM	01/01/2000-06/30/2022	kg m^{-2}	0.125°	1-hour

Table 1. Data sources for all NLDAS-2 climate variables combined with USDM data.

We selected variables from the NLDAS-2 dataset that are strongly linked to hydrological and meteorological drought. Total evapotranspiration and potential latent heat flux provide estimates of how much water is leaving the land surface to the atmosphere, while surface runoff characterizes how much how is being transported out of a given grid cell. Soil moisture quantifies water storage within the grid cell, and soil temperature can indicate heat stress associated with drought. Leaf area index (LAI) describes current plant growth stage, which can be disrupted during severe drought events[35]. LAI predicted from the Noah land-surface model and is spatially and seasonally varying. It depends on land cover type and green vegetation cover fraction[34]. Snow depth, snow melt, snow cover fraction, and the water equivalent of accumulated snow depth are included as many regions in the Western US depend on the slow release of water from snow to sustain periods of low rainfall during the warm season. It is important to note that potential evapotranspiration and potential latent heat flux are essentially the same variable, as latent heat equals evapotranspiration multiplied by the latent heat of vaporization. One would only want to use at most one of these in a statistical model of these data, however we include both as different research communities have different preferences around which variable to use. Similary, snow depth and the water equivalent of accumulated snow depth provide similar information, and only one of these variables should be used in a statistical model.

Hourly NLDAS-2 Forcing File A and Noah land-surface model datasets were downloaded for all of CONUS between January 1, 2000 and June 30, 2022 from the NASA GES DISC. To upscale the data from 0.125° × 0.125° to the common 0.5° spatial support, 16 pixels from the native NLDAS-2 resolution were aggregated to the 0.5 resolution (Fig. 7). For variables that were derived from the Noah land-surface model, water pixels were masked out (as land processes are not valid to approximate over water and not produced by the Noah model). We required that at least 70% of the 0.125° pixels (or 11 pixels total) must be land pixels in order to define and compute an upscaled value at the 0.5° resolution; otherwise, the upscaled grid cell was designated as a water pixel at the coarser scale. Then, each dataset was aggregated to a weekly timestep that aligned with the USDM. For all variables except precipitation, the data were aggregated by taking the average in space and time. For precipitation, the sum is used to aggregate the data in space and in time to the 0.5° and weekly resolutions to aid with interpretability (https://climatedataguide.ucar.edu/climate-tools/regridding-overview).

USGS Streamflow. Gridded data on observed streamflow is not readily available. Streamflow data is point-based and collected from gauges that are unequally distributed in space and have variable temporal coverage. Streamflow percentiles that represent flow magnitude relative to the historical gauge record have been used as a direct measure of hydrologic drought, where thresholds have been used to identify drought severity, duration, and frequency[36] and the extent to which these characteristics have changed over time[37]. Long-term

Data type	Description	Time	Units	Spatial resolution	Temporal resolution
Streamflow (https://waterdata.usgs.gov/nwis/sw)	In-site streamflow measurements obtained fromUSGS gauges	*m*^{3}/*s*	point	daily	
Pacific North American Pattern	A pattern of air pressure anomalies over the Pacific Ocean and North America	01/01/1950 - present	unitless	none	daily
North Atlantic Oscillation[73]	An atmospheric phenomena related to the difference in pressure at sea level between Iceland and the Azores High	01/01/1950 - present	unitless	none	daily
Arctic Oscillation[74–77]	An index computed from differences in sea-level pressure between anomalies in the Arctic and anomalies around 37–45° N	01/01/1950 - present	unitless	none	daily
El Niño Southern Oscillation[78]	An index computed from sea-surface temperature anomalies in a region off the coast of South America. Included in 7-day, 14-day, 29-day, and 84-day averages.	09/02/1981 - present	unitless	none	weekly

Table 2. Data sources for all streamflow and teleconnections climate variables combined with USDM data.

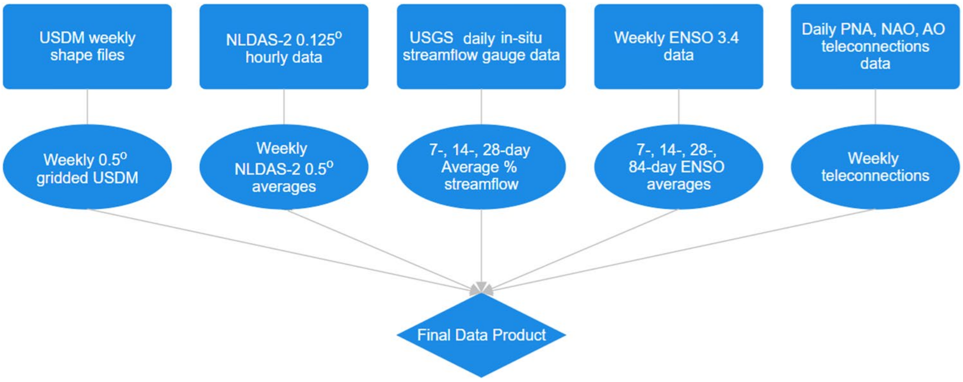
USDM weekly shape files
NLDAS-2 0.125° hourly data
USGS daily in-situ streamflow gauge data
Weekly ENSO 3.4 data
Daily PNA, NAO, AO teleconnections data
Weekly 0.5° gridded USDM
Weekly NLDAS-2 0.5° averages
7-, 14-, 28-day Average % streamflow
7-, 14-, 28-, 84-day ENSO averages
Weekly teleconnections
Final Data Product

Fig. 4 Flowchart of the processing of raw data sources to a combined final data product.

records of streamflow have been related to regional climate and hydrologic processes[38], indicating that this information could also inform more holistic drought assessments, such as the USDM drought severity metric. Many state-level drought advisory councils currently use point-based streamflow percentiles in their manual assessments that inform the USDM[39,40].

We derived 7-day, 14-day, and 28-day streamflow percentiles for each *in-situ* gauge to be consistent with how state-level drought advisory councils manually aggregate streamflow data to inform the USDM[39,40]. Using weekly to monthly averages allows for the incorporation of hydrologic processes that impact drought on different times scales; for example, groundwater infiltration resulting from a storm event is a very slow process compared to surface water runoff. The percentiles represent the magnitude of flow relative to what has been observed in the past and are used to assess drought severity, where areas with lower percentiles are often interpreted to have higher drought risk. Using this relative measure instead of an absolute measure helps managers combine and directly compare streamflow gauges with a wide range of flow magnitudes, enabling drought assessments on large spatial scales. To produce our gridded streamflow percentiles data product, we used the following data processing workflow broadly summarized as follows (and described in greater detail in the subsequent text, tables, and figures):

1. We downloaded all CONUS-scale *in-situ* streamflow data and identified the gauges that fall within each model grid.
2. We calculated the 7-day, 14-day, and 28-day streamflow averages as of each weekly time period in the USDM dataset, and converted each observation from flow (cfs) to a percentile based on its rank within a baseline empirical distribution. All streamflow observations from 1990 to 2020 were used to create the baseline distribution for each gauge. This 30-year time period is consistent with the World Meteorological Organization climatology standard normals.
3. For all grid cells and dates that did contain *in-situ* data, we calculated the arithmetic means of all *in-situ* observations within that grid cell.
4. For all grid cells and dates that did not contain *in-situ* data, we performed the following:

   (a). We identified the Hydrologic Unit Codes (HUC) that specify the watersheds at multiple scales–HUC-8, HUC-6, and HUC-4—associated with each grid.
   (b). For each watershed scale (e.g. HUC-8, HUC-6, or HUC-4) that contained the grid cell with missing data, we identified the gauges within that watershed that contain data. We calculated the arithmetic mean of the gauges within the watershed as well as the the inverse-distance-weighted arithmetic mean, using the distance between the gauges and the grid centroid.

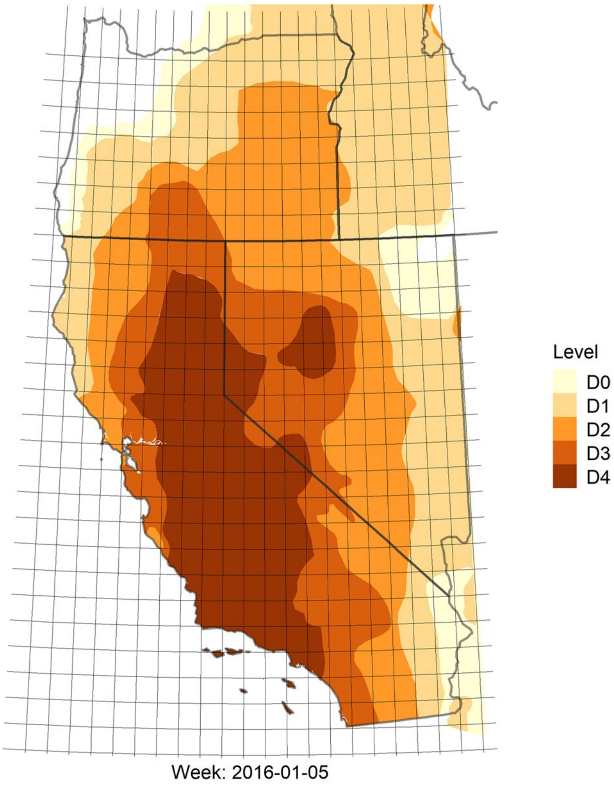

Fig. 5 Overlay of the 0.5 degree grid over a portion of the western United States, showing USDM values for 01/05/2016. Our method selected the centroid of each grid cell. Other regions and times demonstrate a similar correspondence between the centroid and the mode (computed spatially).

Streamflow data obtained in step 1 came from the USGS Water Data for the Nation web interface (https://waterdata.usgs.gov/nwis). Over thirty years of data (1990–2022) were downloaded as text files for all gauges (23,084 total) located in the 18 hydrologic regions within the CONUS. Figure 8 shows the number of gauges that lie within each grid and highlights variations in gauge density across the US. Many of the model grids (13.4%) do not contain a single streamflow gauge, and many of the gauges have significant data gaps within the 2000 to 2022 period of interest. Over this full twenty-two-year period, the average percent of model grids that do not have concurrent gauge data is 25.1%.

For steps 2 and 3, the daily average streamflow observations from each gauge were converted to 7-day, 14-day, and 28-day averages that align with the Tuesday of each week. If less than half of the daily observations were available over the averaging period, then NA (i.e. null) values were applied. Empirical frequency curves were produced for each gauge and for each averaging period by sorting the thirty-year time series, and then each 7-day, 14-day, and 28-day observation was converted to a percentile based on its position within the sorted data and the total number of valid observations. However, as shown in Figs. 8 and 9, there are many grids and time periods where no data is available within a grid.

Step 4 involves infilling missing grid cells whose streamflow averages were recorded as NA in steps 2 and 3. The simplest approach to filling the data gaps would be to search for gauges that are closest to missing grid values; however, streamflow is largely governed by the physical terrain, and so the measure of "closeness" should match physical terrain. To account for physical flow processes in the gap-filling procedure, the Standardized USGS Hydrologic Unit Code (HUC) watersheds were used to identify the gauges "near" a missing grid value. HUC watersheds are standardized and can be downloaded from the National Map (https://hydro.nationalmap.gov/). They are derived from digital elevation models of the landscape and delineate boundaries of land that drain to the same outlet point. The area within a HUC watershed is therefore hydrologically connected, and the associated streamflow magnitudes should be correlated within. Considering the distance between gauges varies substantially across the CONUS, grid centroids were matched with watersheds at three different scales – HUC-4, HUC-6, and HUC-8. Each of these scales is illustrated in relation to the *in-situ* gauges in Fig. 10.

Gaps within the gridded dataset were subsequently filled using the HUC/gauge match-up information according to the procedure outlined in step 4. Distances between each grid centroid and gauges contained within each watershed were calculated in decimal degrees using the latitude and longitude information associated with each gauge. These distances and flow percentiles were then queried to calculate the inverse distance weighted (IDW) mean, defined as

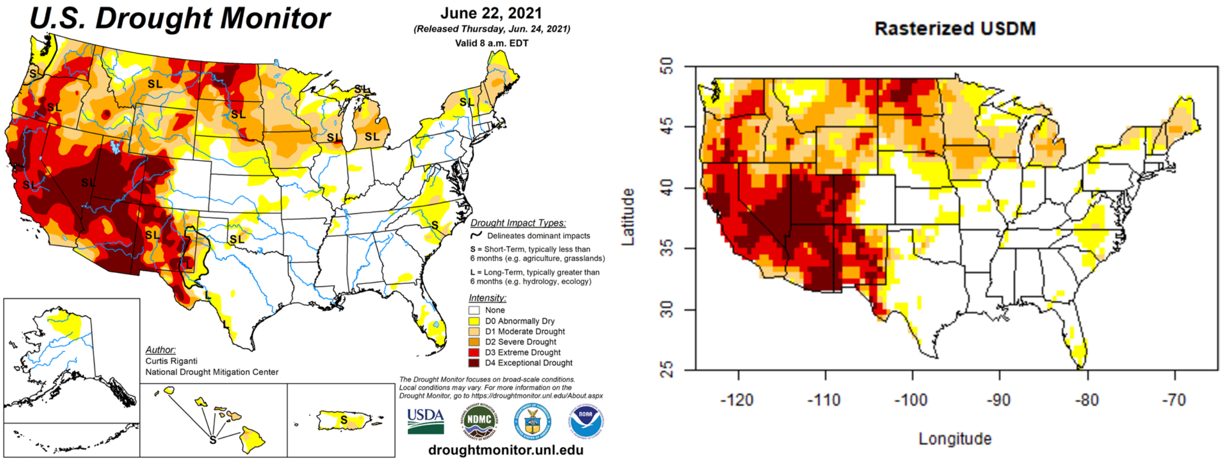
U.S. Drought Monitor
June 22, 2021
(Released Thursday, Jun. 24, 2021)
Valid 8 a.m. EDT
SL
S
L
Drought Impact Types:
Delineates dominant impacts
S = Short-Term, typically less than 6 months (e.g. agriculture, grasslands)
L = Long-Term, typically greater than 6 months (e.g. hydrology, ecology)
Intensity:
None
D0 Abnormally Dry
D1 Moderate Drought
D2 Severe Drought
D3 Extreme Drought
D4 Exceptional Drought
Author:
Curtis Riganti
National Drought Mitigation Center
The Drought Monitor focuses on broad-scale conditions. Local conditions may vary. For more information on the Drought Monitor, go to https://droughtmonitor.unl.edu/About.aspx
USDA
NDMC
NOAA
droughtmonitor.unl.edu
Rasterized USDM
Latitude
50
45
40
35
30
25
-120
-110
-100
-90
-80
-70
Longitude

Fig. 6 Left panel: Raw data for the U.S. Drought Monitor for June 22, 2021. Right panel: Discretized data from the same time period on a 0.5 degree spatial support of 3258 grid cells covering the contiguous US only.

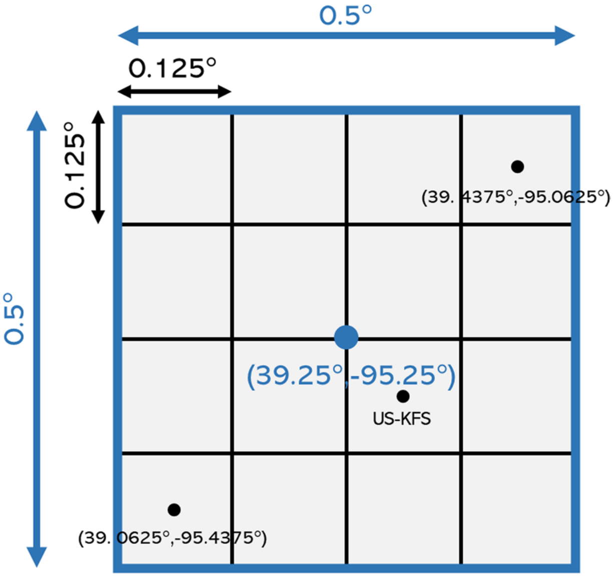

Fig. 7 Graphical representation of the upscaling procedure for the NLDAS-2 data for pixels surrounding the US Kansas Field Station (US-KFS) (39.0561°, −95.1907°). Sixteen pixels from the 0.125° native resolution are aggregated to the 0.5° grid. Specifically, for all variables except for precipitation, the average of the 0.125° pixels is computed and taken as the value across the entire 0.5° pixel. The sum of all 16 pixels is used for precipitation.

$$Q_{avg}=\frac{\sum_{i=1}^{n}Q_i d_i^{-1}}{\sum_{i=1}^{n}d_i^{-1}},$$

where Q_i is the flow percentile for gauge i, and d_i is the distance from the grid centroid for gauge i. We also separately computed an arithmetic mean for each missing value with no distance weighting. The IDW mean allows more weight to be placed on gauges that are closer to the grid centroid, an approach that is commonly applied to interpolate geospatial data[41–43] If this gap-filling procedure is limited to the finest (HUC-8) watershed scale, then the average percentage of missing values drops from 25.1% to 15.6% (based on the 7-day flows). If the HUC-6 and HUC-4 watershed scales are used to search for eligible gauges, then the average percentage of missing values drops further to 1.8% and 1.2%, respectively. The grids that do not have fill values at each of the watershed scales are highlighted in Fig. 9. At the coarsest scale, the only remaining missing grids are located in arid locations, including Southwest Texas along the border with Mexico, a small watershed that lies between the central California and Nevada border, a small region of Southern Colorado.

Teleconnections. Teleconnections data were obtained for four commonly used indices: Niño 3.4 (ENSO), the Pacific/ North American Pattern (PNA), the Arctic Oscillation (AO), and the North Atlantic Oscillation

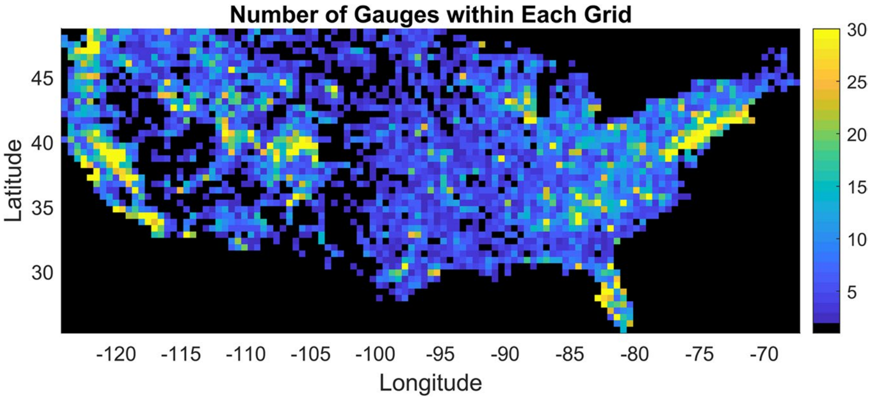

Fig. 8 Map of the common grid system showing the number of *in-situ* stream gauges that lie within each grid cell. Grids that are black do not have any *in-situ* gauges within their extent.

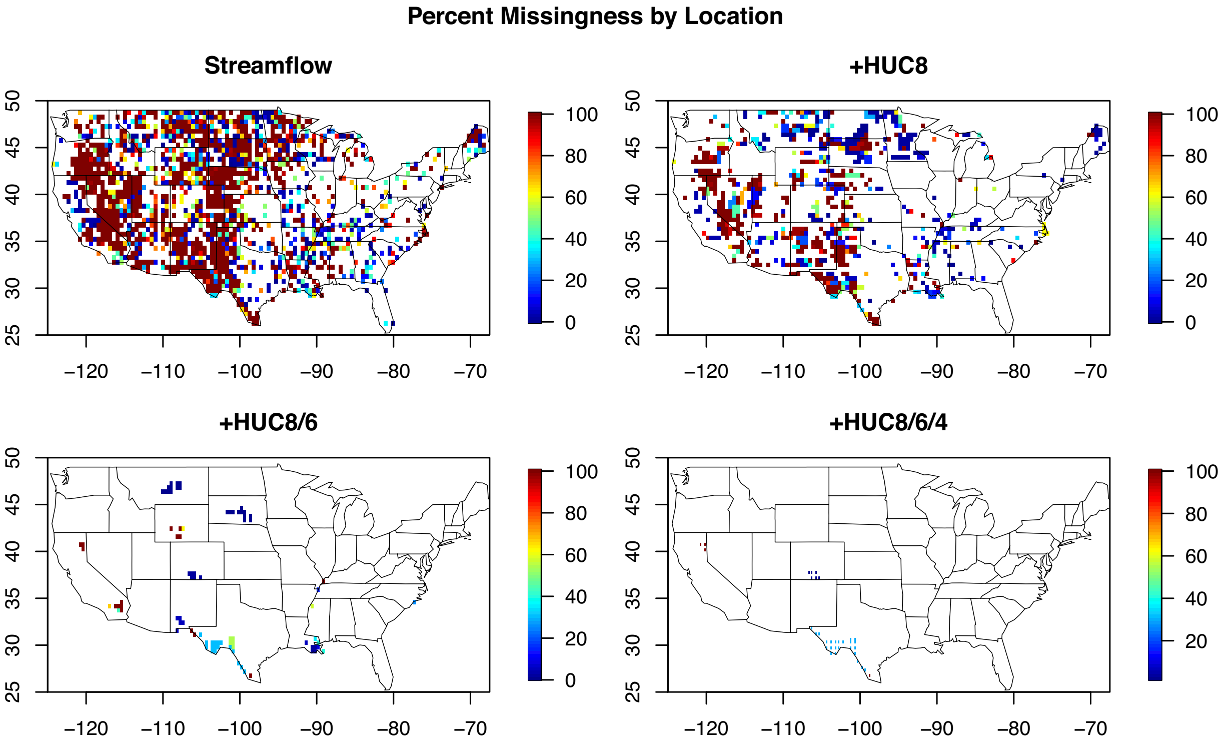

Fig. 9 Percent missingness for 7-day average streamflow data, 2000–2022. Only grid cells with some amount of missing data are shown in color. The top left panel shows baseline missingness for USGS streamflow data only. The other three panels infill missing data by using the first available averages from increasingly larger watersheds (HUC8, HUC6, HUC4) containing a grid cell missing any data.

(NAO). These were selected from those available at National Weather Service Climate Prediction Center (https://www.cpc.ncep.noaa.gov/). These four teleconnections indices are not spatially varying, so their values in these data do not differ by grid cell but only time. PNA, NAO, and AO data are available daily covering the entire study period at the National Weather Service Climate Prediction Center. We took 7-day running averages of each index that are reported on the Tuesday of each week to match the USDM data and align with the NLDAS-2 and streamflow data. Thus, dates are 1/4/2000, 1/11/2000, etc. ENSO was not available at a daily timescale, but was available at a weekly timescale (https://www.cpc.ncep.noaa.gov/data/indices/wksst9120.for). The date of each weekly index is only one day later that the USDM (so, 1/5/2000, 1/12/2000, etc.). Given high temporal autocorrelation of ENSO (see Fig. 12), we disregarded this mismatch by a single day, and attributed each weekly index value to the prior day to match the temporal support of all other variables in the data. Figures 11 and 12 show five years of each of the raw teleconnections indices, as well as autocorrelation functions for each, which are defined as

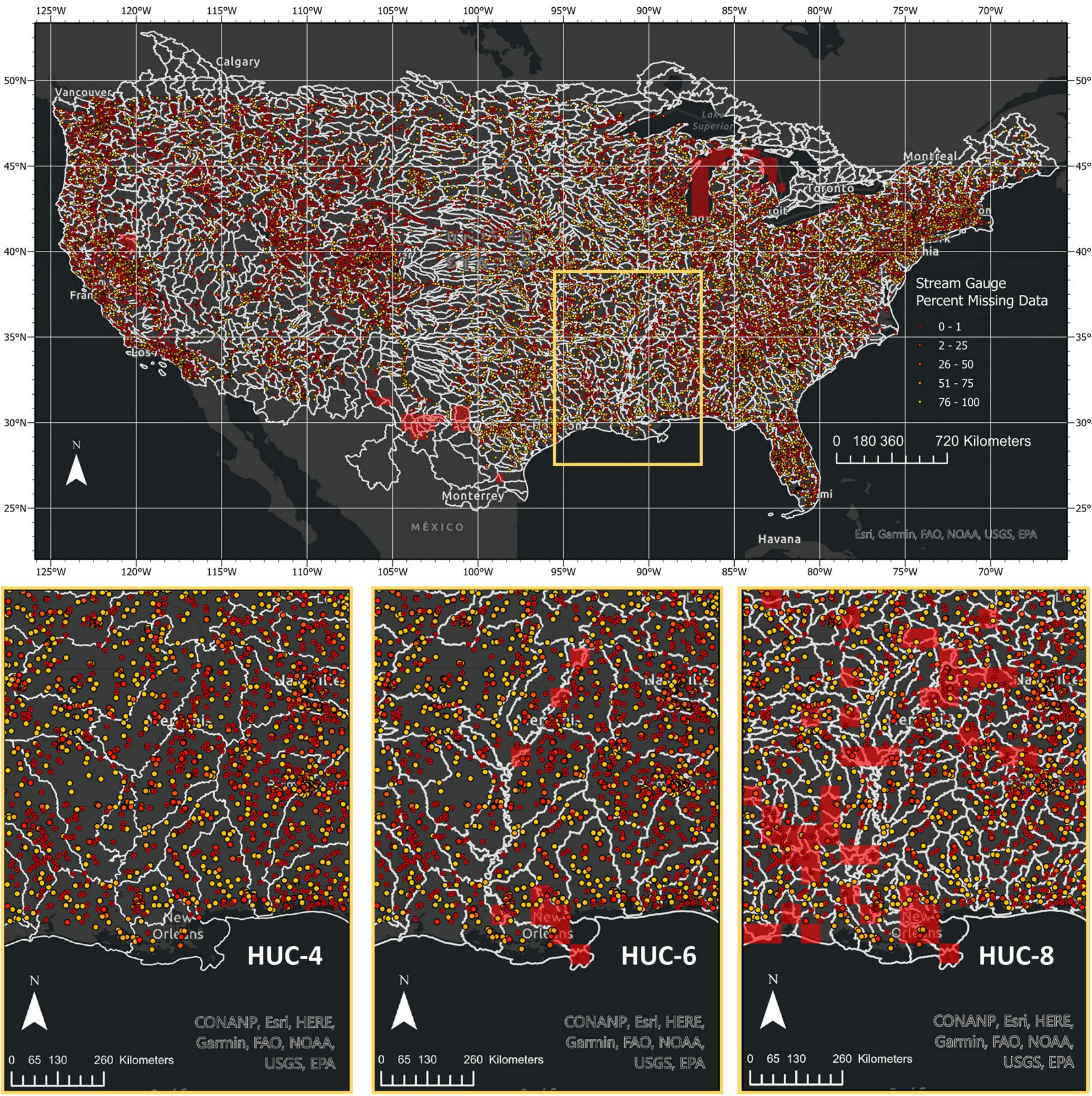

Fig. 10 Top image: All *in-situ* stream gauges within the CONUS with HUC-4 watershed boundaries (light gray). The colors of the stream gauges range from red to yellow, indicating the percentage of valid data during the 2000 to 2022 period. Grids that do not have any *in-situ* data at the HUC-4 scale are highlighted in red. Bottom images: Close-up view of the Southern US that shows stream gauges within HUC-4, HUC-6, and HUC-8 watershed boundaries, and highlights (red) grids where *in-situ* data could not be identified within the watershed using the corresponding HUC scale.

$$\gamma_k = \frac{Cov(X_t, X_{t+k})}{Var(X_t)},$$

where $k = 1, 2, \ldots$ defines the length of the weekly time lag for variable X. Given the prominence of ENSO in climate research in North America, we also provide the 14-day, 28-day and 84-day averages of ENSO to allow the user to consider longer time frames.

Data Records

The resulting data product is available on DataDryad at https://doi.org/10.5061/dryad.g1jwstqw7[44]. These data are stored in a flat csv file, with 45 columns whose names and descriptions are shown in Table 3. This dataset has one row for each grid/time combination taken across all unique 3258 grid cells and 1174 time periods. Sample maps of many variables are shown in Figs. 6, 13, 14 and 15.

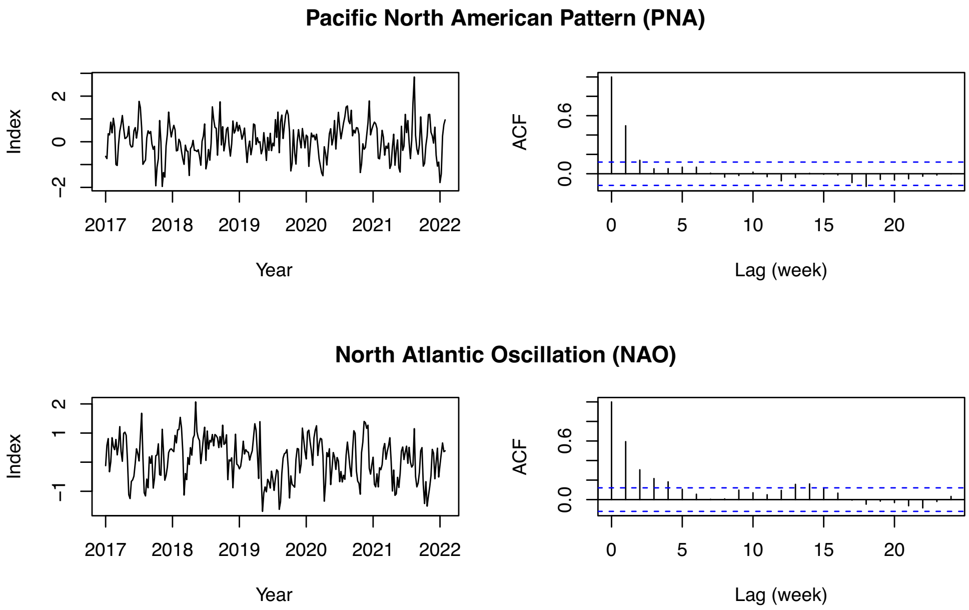

Fig. 11 Left panels: Raw data for the two teleconnections from 2017–2022. Right panel: Autocorrelation functions of the two teleconnections.

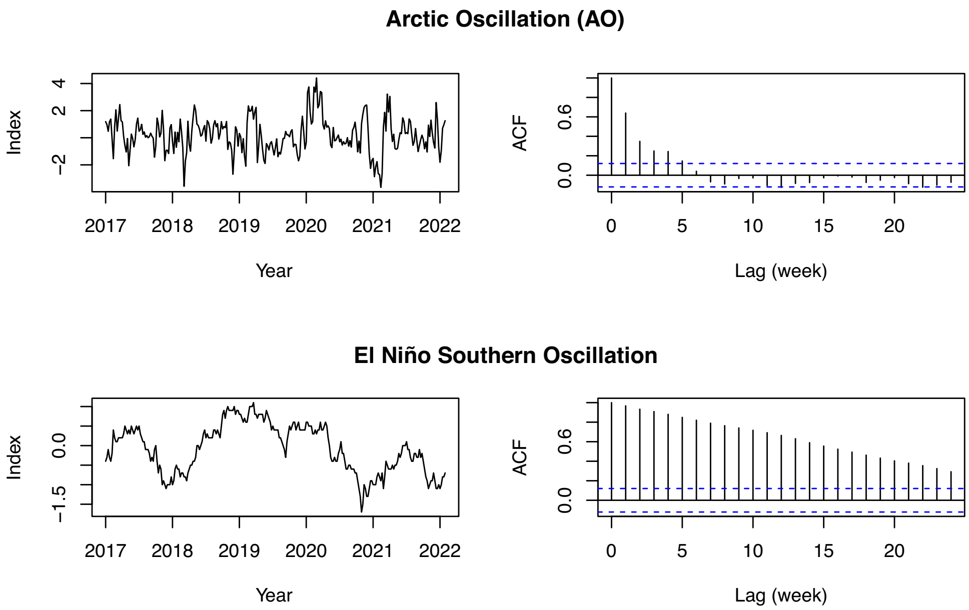

Fig. 12 Left panels: Raw data for the two teleconnections from 2017–2022. Right panel: Autocorrelation functions of the two teleconnections.

NAs are used to indicate any missing or undefined entries throughout, and different variables have different patterns of missingness. In the NLDAS-2 data files, data representing land surface processes were labeled as undefined if they occurred over water. In the upscaled dataset, grid cell R25 at (−112.75, 41.25) covers Great Salt Lake and contains less than 70% land, so it was accordingly removed from the full dataset. There are no missing values for drought. Missing data for the streamflow variables is shown in Fig. 9 and described in detail in the previous section. No teleconnections values are missing for any time period.

Technical Validation

Upscaling of NLDAS-2 Land Surface Variables. We investigate the variables of evapotranspiration (EVP), leaf area index (LAI), precipitation (APCP), and soil moisture (SOILM) to validate the spatial and temporal upscaling of the NLDAS-2 land surface data from its native 0.125° spatial and hourly temporal resolution to the spatially and temporally aggregated weekly grid data. Figures 14 and 15 demonstrate the loss in heterogeneity

Variable Name	Description
time	Time of the record in YYYYMMDD format
grid	Alphanumeric unique ID given to each grid cell
lon	Longitude of the grid centroid, in degrees
lat	Latitude of the grid centroid, in degrees
drought	Ordinal value of drought, with 0, D0, D1, D2, D3, and D4 as possible values
apcp	Accumulated precipitation
evp	Total evapotranspiration
pevap	Potential evapotranspiration
pevpr	Potential Latent Heat Flux
soilm	Soil Moisture
ssrun	Soil Runoff
tsoil	Soil Temperature
lai	Leaf Area Index
snod	Snow Depth
snom	Snow Melt
snowc	Snow Cover Fraction
weasd	Water Equivalent of Accumulated Snow Depth
percFlow7day	7-day average streamflow percentile. NA if missing.
percFlow14day	14-day average streamflow percentile. NA if missing.
percFlow28day	28-day average streamflow percentile. NA if missing.
avg.HUC8.7day	Average of 7-day average streamflow values taken over all gauges in the HUC 8 watershed containing the grid cell. Only computed if percFlow7day is NA.
avgDist.HUC8.7day	Inverse distance-weighted average of 7-day average streamflow values taken over all gauges in the HUC 8 watershed containing the grid cell. Only computed if percFlow7day is NA.
avg.HUC6.7day	Average of 7-day average streamflow values taken over all gauges in the HUC 6 watershed containing the grid cell. Only computed if both percFlow7day and avg.HUC8.7day are NA.
avgDist.HUC6.7day	Inverse distance-weighted average of 7-day average streamflow values taken over all gauges in the HUC 8 watershed containing the grid cell. Only computed if both percFlow7day and avg.HUC8.7day are NA.
avg.HUC4.7day	Average of 7-day average streamflow values taken over all gauges in the HUC 4 watershed containing the grid cell. Only computed if percFlow7day, avg.HUC8.7day and avg.HUC6.7day are all NA.
avgDist.HUC4.7day	Inverse distance-weighted average of 7-day average streamflow values taken over all gauges in the HUC 4 watershed containing the grid cell. Only computed if percFlow7day, avg.HUC8.7day and avg.HUC6.7day are all NA.
avg.HUC8.14day	Analagous to the six variables listed above, but computed using 14-day average streamflow.
avgDist.HUC8.14day	
avg.HUC6.14day	
avgDist.HUC6.14day	
avg.HUC4.14day	
avgDist.HUC4.14day	
avg.HUC8.28day	Aanalagous to the six variables listed above, but computed using 28-day average streamflow.
avgDist.HUC8.28day	
avg.HUC6.28day	
avgDist.HUC6.28day	
avg.HUC4.28day	
avgDist.HUC4.28day	
pna	Pacific North American pattern
nao	North Atlantic Oscillation
ao	Arctic Oscillation
enso, enso14, enso28, enso84	El Niño Southern Oscillation at 7-day, 14-day, 28-day, and 84-day averages

Table 3. Descriptions of the 45 variables in the drought data product.

reflecting complexity in the land surface when upscaling from hourly to weekly (top row to bottom for each variable) as well as 0.125 degrees to 0.5 degrees (left column to right column for each variable). The consequences of this loss of heterogeneity can be minimal if analyses and conclusions are restricted to the coarser weekly and 0.5 degree resolution. Larger regional patterns of low evaporation rates over arid regions of the US and high rates over humid regions remain.

Aggregating the data in time from hourly amounts to weekly averages removes the diurnal variability for some variables (i.e. EVP) that are strongly controlled by incoming solar radiation (Fig. 16). However, the weekly averages follow the expected seasonal variability and are higher in magnitude during the summer, when daily

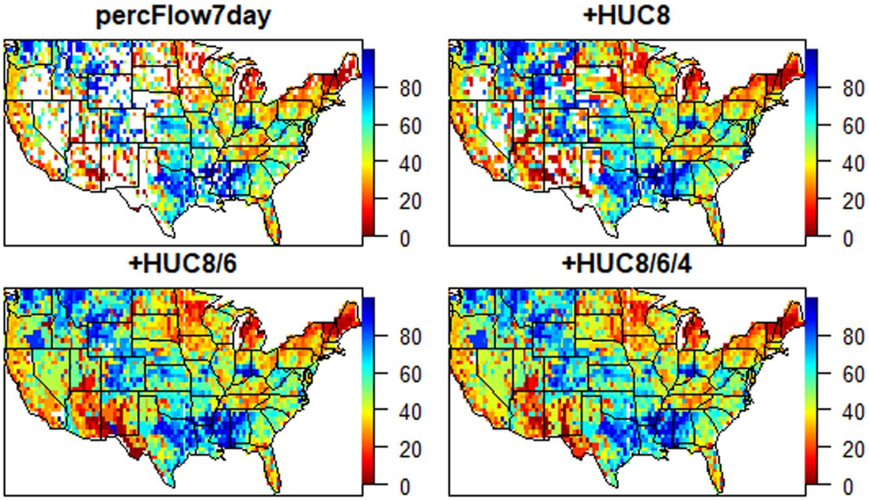

Fig. 13 Example streamflow data from the week of June 22, 2021, which is the same date shown in Fig. 6. The top left panel shows the single variable `percFlow7day`, the percentage of streamflow over 7 days for all gauges within each 0.5 degree grid cell. Since some grid cells have no gauges, these are missing and showin in white in the top left figure. The next three successive panels fill in missing values using values averaged over HUC8 watersheds (`avg.HUC8.7day`), both HUC8 and HUC6 watersheds (`avg.HUC8.7day` and `avg.HUC6.7day`), and HUC8, HUC6 and HUC4 watersheds (`avg.HUC8.7day, avg.HUC6.7day` and `avg.HUC4.7day`).

maximum evapotranspiration is high, and near zero in the winter when little energy is available for evapotranspiration to occur (Fig. 16 a,c). For other variables which do not follow diurnal variability, upscaling can aggregate roughly stationary time series values across space (Fig. 17), or smooth over minor temporal variability in variables with relatively low spatial variation at 0.125 degrees (Fig. 18).

Streamflow. To validate the point-to-grid conversion of the streamflow percentiles, we explored time series of individual grids and quantified the variability between the gridded values and individual gauges that were used to estimate the grid mean. The goal of these analyses was to check that the temporal variations of the grid means are similar to those of the individual gauges, and to assess the variability of the gauge values that were used for gridded estimates. Time series of the 7-day streamflow percentiles for two grids are shown in Fig. 19. The top image shows the values for a grid that overlays the Colorado River near the Palo Verde Dam along the Arizona and California border (gauge site ID 9429100) that contains 5 gauges with valid data. The bottom image shows the values for a grid that overlays the Mississippi River near St. Louis, IL (gauge site ID 7010000) that contains 33 gauges with valid data.

The variability between gauge values for all grids in the CONUS was explored by calculating the mean of the standard deviation of the residuals, where the residuals are the differences between the averaged grid value and each gauge that was used to estimate the grid value. The 7-day flow percentile values were used in this exercise. These calculations result in a single value for each grid that represents the mean variability over time, which can be presented as a gauge data variability map, shown in Fig. 20. If a grid contains a value of zero, then a single gauge was used to calculate the grid mean.

We expected that the data variability would be greater for grids that used the HUC watershed extents as opposed to those that contained *in-situ* data within its extents. This hypothesis was tested by splitting the mean residuals between these two grid types and plotting the resulting distributions, shown in Fig. 21. The mean standard deviation of residuals for the grids that contain gauges within the extents is 4.37%, and 5.54% for the grids that used the HUC watershed extents. The distribution of this metric is somewhat bi-modal for the grids that were filled using the HUC watersheds, with peaks close to 0% and 10%. The data variability is slightly greater for grids that used the watershed-based filling procedure. However, the overall variability across the CONUS is low, with the exception of a few grids that have values between 15% to 30%. This result and the time series plots suggest that the point-to-grid conversion was performed correctly and that the representative grid values are capturing the multi-temporal behavior of the individual gauges.

To conceptualize this variability measure in the context of streamflow magnitudes in major US rivers, it was applied to the gauge data from the Colorado and Mississippi Rivers, shown in Fig. 19. The mean of the residuals for the grid corresponding with the Colorado gauge is 9.75%, on the high end of the distribution. The mean streamflow recorded by this gauge from 2000 to 2021 was 7,200 cfs, so conceptually the 9.75% variability introduces an uncertainty of about 700 cfs, on average. Similarly, the mean flow of the Mississippi River from 2000 to 2021 at St. Louis, IL was 234,418 cfs, and the mean of the residuals is 8.75%, translating to an estimated average uncertainty of 20,512 cfs.

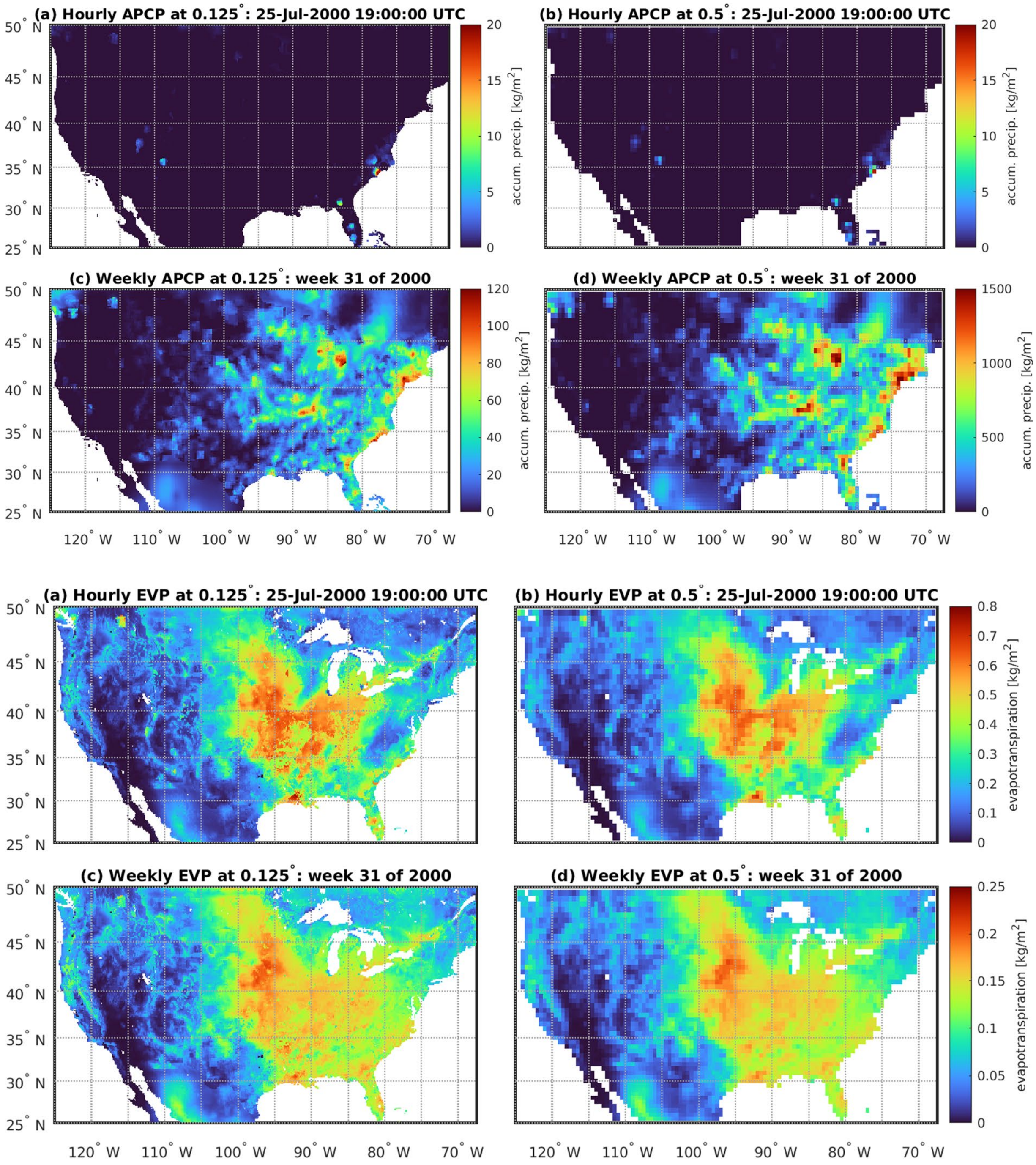
(a) Hourly APCP at 0.125°: 25-Jul-2000 19:00:00 UTC
(b) Hourly APCP at 0.5°: 25-Jul-2000 19:00:00 UTC
(c) Weekly APCP at 0.125°: week 31 of 2000
(d) Weekly APCP at 0.5°: week 31 of 2000
accum. precip. [kg/m^2]
0
5
10
15
20
0
20
40
60
80
100
120
500
1000
1500
50° N
45° N
40° N
35° N
30° N
25° N
120° W
110° W
100° W
90° W
80° W
70° W
(a) Hourly EVP at 0.125°: 25-Jul-2000 19:00:00 UTC
(b) Hourly EVP at 0.5°: 25-Jul-2000 19:00:00 UTC
(c) Weekly EVP at 0.125°: week 31 of 2000
(d) Weekly EVP at 0.5°: week 31 of 2000
evapotranspiration [kg/m^2]
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.05
0.1
0.15
0.2
0.25

Fig. 14 Comparison of raw and upscaled data for APCP (top 4 images) and EVP (bottom 4 images). For each variable, images show: (**a**) raw NLDAS-2 data at the hourly and 0.125 degree resolution; (**b**) upscaled hourly data at 0.5 deg resolution; (**c**) upscaled weekly data at the 0.125 degree resolution; and (**d**) upscaled weekly data at the 0.5 degree resolution. A threshold of 70% land pixels is used at the native resolution to define a land pixel at the upscaled resolution. White areas indicate water pixels.

Usage Notes

Time series models for ordinal drought. These data are indexed in three dimensions—space (grid cells with centroids indexed by latitude and longitude) as well as time. If one considers a single fixed grid cell, the drought status at that location can be represented as Y_t, $t = 1, \ldots, T$ for weekly drought status, where $Y_t \in 1, \ldots, 6$ representing each of the six USDM drought levels. All covariates measured at the same location are $\mathbf{X}_t = (X_1, \ldots, X_j)_t$, $t = 1, \ldots, T$. This setup permits time series modeling of ordinal drought level in terms of all previous drought history and all previous covariate history. If we define a vector $\mathbf{W}_t$ which contains functions of some subset of prior response values Y_{t-1}, Y_{t-2}, … and covariates values known through time t, $(\mathbf{X}_t, \mathbf{X}_{t-1}, \ldots)$, one can set up time series models for ordinal data such as the cumulative odds time series model[45],

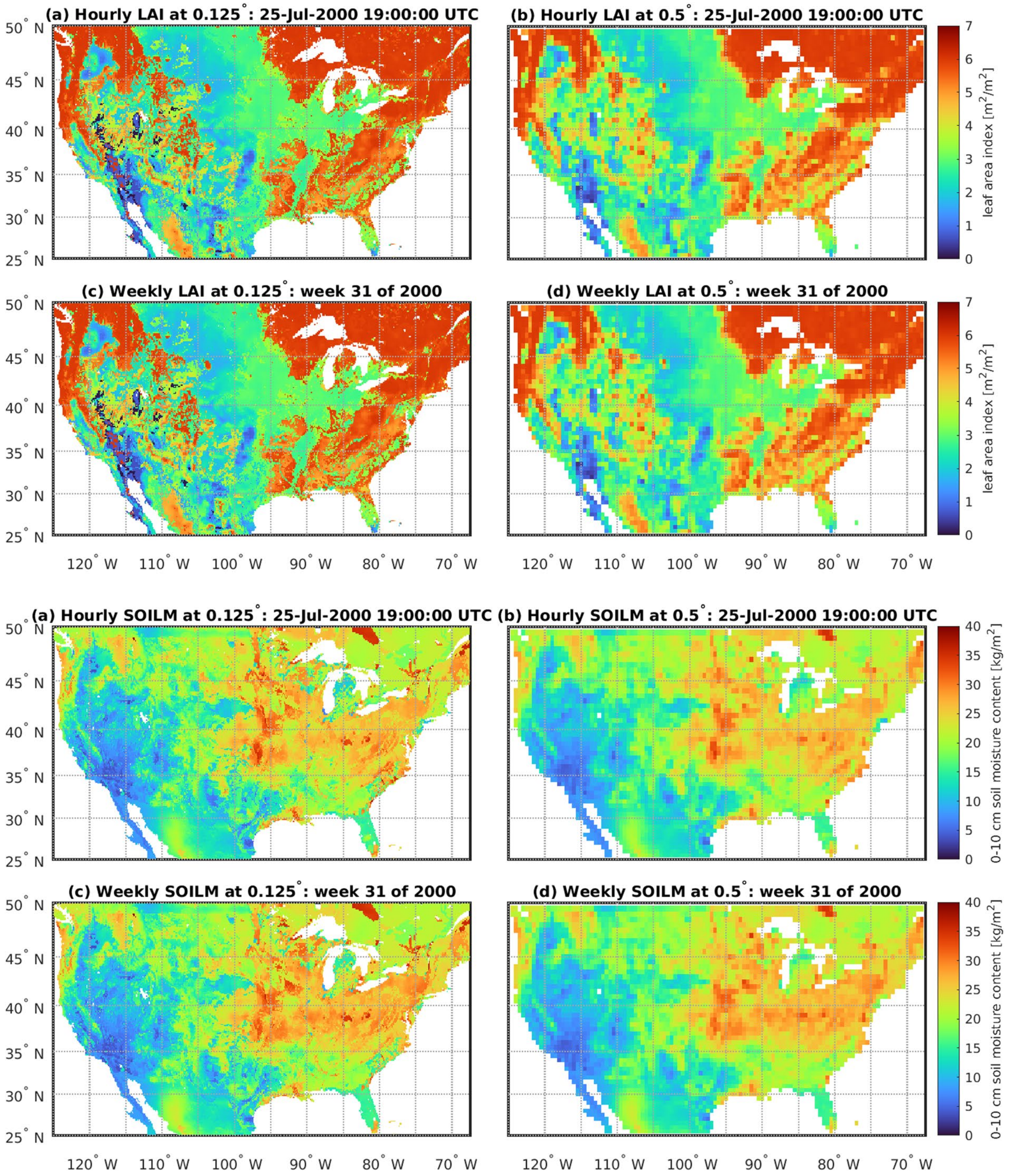
(a) Hourly LAI at 0.125°: 25-Jul-2000 19:00:00 UTC
(b) Hourly LAI at 0.5°: 25-Jul-2000 19:00:00 UTC
(c) Weekly LAI at 0.125°: week 31 of 2000
(d) Weekly LAI at 0.5°: week 31 of 2000
leaf area index [m²/m²]
0
1
2
3
4
5
6
7
(a) Hourly SOILM at 0.125°: 25-Jul-2000 19:00:00 UTC
(b) Hourly SOILM at 0.5°: 25-Jul-2000 19:00:00 UTC
(c) Weekly SOILM at 0.125°: week 31 of 2000
(d) Weekly SOILM at 0.5°: week 31 of 2000
0-10 cm soil moisture content [kg/m²]
0
5
10
15
20
25
30
35
40
50° N
45° N
40° N
35° N
30° N
25° N
120° W
110° W
100° W
90° W
80° W
70° W

Fig. 15 Comparison of raw and upscaled data for LAI (top 4 images) and SOILM (bottom 4 images). For each variable, images show: (**a**) raw NLDAS-2 data at the hourly and 0.125 degree resolution; (**b**) upscaled hourly data at 0.5 deg resolution; (**c**) upscaled weekly data at the 0.125 degree resolution; and (**d**) upscaled weekly data at the 0.5 degree resolution. A threshold of 70% land pixels is used at the native resolution to define a land pixel at the upscaled resolution. White areas indicate water pixels.

$$\log\left(\frac{P(Y_t \leq j)}{P(Y_t > j)}\right) = \beta_{0,j} + \mathbf{W}\beta, \; j = 1, \ldots, J.$$

One could also choose the standard normal cumulative distribution function $\Phi(\cdot)$ to yield a familiar probit link between cumulative probabilities and covariates as

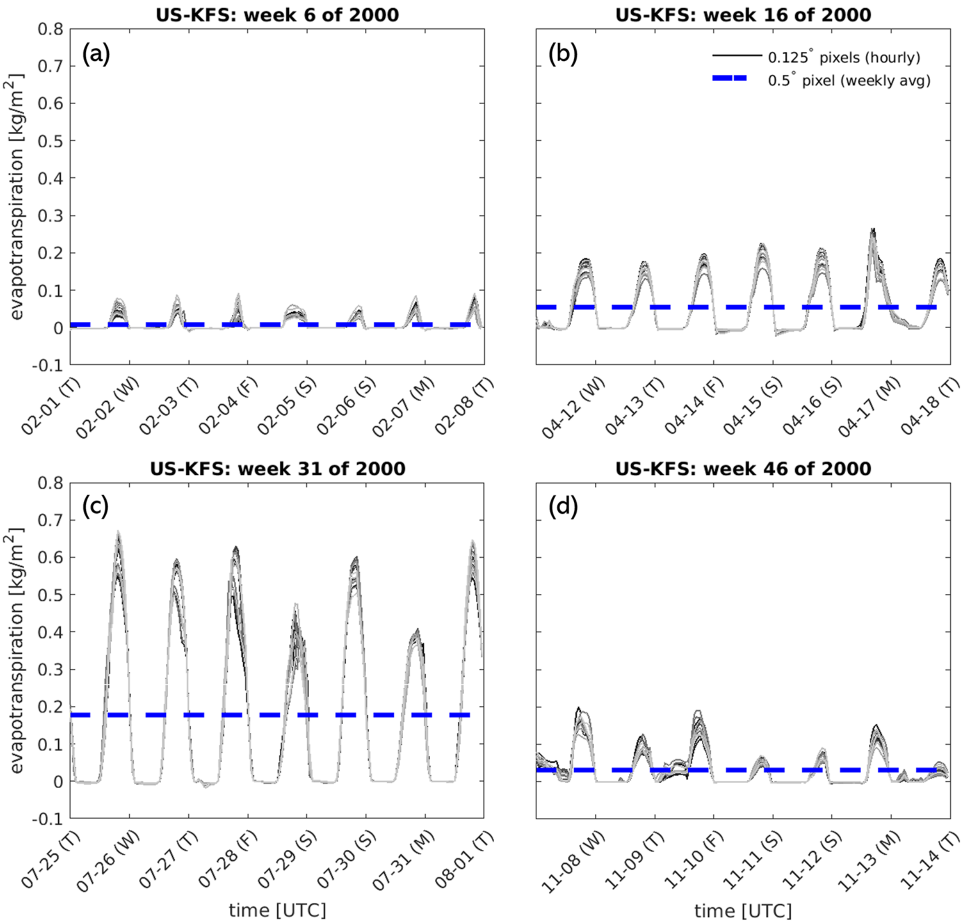

Fig. 16 Comparison of time series of raw EVP data at 0.125 deg data to upscaled 0.5 deg data for a single pixel representing the Kansas Field Station (39.0561°, −95.1907°) during: (**a**) winter, (**b**) spring, (**c**) summer, and (**d**) fall. All 16 pixels at the native spatial resolution that contribute to the upscaled 0.5 deg grid cell are shown as individual gray lines. The weekly average for the 0.5 deg grid cell is shown as a dashed blue line.

$$\text{probit}(P(Y_t \le j)) = \beta_{0,j} + \mathbf{W}\beta, \; j = 1, \ldots, J.$$

These time series models could be used to forecast future ordinal drought levels with full accounting of uncertainty. The R package `ordinal` can be used to fit cumulative link models for ordinal regression with both fixed and random effects[46]. For more detail on fitting such models, see Fokianos *et al.*[47]. or Weiss *et al.*[48].

As an extension, one can also introduce a latent continuous variable $Z_t = \mu_t + \varepsilon_t$, where ε_t is a sequence of independent and identically distributed random variables with continuous cumulative distribution function F, and μ_t can be parameterized in terms of explanatory variables, random effects, and past values of the process $Z_{t-1}, Z_{t-2}, \ldots$ Partitioning the support of the latent variable with cutoffs $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_{J-1} < \alpha_J = \infty$ can facilitate modeling the ordinal data. One can recover the observed ordinal variable as

$$Y_t = \sum_{j=1}^{J} j \cdot I(\alpha_{j-1} < Z_t \le \alpha_j),$$

where $I(\cdot)$ denotes the indicator function, and j indexes the J possible ordered categories of drought. Marginalizing over Z, some of these models are identical to the cumulative odds or probit models described above. However this latent variable formulation also permits a wider class of time series models and can have computational advantages with model fitting[49].

Computationally efficient spatio-temporal drought modeling. The prepared data are also indexed by latitude and longitude and therefore hold spatial dependence in addition to temporal dependence described above. Statistical modeling of such spatio-temporal ordinal data is challenging due to the computational expense required with a large number of locations and/or time periods. Feng *et al.*[50]. proposed a composite likelihood estimate approach as one way to address this computational expense, and a spatial-only model of ordinal data which was implemented in the `clespr` package in R[50,51]. Brewer *et al.* (2014) and Higgs and Hoeting[52] implemented Bayesian versions of the spatial ordinal model[53], which can require a large computational cost to achieve

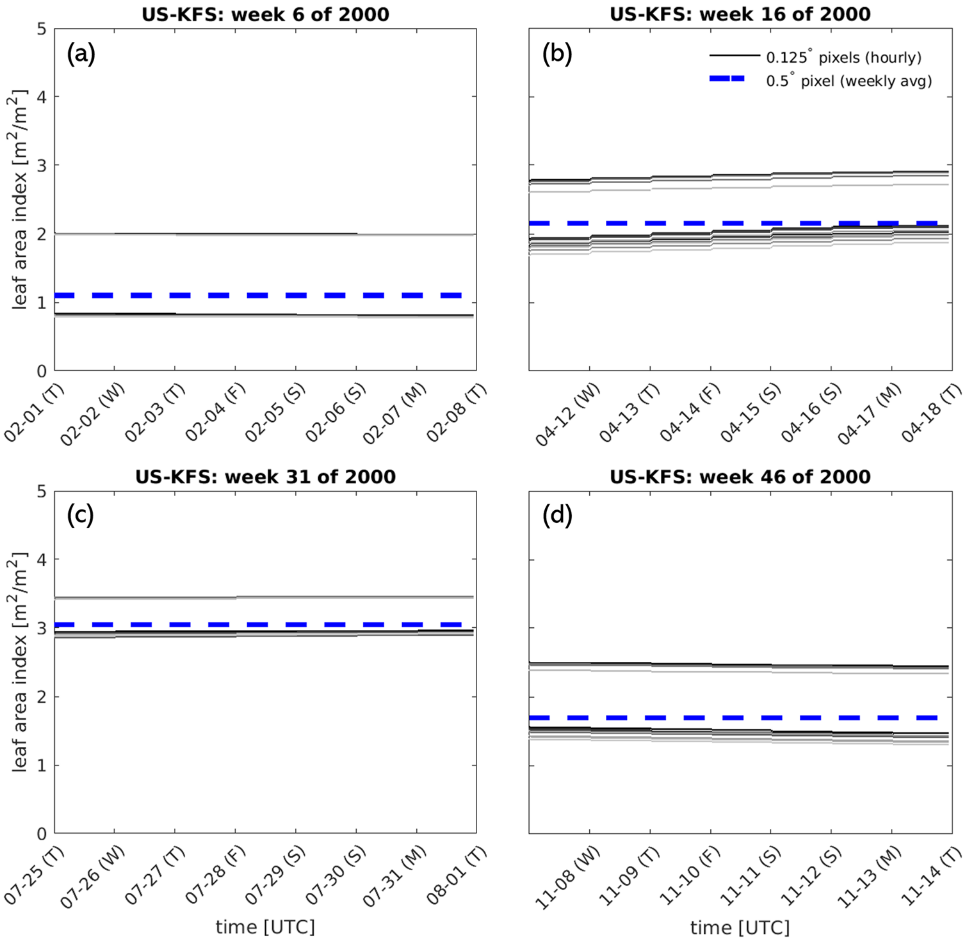

Fig. 17 Comparison of time series of raw LAI data at 0.125 deg data to upscaled 0.5 deg data for a single pixel representing the Kansas Field Station (39.0561°, −95.1907°) during: (**a**) winter, (**b**) spring, (**c**) summer, and (**d**) fall. All 16 pixels at the native spatial resolution that contribute to the upscaled 0.5 deg grid cell are shown as individual gray lines. The weekly average for the 0.5 deg grid cell is shown as a dashed blue line.

mixing. Schliep and Hoeting[54] proposed a data augmentation approach to improve the performance of MCMC algorithms for fitting Bayesian versions of spatial ordinal models[54]. Schliep *et al.* (2016) extended these models to the spatio-temporal ordinal data setting[49].

Each of the approaches mentioned above is limited by the computational complexity of spatio-temporal ordinal data, but this can be partially reduced through a common spatial support and parsimonious statistical modeling. In this data set, all variables have been discretized to the same spatio-temporal grid which enables more parsimonious statistical modeling. While change of support methods exist for situations where the response and explanatory variables are misaligned, these approaches add computational complexity. Spatio-temporal statistical modeling of this full data set is already computationally challenging due to the size of the data, so it is advantageous to work with aligned data.

The aforementioned approaches to modeling spatial or spatio-temporal ordinal data can be implemented for continuous space, or geostatistical data, and also for discrete space, or areal data. Our final gridded data set contains the latitude/longitude coordinates of the grid centroids. However, these data are suited for areal spatial modeling techniques and not geostatistical modeling as the values have been processed as previously described so that the value is a representation of the variable across the entire grid cell and not the measured value at the exact point. Areal spatial models, such as conditional autoregressive models or simultaneous autoregressive models, require the specification of a neighborhood weight matrix, or a known matrix that assigns a numeric value to all pairs of locations[55]. A common choice for the weight matrix is to assign the value of 1 if two areal units share a border and a value of 0 otherwise, thus indicating whether or not two locations are "neighbors." One attractive feature of areal models is that the precision matrix is sparse, which significantly eases computational expense generally associated with large spatial data sets. More specifically, statistical inference with a geostatistical spatio-temporal model has computational complexity of $\mathcal{O}(N^3T)$, where N is the number of spatial locations and T the time periods; whereas approaches based on first-order Markov random fields, such as the conditional autoregressive model, require $\mathcal{O}(N^{3/2}T)$, operations[56].

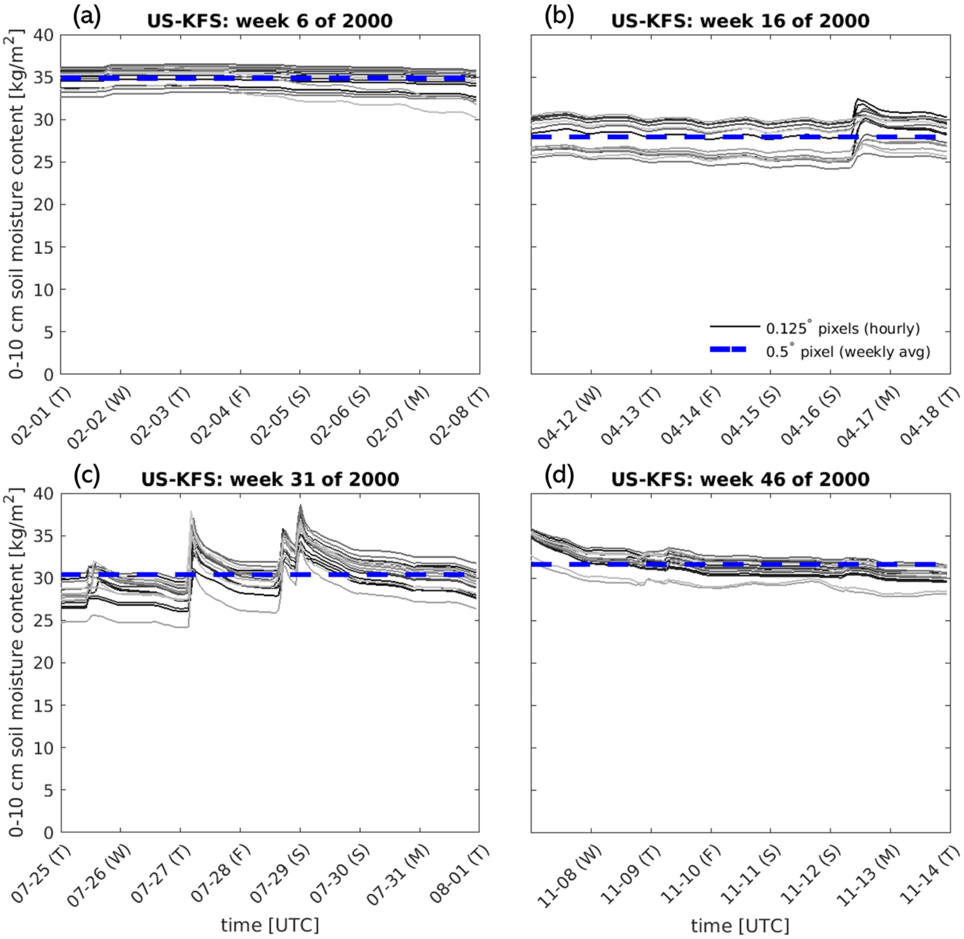

Fig. 18 Comparison of time series of raw SOILM data at 0.125 deg data to upscaled 0.5 deg data for a single pixel representing the Kansas Field Station (39.0561°, −95.1907°) during: (**a**) winter, (**b**) spring, (**c**) summer, and (**d**) fall. All 16 pixels at the native spatial resolution that contribute to the upscaled 0.5 deg grid cell are shown as individual gray lines. The weekly average for the 0.5 deg grid cell is shown as a dashed blue line.

Alternative drought classification. While the US Drought Monitor is a widely-used definition of drought, there remains an active stream of research into refining the definition of drought. In recent work[11], Hobeichi *et al.* used a set of climate phenomena mostly measured at 0.5 degrees in a random forest algorithm to classify a location and time period as being in drought or no drought. Labels for drought were based on the Drought Impacts Reporter (DIR), a database of drought impacts housed at the U.S. National Drought Mitigation Center[57]. The resulting random forest method for classification was tested on out-of-sample data, and showed comparatively strong statistical performance scores as compared to other commonly used classifiers. These data provided here could be used in a similar fashion to produce labels for drought that differ from the USDM.

Other possible uses. One could consider the construction of clusters of locations based on similar drought profiles, using techniques such as k-means clustering, hierarchical clustering, fitting a finite mixture model for probabilistic clustering, and other techniques. Such methods could define zones or regions of similar drought experience. These spatio-temporal data are also a collection of 4-way tensors (across lat/lon/time/type), and the decomposition and analysis of such data is an active area of research[58]. These data could serve as a valuable test case for tensor algorithmic development and interpretability.

The gridded representation of measured streamflow is a novel product that offers new opportunities to explore relationships and trends on the CONUS scale. The USGS monitors streamflow trends for select gauges (https://iwaas.wim.usgs.gov/sw-flow-trends/) that are often spaced far apart, and individual studies analyze regional-scale trends[59,60]. However, this new streamflow product enables comprehensive and computationally efficient trend analyses. Furthermore, streamflow can now be easily compared to the other drought-related covariates using the common grid structure, creating opportunities to complete multivariate analyses that can help explain relationships and their variability in time and space. This synoptic set of covariates could also be grouped by their association with different types of drought (meteorological, hydrological, and agricultural) and combined with the gridded USDM to explore the relative influence each drought type on the lumped drought severity index.

Drought is known to affect water supplies, energy production, public health, agriculture, and wildfire potential. Therefore, this new synchronized dataset can be combined with satellite-based observations of fires, such

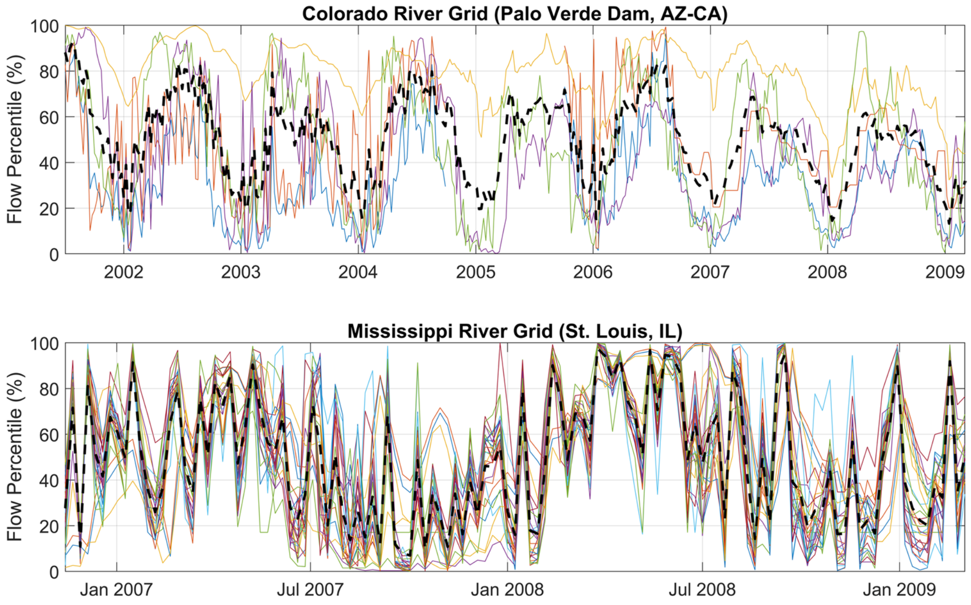

Fig. 19 Top image: streamflow percentiles within a grid that contains a gauge on the Colorado River. Individual gauges are represented with the thin solid colored lines, and the grid mean is represented with the thick dashed black line. Bottom: streamflow percentiles within a grid that contains a gauge on the Mississippi River. The time periods were selected to enhance visibility of the plots.

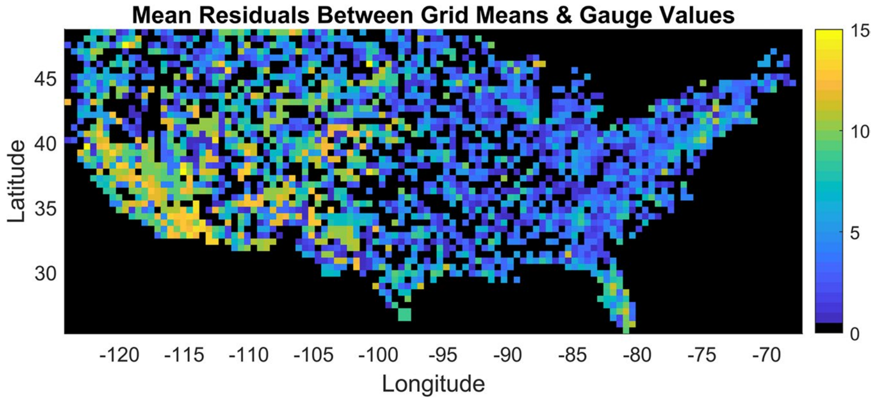

Fig. 20 Mean standard deviation of residuals between gauage data and grid cell average.

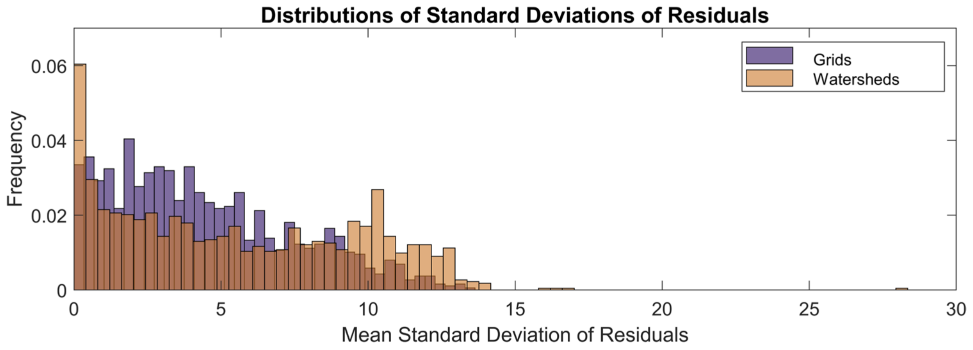

Fig. 21 Histogram plots that represent the distribution of the gridded streamflow variability measure. The grids that contain data within their extents are compared to the grids that used the HUC watershed extents to identify gauges.

as the MODIS burned area[61] and active fire[62] products, to develop improved approaches to fire risk assessments. A primary application of the USDM is to assess risks to croplands and help farmers obtain financial support during drought period. The US Department of Agriculture provides cropland datasets in raster formats (https://croplandcros.scinet.usda.gov/) that can be combined with this new product to study the impact of drought on agricultural productivity, and identify the key hydrological and meteorological drivers of cropland successes and failures across space and time.

There is increasing interest in the relationship between the environment and public health outcomes. The National Center for Environmental Health has noted that drought has both short-term and long-term health effects related to, e.g. the impact on air quality and increased incidence of illness and disease https://www.cdc.gov/nceh/features/drought/index.html. Berman *et al.*[63] used data from the USDM for counties in the western United States and found that high severity worsening drought was associated with increased risk of respiratory-related mortality among adults ages 65 and older on Medicare[63]. Paull *et al.*[64] found the drought was the primary climatic driver of increased rates of West Nile Virus[64]. Head *et al.*[65] found that drought was associated with increased transmission of Valley Fever[65]. This data set can enable further research on the relationship between drought and its related variables and public health outcomes.

Code availability

The data can be found at https://datadryad.org/stash/dataset/doi:10.5061/dryad.g1jwstqw7[44]. All code is freely available at https://github.com/heplersa/USDMdata.

Received: 20 October 2023; Accepted: 28 March 2024;
Published online: 12 April 2024

References

1. Beguería, S., Vicente-Serrano, S. M. & Angulo-Martínez, M. A multiscalar global drought dataset: the speibase: a new gridded product for the analysis of drought variability and impacts. *Bulletin of the American Meteorological Society* **91**, 1351–1354 (2010).
2. Vicente-Serrano, S. M., Beguería, S., López-Moreno, J. I., Angulo, M. & El Kenawy, A. A new global 0.5 gridded dataset (1901–2006) of a multiscalar drought index: comparison with current drought index datasets based on the palmer drought severity index. *Journal of Hydrometeorology* **11**, 1033–1043 (2010).
3. Ziese, M. *et al*. The gpcc drought index–a new, combined and gridded global drought index. *Earth System Science Data* **6**, 285–295 (2014).
4. Zhang, R. *et al*. The first high spatial resolution multi-scale daily spi and spei raster dataset for drought monitoring and evaluating over china from 1979 to 2018. *Big Earth Data* **7**, 860–885 (2023).
5. Ryu, J. H., Sohrabi, M. & Acharya, A. Toward mapping gridded drought indices to evaluate local drought in a rapidly changing global environment. *Water resources management* **28**, 3859–3869 (2014).
6. Svoboda, M. *et al*. The drought monitor. *Bulletin of the American Meteorological Society* **83**, 1181–1190 (2002).
7. Alley, W. M. The palmer drought severity index: limitations and assumptions. *Journal of climate and applied meteorology* **23**, 1100–1109 (1984).
8. Fan, Y., Van den Dool, H., Mitchell, K. & Lohmann, D. Nws-cpc's monitoring and prediction of us soil moisture and associated land surface variables: Land data renalysis. In *Proceedings of the Climate Diagnostics Workshop, 21–25 October 2002* (2002).
9. Kirtman, B. P. *et al*. The north american multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bulletin of the American Meteorological Society* **95**, 585–601 (2014).
10. Hao, Z. *et al*. Toward a categorical drought prediction system based on us drought monitor (usdm) and climate forecast. *Journal of Hydrology* **551**, 300–305 (2017).
11. Hobeichi, S., Abramowitz, G., Evans, J. P. & Ukkola, A. Toward a robust, impact-based, predictive drought metric. *Water Resources Research* **58**, e2021WR031829 (2022).
12. Pu, B., Fu, R., Dickinson, R. E. & Fernando, D. N. Why do summer droughts in the southern great plains occur in some la niña years but not others? *Journal of Geophysical Research: Atmospheres* **121**, 1120–1137 (2016).
13. Schubert, S. D., Suarez, M. J., Pegion, P. J., Koster, R. D. & Bacmeister, J. T. Causes of long-term drought in the us great plains. *Journal of Climate* **17**, 485–503 (2004).
14. Seager, R., Goddard, L., Nakamura, J., Henderson, N. & Lee, D. E. Dynamical causes of the 2010/11 texas–northern mexico drought. *Journal of Hydrometeorology* **15**, 39–68 (2014).
15. Hoerling, M. *et al*. Causes and predictability of the 2012 great plains drought. *Bulletin of the American Meteorological Society* **95**, 269–282 (2014).
16. Mo, K. C. Drought onset and recovery over the united states. *Journal of Geophysical Research: Atmospheres* **116** (2011).
17. Schubert, S. D. *et al*. Global meteorological drought: a synthesis of current understanding with a focus on sst drivers of precipitation deficits. *Journal of Climate* **29**, 3989–4019 (2016).
18. Wang, H. & Kumar, A. Assessing the impact of enso on drought in the us southwest with ncep climate model simulations. *Journal of Hydrology* **526**, 30–41 (2015).
19. Bonaccorso, B., Cancelliere, A. & Rossi, G. Probabilistic forecasting of drought class transitions in sicily (italy) using standardized precipitation index and north atlantic oscillation index. *Journal of Hydrology* **526**, 136–150 (2015).
20. Pendergrass, A. G. *et al*. Flash droughts present a new challenge for subseasonal-to-seasonal prediction. *Nature Climate Change* **10**, 191–199 (2020).
21. Yuan, X. *et al*. A global transition to flash droughts under climate change. *Science* **380**, 187–191 (2023).
22. Christian, J. I. *et al*. Global distribution, trends, and drivers of flash drought occurrence. *Nature communications* **12**, 6330 (2021).
23. Christian, J. I. *et al*. A methodology for flash drought identification: Application of flash drought frequency across the united states. *Journal of Hydrometeorology* **20**, 833–846 (2019).
24. Christian, J. I., Basara, J. B., Otkin, J. A. & Hunt, E. D. Regional characteristics of flash droughts across the united states. *Environmental Research Communications* **1**, 125004 (2019).
25. Lorenz, D. J. *et al*. Predicting the us drought monitor using precipitation, soil moisture, and evapotranspiration anomalies. part ii: Intraseasonal drought intensification forecasts. *Journal of Hydrometeorology* **18**, 1963–1982 (2017).
26. Lowman, L. E., Christian, J. I. & Hunt, E. D. How land surface characteristics influence the development of flash drought through the drivers of soil moisture and vapor pressure deficit. *Journal of Hydrometeorology* (2023).
27. Christian, J. I. *et al*. Global projections of flash drought show increased risk in a warming climate. *Communications Earth & Environment* **4**, 165 (2023).

28. Hao, Z. *et al*. A statistical method for categorical drought prediction based on nldas-2. *Journal of Applied Meteorology and Climatology* **55**, 1049–1061 (2016).
29. Hao, Z. *et al*. Probabilistic drought characterization in the categorical form using ordinal regression. *Journal of Hydrology* **535**, 331–339 (2016).
30. Mitchell, K. E. *et al*. 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** (2004).
31. Xia, Y. *et al*. 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** (2012).
32. Cosgrove, B. A. *et al*. Real-time and retrospective forcing in the north american land data assimilation system (nldas) project. *Journal of Geophysical Research: Atmospheres* **108** (2003).
33. Livneh, B., Xia, Y., Mitchell, K. E., Ek, M. B. & Lettenmaier, D. P. Noah lsm snow model diagnostics and enhancements. *Journal of Hydrometeorology* **11**, 721–738 (2010).
34. Wei, H., Xia, Y., Mitchell, K. E. & Ek, M. B. Improvement of the noah land surface model for warm season processes: Evaluation of water and energy flux simulation. *Hydrological Processes* **27**, 297–303 (2013).
35. Lowman, L. E. & Barros, A. P. Predicting canopy biophysical properties and sensitivity of plant carbon uptake to water limitations with a coupled eco-hydrological framework. *Ecological Modelling* **372**, 33–52 (2018).
36. Sung, J. H. & Chung, E. S. Development of streamflow drought severity–duration–frequency curves using the threshold level method. *Hydrol. Earth Syst. Sci.* **18**, 3341–3351, https://doi.org/10.5194/hess-18-3341-2014 (2014).
37. Patterson, L. A., Lutz, B. D. & Doyle, M. W. Characterization of drought in the south atlantic, united states. *Journal of the American Water Resources Association (JAWRA)* **49**, 1385–1397, https://doi.org/10.1111/jawr.12090 (2013).
38. Barros, A. P., Hodes, J. L. & Arulraj, M. Decadal climate variability and the spatial organization of deep hydrological drought. *Environ. Res. Lett.* **12**, 104005, https://doi.org/10.1088/1748-9326/aa81de (2017).
39. of Environmental Quality, N. C. D. North carolina drought management advisory council report (2022).
40. of Water Resources, C. D. California data exchange center - river stages / flow (2023).
41. Blanco, I. *et al*. Geographical distribution of copd prevalence in europe, estimated by an inverse distance weighting interpolation technique. *International Journal of Chronic Obstructive Pulmonary Disease* **13**, 57–67, https://doi.org/10.2147/COPD.S150853 (2018).
42. Stachelek, J. & Madden, C. J. Application of inverse path distance weighting for high-density spatial mapping of coastal water quality patterns. *International Journal of Geographical Information Science* **29**(7), 1240–1250, https://doi.org/10.1080/13658816.2015.1018833 (2018).
43. Wang, R., Zou, R., Liu, J., Liu, L. & Hu, Y. Spatial distribution of soil nutrients in farmland in a hilly region of the pearl river delta in china based on geostatistics and the inverse distance weighting method. *Agriculture* **11**(1), 50, https://doi.org/10.3390/agriculture11010050 (2018).
44. Erhardt, R., Di Vittorio, C., Lowman, L., Hepler, S. & Wei, W. Discretized u.s. drought data to support statistical modeling, *Dryad* https://doi.org/10.5061/dryad.g1jwstqw7 (2024).
45. McCullagh, P. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)* **42**, 109–127 (1980).
46. Christensen, R. H. B. ordinal–regression models for ordinal data. *R package version* **28**, 2015 (2015).
47. Fokianos, K. & Kedem, B. Regression theory for categorical time series. *Statistical science* **18**, 357–376 (2003).
48. Weiß, C. H. Distance-based analysis of ordinal data and ordinal time series. *Journal of the American Statistical Association* **115**, 1189–1200 (2020).
49. Schliep, E. M. *et al*. Joint species distribution modelling for spatio-temporal occurrence and ordinal abundance data. *Global Ecology and Biogeography* **27**, 142–155 (2018).
50. Feng, X., Zhu, J., Lin, P.-S. & Steen-Adams, M. M. Composite likelihood estimation for models of spatial ordinal data and spatial proportional data with zero/one values. *Environmetrics* **25**, 571–583 (2014).
51. Ma, T. F. R., Wu, W., Zhu, J. & Feng, X. *clespr: Composite Likelihood Estimation for Spatial Data*. R package version 1.1.2 (2018).
52. Higgs, M. D. & Hoeting, J. A. A clipped latent variable model for spatially correlated ordered categorical data. *Computational Statistics & Data Analysis* **54**, 1999–2011 (2010).
53. Brewer, M. J. *et al*. A spatial model with ordinal responses for grazing impact data. *Statistical Modelling* **4**, 127–143 (2004).
54. Schliep, E. M. & Hoeting, J. A. Data augmentation and parameter expansion for independent or spatially correlated ordinal data. *Computational Statistics & Data Analysis* **90**, 1–14 (2015).
55. Banerjee, S., Carlin, B. P. & Gelfand, A. E. *Hierarchical modeling and analysis for spatial data* (CRC press, 2014).
56. Simpson, D., Lindgren, F. & Rue, H. In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics* **23**, 65–74 (2012).
57. Wilhite, D. A., Svoboda, M. D. & Hayes, M. J. Understanding the complex impacts of drought: A key to enhancing drought mitigation and preparedness. *Water resources management* **21**, 763–774 (2007).
58. Kolda, T. G. & Bader, B. W. Tensor decompositions and applications. *SIAM review* **51**, 455–500 (2009).
59. Diem, J. E., Hill, T. C. & Milligan, R. A. Diverse multi-decadal changes in streamflow within a rapidly urbanizing region. *Journal of Hydrology* **556**, 61–71 (2018).
60. Fang, B. *et al*. The recent decline of apalachicola–chattahoochee–flint (acf) river basin streamflow. *Hydrology* **9**, 140–151 (2022).
61. Giglio, L., Boschetti, L., Roy, D., Humber, M. & Justice, C. The collection 6 modis burned area mapping algorithm and product. *Remote Sensing of Environment* **217**, 72–85 (2018).
62. Giglio, L., Schroeder, W. & Justice, C. O. The collection 6 modis active fire detection algorithm and fire products. *Remote Sensing of Environment* **178**, 31–41 (2016).
63. Berman, J. D., Ebisu, K., Peng, R. D., Dominici, F. & Bell, M. L. Drought and the risk of hospital admissions and mortality in older adults in western usa from 2000 to 2013: a retrospective study. *The Lancet Planetary Health* **1**, e17–e25 (2017).
64. Paull, S. H. *et al*. Drought and immunity determine the intensity of west nile virus epidemics and climate change impacts. *Proceedings of the Royal Society B: Biological Sciences* **284**, 20162078 (2017).
65. Head, J. R. *et al*. Effects of precipitation, heat, and drought on incidence and expansion of coccidioidomycosis in western usa: a longitudinal surveillance study. *The Lancet Planetary Health* **6**, e793–e803 (2022).
66. Xia, Y. *et al*. 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** (2012).
67. Mahrt, L. & Ek, M. The influence of atmospheric stability on potential evaporation. *Journal of Applied Meteorology and Climatology* **23**, 222–234 (1984).
68. Penman, H. L. Natural evaporation from open water, bare soil and grass. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **193**, 120–145 (1948).
69. Mo, K. C. *et al*. Drought indices based on the climate forecast system reanalysis and ensemble nldas. *Journal of Hydrometeorology* **12**, 181–205 (2011).
70. Xia, Y. *et al*. Evaluation of multi-model simulated soil moisture in nldas-2. *Journal of Hydrology* **512**, 107–125 (2014).

71. Xia, Y. *et al*. Validation of noah-simulated soil temperature in the north american land data assimilation system phase 2. *Journal of applied meteorology and climatology* **52**, 455–471 (2013).
72. Ek, M. B. *et al*. Implementation of noah land surface model advances in the national centers for environmental prediction operational mesoscale eta model. *Journal of Geophysical Research: Atmospheres* **108**, https://doi.org/10.1029/2002JD003296 (2003).
73. Wallace, J. M. & Gutzler, D. S. Teleconnections in the geopotential height field during the northern hemisphere winter. *Monthly weather review* **109**, 784–812 (1981).
74. Thompson, D. W. & Wallace, J. M. Regional climate impacts of the northern hemisphere annular mode. *Science* **293**, 85–89 (2001).
75. Thompson, D. W., Lee, S. & Baldwin, M. P. Atmospheric processes governing the northern hemisphere annular mode/north atlantic oscillation. *The North Atlantic Oscillation: climatic significance and environmental impact* **134**, 81–112 (2003).
76. Thompson, D. W. & Wallace, J. M. Annular modes in the extratropical circulation. part i: Month-to-month variability. *Journal of climate* **13**, 1000–1016 (2000).
77. Thompson, D. W. & Wallace, J. M. The arctic oscillation signature in the wintertime geopotential height and temperature fields. *Geophysical research letters* **25**, 1297–1300 (1998).
78. McPhaden, M. J., Zebiak, S. E. & Glantz, M. H. Enso as an integrating concept in earth science. *science* **314**, 1740–1745 (2006).

Acknowledgements

The authors acknowledge support from NSF award #2151881.

Author contributions

R.E. and S.H. conceived of the original statistical research which necessitated these data. R.E., L.L. and C.D. collected and processed all raw data to form the final data product. W.W. processed data and made several figures. R.E., C.D., S.H. and L.L. wrote the manuscript, and all authors reviewed the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.