scientific data

OPEN

DATA DESCRIPTOR

Restructuring and serving web-accessible streamflow data from the NOAA National Water Model historic simulations

J. Michael Johnson[1,2] ✉, David L. Blodgett[3], Keith C. Clarke[2] & Jon Pollak[4]

In 2016, the National Oceanic and Atmospheric Administration deployed the first iteration of an operational National Water Model (NWM) to forecast the water cycle in the continental United States. With many versions, an hourly, multi-decadal historic simulation is made available to the public. In all released to date, the files containing simulated streamflow contain a snapshot of model conditions across the entire domain for a single timestep which makes accessing time series a technical and resource-intensive challenge. In the most recent release, extracting a complete streamflow time series for a single location requires managing 367,920 files (~16.2 TB). In this work we describe a reproducible process for restructuring a sequential set of NWM steamflow files for efficient time series access and provide restructured datasets for versions 1.2 (1993–2018), 2.0 (1993–2020), and 2.1 (1979–2022). These datasets have been made accessible via an OPeNDAP enabled THREDDS data server for public use and a brief analysis highlights the latest version of the model should not be assumed best for all locations. Lastly we describe an R package that expedites data retrieval with examples for multiple use-cases.

Background & Summary

Streamflow records provide information for a range of people including emergency responders, water managers, environmental and transportation agencies, researchers, utility companies, and consulting firms[1–3]. Specific needs might include short- and long-range planning[4,5], warning about floods and droughts[6–8]; managing water rights, regulating and monitoring environmental impacts[9,10]; operating waterways for commerce; and designing flood frequency curves[11,12].

Despite the vast utility of streamflow records, there is a divergence between where water exists and where it is measured even in densely gaged countries[13,14]. Over the last few decades, continental to global scale hyper-resolution hydrologic prediction has been dubbed a "grand challenge" within the hydrology community to address this shortcoming[15–17]. Although many have praised the scientific and societal advantages these massive models offer, there is a steep set of technical, conceptual, and practical hurdles[18,19].

In the United States, the National Oceanic and Atmospheric Administration (NOAA) undertook the development of an operational, high-resolution model as part of the strategic mission of building a 'weather-ready' nation[20]. In 2016, version 1.0 of the National Water Model (NWM) was put into operation[21,22] to provide the first-ever continental United States-wide modelling capability using real-time weather forecasts, a high resolution (1 km^2) land surface model[23,24], and a multi-resolution surface routing model[25,26]. Shortly following the initial release, version 1.2 expanded the calibration basins from 40 to more than 1,000 and improved parameter regionalization[27]. Version 2.0 expanded the NWM domain further to include Hawaii and added new configurations (medium-range ensemble); out-of-bank compound channels parameterizations; and improved snow physics. Many land surface and hydrologic parameters were further refined by expanding the calibration basin set to approximately 1,400 basins[28]. Version 2.1 saw the expansion to Puerto Rico, the U.S. Virgin Islands, and the Great Lakes region and included improved reservoir treatment and modifications to the model's snow and

[1]Lynker, Fort Collins, CO, USA. [2]University of California, Santa Barbara, USA. [3]U.S. Geological Survey, Reston, USA. [4]Consortium of Universities for the Advancement of Hydrologic Science, Inc, Cambridge, USA. ✉e-mail: jjohnson@lynker.com

A. User0defined variables		
FILELIST = collection of sequential NetCDF model output files		
OUTFILE = desired output file to create (path)		
VAR = variable name (e.g. streamflow)		
B. Concatenate & Pivoting Operations Step-by-Step		
1	ITERATE (Over *i*)	*ncks -O -4 -L 1 --cnk_plc = all --cnk_map = dmn -C -v feature_id,time*, **VAR FILELIST[i] FILELIST[i]**
	|	*ncatted -O -a "scale_factor*,**VAR**,*d,," -a "add_offset*,**VAR**,*d,,"* **FILELIST[i] FILELIST[i]**
	|	*ncap2 -O -s VAR [time,feature_id] = **VAR*** **FILELIST[i] FILELIST[i]**
	| STOP	*ncks -O --mk_rec_dmn time* **FILELIST[i] FILELIST[i]**
2		*ncrcat -O -6* **FILELIST OUTFILE**
3		*ncpdq -O -a feature_id,time* **OUTFILE OUTFILE**
4		*ncatted -h -O -a "scale_factor*,**VAR**,*o,f,0.01"* **OUTFILE OUTFILE**
5		*ncks -O --cnk_plc = g2d --cnk_dmn feature_id,10000 --cnk_dmn time*,**#FILELIST OUTFILE OUTFILE**
6		*ncks -4 -L 3 -O* **OUTFILE OUTFILE**
C. Description		
Plain language operations:		
1. Start Iterating over files and for each:		
a. Extract the desired **VAR**, feature_id, and time 1D variables		
b. Delete (d) the scale factor (double) and add_offset(double) attribute in each file		
c. Reshape the 1D **VAR** to be [time,feature_id]		
d. Set the time dimension to the record dimension allowing it to 'grow'		
Stop Iteration		
2. Concatenate files across the time dimension		
3. Re-shape **VAR** such that the dimensions are [feature_id, time]		
4. Overwrite (o) the scale_factor of **VAR** to be a float (f) of value 0.01		
4. Rechunk the time dimension (**#FILELIST**) and feature_id dimension (10,000)		
4. Compress the file		
NCO utilities:		
1. https://linux.die.net/man/1/ncks		
2. https://linux.die.net/man/1/ncatted		
3. https://linux.die.net/man/1/ncap2		
4. https://linux.die.net/man/1/ncrcat		
5. https://linux.die.net/man/1/ncpdq		

Table 1. The process for concatenating and pivoting a collection of NWM channel files leveraging the NCO[61] tool set. Panel A outlines the user-defined variables which are bolded throughout the table. Panel B describes the steps in terms of NCO pseudo code. Panel C provides a plain language summary of the steps demonstrated in panel B along with references to the appropriate NCO utilities.

runoff parameters[29]. This version also began using the Analysis of Record for Calibration[30,31] (AORC) dataset to enhance calibration and improve the estimation and regionalization of hydrologic parameters[32].

Although the timesteps, horizons, ensemble members, configurations, and file names have evolved with each release, the model has consistently produced 1 km^2 gridded files of land surface and forcing states, a 250 m^2 gridded file of the terrain conditions (ponded water), and point files containing the stream routing and reservoir variables for the entire domain. Starting with version 1.2, a multi-decadal, hourly, historical simulation has been released with most versions providing a resource for better understanding the NWM and the earth system it simulates. Versions 1.2 and 2.0 of the historical simulations used the NLDAS/NARR forcings[33–36] whereas v2.1 used the AORC dataset. These historic simulations provide an unprecedented resource that can spur research and understanding about the model, its evolution, and its applications[37–40], and can help better understand what improvements were made version to version.

Despite all historic simulation data being available on on Amazon Web Services (AWS) Registry of Open Data[41], the data structure can be hard to use for specific use cases. In the case of long-term streamflow records there are three primary challenges to overcome. First, the point files contain a snapshot of conditions for a given domain (CONUS) and timestep (1 hour). As a result, extracting a single time series for a location of interest would require managing anywhere from 4 (for v1.2) to 16 (v21.1) terabytes of data.

Second, the hourly point files are structured to prioritize space-based, rather than time-based subsetting even when concatenated with common open-source tools. Although this can be a typical design pattern for spatial grids, it is limiting when trying to extract time oriented data.

Version	Root	NetCDF Naming	NcML Name
1.2	https://thredds.hydroshare.org/thredds/catalog/nwm/retrospective/		
		nwm_retro_full/nwm_retro_XXX.nc	*nwm_retro_full.ncml*
2.0	https://thredds.hydroshare.org/thredds/catalog/nwm/retrospective/		
		nwm_v2_retro_full/nwm_retro_v2_XXX.nc	*nwm_v2_retro_full.ncml*
2.1	https://cida.usgs.gov/thredds/catalog/demo/morethredds/nwm/		
		v21_reshape/nwm_retro_v2_XXX.nc	*nwm_v21_retro_full.ncml*

Table 2. This table lists the THREDDS catalog root directory and file naming convention for each version of the data. The XXX represents a three-digit number, padded with leading zeroes ranging from 1 to 273.

Lastly, the point datasets index 1D variables using a non-spatial, non-sequential, identifier (feature_id) that is adopted from the common identifiers (COMID) associated with stream reaches in the NHDPlusV2[42]. This requires users to first find the identifiers of interest, then use the position of that identifier in the dataset to extract the needed records. In the operational data, there is no spatial information associated with these feature_ids and NOAA states that "due to storage space limitations, the latitude and longitude of each point are stored in an external Esri file geodatabase…"[30]. In the historic data (v2.0 and v2.1) coordinate data were added to the channel files at the expense of increasing the average file size from 6.6 MB to 47.6 MB. While adding the capacity for pseudo spatial subset (given the streamflow variable is not indexed to theses coordinate dimensions), it exacerbates the amount of data that needs to be managed.

When looking at common use patterns for optimal streamflow dissemination we focused on the design of the U.S. Geologic Survey (USGS) National Water Information System (NWIS) which delivers data by site through time, rather than by time across all sites[43,44]. During the 2017 water year alone more than 640 million requests for streamflow data were fulfilled by NWIS, with 98% being fulfilled by webservices[2]. Thus, we believe NWIS can serve as a guide for an optimized streamflow dataset located in a centralized web-accessible resource.

In this data descriptor we highlight the approach developed for restructuring NWM point files for time-based access and restructure the streamflow records from v1.2, v2.0, and v2.1. The data are served by an OPeNDAP-enabled THREDDS data server, and we illustrate how data can be discovered programmatically using combinations of publicly available tools and a corresponding R package called nwmTools.

Methods

The key technologies used to aggregate and reshape the NWM channel output files include the Network Common Data Form (NetCDF) data model[45] (https://docs.unidata.ucar.edu/netcdf-c/current/netcdf_data_model.html) and a THREDDS data server configured with OPeNDAP access[46]. NetCDF files are a common, platform independent, and self-describing data format used to store multi-dimensional, array-based variables. The explicit entities in any NetCDF file include the *dimensions*, *variables*, and *data*[47].

Dimensions are one-dimensional arrays with a name and size. They may represent a physical property, like time and latitude, or abstract values like the unique identifiers in the channel files. Record dimensions are defined with a length of "unlimited," allowing them to grow with data being written to the file, or through the concatenation of multiple files. Critically, non-record dimensions cannot grow in this way.

Variables are defined by a name and shape determined by an ordered set of dimensions. For example, *rainfall(latitude, longitude, time)* is a three-dimensional variable with an X,Y, and T dimension. For any variable, the last dimension in the above syntax varies the fastest, while the first varies the slowest. The implications of this are that the shape of a variable directly impacts the performance for a specific use case.

In the NWM point files, streamflow is shaped as *streamflow(feature_id)* with an "unlimited" global record time dimension. If two files were concatenated along the time dimension, the variable would become a 2D variable *streamflow(time, feature_id)* which prioritizes extraction along the feature_id. Changing the order of dimensions requires pivoting the variable array to achieve *streamflow(feature_id, time)* which again can be quite intensive for large arrays.

With respect to sharing data, its long been understood that many NetCDF datasets are too big to be efficiently shared via regular upload and download. This has spurred the development of remote access technologies. One of these options is a THREDDS Data Server (TDS) which uses the NetCDF-Java/CDM library to read multidimensional data into a Common Data Model (CDM)[45,46,48]. Once in a TDS, any CDM resources can be accessed using a uniform resource locator (URL). THREDDS catalogues allow users to navigate what data are available in a browser and can leverage the NetCDF Markup Language (NcML) to modify and aggregate multiple CDM datasets into an entity that acts as a single resource.

An integrated TDS server can also provide OPeNDAP access which extends the HTTP protocol to allow CDM subsets to be requested via URL. Subsets are defined by appending constraint expressions to the CDM URL in the form of *?variable[X:Y:Z]*. This expression will return the data from the named variable array at index X to index Z with an interval of Y along the first dimension. OPeNDAP arrays are zero indexed, thus *?time[3:1:3]* would return the fourth value in the 1D time dimension and *?streamflow[0:1:0][0:1:9]* would return the first ten streamflow records for the first *feature_id* in the streamflow array organized as (feature_id, time). OPeNDAP requests can be submitted through any Data Access Protocol (DAP) client (including a standard web browser) and any tool that uses the C-based NetCDF Application Programming Interface (API) acts as a DAP client making OPeNDAP available in most common programming languages.

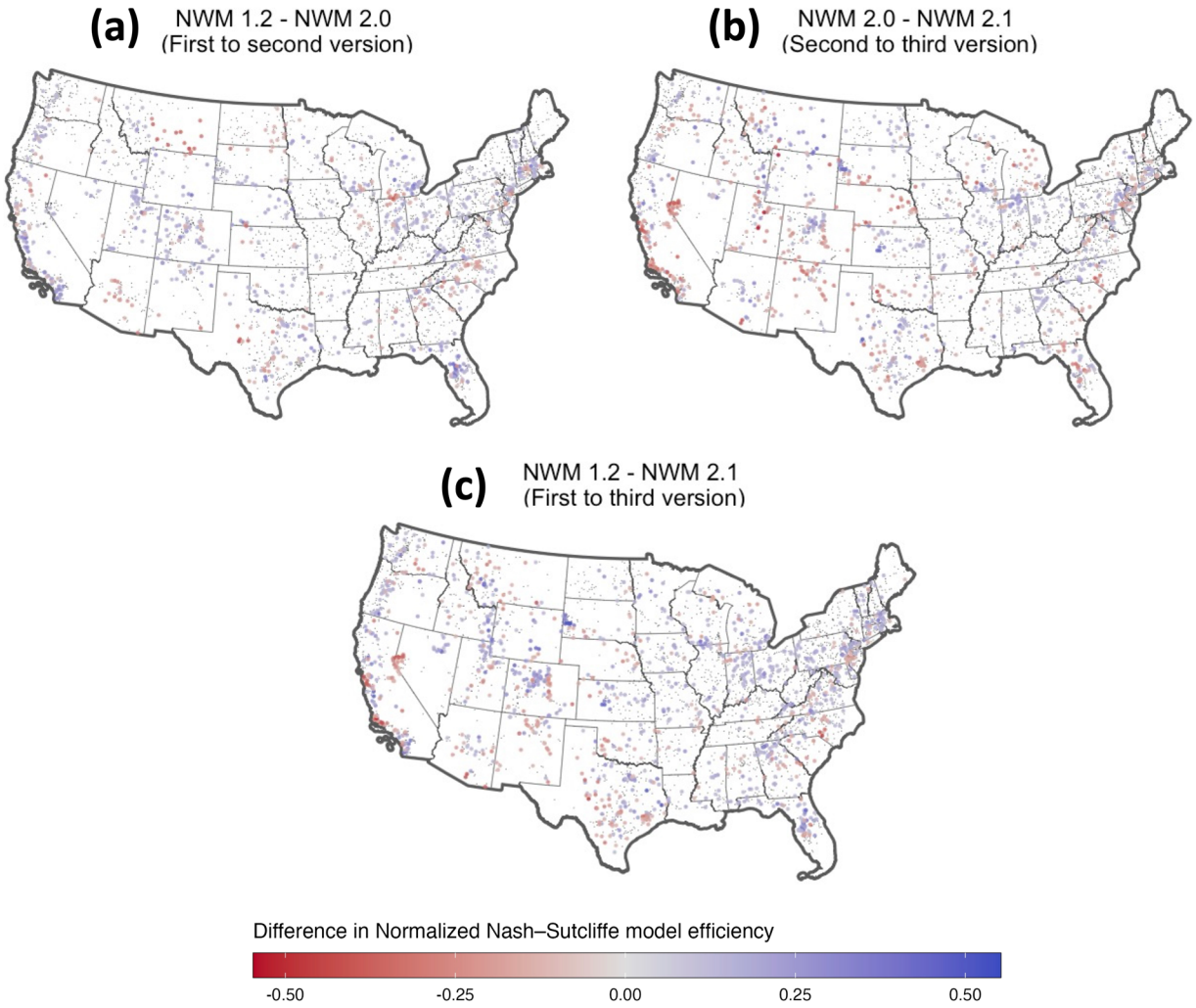

Fig. 1 The difference in NNSE between National Water Model versions. Here blue dots indicate an improvement in NNSE at a given site in the new model, and a red location indicates the model degraded in the new version. Panel (**a**) shows the change from version 1.2 to 2.0, panel (**b**) the change from version 2.0 to 2.1, and panel (**c**) the change from version 1.2 to 2.1.

With this background, we can leverage these technologies to develop a performant, time series-oriented channel output file. To do this, a set of sequential NWM files must be identified and an iterative process applied in which: (1) all variables except feature_id, time, and variable of interest (e.g. streamflow) are dropped; (2) the scale factor and offsets are removed; (3) the dimensions of the variable are redefined to be *streamflow*[*time, feature_id*]; and (4) the time dimension is set to the record dimension. Once complete, all files can be merged along the time dimension creating a single 2D variable. To prioritize time series access, the variable is reshaped as *streamflow*(*feature_id*, *time*) after which the scale factor and offset are reintroduced. Optionally the file can be rechunked and compressed for performance. Table 1 provides pseudo-NCO code of this process and a plain language explanation.

Operational NWM outputs are publicly available for a 48-hour rolling window on the NOAA Operational Model Archive and Distribution System (NOMADs)[49]. Since the release of v2.0 (2018), these have been archived in Google Cloud Platform (GCP) with a minimal lag. The method proposed here works on any sequential set of NWM data from any of these sources which have small to modest file lists that can be downloaded, merged, and reshaped on local hardware and memory. Although this responsibility could reside with an authoritative organization (either natively in the model, or as a post-processing step), it is feasible for users to execute themselves as new forecasts become available.

In contrast, the historical simulations on the AWS Registrary of Open Data[41] constitute a much larger file set (up to 16 TB for v2.1) which require either massive hardware, or a divide and conquer approach to execute. We elected the latter approach and each month in the historic record was treated as a unique file list (n = ~744 hourly files). Steps 1 (file processing) and 2 (concatenation) were completed to create a set of monthly files. Each monthly file was split into 273 new files containing ~10,000 *feature_ids* each. The respective subsets from each month were then merged with their counterparts along the time dimension producing 273 files with the entire

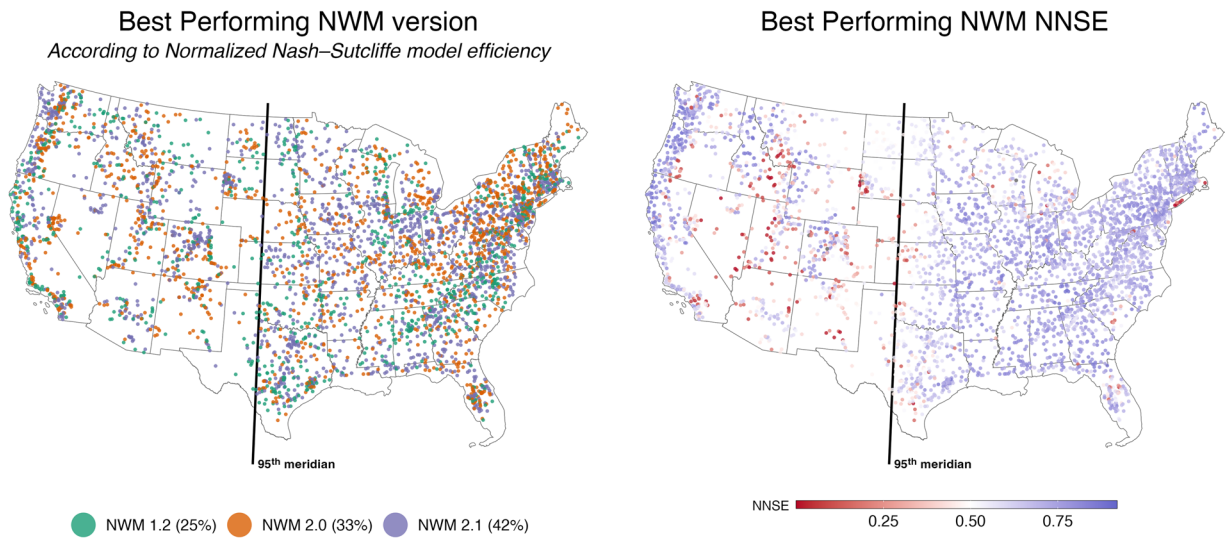

Fig. 2 (**a**) Each evaluated NWIS gage is coloured by the best performing NWM historic simulation (**b**) the NNSE of the best performing historic simulation is mapped.

streamflow[time, feature_id] variable for ~10,000 feature_ids. This space and time divide-and-conquer approach was designed to optimize speed while allowing the processing to remain in memory.

The files were then joined with an NcML aggregation and pivoted within a local THREDDS workspace using the *nccopy* utility (see code at https://linux.die.net/man/1/nccopy)[50]. The result is 273 time-optimized files that act as a single logical entity via a NcML file.

Data Records

The complete dataset for the three version of the NWM are distributed across servers. The composite resource for all versions has been documented in HydroShare[51]. Version 1.2 and 2.0 are available on the HydroShare partition of the Renaissance Computing Institute at the University of North Carolina at Chapel Hill, and version 2.1 is hosted by the USGS Center for Integrated Data Analytics (see Table 2 for key web addresses).

Each version of the dataset is structured in the same way regardless of server and includes a top level THREDDS directory containing a subdirectory of 273 time-optimized NetCDF files and a NcML file that allows that directory to act as a single resource. Table 2 documents the root path to each version along with the naming convention of the NetCDF and NcML files. Each individual file (and resulting NcML file) has a 1D latitude, longitude, time, and feature_id variable along with a 2D streamflow array. The content and associated metadata can be viewed at the NcML catalog page and can be accessed by any DAP client.

Technical Validation

Appropriate checks have been made to ensure the correct data are returned from OPeNDAP queries and that nothing was misaligned in the data transformation. Beyond that, the data are exactly as output from the historical simulations.

As a preliminary evaluation of the data quality, we extracted the daily observed streamflow records for the 4,780 USGS NWIS sites from the GAGES-II dataset[25] with at least 10 years of flow between 1993 and 2018. These records were then compared to the daily mean flows from all three historic versions using the Nash Sutcliffe Efficiency (NSE) metric[52]. NSE provides a normalized goodness-of-fit statistic commonly used in hydrologic model evaluation. However the nature of NSE implies a lower limit of $-\infty$ which can create problems in multi-site model evaluation. As such, Nossent & Bauwens proposed a Normalized NSE (NNSE)[53] following Eq. 1.

$$NNSE = \frac{1}{2 - NSE} \tag{1}$$

An NNSE of 1 indicates a "perfect model", an NNSE = 0.5 indicates the model has the same predictive skill as the mean of the observed time-series, and anything less than 0.5 suggests a user should use the mean of the observed values rather than the model. Although issues of model diagnostics, error evaluation, and interpretation are beyond the scope of this data descriptor, our dataset provides the ability to take a high level look at how model skill evolved spatially from version to version.

Figure 1 maps the locations where the change in NNSE was larger than ± 0.1 from version to version. Blue locations represent sites that improved with the model advancements, and red locations are those that degraded. This analysis focuses on a single error metric across the entire time series and similar plots focusing on seasonal performance, high and or low flows, or different basin types, may look different.

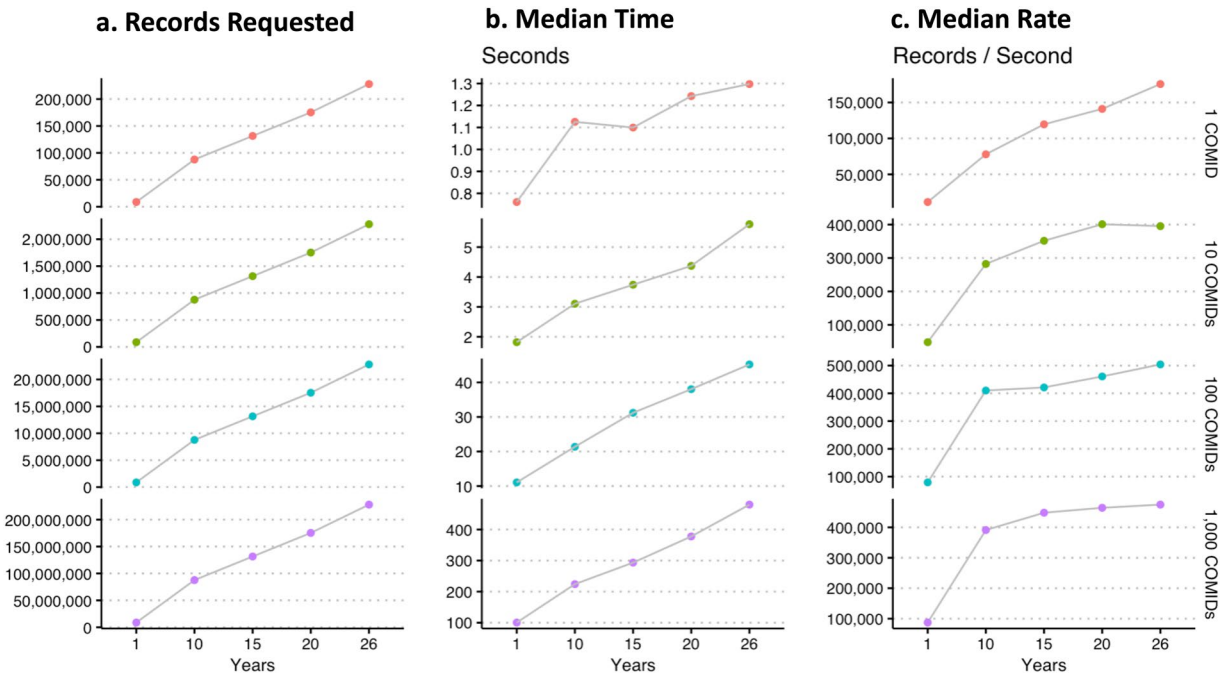

Fig. 3 Benchmark performance tests. (**a**) the number of streamflow records requested based on the number of COMID(s) and years requested. (**b**) the median time (seconds) based on the number of COMID(s) and years requested. (**c**) the median extraction rate (records/second) based on the number of COMID(s) and years requested.

Starting with panel (1a) we see that version 2.0 increased performance in many areas (specifically the West Coast, the Rocky Mountain Range, Florida and much of the northeastern United States). It was able to do this at the expense of degrading performance in New Mexico, southern Montana, southern Michigan, and areas in the southeastern United States. Overall, 15% of the sites saw NNSE improve by more than 0.1, and 6% degraded by more than 0.1. Looking at panel (1b), we see that version 2.1 improved some of the areas degraded by v2.0 (e.g., southern Montana, southern Michigan, and areas in the southeast) at the expense of degrading performance in coastal California, the Lake Tahoe region, and much of the southwestern United States and the southern Atlantic coast and the eastern seaboard. Overall, version 2.1 saw 13% of the locations increase NNSE by more than 0.1 compared to version 2.0, while 11% of the sites decreased by more than 0.1 NNSE. Lastly, panel (1c) shows the overall improvement across the life of the NWM historic datasets. From version 1.2 to 2.1 the general Rocky Mountains, Midwest, and Florida saw improvement while the Bay Area, Lake Tahoe regions, and parts of the southwestern United States, south Atlantic Coast and eastern seaboard saw degradation. In total 20% of the sites saw improvement larger than 0.1 between these models while 10% saw degradation larger than 0.1.

Given the mixed improvement through versions, Fig. 2a colors each location according to the version that performed best. When looking at this map, there are no discernible patterns or clusters indicating one version did better in specific regions. This is notable as a large part of NWM improvement comes from calibration and parameter regionalization[28–30], and NOAA has acknowledged that improvements in NWM performance, by calibration alone, are beginning to plateau[21]. This realization has spurred the development of the Next Generation Water Modeling Framework (NextGen) that allows for heterogeneous models to be run in different parts of the county within a single platform[21]. Figure 2b looks at the maximum performance that can be achieved if the best version is chosen at each location in a pseudo NextGen representation. In general, the NWM skill is poor west of the 95th meridian (~ central Texas) until the more humid west coast is reached. There are versions that can achieve decent results in the eastern and western slopes of the Rockies, and no models are able to adequately capture the areas surrounding New York City. When using a combination of all models, 83% of the locations can achieve an NNSE greater than 0.5 compared to 71% in version 1.2, 75% in version 2.0, and 77% in version 2.1.

The takeaway for data users is that decent performance can be seen in most places, however the assumption that the latest model is the best, simply because it's the most recent, is not accurate. and model choice should depend on the research question and area of study.

Ultimately, this data release describes a new access pattern for the NWM historical streamflow data based on restructuring and serving a time oriented version of the simulated records. The specific data are associated with versions 1.2, 2.0, 2.1 of the NWM but inevitably, a newer version of the NWM will be developed and with it, a new historical product. Although we use this study to provide access to the current simulations, the software code is the reproducible process for formatting the historic product for public consumption and increasing the accessibility that can be used on future NWM historic simulations.

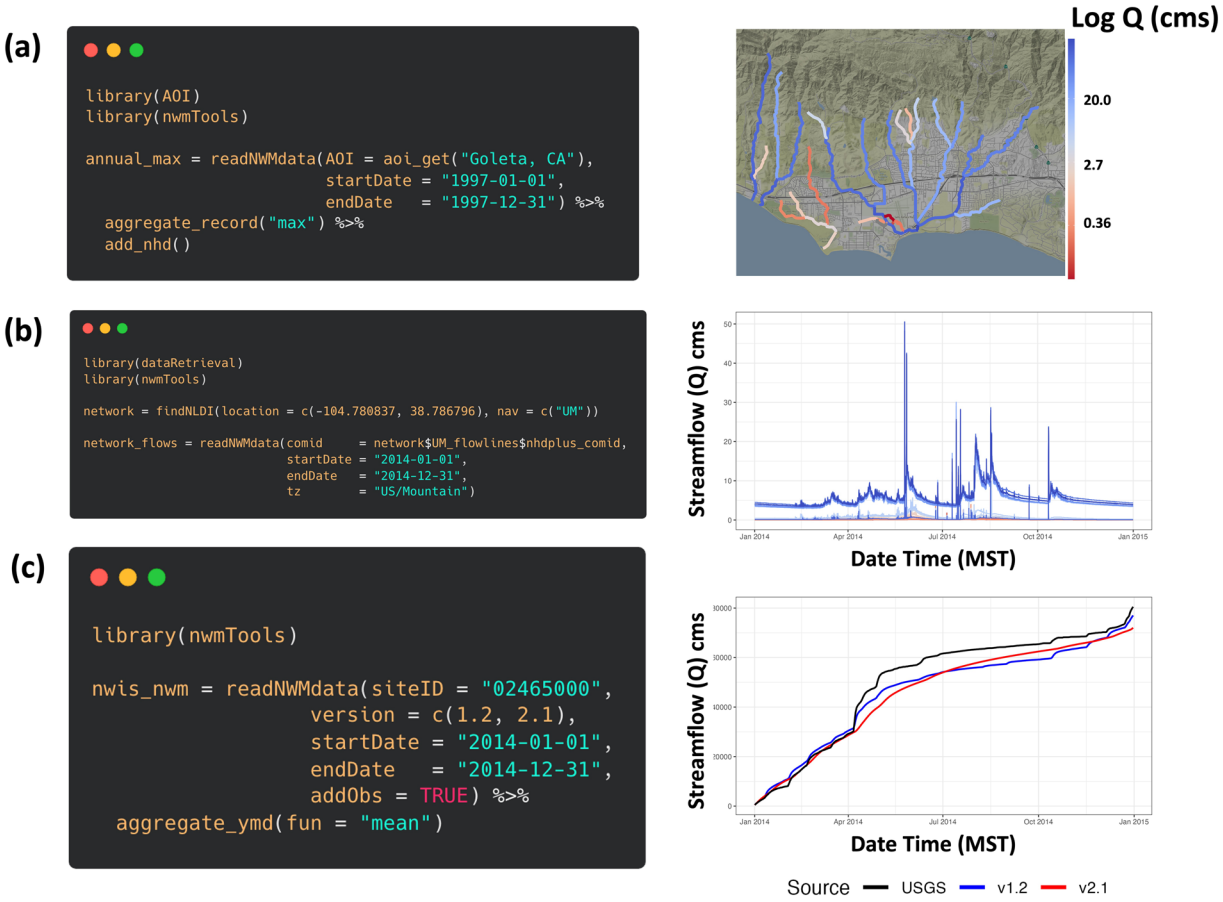

Fig. 4 (**a**) Maximum annual NWM streamflows for the areas of Goleta, California found by integrating *AOI* and *nwmTools*. The data are plotted on the spatial NHDPlus flowlines (**b**) Annual hourly flows for the complete upstream mainstem of a defined point found by integrating *dataRetrieval* and *nwmTools*. The data for all reaches are plotted with red colors near the headwaters graduating to blue as they reach the outlet (**c**) Multi-versions of the NWM compared to observations at a gaged location using *nwmTools*. Cumulative flow plots are shown to see the difference in records.

Usage Notes

Restructuring the historic NWM simulations and serving it through the Consortium of Universities for the Advancement of Hydrologic Science, Inc.'s (CUAHSI) and USGS's infrastructure allows users to extract a time series for any feature_id of interest without individually managing raw data files. These datasets are open to the public and can be accessed with any DAP enabled software.

With any new web-based dataset, questions about performance and scalability arise. Thus, we benchmark data extraction for 1, 5, 10, 20, and 26 years of hourly data for 1, 10, 100, and 1,000 random feature_ids (these are the same as NHDPlus COMIDs). In total, these requests range from 8,760 to 227,904,000 values (Fig. 3a). Tests were made on a personal laptop with internet speeds approximating 40.8 Mb/s download. Each combination of requests was run five times and the median elapsed time is shown in Fig. 3b. Figure 3c shows the extraction rate of each query (requested records/median elapsed time) which ranged between 2,845 records/second and 515,830 records/second. There is an evident plateau, starting around 400,000 records/second and a maximum performance occurring around 500,000 records/second. Overall, these tests give us confidence the new data sources provide the intended capabilities that allow users to access the historic NWM streamflow dataset with a relatively low barrier to entry and low computational requirements.

Building OPeNDAP queries is not a trivial task, and identifying the appropriate feature_id and time ID, and its positions in the aggregated resource is a repetitive process prone to error. To further improve accessibility, we offer the *nwmTools* R package[54] which uses the RNetCDF[55] NetCDF client for sending/retrieving OPeNDAP requests. The primary package function for historical data is *readNWMdata(…)*, which intentionally mirrors functionality provided in the USGS *dataRetrieval* package[44].

By default, the function returns all hourly data from version 2.1 in UTC for a user supplied area of interest (AOI), COMID(s), or USGS streamgage ID(s). Additional parameters allow users to narrow the start and end date, adjust the time zone, and specify the model version desired. The package also provides a family of functions for aggregating streamflow records to other time chunks (e.g., monthly, water year, or season) and for appending the spatial NHDPlus features to the output data.frame (based on webservices rather than the preceeding the Esri geodatabase). Here it should be noted that data can only be returned for NHD features included in the NWM

(e.g., the model does not include NHD flowlines associated with waterbodies and thus only NA values would be returned).

To highlight these tools, we present three use cases (Fig. 4) that show how to find NWM data by AOI, by COMID, and by USGS streamgage number. Indirectly we show ways for finding this information using open-source packages.

The first use case (4a) illustrates how the *nwmTools* can find NWM data for a given area and time range (*readNWMdata*), summarize the hourly data to an annual mean (*aggregate_record*), and append NHDPlus geometries to the forecast (*add_nhd*). A key feature of the NHD data model is the ability to traverse the hydrographic flow network and modern data systems like the Network Linked Data Index (NLDI) are capitalizing on the graph nature of the hydrographic networks to facilitate feature discovery and indexing[56,57]. The NLDI has multiple programmatic interfaces available in R (dataRetrieval[44], nhdplusTools[58]) and Python (HyRiver)[59] that are supported by the web framework developed as part of the Open Water Data Initiative[56,60]. Panel 4b illustrates a use case to find the annual flow records upstream of a known location in Colorado Springs, Colorado. Starting from this point, the NLDI can return the COMIDs associated with the upstream mainstem (UM) which can be passed directly to *readNWMdata* to get all flow records for 2014 in Mountain Standard Time zone. These time series are colored using a red to blue palette where the darkest reds are the headwater reaches and the darkest blue is the outlet reach. Lastly, many efforts aimed at model evaluation, calibration, and improvement need to identify simulation records that match some gaged record. We can use a USGS site number to query the historic NWM simulation *and* the USGS observations. Panel 4c shows how to find two versions of the NWM and the observed record at USGS site number 02465000. The data are returned in a way that is directly comparable, for example, as a cumulative flow plot.

In all three examples the underlying data records are what allow these discovery workflows to succeed. To this end, these use cases could prompt the consideration of distributing the operational products in a similar way which would allow any variation of the workflows illustrated here to be used with the operational forecasts.

Code availability

All code for data download and reformatting can be found in the appropriate USGS repository[50]. The *nwmTools* R package is available on GitHub and the dataset is documented and published via HydroShare[51]. All the data are currently open, and publicly available at this URL: https://www.hydroshare.org/resource/84c2b029f97343a59d0739115d4087f1/.

Received: 23 October 2020; Accepted: 19 June 2023;
Published: 20 October 2023

References

1. Olson, S. A. & Norris, J. M. *US Geological Survey Streamgaging… from the National Streamflow Information Program*. (2007).
2. USGS. *Monitoring the Pulse of Our Nation's Rivers and Streams: The U.S. Geological Survey Streamgaging Network*. https://pubs.usgs.gov/fs/2018/3081/fs20183081.pdf (2018).
3. USGS. USGS Streamgaging Network. (2021).
4. Vogel, R. M. & Fennessey, N. M. Flow-Duration Curves. I: New Interpretation and Confidence Intervals. *Journal of Water Resources Planning and Management* **120**, 485–504 (1994).
5. Vorosmarty, C. J. *et al*. Global Threats to Human Water Security and River Biodiversity. *Nature* **467**, 555–561 (2010).
6. Johnson, J. M. *et al*. Knowledge graphs to support real-time flood impact evaluation. *AI Magazine* **43**, 40–45 (2022).
7. Johnson, J. M., Coll, J. M., Ruess, P. J. & Hastings, J. T. Challenges and Opportunities for Creating Intelligent Hazard Alerts: The "FloodHippo" Prototype. *Journal of the American Water Resources Association*, (2018).
8. Adams, T. III Flood forecasting in the United States NOAA/National Weather Service. in *Flood Forecasting* 249–310 (Elsevier, 2016).
9. Grantham, T. E., Merenlender, A. M. & Resh, V. H. Climatic influences and anthropogenic stressors: an integrated framework for streamflow management in Mediterranean-climate California, USA. *Freshwater Biology* **55**, 188–204 (2010).
10. Patterson, L., Phelan, J., Goudreau, C. & Dykes, R. Flow-Biology Relationships Based on Fish Habitat Guilds in North Carolina. *JAWRA Journal of the American Water Resources Association* **53**, 56–66 (2017).
11. Cunnane, C. Methods and merits of regional flood frequency analysis. *Journal of Hydrology* **100**, 269–290 (1988).
12. Rao, A. R. & Hamed, K. H. *Flood frequency analysis*. (CRC press, 2019).
13. Krabbenhoft, C. A. *et al*. Assessing placement bias of the global river gauge network. *Nature Sustainability* 1–7 (2022).
14. Beran, B. & Piasecki, M. Availability and coverage of hydrologic data in the US geological survey National Water Information System (NWIS) and US Environmental Protection Agency Storage and Retrieval System (STORET). *Earth Science Informatics* **1**, 119–129 (2008).
15. Wood, E. F. *et al*. Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research* **47**, 54–10 (2011).
16. Bierkens, M. F. P. Global hydrology 2015: State, trends, and directions. *Water Resources Research* **51**, 4923–4947 (2015).
17. Archfield, S. A. *et al*. Accelerating advances in continental domain hydrologic modeling. *Water Resources Research* **51**, 10078–10091 (2015).
18. Beven, K. J. & Cloke, H. L. Comment on "Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water" by Eric F. Wood et al. *Water Resources Research* **48** (2012).
19. Beven, K., Cloke, H., Pappenberger, F., Lamb, R. & Hunter, N. Hyperresolution information and hyperresolution ignorance in modelling the hydrology of the land surface. *Science China Earth Sciences* **58**, 25–35 (2015).
20. Uccellini, L. W. & Ten Hoeve, J. E. Evolving the National Weather Service to Build a Weather-Ready Nation: Connecting Observations, Forecasts, and Warnings to Decision-Makers through Impact-Based Decision Support Services. *Bulletin of the American Meteorological Society* **100**, 1923–1942 (2019).
21. Office of Water Prediction, N. The National Water Model. (2022).
22. NOAA. Implement National Water Model: New implementation of the NWM. (2016).
23. Salas, F. R. *et al*. Towards Real-Time Continental Scale Streamflow Simulation in Continuous and Discrete Space. *JAWRA Journal of the American Water Resources Association* **51**, 10078–21 (2017).
24. Niu, G.-Y. *et al*. The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research* **116**, 1381–19 (2011).

25. Yang, Z.-L. *et al*. The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins. *Journal of Geophysical Research* **116**, 4257–16 (2011).
26. Gochis, J. & Chen, F. Hydrological enhancements to the community Noah land surface model. (2003).
27. Gochis, D. *et al*. *Technical Description of the National Water Model Implementation of WRF-Hydro*. (2016).
28. NOAA. NWM Upgrade: Upgrade to V1.2 of National Water Model. (2018).
29. NOAA. *NWM Upgrade: Update to the National Water Model Version 2.0.*, (2019).
30. NOAA. *NWM Upgrade: Upgrade NCEP National Water Model v2.1.*, (2021).
31. Kim, H. & Villarini, G. Evaluation of the Analysis of Record for Calibration (AORC) rainfall across Louisiana. *Remote Sensing* **14**, 3284 (2022).
32. Kitzmiller, D. H., Wu, W., Zhang, Z., Patrick, N. & Tan, X. The analysis of record for calibration: a high-resolution precipitation and surface weather dataset for the united states. in vol. 2018 H41H-06 (2018).
33. Cosgrove, B. A., Gochis, D. J., Clark, E. P. & Flowers, T. NOAA's National Water Model: A Dynamically Evolving Operational Hydrologic Forecasting Framework. (2020).
34. Cosgrove, B. A. *et al*. Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *J. Geophys. Res.* **108**, 2002JD003118 (2003).
35. Mo, K. C., Chen, L.-C., Shukla, S., Bohn, T. J. & Lettenmaier, D. P. Uncertainties in North American Land Data Assimilation Systems over the Contiguous United States. *Journal of Hydrometeorology* **13**, 996–1009 (2012).
36. Berg, A. A. Impact of bias correction to reanalysis products on simulations of North American soil moisture and hydrological fluxes. *J. Geophys. Res.* **108**, 2–15 (2003). ACL 2-1-ACL.
37. Jachens, E. R., Hutcheson, H., Thomas, M. B. & Steward, D. R. Effects of Groundwater-Surface Water Exchange Mechanism in the National Water Model over the Northern High Plains Aquifer, USA. *JAWRA Journal of the American Water Resources Association* (2020).
38. Hansen, C., Shafiei Shiva, J., McDonald, S. & Nabors, A. Assessing Retrospective National Water Model Streamflow with Respect to Droughts and Low Flows in the Colorado River Basin. *JAWRA Journal of the American Water Resources Association* **55**, 964–975 (2019).
39. Viterbo, F. *et al*. General Assessment of the Operational Utility of National Water Model Reservoir Inflows for the Bureau of Reclamation Facilities. *Water* **12**, 2897 (2020).
40. Johnson, J. M., Munasinghe, D., Eyelade, D. & Cohen, S. An integrated evaluation of the National Water Model (NWM)–Height Above Nearest Drainage (HAND) flood mapping methodology. *Natural Hazards and Earth System Sciences* **19**, 2405–2420 (2019).
41. NOAA National Water Model Reanalysis Model Data on AWS. https://docs.opendata.aws/nwm-archive/readme.html.
42. McKay, L. *et al*. NHDPlus Version 2: User Guide, 2012.
43. U.S. Geological Survey. 2023. USGS water data for the Nation: U.S. Geological Survey National Water Information System database, at https://doi.org/10.5066/F7P55KJN accessed 2023-10-05.
44. De Cicco, L. A., Lorenz, D., Hirsch, R. M., Watkins, W. & Johnson, M. *dataRetrieval: R Packages for Discovering and Retrieving Water Data Available from U.S. Federal Hydrologic Web Services*. https://doi.org/10.5066/P9X4L3GE (2018).
45. Unidata, (2023): NetCDF [software]. Boulder, CO: UCAR/Unidata. https://doi.org/10.5065/D6H70CW6.
46. Unidata, (2023): THREDDS Data Server [software]. Boulder, CO: UCAR/Unidata. https://doi.org/10.5065/D6N014KG.
47. Unidata, (2023): NetCDF User's Guide (NUG). Boulder, CO: UCAR/Unidata. https://doi.org/10.26024/nw73-vm64.
48. Unidata, (2023): NetCDF-Java [software]. Boulder, CO: UCAR/Unidata. https://doi.org/10.5065/DA15-J131.
49. Rutledge, Glenn K., Jordan Alpert, and Wesley Ebisuzaki. NOMADS: A climate and weather model archive at the National Oceanic and Atmospheric Administration. *Bulletin of the American Meteorological Society* **87.3**, 327–342 (2006).
50. Blodgett, D. L. *NWM V2 Processing Steps*. https://code.usgs.gov/water/nwm_subset (2020).
51. Johnson, JM. & Blodgett, DL. NOAA National Water Model Reanalysis Data at RENCI, *HydroShare*, https://doi.org/10.4211/hs.a1e329ad20654e72b7b423f991bf9251 (2023).
52. Nash, J. E. & Sutcliffe, J. V. River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*. **10**(3), 282–290, https://doi.org/10.1016/0022-1694(70)90255-6 (1970).
53. Nossent, J. & Bauwens, W. Application of a normalized Nash-Sutcliffe efficiency to improve the accuracy of the Sobol' sensitivity analysis of a hydrological model. *EGUGA* 237 (2012).
54. Johnson, J. M. *nwmTools*. https://github.com/mikejohnson51/nwmTools/ (2020).
55. Michna, P. & Woods, M. RNetCDF–A package for reading and writing NetCDF datasets. *The R Journal* **5**, 29–36 (2013).
56. USGS. Network Linked Data Index API. (2022).
57. Blodgett, D., Johnson, J. M., Sondheim, M., Wieczorek, M. & Frazier, N. Mainstems: A logical data model implementing mainstem and drainage basin feature types based on WaterML2 Part 3: HY_Features concepts. *Environmental Modelling & Software* 104927 (2020).
58. Blodgett, D. & Johnson, J. nhdplusTools: Tools for Accessing and Working with the NHDPlus. Avaiable from https://code.usgs.gov/water/nhdplusTools (2018).
59. Chegini, T., Li, H. Y. & Leung, L. R. HyRiver: Hydroclimate Data Retriever. Journal of Open Source Software, **6**(66), 3175 (2021).
60. Federal Geographic Data Committee. Open Water Data Initative. (2022).
61. Zender, C. S. Analysis of self-describing gridded geoscience data with netCDF Operators (NCO). *Environmental Modelling & Software* **23**, 1338–1342 (2008).

Acknowledgements

This material is funded by the 2020 HydroInformatics Innovation Fellowship provided by CUAHSI with support from the National Science Foundation (NSF) Cooperative Agreement No. EAR-1849458. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Author contributions

The historic simulation was executed by the NCAR Research Applications Laboratory and NOAA Office of Water Prediction. The data were restructured at the USGS. The data are hosted - and this project funded - by CUAHSI. The project was led by, and the *nwmTools* package written at UCSB and Lynker. The first author created the manuscript with input from all involved parties.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.M.J.

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.