Environmental Modelling and Software 165 (2023) 105726

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft

Generating a reference flow network with improved connectivity to support durable data integration and reproducibility in the coterminous US

David Blodgett [a,*], J. Michael Johnson [b], Andy Bock [a]

[a] U.S. Geological Survey, 12201 Sunrise Valley Dr, Reston, VA, 20192, USA
[b] NOAA Office of Water Prediction Affiliate, Lynker, 5445 Conestoga Ct Ste 100, Boulder, CO, 80301, USA

ARTICLE INFO

Handling Editor: Daniel P Ames

Keywords:
Hydrofabric
Flow network
Catchment
Hydrography

ABSTRACT

This report presents a *reference flow network* for the conterminous United States that is built from the best available information from the U.S. Geological Survey, the National Oceanic and Atmospheric Administration National Weather Service, and the U.S. Environmental Protection Agency. The work is intended to support durable data integration and reproducibility. Originating from the National Hydrography Dataset Plus (NHDPlus) V2.1, the *reference flow network* incorporates network connectivity enhancements from federal agency efforts. After incorporating these network improvements, many original NHDPlus attributes were regenerated to enable network navigation and related operations. After introducing the motivation and background for this work, this report describes the attribute generation workflow and data quality checks that were performed in preparation of the dataset. The *reference flow network* follows the NHDPlus data model and is described using terms defined in the *Mainstem and Drainage Basin* logical model and *WaterML2 Part3: Surface Hydrology Features* conceptual model.

1. Introduction

The route water follows from its source on the landscape to an inland sink or the ocean is important to many domains of environmental science including hydrology, hydrodynamics, geomorphology, water quality, limnology, aquatic ecology, water availability, disaster management, and others. Many water quantity and water quality models adopt network connectivity as a given that is not altered and has strong impacts on model performance. For this reason, improving network connectivity should help minimize fundamental errors in our models.

Models that simulate flowing water, or that use data located on a flow network, require the best network representation possible. The connectivity of these flow networks can be very complex, and ever changing. As a result, even when authoritative, quality-controlled networks are used, it is common for projects to apply changes based on local knowledge. While many modeling groups start with the same source flow network, changes to improve network connectivity rarely make it back into an integrated, updated network. Systems to update the geometry of a latest available hydrographic dataset in the United States (US) (the National Hydrography Dataset) (U.S. Geological Survey, 2022) have been used to capture updates and fixes in some cases. However, fixes to the flow network of source datasets for modeling made by modeling projects have, by in large, not been incorporated back into a *reference flow network* for use in future modeling. Given this common occurrence, interoperability and reproducibility become a challenge.

In the US, digital representation of the national flow network has evolved for three decades (Horn, 1994; Bondelid et al., 2010; McKay et al., 2015; Brakebill et al., 2020; National Oceanic and Atmospheric Administration, 2021). Since its release in 2015, the U.S. Geological Survey, the National Oceanic and Atmospheric Administration National Weather Service, and the U.S. Environmental Protection Agency have all worked to improve the network (Dewald, 2017) of the National Hydrography Dataset Plus V2.1 (NHDPlusV2) (McKay et al., 2015) The identifiers used to integrate data with these datasets (the "comid", which is a per-line identifier and the Reachcode, which aggregates one or more "comid"s) do not persist across datasets and have changed through time as data improvements have been made. The focus of the work described here is incorporation of this legacy of changes into a central *reference flow network* (italicized throughout this report) that can be updated regularly but use persistent identifiers and still support reproducibility

Abbreviations: Findable Available Interoperable and Reusable, (FAIR); National Hydrography Dataset, (NHD); WaterML2 Part 3: Surface Hydrology Features, (HY_Features).

* Corresponding author. 1 Gifford Pinchot Dr, Madison, WI, 53726, USA.
E-mail addresses: dblodgett@usgs.gov (D. Blodgett), jjohnson@lynker.com (J.M. Johnson), abock@usgs.gov (A. Bock).

https://doi.org/10.1016/j.envsoft.2023.105726
Received 12 May 2023; Accepted 16 May 2023
Available online 17 May 2023

and durable data integration (See Blodgett, 2023; Blodgett, 2023-1 data releases).

Persistent (or durable) identifiers and appropriate representation of flow networks across scale support continental-scale, nationally consistent, and locally relevant modeling. Automation of data quality checks facilitates work with national or continental scale datasets that may be too large for review by a domain expert. Reproducibility and comparability can be difficult with a changing (even if improving) flow network if the connectivity and representation of flowing water bodies changes. For example, associations between stream gages or water quality sample locations (sources of truth) and rivers (subjects of reproducible science) must remain consistent if we are to expect reproducible outcomes. However, if identifiers or representations of the flow network change, we will get different results. The *reference flow network* aims to resolve the problem of durable identification and network representation. It improves our collective ability to build representative modeling frameworks and integrate relevant landscape and observational data by simultaneously addressing the needs of continental hydrologic modeling and Findable, Available, Interoperable, and Reusable (FAIR) (Wilkinson et al., 2016) environmental data. To facilitate these goals, the *reference flow network* is based on the logical data model presented in Blodgett et al. (2021) and the more general conceptual data model in Blodgett (2020). Key terms used in this paper that are defined in these reports are introduced when first used and, in addition to *reference flow network*.

The *reference flow network* is part of what is referred to here as a “reference fabric”. The concept of “reference fabric”, as introduced here, is intended to support collaborative inter-agency hydrologic modeling (Fig. 1). A “reference fabric” is an integrated collection of data that is both a reference system to which information can be addressed and a reference dataset with which to create baseline representations of hydrologic systems. A “reference fabric” includes a non-spatial *reference flow network*, line and polygon geometries, and community-recognized hydrologic locations (Points of Interest) that are integral to the flow network (e.g. stream gages, dams). This report describes the creation of the *reference flow network* only.

US initiatives that may benefit from this work include the U.S. Geological Survey National Water Census (Michelsen et al., 2016; Miller et al., 2020) and the National Weather Service National Water Model (National Oceanic and Atmospheric Administration, 2016). Within these, the National Oceanic and Atmospheric Administration (NOAA) Next Generation Water Resource Modeling Framework (Ogden et al., 2021) and the U.S. Geological Survey National Hydrologic Model (Bock et al., 2021; Regan et al., 2019) are the work’s initial focus. There have been calls in the hydrologic and Earth System Modeling science communities for general advancement in continental-domain hydrologic modeling which this work seeks to advance (Archfield et al., 2015; Clark et al., 2015).

Treating public data as a universally available and useable strategic asset is widely understood to be in the public interest (OECD, 2021). Recognizing this, the development of the Reference Fabric is also motivated by the Go FAIR initiative and the principles of the Internet of Water (Internet of Water Organization, 2021). The work has specific focus on public domain Web identifiers for environmental features (Blodgett, 2020) that allow all data providers to contribute to a well understood system of linked environmental data. These FAIR public data aspects of the motivation for this work are realized through a community (Internet of Water) data indexing system known as the Network Linked Data Index (Blodgett, 2023-2) and the geoconnex.us identifier registry and knowledge graph (Internet of Water Organization, 2023). Objectives of modeling and FAIR data work together naturally. Models benefit from additional data brought to bear while FAIR data can be made more capable by leveraging and incorporating data improvements and outputs of models.

1.1. Background

The HY_Features conceptual data model standard (Blodgett and Dornblut, 2018) provides a set of concepts for formalizing hydrologic data integration and reproducibility in the hydrosciences. It introduces a wholistic conceptual definition of *catchment* (italicized when used in the HY_Features sense) that serves two roles: conveying water from the land into a flow network and conveying water from an inlet to an outlet. These two roles are fulfilled by the “*flowpath*” and “*catchment area*” conceptual realizations of the catchment concept respectively. *Flowpath* and *catchment area* are italicized to draw attention to their specific definition from the HY_Features conceptual data model.

The *catchment* concept is wholistic in that it includes all aspects of hydrologic function within its divide and it can be applied at any scale. A special case (but still wholistic and applicable at any scale) of the *catchment* concept is “*drainage basin*”. A *drainage basin* is a *catchment* that is the total upstream area draining to an outlet (Blodgett & Johnson, 2022-1) (*drainage basin* is italicized to draw attention to its specific definition). As such, a *drainage basin* can be used to define nested hierarchies across a wide range of spatial scales. Given that a *drainage basin* has a primary *flowpath* (a “*mainstem*”), nested hierarchies of *drainage basins* are connected by a directed acyclic graph or tree of *mainstems* (Blodgett et al., 2021) (*mainstem* is italicized to draw attention to its specific definition). The *reference flow network* depends heavily on the *mainstem* concept and its implementation in the NHDPlus data model, “level path”. In NHDPlus, the level path attribute is a unique identifier for the primary upstream path from anywhere in the network that is described in detail below. Level paths generally follows a river’s name but are defined for features without names as well. Fig. 2 shows the relationship between incremental *catchment areas* surrounded by their divides, *flowpaths*, a scale-dependent *drainage basin* surrounded by its divide and its respective *mainstem.* This framework is intended to support integration of multi-scale hydrologic process investigation from zero-order to large *mainstems* using common identifiers and feature linkages.

Previous works on national hydrographic data have been distributed as static snapshots that do not evolve with time. The source datasets and

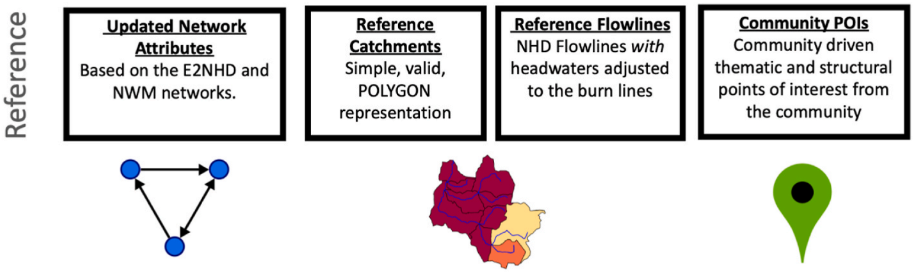
Reference
Updated Network Attributes
Based on the E2NHD and NWM networks.
Reference Catchments
Simple, valid, POLYGON representation
Reference Flowlines
NHD Flowlines *with* headwaters adjusted to the burn lines
Community POIs
Community driven thematic and structural points of interest from the community

Fig. 1. Components of the overall “reference fabric”. Updated network attributes from Enhanced NHDPlusV2 (e2NHDPlus) (Brakebill et al., 2020), National Water Model (NWM) (National Oceanic and Atmospheric Administration, 2016). Reference catchments and flowlines from NHDPlusV2 (McKay et al., 2015) Community Points of Interest (POIs) from varied sources.

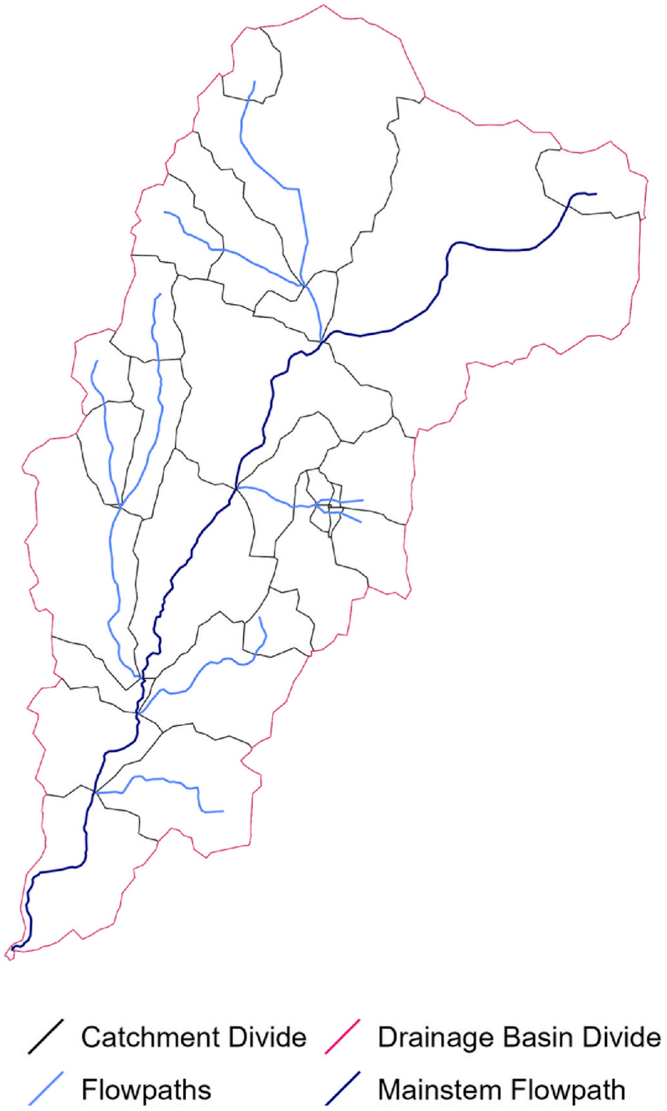

Fig. 2. A set of incremental catchment areas surrounded by their divides (black) each have a flowpath (sky blue). The incremental catchment areas constitute a drainage basin boundary (red) that has a predominate mainstem (dark blue).

software used to create the hydrographic datasets have not been made available for inspection, reuse, or enhancement (as with Horn, 1994; Bondelid et al., 2010; and McKay et al., 2015). This, in addition to diversity in manually created and edited content, has posed challenges to data curation, maintenance, and general community involvement. In contrast, the *reference flow network* has been developed as a reproducible and open-source workflow based on publicly available input datasets and software. While updating such a dataset is a non-trivial process, these practices aim to make that process more efficient and foster better community understanding and involvement as the network is updated. The dataset will evolve in three ways: 1) its resolution will improve, 2) its representation of the network will improve, and 3) the flaws in its source data will be identified and fixed. These changes will be incorporated in an open and reproducible process where identifiers are maintained, and backward compatibility is established explicitly. The *mainstem* identifier system at the core of the *reference flow network* facilitates incorporation of these updates and backward compatibility into the future.

1.2. Network representation and attributes

The NHDPlusV2 dataset (McKay et al., 2015) is used as the base data source and base data model for the *reference fabric* (Dewald, 2017; Brakebill et al., 2020). This ensures that the work can capitalize on the efforts that came before and that this work is compatible with applications using the NHD Data model (U.S. Geological Survey, 2022) and associated features. The NHDPlus data model includes what it calls ‘value added attributes’ that are documented in the NHDPlusV2 manual (McKay et al., 2015) and are also implemented in the *reference flow network*.

All flow networks can be represented as an edge-to-edge (edge list) or edge-to-node topology (node topology) (Fig. 3). An edge list only expresses the connectivity between edges (*flowpaths* in the context of rivers), requiring nodes (confluences in the context of rivers) to be inferred. Both of these schemes are present in the NHDPlus data model. Specifically, the “hydroseq”/“dnhydroseq” (hydrosequence/down hydrosequence) relationship expresses the network as a dendritic edge list, and the fromnode/tonode relationship expresses it as a node topology (these attributes are described in detail in Table 2). This initial work on the *reference flow network* did not require representation of diverted flow. Given this, the current *reference flow network* only includes the edge list representation of a network and does not include diverted flow.

By treating the network as a dendritic tree of “primary” downstream paths, headwaters and diverted paths are treated similarly. Practically, this means that the diverted fraction is always 0 at a divergence. In practice, a flow routing algorithm could change this assumption as source dataset diversion information is not lost. For example, Fig. 4 illustrates two paths, one that is diverted from the other. The path through edges 1, 4, and 5 would be considered the main path. The diversion at N2 is only represented in the edge-node version of the flow network. Edge 2 is treated as if it has no inflow in the edge-to-edge version of the flow network and the path through edges 2 and 3 is treated as a tributary path. Explicit diversion handling could be introduced by reintroducing edge-node connections but was not required for the initial needs of the dataset so was not included.

While the current version of *the reference flow network* focuses on the NHDPlus, it is important to note that the methods and software developed for it are applicable to any hydrologic network that contains a set of key base attributes (Table 1). These attributes are used to generate two key network attributes, hydrosequence and level path.

Any algorithm that uses a flow network and requires understanding the upstream to downstream relationship of network elements requires a sorted version (or attribute that facilitates upstream downstream sorting) of the network. The NHDPlus data model attribute, “hydrosequence” is functionally a topological sort f the *flowpath* network (Cormen and Leiserson, 2022). An attribute functionally equivalent to hydrosequence has been used since the earliest digital hydrographic datasets (Horn, 1994). It is an integer identifier that is guaranteed to decrease in the downstream direction. For *flowpaths* that are not connected by a single-direction navigation (e.g. parallel tributaries) the hydrosequence has no significance. In other words, when two *flowpaths* have a single-direction navigable connection, the downstream *flowpath* will always have the smaller hydrosequence attribute. Fig. 5D shows the hydrosequence attribute visually.

Level path is derived from “stream level” which is a constant integer attribute along a *mainstem* rivers from outlet to headwater. “Stream leveling” is the process of establishing level paths through a stream network. This is accomplished with a set of rules that determine which tributary should be considered dominant at every confluence and establishes the *mainstem* for each *drainage basin* in a network. In the stream level algorithm, rivers terminating to the ocean are given level 1; this level extends all the way to the headwater. Rivers terminating into level 1 rivers are given level 2, and so on. Fig. 5 illustrates stream level (5B) and level path (5C). As a point of reference, the NHDPlusV2 has about

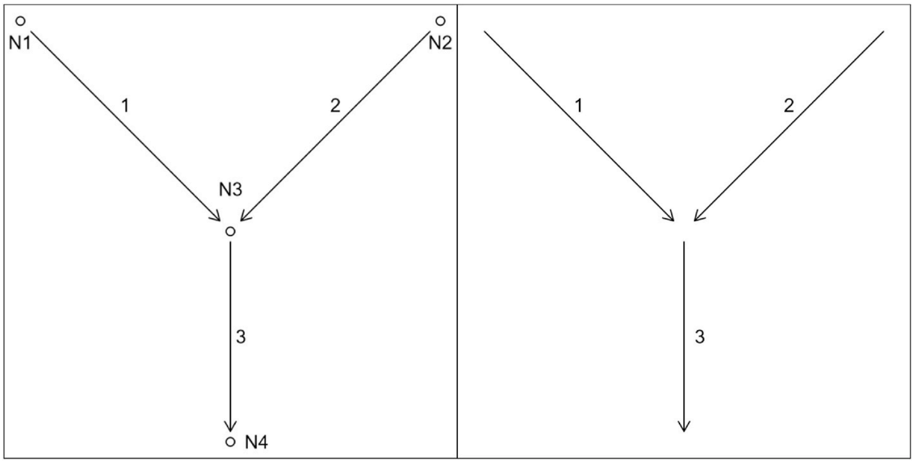

Fig. 3. In an edge-node topology (left), edges are directed to nodes which are then directed to other edges. An edge-to-edge topology (right) does not have intervening nodes.

Table 1
Base Attributes needed to derive hydrosequence and level path.

Attribute	Description	Purpose	Notes
fromnode/ tonode	Node topology	Must have this or the edge list topology for any network calculations.	From and to nodes can be used to generate an edge list *flowpath* topology.
id/toid	Dendritic edge list topology.	All network traversal is based on the dendritic edge list topology.	Secondary divergent paths are represented as network initiation *flowpaths*.
divergence[a]	Diverted path attribute.	Used with nodes to create a dendritic upstream to downstream edge list topology.	This attribute is 0 for normal (already [many:1] dendritic) connections, 1 for the main path through a divergence, and 2 for any diverted path.
Length	A length for each *flowpath* in the network	Determine a flow distance, and, if using the arbolate sum, stream leveling.	Can be derived from *flowpath* geometry
area	The drainage area of each *flowpath's catchment area.*	Primarily used to calculate total drainage area.	Can be derived from *catchment area* geometries
weight[a]	Indication of how large a flowpath is relative to others in the network.	A weight metric is required to determine the dominant upstream *flowpath*	In *reference flow network*, the arbolate sum (the sum of all upstream *flowpath lengths*) is used.
nameid	Stream Name	It is often preferable to follow a named path rather than a weight.	Optional

[a] Can be derived from other input sources, or are optional inputs.

2.7 million flowlines and includes about 1 million unique level paths. The longest level path (the Missouri River) is over 2400 individual flowlines.

In the NHDPlus data model, level path identifiers are set to the same value as the hydrosequence of the *flowpath* at the level path's outlet. See Blodgett et al., 2021 for a more in-depth discussion of these concepts.

A detail worth illustrating here relates to durability of identifiers. Given that the value of hydrosequence attributes (the sort order) will change if the number of features in the network changes, using the outlet hydrosequence value as a level path's value results in unstable level path identifiers. As a result, use of the hydrosequence-based level path

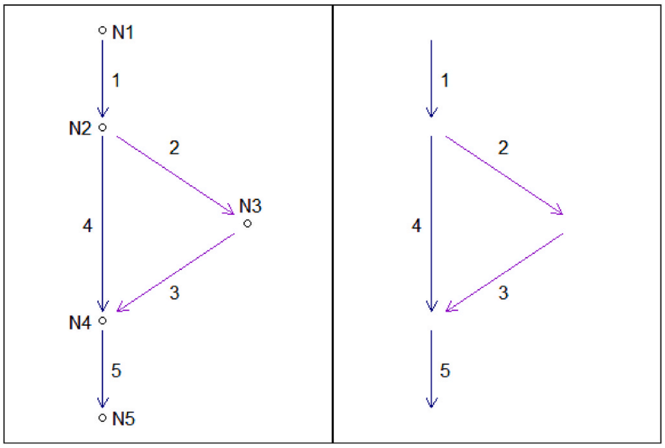

Fig. 4. In an edge-node topology (left) divergences can be represented as multiple edges emerging from a single node. In an edge-to-edge topology (right) diversions require one to many relationships that can be difficult to work with in practice.

Table 2
Attributes generated from the hydrosequence and level path attributes.

Attribute	Description	Purpose
terminal path	The hydrosequence identifier of the terminal *flowpath* of network.	Identifies the terminus of a dendritic network.
up hydrosequence:	The identifier of the upstream *flowpath* along the *mainstem.*	Identifies the upstream *flowpath* along the *mainstem*
down hydrosequence	The identifier of the downstream *flowpath* along the *mainstem.*	Identifies the downstream *flowpath* along the *mainstem*
up level path	The identifier of the upstream level path along the *mainstem.*	Identifies the upstream *mainstem*
down level path	The identifier of the downstream level path along the *mainstem.*	Identifies the downstream *mainstem*
path length	The distance to the network outlet downstream along the main path.	Understanding the routed length or river network to the terminus
total drainage area	The total accumulated area from all upstream *catchment areas.*	The complete upstream area of a *flowpath* outlet
arbolate sum	The total accumulated length of upstream *flowpaths.*	Used to determine primary upstream path.

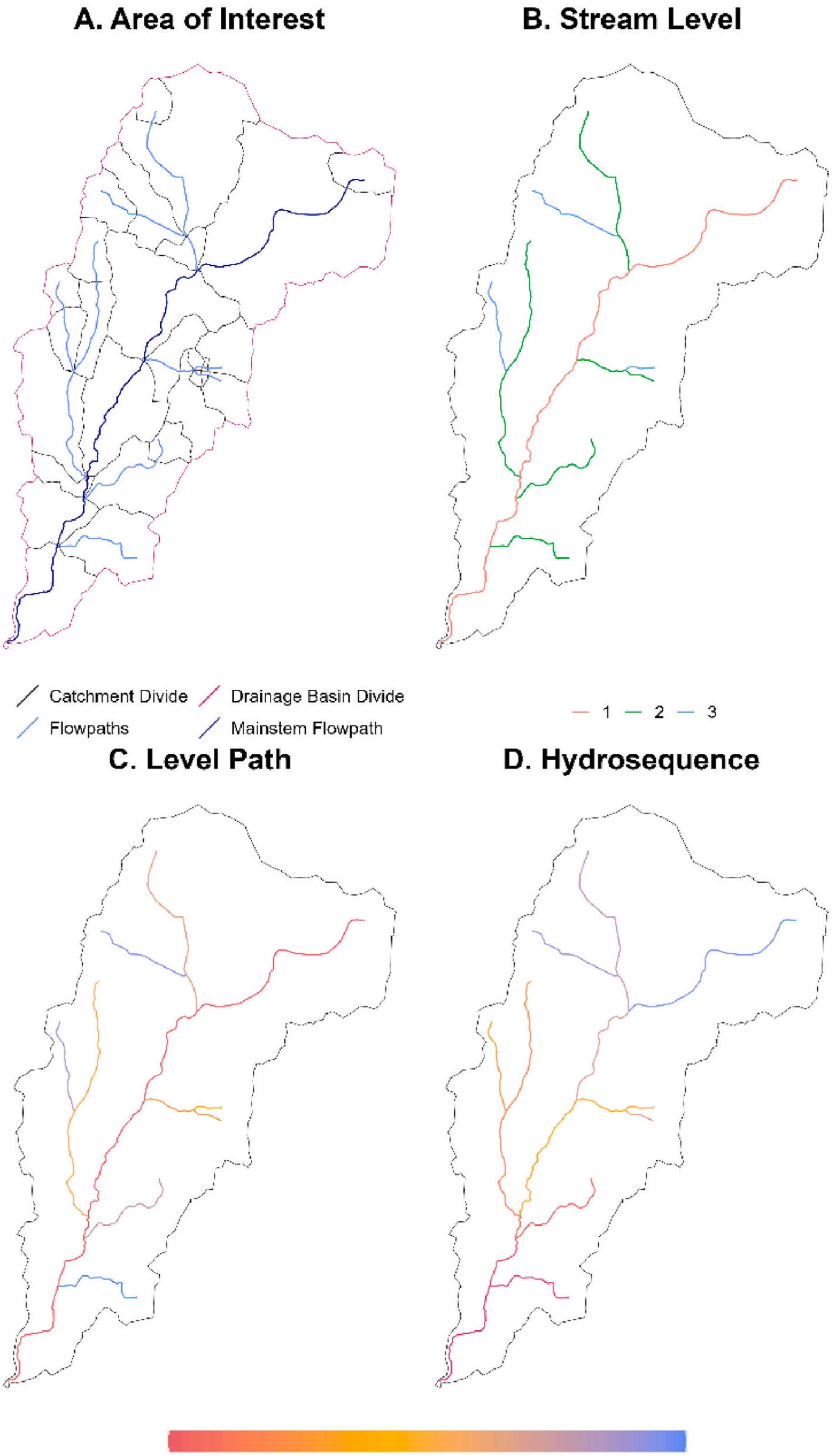

Fig. 5. (A) A set of incremental catchment areas surrounded by their divides (black) each have a flowpath (sky blue). The incremental catchment areas constitute a drainage basin boundary (red) that has a predominate mainstem (dark blue).
(B) Stream level values are constant along mainstem paths and express how deeply nested the drainage basin network is.(C) Level path values are constant along mainstem paths and are derived from the hydrosequence of their outlet flowpath. (D) Smaller 'hydrosequence' values are guaranteed to be downstream of larger values along connected paths.

identifier for cross-dataset integration is impossible. Level path and hydrosequence form the basis for a number of additional attributes useful for hydrologic networking. Table 2 summarizes these attributes' definitions and their purposes for hydrologic network operations.

2. Methods

The main improvements made to the NHDPlusV2 network have been to the network topology. Anytime there is a change to the topology, there are cascading impacts to the base attributes (Table 1), the resulting hydrosequence, level path, and their derived values (Table 2). A robust, open-source workflow for regenerating the derived values from the base attributes should help meet the objectives of a “reference fabric” that represents the legacy of improvements and is able to adapt to future change.

2.1. Source data

Processing on the *reference flow network* started by combining the “Value Added Attribute” table of the NHDPlusV2 (McKay et al., 2015), with the network improvements contained in the e2NHDPlus (Brakebill et al., 2020), and NWM (National Oceanic and Atmospheric Administration, 2021) applications.

The first challenge in combining the networks is associating their respective network topologies. The NHDPlusV2 and e2NHDPlus representation of the network are designed to account for divergences and thus include nodes in the network topology representation. As noted in Table 1, this node topology can be used to derive an edge list. A node topology is included in the NHDPlusV2 and e2NHDPlus, but not the NWM. The NWM only includes an edge list topology (defined as “link” and “to” in the NWM “RouteLink” file). The respective topologies are shown in Table 3.

In general, modifications to network connectivity in the NWM network established connections where the NHDPlusV2 indicated disconnected network. Modifications in the e2NHDPlus are generally corrections to the “main path” where a divergence occurs in the NHDPlusV2. Figs. 6 and 8 show the spatial distribution of changes to the network. While these are the initial improvements integrated with the network, they are not exhaustive. As new network improvements are identified, the design of the *reference flow network* could support incorporation of the updates.

2.2. nhdplusTools R package

The nhdplusTools R package (Blodgett and Johnson, 2022) houses the majority of the hydrologic network and hydrographic data functionality used in development of the *reference flow network*. Workflow repositories associated with the Mainstem Rivers of the Conterminous United States data release (Blodgett, 2023; Blodgett, 2023-1) and the broader Geospatial Fabric for National Hydrologic Modeling data release (provisional at the time of writing as described in https://waterdata.usgs.gov/blog/nldi-intro/(Blodgett and Johnson, 2023)) contain additional data manipulation logic. The following sections describe hydrologic network attribute generation functionality built into nhdplusTools and used in the workflows.

All nhdplusTools functions are implemented using a rigorous test-driven development style. That is, tests were developed as part of the software development process verifying that results match expectations precisely from initial implementation into the future. Most functions have associated tests that reproduce NHDPlusV2 attributes and verify that results are equivalent and tests that verify fine grained details and edge cases. Additionally, as bugs and edge cases were identified, tests to reproduce the bug were implemented as part of the fix, ensuring the issue does not reappear as the code base is modified in the future. This testing approach is critical to ensure accurate results in future use of workflows with new and more regularly updated data. nhdplusTools is the first open-source implementation of the NHDPlus data model attributes that uses this test-driven development style.

Table 3
Topology types and attributes in three input datasets. (NHDPlusV2 McKay et al., 2015), (NWM National Oceanic and Atmospheric Administration, 2021), (e2NHDPlus Brakebill et al., 2020).

Dataset	Edge list Topology	Node Topology
NHDPlusV2	comid → tocomid	fromnode → tonode
NWM	link → to	
e2NHDPlus		fromnode → tonode

2.3. Workflow

The workflow described here is in two parts, (1) the logic for combining NHDPlusV2, NWM, and e2NHDPlus networks. (2) the application of nhdplusTools to regenerate network attributes.

2.3.1. Network combination

The three input datasets were developed from the NHDPlusV2 originally and all use the same common identifier (“comid”). As a first step, the NHDPlusV2 and e2NHDPlus tables were joined based on “comid”. In places where there was disagreement in the node topology, the e2NHDPlus modifications were kept. Once changed, a new “tocomid” (edge list) field was created based on the updated node topology. Out of a total 2.7 million features in the NHDPlusV2, 2151 “tocomid” values and 2,769 divergence indicators were changed.

With this network, connections from the NWM could be added. The following prefilters were applied to avoid changes that could not be used for the *reference flow network*:

1) Where a given feature had more than one “tocomid”, the divergence indicator was used to limit network connectivity to one and only one downstream feature. There were about 73,000 diversions avoided here.
2) No changes were made to relationships involving features with a “coastal” feature type code even though some NWM changes were directed to coastline features.
3) Changes were not applied where more than ten features were upstream of one downstream feature. Some NWM changes were directed through large lakes and could not be used.
4) Only changes directing flow to features originally part of the NHDPlusV2 were considered for inclusion. Some NWM changes went to features not included in the NHDPlusV2 domain and could not be used.
5) After release of the first version of the *reference flow network,* an issue was found such that changes to primary downstream connectivity applied in e2NHDPlus were being reverted when applying NWM changes. To fix the issue, any disagreement in downstream connectivity where the e2NHDPlus made a change were not altered when incorporating the NWM network. The *reference flow network* was released as a version 2 (Blodgett, 2023-1)

With these caveats applied, 201 features (locations shown in Fig. 6) had network connectivity in the NWM that did not exist in the e2NHDPlus update of NHDPlusV2. For these, the divergence and “tocomid” attributes were updated. If what was a divergence was to become a main path or vice versa, the divergence indicator was switched. The “tocomid” attribute was then switched to that indicated by the NWM. With these applications, the resulting network retained the information in Table 4 from which NHDPlus network attribute calculations (Table 2) could be made.

2.3.3. Application of nhdplusTools

With the completed network defined in Table 4, one remaining attribute is needed per Table 1, a weight for determination of primary upstream paths. For this application, the arbolate sum was calculated using the calculate_arbolate_sum nhdplusTools (Blodgett and Johnson, 2022) function. Once added, a series of nhdplusTools functions, shown in Fig. 7, were executed sequentially to generate the hydrosequence and level path identifiers and the NHDPlus data model attributes in Table 2.

- **calculate_arbolate_sum** calculates the sum of all upstream *flowpath* lengths for each *flowpath* outlet.
- **get_sorted** walks the network of features from outlets to headwaters, returning data in the order it was encountered. The row number of

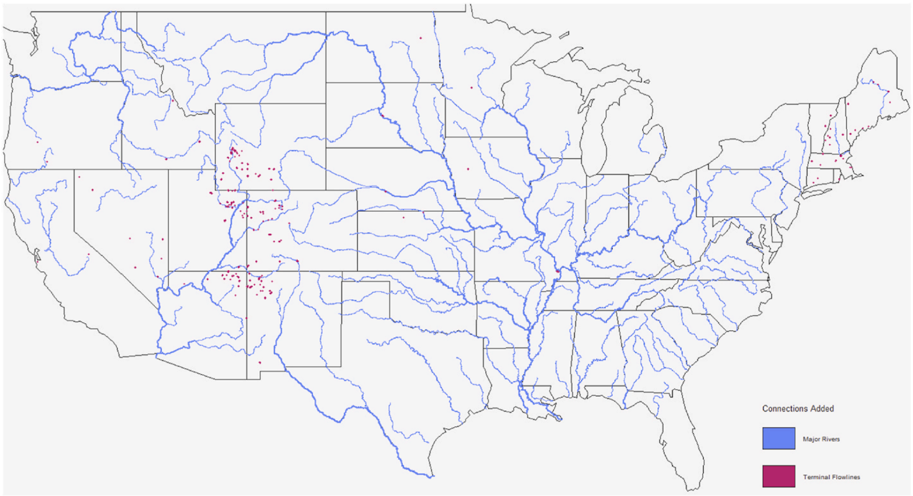

Fig. 6. Shows locations where terminal flowlines in NHDPlusV2 (McKay et al., 2015) were connected and are no longer terminal in the *reference flow network*.

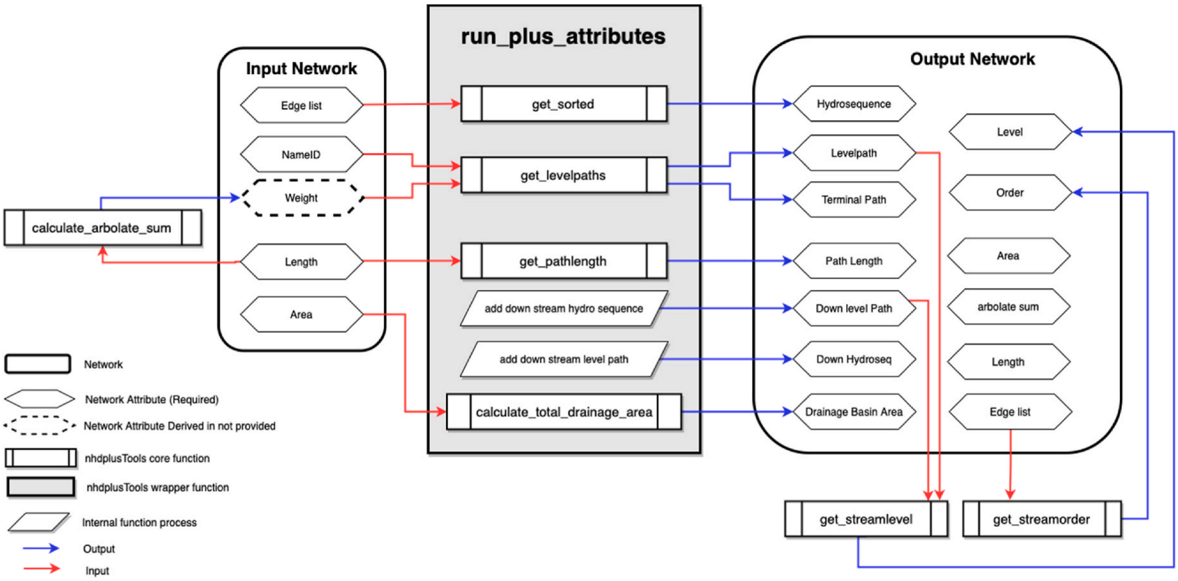

Fig. 7. Shows the input networks, functions that require them, and output variables.

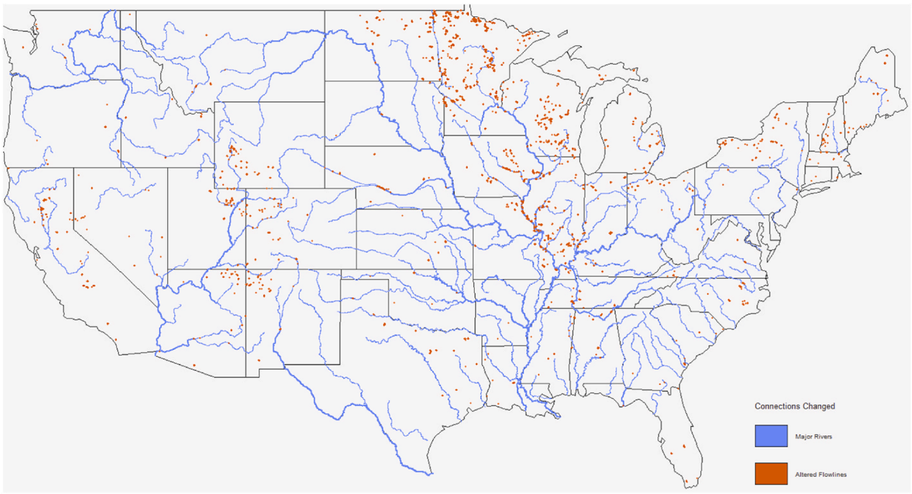

Fig. 8. Shows locations where downstream connections from NHDPlusV2 (McKay et al., 2015) were altered such that a given flowline is now connected to a different downstream flowline in the *reference flow network*.

Table 4
Attributes included in initial network used in creation of the *reference flow network* (Blodgett, 2023-1).

Purpose	Attribute names
Flowpath edge list	comid, tocomid
Flowpath node topology	fromnode, tonode, divergence
Identifiers	feature type, level path ID (used as nameid)
Measures	Length (of *flowpath*), area of *catchment area*

the returned features is a topological sort for the network that can be used as a hydrosequence.

- **get_levelpaths** uses a name identifier (in this application, the name is the NHDPlus (McKay et al., 2015) levelpathid which corresponds to the NHD Geographic Names Information System (U.S. Geological Survey, 2021) name) and weight to identify dominant level paths through the network. The upstream tributary with the same name is considered dominant unless the weight is X times larger (X is defined by the user) for the unnamed or differently named upstream feature. For the work here, arbolate sum was used for the weight and the override factor was set to 5.
- **get_pathlengths** walks the network of features from outlets to headwaters adding the length of a feature to the already visited pathlength of its downstream neighbor.
- **calculate_total_drainage_area** accumulates incremental drainage area starting at headwaters working downstream summing upstream neighbor's total drainage area.
- **run_plus_attributes** function calls get_sorted, get_levelpaths, get_pathlengths, and calculate_total_drainage_area and adds down level path and down hydrosequence with a table join post process.

As noted early in this process summary section, the name identifier used in get_levelpaths was actually the level path from NHDPlusV2. Typically, this would be a name from a source of geographic names. In the NHD, this is derived from the U.S. Geographic Names Information System (U.S. Geological Survey, 2021). This was a convenient way to ensure the new level path identifiers would correspond to the NHDPlusV2 source except where the weight indicated is 5 times greater than the arbolate sum along the name indicated path. The original NHDPlusV2 level paths followed names without this override, so by using that level path here, names are generally followed while major outliers could be fixed. Notable outliers were related to artificial paths that form connections in lakes and rivers that did not have the name of a major river associated with them. In this case, a named tributary would be followed rather than the major upstream path that would have otherwise been the obvious option to follow. The selection of 5 for the override was based on evaluation of instances where following the name would cause issues but not set so high as to change important confluences such as the Missouri river and Mississippi river (arbolate sum 940,400 km vs 302,300 km).

Once processing described just above was complete, the NHDPlusV2 level path identifier (i.e., nameid) was dropped and the "weight" attribute renamed to arbolate sum. A stream order and stream level were then added to the output network by calling the get_streamlevel and get_streamorder functions. While not applicable to modeling and data integration applications, stream order and stream level were added for completeness relative to NHDPlusV2 value added attributes. As a last step, useful NHDPlusV2 attributes that were not affected by network topology were joined to the output network and saved.

3. Results

The *reference flow network* (Blodgett, 2023-1) has been checked and reviewed using automated and manual methods. Automated checks ensured that expected relationships between identifiers such as hydrosequence, level path, and terminal path were correct, and that all attributes were fully populated and in their expected range. Due to the size of the dataset, manual checks were not comprehensive but were used to identify some systematic issues (specifically with initializing some coastal outlets) that were addressed prior to release of the dataset. Manual checks also cross validated attributes such as total drainage area between NHDPlusV2 and this new network. Due to the addition of previously disconnected areas and alteration of primary vs diverted path, fully automated checks of accumulated network attributes were not possible as no one "accurate" value is available for all feature attributes. However, some wholistic validation checks were performed. For example the only places with more than 10% difference in total drainage area (Fig. 9) were are associated with nearby alteration of network connections (Fig. 8). Note that the absence of yellow overlay in Fig. 9 indicates agreement in total drainage area.

In addition to automated and manual checks, the *reference flow network* (Blodgett, 2023-1) has been tested through its use as a primary input to the ongoing National Reference Fabric development (Fig. 1) (Blodgett and Johnson, 2023). One of these derived products is a CONUS wide hydrofabric created for development of the NOAA NextGen National Water Resource Modeling Framework (Johnson, 2022). For this initial proof of concept, a complete 2-month, hourly, CONUS scale hydrologic simulation has been executed on this dataset. Each of these modeling applications has been a rigorous test of the dataset's continuity and quality.

In the interest of FAIR data and the Internet of Water, the *reference flow network* has also been tested through its use as the primary input to the "*Mainstem Rivers of the Conterminous United States.*" data release (Blodgett, 2023). Through development of that data release, the *reference flow network* has been subjected to in depth verification. The workflow that created the "*Mainstem Rivers of the Conterminous United States*" data release, identified matching level paths between the *reference flow network* and several other hydrographic networks. Manual inspection (e.g., not automated) conducted during dataset development and the fact that strong correspondence was found between many networks is further evidence of the quality of this network.

4. Discussion

While implemented at a national scale, this work is also intended to serve the needs of regional and local applications. The national model applications this work supports may provide local-scale information and multi-purpose local modeling capacity. Regional and local studies may be able to conduct research and assessment activities reusing the network data, data integration it enables, and/or the tool chain used in its creation. These three modes of reuse illustrate the value that could be derived from this work:

1) adoption of network features and attributes,
2) use of the network features as a data integration aid, and
3) use of the tool chain developed in support of 1 and 2 on various hydrographic datasets meeting minimum data requirements.

US Federal water resources research, assessment, and forecasting applications (Michelsen et al., 2016; Miller et al., 2020; Ogden et al., 2021) were the primary motivation for this work and the ability for local and regional stakeholders to utilize it for their own context is important, but the work also has global significance. The tools developed for this work concerning US national hydrographic datasets (Blodgett and Johnson, 2022) have also been used with a global network of river features. For the National Hydrologic Model Geospatial Fabric, the global "MERIT Basins" dataset (Lin et al., 2019; Yamazaki et al., 2019) was used in the Alaska domain. The tools for creation of NHDPlus network attributes such as level path were also used for a complete global river network web visualization (Learner, 2023). These data are available in *Mainstem rivers of the world based on MERIT hydrography and natural Earth names* (Blodgett, 2022) and could support a global dataset

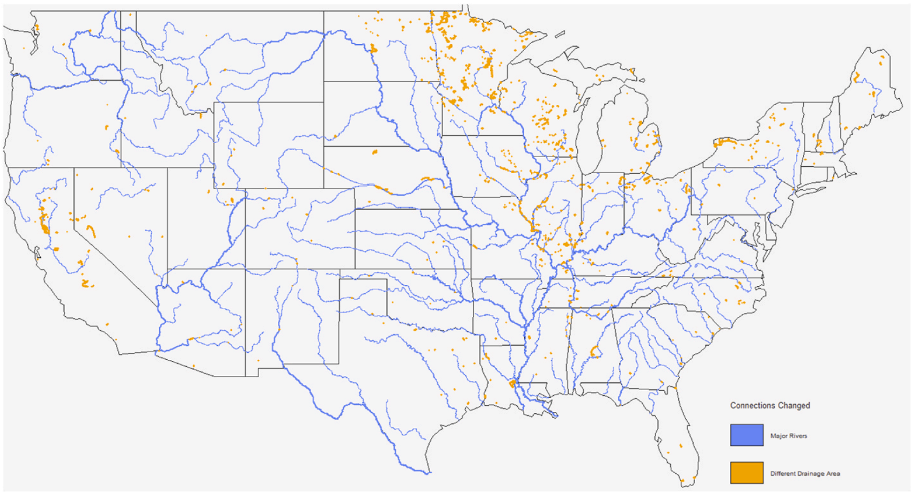

Fig. 9. Shows locations where total drainage area is more than 10% different from the NHDPlusV2 (McKay et al., 2015) and the *reference flow network (Blodgett 2023-1)*. This comparison uses dendritic total drainage area assuming 0 drainage area at the top of any secondary (diverted) flowline. Note that the absence of yellow overlay indicates agreement in total drainage area.

of flow network features as has been done for the continental US. Additionally, the idea of a *reference flow network* could be extended to include other kinds of reference features, like coastal zones and hydrogeology.

The specificity of the data models and implementation of this work are intentionally limited in scope with an eye on incremental advancement. Divergences and "divergence groups" is a topic that was out of scope for this work but inclusion may be required support some applications. Referred to as "diversion groups" in the report accompanying the e2NHDPlus dataset (Brakebill et al., 2020), this concept recognizes that some sets of diversions represent a group of diverted paths whose flow recombines some distance downstream. These groups of diversions result from many physical phenomena but are most commonly found where rivers combine in a complex system of low slope channels. In contrast to "hydrologic diversions", where a diversion forms the *mainstem* of its own *drainage basin*, the combination of channels that make up a divergence group can be treated as a single *flowpath* relative to overall network connectivity. Representing the distinction between groups of divergent paths that recombine and more significant "hydrologic diversions" is challenging due to many factors and is an important topic for future efforts of this overall body of work.

A hydrologic flow network, as presented in this report, is a dendritic tree represented as a directed acyclic graph. In reality, each element of the network is, at some time, a flowing body of water. Representation of flowing waterbodies as linestrings as has been presented here is common and useful, but not complete by any means. Visually, a line representation of a river has some nominal width that, depending on map scale, represents some real-world waterbody width. The datasets discussed here do not include a physically based attribute to tie to this nominal line width. Taking this one step further, a given body of water has some depth, which could be expressed simply as a "representative depth" or more precisely with a collection of cross sections or other physical representations. This problem of waterbody integration with flow network representation is confounded by large waterbodies that, especially when flooded, inundate parts of the hydrologic landscape and associated flow network. The joint channel (as container) and waterbody (as contained fluid body) concepts presented in HY_Features provide a framework for this problem. Research into how to separate the topo-bathymetric channel and hydrodynamic waterbody representation may allow necessary integrations of waterbody, in channel, and river corridor models.

5. Conclusions

The *reference flow network* is built on NHDPlusV2 (McKay et al., 2015), an improved version of NHDPlusV1, which itself was based on best available inputs and was the product of significant quality control (Bondelid et al., 2010). Additional network improvements from the e2NHDPlus dataset, (Brakebill et al., 2020), and the NWM hydrofabric (National Oceanic and Atmospheric Administration, 2021) were then applied in an open-source reproducible workflow. National scope integration and network manipulation work have successfully used this network with as good or better results than previous efforts (Johnson, 2022). Combined, these advances justify application of the *reference flow network* in local to national scale hydrologic modeling applications. However, representation of hydrologic features is imperfect with numerous sources of ambiguity and inaccuracy. It is because of this that a second, equally critical part of this work is an automated and rigorously tested workflow that generates the attributes based on changes in topology (Blodgett and Johnson, 2022).

It should be expected that, while rare, there will be cases where the network connectivity in the *reference flow network* does not match reality. Some will be instances where the scale at which these data resolve features is too coarse, others will be cases where on-the-ground conditions are misrepresented or have changed. Given this, updates should be expected when issues are encountered. However, the design of this dataset and system surrounding it is intended for this and capable of rapidly integrating changes to serve the needs of the hydroscience community.

Software and data availability

All software and data used in preparation of this report can be found at:

Blodgett, D.L., 2022, Mainstems Workflow: HU12 NHDPlusV2 NHDPlus HiRes Matching, https://doi.org/10.5066/P9H0PTRH.

Blodgett, D.L., Johnson, J.M., 2022, nhdplusTools: Tools for Accessing and Working with the NHDPlus, https://doi.org/10.5066/P97AS8JD.

Blodgett, D.L., 2023, Updated CONUS river network attributes based on the E2NHDPlusV2 and NWMv2.1 networks (ver. 2.0, February 2023): U.S. Geological Survey data release, https://doi.org/10.5066/P976XCVT.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data is referenced in the paper

Acknowledgements

All U.S. Geological Survey work documented in this report was funded by the U.S. Geological Survey Water Mission Area Water Availability and Use Science Program. National Oceanic and Atmospheric Administration National Weather Service contributions are part of the Next Generation Water Modeling Framework initiative of the Office of Water Prediction.

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Archfield, S.A., Clark, M., Arheimer, B., Hay, L.E., McMillan, H., Kiang, J.E., Seibert, J., Hakala, K., Bock, A., Wagener, T., Farmer, W.H., Andréassian, V., Attinger, S., Viglione, A., Knight, R., Markstrom, S., Over, T., 2015. Accelerating advances in continental domain hydrologic modeling. Water Resour. Res. 51, 10078–10091. https://doi.org/10.1002/2015WR017498.

Blodgett, D., 2020. Second Environmental Linked Features Experiment. http://www.opengis.net/doc/PER/SELFIE-ER.

Blodgett, D.L., 2022. Mainstem Rivers of the World Based on MERIT Hydrography and Natural Earth Names. U.S. Geological Survey data release. https://doi.org/10.5066/P9O15C70.

Blodgett, D., 2023. Mainstem Rivers of the Conterminous United States (Ver. 2.0, February 2023). *U.S. Geological Survey Data Release*. U.S. Geological Survey. https://doi.org/10.5066/P92U7ZUT.

Blodgett, D., 2023-1. Updated CONUS River Network Attributes Based on the E2NHDPlusV2 and NWMv2.1 Networks (Ver. 2.0, February 2023) in U.S. Geological Survey Data Release. U.S. Geological Survey. https://doi.org/10.5066/P976XCVT.

Blodgett, D., 2023-2. The hydro network-linked data Index. U.S. Geological Survey Water Data for the Nation blog. U.S. Geological Survey. https://waterdata.usgs.gov/blog/nldi-intro/.

Blodgett, D., Dornblut, I., 2018. OGC® WaterML 2: Part 3 - Surface Hydrology Features (HY_Features) - Conceptual Model. https://docs.opengeospatial.org/is/14-111r6/14-111r6.html. (Accessed 4 June 2018).

Blodgett, D., Johnson, J.M., 2022. *nhdplusTools: Tools For Accessing and Working with the NHDPlus* (0.5.1). U.S. Geological Survey. https://doi.org/10.5066/P97AS8JD.

Blodgett, D., Johnson, J.M., 2022-1. Hydrologic Modeling and River Corridor Applications of HY_Features Concepts. http://www.opengis.net/doc/PER/22-040.

Blodgett, D., Johnson, J.M., 2023. Progress toward a Reference Hydrologic Geospatial Fabric for the United States *U.S. Geological Survey Water Data For the Nation blog.* https://waterdata.usgs.gov/blog/hydrofabric/.

Blodgett, D., Johnson, J.M., Sondheim, M., Wieczorek, M., Frazier, N., 2021. Mainstems: a logical data model implementing mainstem and drainage basin feature types based on WaterML2 Part 3: HY Features concepts. Environ. Model. Software 135. https://doi.org/10.1016/j.envsoft.2020.104927.

Bock, A.E., Santiago, M., Wieczorek, M.E., Foks, S.S., Norton, P.A., Lombard, M.A., 2021. GIS features of the geospatial fabric for the national hydrologic model, version 1.1 (ver. 3.0, november 2021). In: *U.S. Geological Survey data Release.* U.S. Geological Survey.. https://doi.org/10.5066/P971JAGF.

Bondelid, T., Johnston, J., McKay, L., Moore, R., Rea, A., 2010. NHDPlus Version 1 (NHDPlusV1) User Guide, vol. 126. ftp://ftp.horizon-systems.com/NHDPlus/NHDPlusV1/documentation/NHDPLUSV1_UserGuide.pdf.

Brakebill, J.W., Schwarz, G.E., Wieczorek, M.E., 2020. An enhanced hydrologic stream network based on the NHDPlus medium resolution dataset. In: Scientific Investigations Report. https://doi.org/10.3133/sir20195127.

Clark, M.P., Fan, Y., Lawrence, D.M., Adam, J.C., Bolster, D., Gochis, D.J., Hooper, R.P., Kumar, M., Leung, L.R., Mackay, D.S., Maxwell, R.M., Shen, C., Swenson, S.C., Zeng, X., 2015. Improving the representation of hydrologic processes in Earth system models. Water Resour. Res. 51 (8), 5929–5956. https://doi.org/10.1002/2015WR017096.

Cormen, T.H., Leiserson, C.E., 2022. Introduction to Algorithms, fourth ed. MIT Press.

Dewald, T.G., 2017. Making the Digital Water Flow the Evolution of Geospatial Surface Water Frameworks. USEPA, Office of Water, Washington, DC. https://s3.amazonaws.com/nhdplus/NHDPlusV21/Documentation/History/Making_the_Digital_Water_Flow.pdf.

Horn, C.R., 1994. Appendix A to Metadata for RF1 USEPA Reach File 1. https://water.usgs.gov/GIS/browse/rf1_appA.HTML.

Internet of Water Organization, 2021. Internet of Water Principles. *internetofwater.org/internet-of-water-principles/*.

Internet of Water Organization, 2023. *geconnex.us. geoconnex.us*.

Johnson, J.M., 2022. National Hydrologic Geospatial Fabric (Hydrofabric) for the Next Generation (NextGen) Hydrologic Modeling Framework. HydroShare. http://www.hydroshare.org/resource/129787b468aa4d55ace7b124ed27dbde.

Learner, S., 2023. River Runner Global. https://river-runner-global.samlearner.com/.

Lin, P., Pan, M., Beck, H.E., Yang, Y., Yamazaki, D., Frasson, R., David, C.H., Durand, M., Pavelsky, T.M., Allen, G.H., Gleason, C.J., Wood, E.F., 2019. Global reconstruction of naturalized river flows at 2.94 million reaches. Water Resour. Res. 55 (8), 6499–6516. https://doi.org/10.1029/2019WR025287.

McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., Rea, A., 2015. NHD Plus Version 2 : User Guide. https://www.epa.gov/system/files/documents/2023-04/NHDPlusV2_User_Guide.pdf.

Michelsen, A.M., Jones, S., Evenson, E., Blodgett, D., 2016. The USGS water availability and use science program: needs, establishment, and goals of a water Census. J. Am. Water Resour. Assoc. 52 (4) https://doi.org/10.1111/1752-1688.12422.

Miller, M.P., Clark, B.R., Eberts, S.M., Lambert, P.M., Toccalino, P., 2020. Water priorities for the nation—U.S. Geological Survey integrated water availability assessments. In: U.S. Geological Survey Fact Sheet. https://doi.org/10.3133/fs20203044.

National Oceanic and Atmospheric Administration, 2016. National Water Model Improving NOAA's Water Prediction Services. https://water.noaa.gov/documents/wrn-national-water-model.pdf.

National Oceanic and Atmospheric Administration, 2021. National water model (NWM) parameter files V2.1. In: National Oceanic and Atmospheric Administration Public Data. U.S. National Oceanic and Atmospheric Administration. https://www.nohrsc.noaa.gov/pub/staff/keicher/NWM_live/NWM_parameters/.

OECD, 2021. Data as a Strategic Asset for the Public Sector. OECD Publishing, Paris. https://doi.org/10.1787/785cb67f-en. *Government at a Glance 2021*.

Ogden, F., Avant, B., Bartel, R., Blodgett, D., Clark, E., Coon, E., Cosgrove, B., Cui, S., Kindl da Cunha, L., Farthing, M., Flowers, T., Frame, J., Frazier, N., Graziano, T., Gutenson, J., Johnson, D., McDaniel, R., Moulton, J., Loney, D., et al., 2021. The Next Generation Water Resources Modeling Framework: Open Source, Standards Based, Community Accessible, Model Interoperability for Large Scale Water Prediction, 2021, AGU Fall Meeting, 2021 H43D-01. https://ui.adsabs.harvard.edu/abs/2021AGUFM.H43D..01O.

Regan, R.S., Juracek, K.E., Hay, L.E., Markstrom, S.L., Viger, R.J., Driscoll, J.M., LaFontaine, J.H., Norton, P.A., 2019. The U. S. Geological Survey National Hydrologic Model infrastructure: rationale, description, and application of a watershed-scale model for the conterminous United States. Environ. Model. Software 111, 192–203. https://doi.org/10.1016/j.envsoft.2018.09.023.

U.S. Geological Survey, 2021. Geographic names information system (GNIS). U. S. Jpn. Outlook. *Geological Survey*. https://www.usgs.gov/tools/geographic-names-information-system-gnis.

U.S. Geological Survey, 2022. National Hydrography Dataset Data Model Poster, 2.3.1 V2. U.S. Geological Survey. https://www.usgs.gov/media/files/national-hydrography-dataset-data-model-poster-231-v2.

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3 (1), 160018. https://doi.org/10.1038/sdata.2016.18.

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P.D., Allen, G., Pavelsky, T., 2019. MERIT Hydro: A High-resolution Global Hydrography Map Based on Latest Topography Datasets. Water Resources Research. https://doi.org/10.1029/2019WR024873, 2019WR024873.