Discrete Global Grid System-based Flow Routing Datasets in the Amazon and Yukon Basins

Chang Liao[1], Darren Engwirda[2], Matthew G Cooper[1], Mingke Li[3], and Yilin Fang[4]

[1]Atmospheric, Climate, and Earth Sciences, Pacific Northwest National Laboratory, Richland, WA, USA
[2]T-3 Fluid Dynamics and Solid Mechanics Group, Los Alamos National Laboratory, Los Alamos, NM, USA
[3]Department of Geomatics Engineering, Schulich School of Engineering, University of Calgary, Calgary, Canada
[4]Hydrology Group, Pacific Northwest National Laboratory, Richland, WA, USA

Correspondence: Chang Liao (chang.liao@pnnl.gov)

Abstract.

Discrete Global Grid systems (DGGS) are emerging spatial data structures widely used to organize geospatial datasets across scales. While DGGS have found applications in various scientific disciplines, including atmospheric science and ecology, their integration into physically based hydrologic models and Earth System Models (ESMs) has been hindered by the lack of flow-routing datasets based on DGGS. In response to this gap, this study pioneers the development of new flow routing datasets using Icosahedral Snyder Equal Area (ISEA) DGGS and a novel mesh-independent flow direction model. We present flow routing datasets for two large basins, the tropical Amazon River Basin and the Arctic Yukon River Basin. These datasets (1) facilitate the adoption of DGGS for hydrologic models and (2) provide flow routing inputs for evaluation of DGGS-based flow routing in the Amazon and Yukon River Basins. The data are available at https://doi.org/10.5281/zenodo.8377765 (Liao, 2023).

1 Introduction

Discrete Global Grid systems (DGGS) are emerging spatial data models that use hierarchical tessellations of cells to partition and address the Earth's surface. DGGS have been widely adopted as a standard data fabric to organize geospatial datasets across various granularities (Goodchild, 1994; Kimerling et al., 1999; Sahr, 2024; Purss et al., 2016). DGGS, such as the Icosahedral Snyder Equal Area (ISEA) aperture 3 Hexagon (3H), are used in many disciplines, including Geographic Information System (GIS) (Sahr, 2019), hydrology (Li et al., 2022), atmospheric science (Randall et al., 2002), and ecology (Ellis et al., 2021; Mechenich and Žliobaitė, 2023). However, DGGS have seen limited adoption in physically-based, spatially distributed hydrologic models and Earth System Models (ESMs) (Li et al., 2022), mainly because ready-for-use flow routing datasets based on DGGS are unavailable.

Flow routing datasets are essential for spatially distributed hydrologic models, and they typically rely on two data model paradigms. The first one is the rectangular mesh-based grids, also known as rasters (Esri Water Resources Team, 2011; Wu et al., 2012). This method often requires high-quality digital elevation model (DEM) rasters such as METIR hydro (Yamazaki et al., 2019; Amatulli et al., 2022). It is also subject to several other limitations, including the challenge of coupling with other unstructured mesh-based numerical models. The second one is the vector-based polylines (Lin et al., 2021), which are often

produced through the combination of high-resolution raster-based and remote sensing product-based methods. However, these polylines often contain various artifacts, including disconnected segments, and thus cannot be directly used across different spatial scales (Huang and Frimpong, 2016). Another limitation of this method is the lack of communication between the river and its adjacent riparian zones.

We recently pioneered the ability to generate flow routing datasets using unstructured Model for Prediction Across Scales (MPAS) meshes (Engwirda and Liao, 2021; Liao et al., 2023b). Although MPAS meshes have gained traction in the oceanographic and atmospheric modeling communities, DGGS meshes are also widely adopted across the Earth Sciences and GIS communities. To date, there are no available DGGS-based flow routing datasets that include flow direction information. This is because existing DGGS-based hydrology datasets are often derived by resampling from existing raster-based datasets, which does not support vector-based datasets such as flow direction (Chaudhuri et al., 2021). Besides, most traditional flow direction models in various GIS software only support raster datasets. This highlights the need for a method to natively generate flow direction datasets within a DGGS-based framework.

Compared to structured rectangular meshes, including latitude-longitude geographic coordinate systems (GCS) and projected coordinate systems, DGGS have several advantages. These include 1) better spatial coverage and consistent spatial resolution for the high latitudes (Liao et al., 2020); 2) potential numerical performance improvement for coupled surface and subsurface hydrologic models (Liao et al., 2022); and 3) more flexibility in spatial resolution due to their hierarchical data structure. Specifically, the ISEA DGGS projection stands out for its benefits to hydrologic models and ESMs. As an equal-area icosahedral DGGS projection, ISEA eliminates the need for equal-area projection. In addition, the hexagonal grid geometry resolves ambiguity among cell neighborhoods by ensuring uniform adjacency, thereby offering significant advantages in the domain of hydrology.

This study breaks new ground by developing new flow routing datasets using the ISEA3H DGGS and our innovative mesh-independent flow direction model. We present flow routing datasets for the Amazon and Yukon Basins, which are among the world's largest river basins in the tropics and Arctic, respectively, and play important roles in local, regional, and global climate and ecosystems. These datasets can (1) facilitate the adoption of DGGS for hydrologic model development and (2) provide flow routing inputs for evaluating DGGS-based flow routing in the Amazon and Yukon River Basins.

2 Method

A list of datasets and models used in our workflow to produce DGGS-based flow routing datasets is depicted in Figure 1.

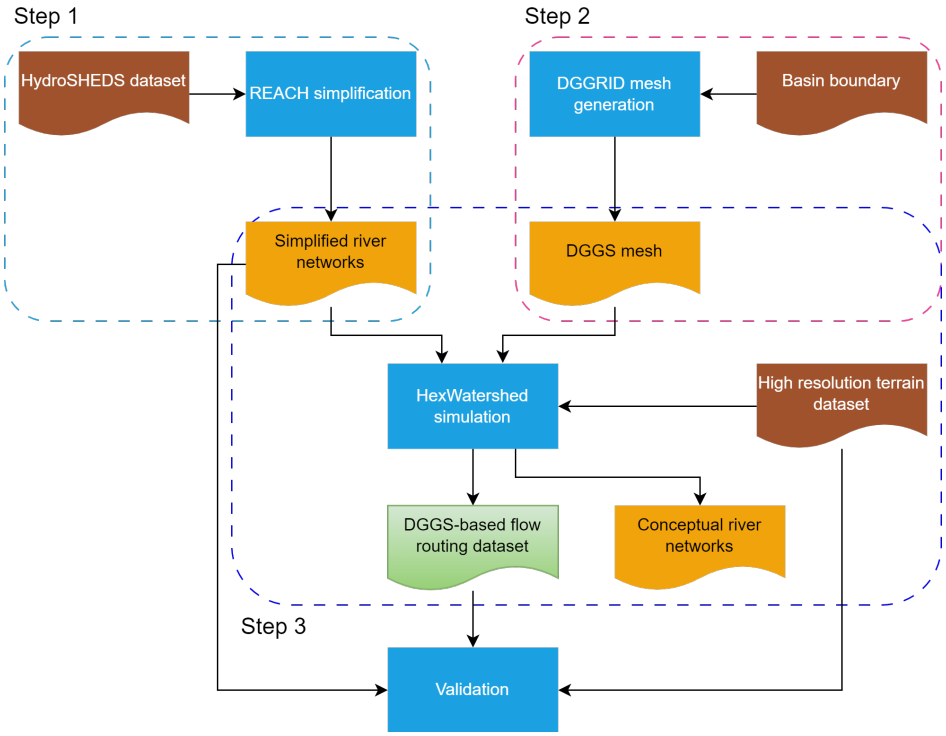

Figure 1. Workflow diagram demonstrating the DGGS-based flow routing dataset generation process using three steps (dashed boxes). Step 1: river network simplification using REACH; Step 2: DGGS mesh generation using DGGRID; Step 3: Flow direction modeling using HexWatershed (including PyFlowline). Brown boxes are user-provided datasets. Orange boxes are intermediate results (e.g., conceptual river networks modeled by PyFlowline). The light green box is the final data product.

We first introduce the input datasets used in each step and then the models used. Additional information is provided in the Supplementary Information.

2.1 Input datasets

2.1.1 Vector river networks

The vector river network datasets of the Amazon and Yukon Basins were obtained from the HydroSHEDS database (Lehner et al., 2008). HydroSHEDS v1 dataset is derived primarily based on elevation data obtained in 2000 by the United States National Aeronautics and Space Administration's (NASA) Shuttle Radar Topography Mission (SRTM). Specifically, we obtained the HydroRiver datasets of South America and the Arctic (https://www.hydrosheds.org/products/hydrorivers). While HydroSHEDS products, including the river networks, may not provide the highest level of accuracy in depicting river and basin maps, they are widely acknowledged and evaluated for applications at regional and global scales. This datasets are used in Step 1.

2.1.2 Vector watershed boundary

The vector Amazon Basin boundary was obtained through NASA's Oak Ridge National Laboratory (ORNL) Distributed Active Archive Center (DAAC) (Mayorga et al., 2012). The vector Yukon Basin boundary was obtained through HydroBASINS, which is part of the HydroSHEDS product. These datasets are used in Step 2.

2.1.3 Raster terrain datasets

High spatial resolution DEM datasets of the Amazon Basin at 30-arc-second ($\sim 1\,\text{km}$ at the equator) were obtained through NASA's ORNL DAAC (Saatchi, 2013). Similar to HydroSHEDS, this DEM was produced as a subset of the SRTM DEM. The flow accumulation and length datasets at the same spatial resolution are used for data validation. Similarly, void-filled DEMs and flow accumulation datasets of the Yukon Basin at 15-arc-second ($\sim 500\,\text{m}$ at the equator, $\sim 250\,\text{m}$ in Alaska) resolution were obtained from the HydroSHEDS. These datasets are used in Step 3.

2.2 Models

Our workflow primarily leverages three software models to produce the DGGS-based flow routing datasets. The models are run in sequence in three steps: 1) the REACH model pre-processes the vector river networks, i.e., HydroSHEDS, to produce the simplified river networks; 2) the DGGRID model generates the DGGS mesh using the basin boundaries; and 3) the HexWatershed model generates the flow routing datasets using outputs from Step 1 and 2. Descriptions of each model and step are provided below.

2.2.1 HydroSHEDS river network simplification using REACH

Because the full HydroSHEDS river network dataset contains millions of river channels that range between a few to thousands of kilometers, they cannot be represented equally in hydrologic and Earth system models. For example, any river channel less than $10\,\text{km}$ in length cannot be represented well if the mesh cell resolution (in length) is also $10\,\text{km}$. To address this challenge, we used the REACH library (Engwirda, 2023) to pre-process (simplify) the HydroSHEDS river network. In this step, only major river channels and tributaries resolvable at scales of interest are preserved. REACH employs a greedy network simplification algorithm in which the maximal set of river reaches is processed in priority order of increasing upstream catchment area. River reaches are removed incrementally if they meet the following criteria: 1) their length is shorter than a user-defined tolerance, or 2) they are geometrically closer to another higher priority reach segment than a user-defined tolerance. Upon removal of a given river reach, the downstream network is simplified — merging any newly contiguous segments into 'super-reaches' and updating their associated priorities. While heuristic in nature, this greedy approach leads to simplified river vector networks that are appropriate for both hydrological analysis and unstructured mesh generation, with the network pruned in a least-catchment-area-first manner. This retains hydrologically important reaches while removing geometrical features smaller than the desired mesh scale to ensure compatibility between the flow network and the computational grid.

The user-defined tolerance is a measure of how much detail the mesh should preserve and distinguish river channels because a mesh cell conceptually can only represent one main channel unless it is a river confluence. In practice, the user-defined tolerance is often set as the mesh cell's spatial resolution (in length), which can vary in space as well. In this study, because we use four resolution levels from the DGGRID model to generate the flow routing datasets, the corresponding resolutions in length (square root of area) are used as the user-defined tolerance parameters (Table 1).

Resolution level	Internode spacing (km)	Resolution in area (km^2)
10	31.7596	863.80061
11	18.341	287.93354
12	10.5871	95.97785
13	6.11367	31.99262
14	3.52911	10.66421

Table 1. The DGGRID mesh generation resolutions used to produce the flow routing datasets for Amazon and Yukon. These resolutions are chosen to accommodate major existing large-scale hydrologic and Earth System Models. The level 14 resolution is only used for mesh generation.

Figure 2 illustrates the simplified HydroSHEDS river networks in the Amazon Basin.

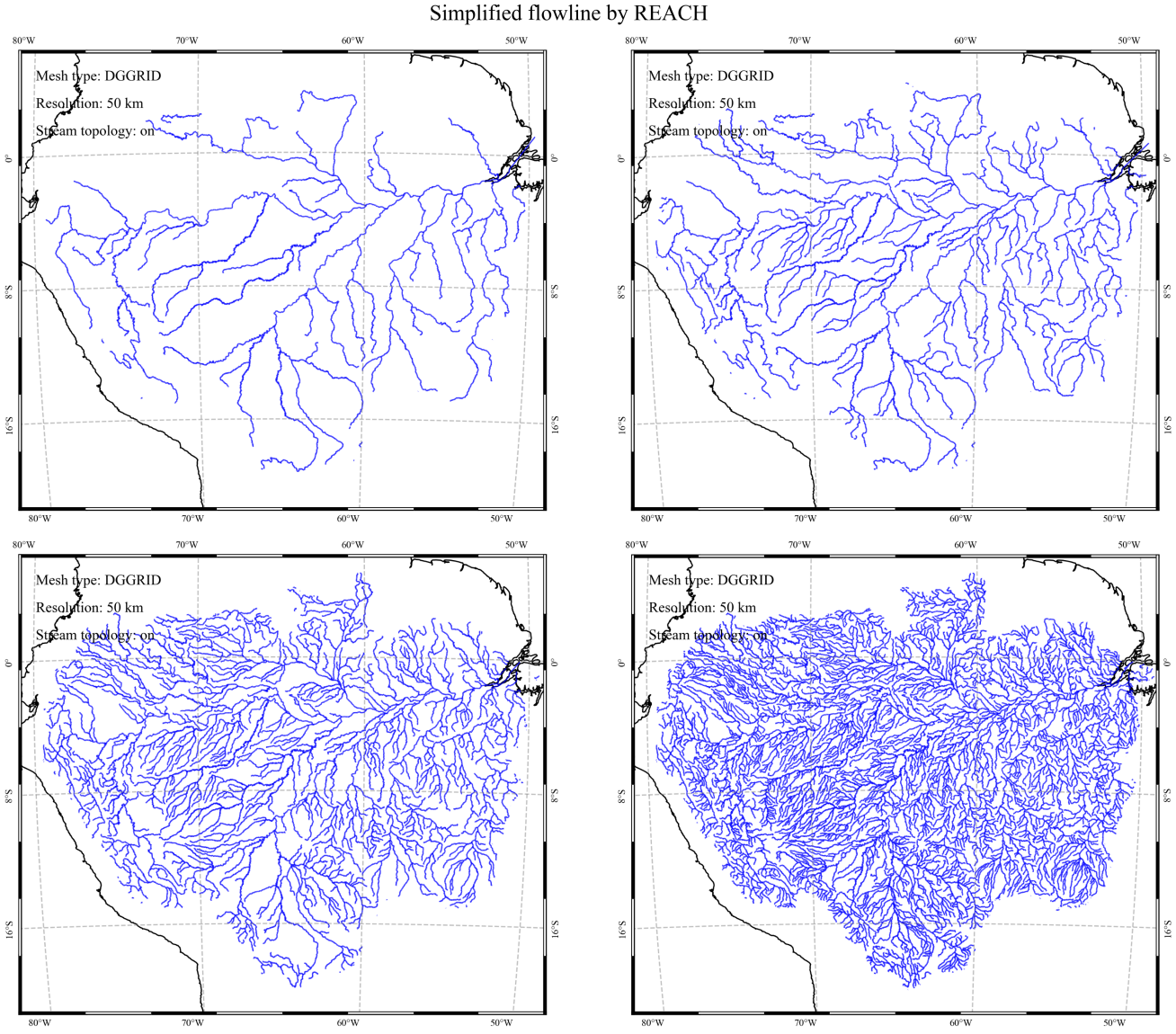

Figure 2. Simplified river networks using the REACH library at DGGRID ISEA3H level 10 to 13 resolutions in the Amazon Basin (Table 1). Black and blue lines are coastlines and river networks, respectively. As these datasets are extracted using the Amazon watershed boundary from the global HydroSHEDS river networks, there may be isolated river segments near the boundary of the basin, which are automatically excluded in the HexWatershed model.

2.2.2 Mesh generation using DGGRID

DGGRID is an actively maintained, open-source library, initially developed by Kevin Sahr in 2003, mainly used for generating and manipulating DGGS with diverse configurations (Sahr, 2024). The DGGRID library provides various grid geometry

options, including triangles, diamonds, and hexagons. It also allows for specifying multiple refinement ratios between successive resolutions, customized orientation relative to the Earth's surface, and different projection methods when generating the grids such as the ISEA and FULLER projections (Sahr, 2024). A list of parameters to define the mesh generation process is summarized in (Table A1). For the complete list of parameters, please refer to the DGGRID user manual (Sahr, 2024).

The DGGRID version 8.3 was used in our study to generate the ISEA Aperture 3 Hexagonal (ISEA3H) meshes with the default orientation. A total of five resolution levels from 10 to 14 are defined (Table 1). The level 14 mesh was used for validation only.

The four spatial resolutions (levels 10-13) were selected because most large-scale hydrologic models and Earth System Models run at approximately 0.5-degree (~ 50 km at the equator) spatial resolution, which is similar to resolution (in length) level 10, while many large-scale hydrologic models run at spatial resolutions of ~ 5 km, similar to resolution (in length) level 13. These four spatial resolutions, therefore, cover a wide range of hydrologic model applications.

Once the DGGRID model is built from its C/C++ source code, it can be called through several application programming interfaces (APIs), which have been implemented within the HexWatershed model (Liao, 2022a). As a result, Step 2 can be run as part of Step 3.

2.2.3 Flow direction modeling using HexWatershed

HexWatershed is a mesh-independent flow direction model for hydrologic models. Unlike most flow direction models that only support structured rectangle meshes, HexWatershed supports both structured and unstructured meshes. HexWatershed includes the state-of-science topological relationship-based river network representation and depression removal methods to generate high-quality flow routing datasets across scales (Liao and Cooper, 2023; Liao, 2022a). These methods allow the embedding of river networks and other hydrologic features within the flow routing map from regional to global scales. To achieve this, HexWatershed uses a two-step approach to model flow direction. First, it uses the mesh-river network intersection to build the topological relationship between mesh cells and river channels (e.g., upstream-downstream channel cells). Next, it uses a hybrid stream burning-depression filling algorithm to generate the flow direction between all the mesh cells. This step will first define the elevation and flow direction of the river channels and then process the remaining mesh cells. Additional explanations of these techniques are provided in the Supplementary Information and can be found in our two-part series of studies (Liao et al., 2023a, b).

The computational geometry algorithms within HexWatershed accept all types of mesh cells (e,g., rectangle, hexagon, triangle), and the depression removal algorithms automatically consider different numbers of neighbors when defining flow directions. Therefore, HexWatershed is mesh-independent and supports both structured and unstructured meshes.

In this study, we extended HexWatershed to support the DGGRID mesh type. Specifically, we implemented several APIs to set up a DGGRID model run and convert the DGGRID outputs to the HexWatershed model data structure (Step 2). Then we run HexWatershed v3.0 to generate flow routing datasets using the DGGRID-generated ISEA3H meshes at four different spatial resolutions (Table 1). For each spatial resolution, the HexWatershed model simulation includes the following steps:

a Prepare all the input datasets (outputs from Step 1) and binaries (DGGRID and HexWatershed C++ binaries) into a workspace folder;

b Call the PyFlowline Python package (Liao et al., 2023a; Liao and Cooper, 2023) to generate the conceptual river networks (Figure 1). PyFlowline is a core component in the HexWatershed model. This step includes three sub-steps:

- Pre-process the vector river network datasets, i.e., simplified HydroSHEDS river networks from Step 1. This step further processes the river networks, including re-building the stream segment indices and (Strahler) orders;
- Generate the DGGRID configuration file and run the DGGRID model to generate the DGGS mesh file. This is also the Step 2 in Figure 1;
- Model the conceptual river networks using the topological relationship-based reconstruction method.

c Assign elevation to the mesh cells based on raster DEM and each mesh cell boundary (Liao et al., 2022). A zonal mean resampling method is used by default;

d Conduct the depression removal. This step includes two sub-steps (Liao et al., 2023b):

- Run the topological relationship-based stream burning on the river cells and their riparian zone cells using outputs from Step 3b and 3c;
- Run the revised priority-flood depression filling for the remaining mesh cells.

e Export and visualize the model outputs, including the flow direction map and other flow routing parameters (Liao, 2022b).

Last, spatial visualizations were produced using Python packages, including Geospatial Data Abstraction Library (GDAL) (GDAL/OGR contributors, 2019) and PyEarth (Liao, 2022b).

3 Data record

These datasets contain four collections of flow-routing datasets corresponding to four spatial resolutions for both the Amazon and Yukon Basins (Liao, 2023). Within each collection, several files are provided, with a README file explaining each file. The results from resolution level 10 are used here for illustration purposes. The Supplementary Information provides visualizations of all four dataset collections with zoom-in views. All the spatial datasets are provided using GIS file formats (e.g., GeoJSON) with the WGS84 EPSG:4326 spatial reference.

3.1 Surface elevation

The **variable_polygon.geojson** file is a polygon-based GeoJSON data file. The attribute "elevation" stores the modeled mean surface elevation for each DGGRID mesh cell after the depression removal (Figures 3 and 4).

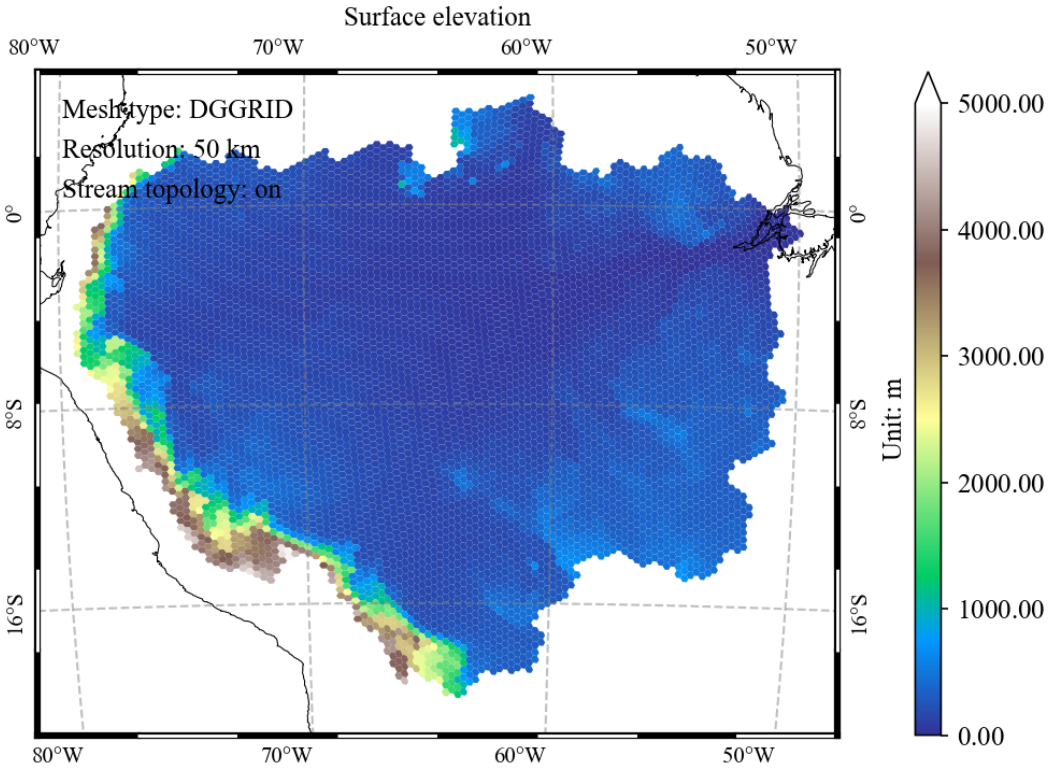

Figure 3. Spatial distribution of modeled surface elevation at DGGRID ISEA3H level 10 resolution in the Amazon Basin (unit: m).

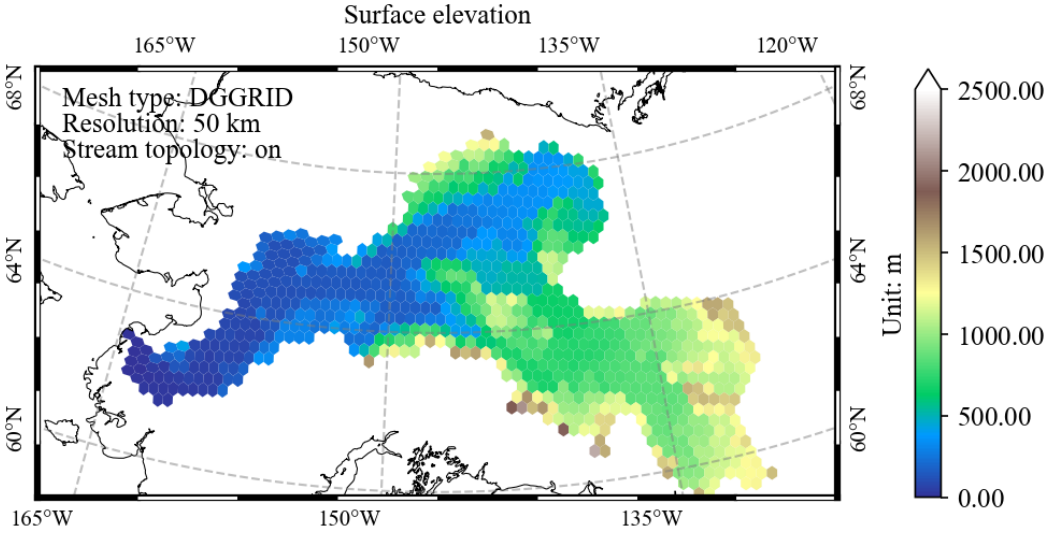

Figure 4. Spatial distribution of modeled surface elevation at DGGRID ISEA3H level 10 resolution in the Yukon Basin (unit: m).

3.2 Surface slope

In the **variable_polygon.geojson** data file, the attribute “slope” stores the modeled between-cell surface slope based on the depression-free elevation difference and cell center-to-cell center (great circle) distance (Figures 5 and 6).

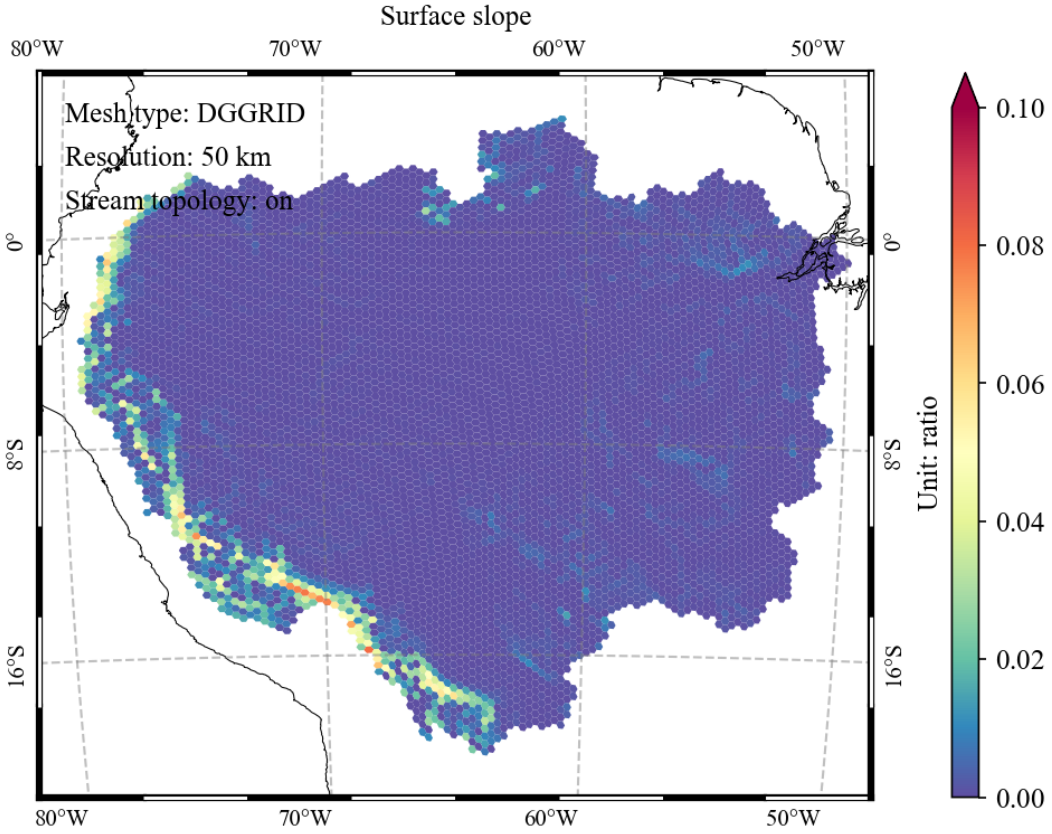

Figure 5. Spatial distribution of modeled mesh cell center to cell center slope at DGGRID ISEA3H level 10 resolution in the Amazon Basin (unit: ratio).

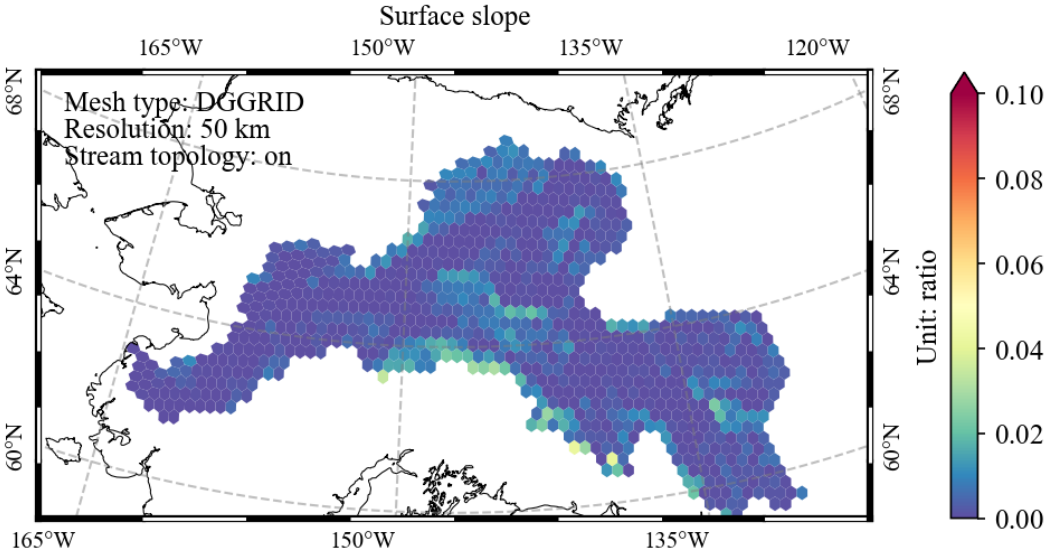

Figure 6. Spatial distribution of modeled mesh cell center to cell center slope at DGGRID ISEA3H level 10 resolution in the Yukon Basin (unit: ratio).

3.3 Flow direction

The **flow_direction.geojson** is a polyline-based GeoJSON data file. Each polyline feature defines the single flow direction (the steepest slope) from one DGGRID mesh cell center to its downslope/downstream mesh cell center (Figures 7 and 8).

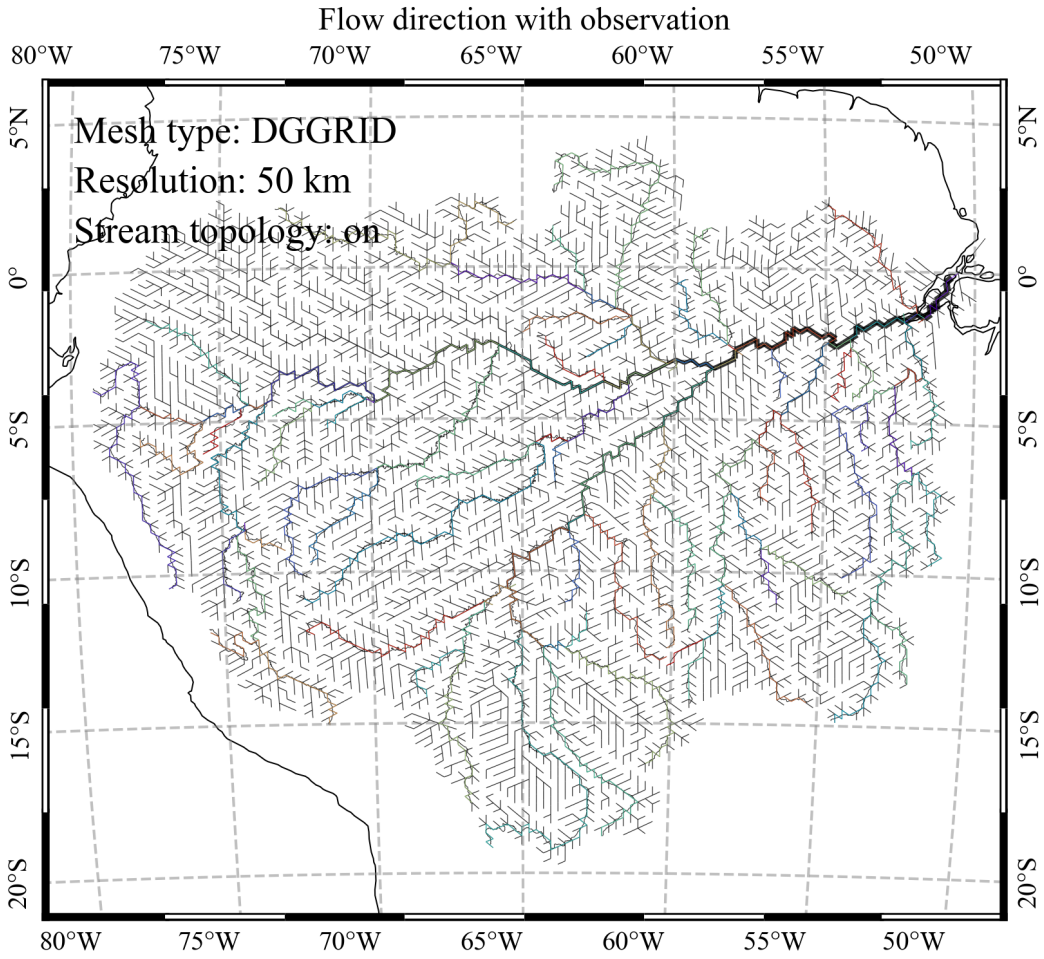

Figure 7. Modeled flow direction at DGGRID ISEA3H level 10 resolution in the Amazon Basin. Black straight lines are cell-to-cell conceptual flow direction. Line thickness is scaled with drainage area. Colored and curved black lines are conceptual (by PyFlowline) and simplified (by REACH) HydroSHEDS river networks.

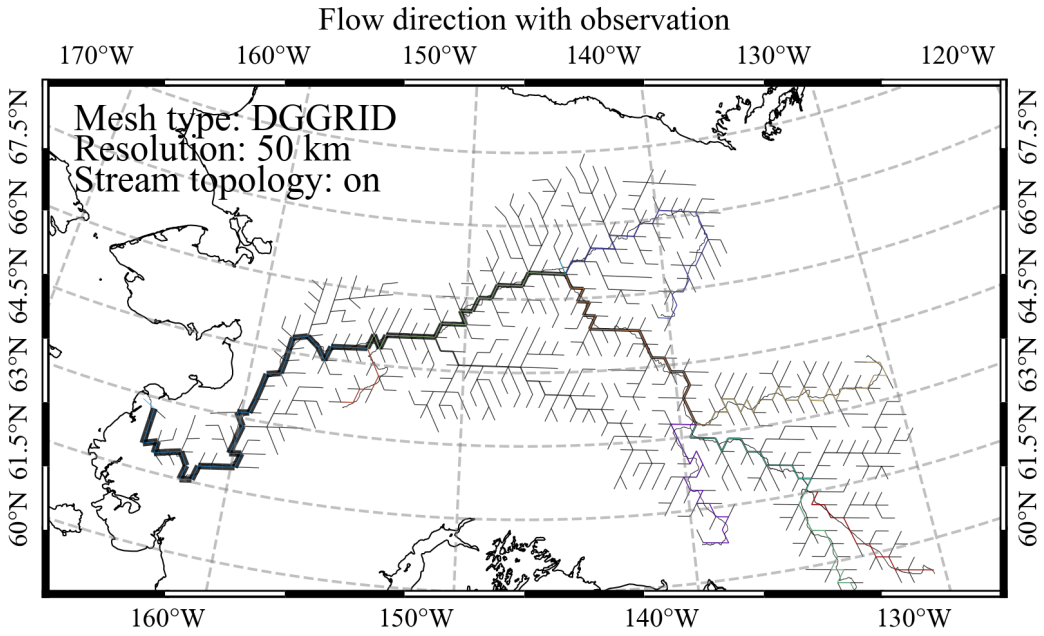

Figure 8. Modeled flow direction at DGGRID ISEA3H level 10 resolution in the Yukon Basin. Black straight lines are cell-to-cell conceptual flow direction. Line thickness is scaled with drainage area. Colored and curved black lines are conceptual and simplified HydroSHEDS river networks.

3.4 Drainage area

In the **variable_polygon.geojson** data file, the attribute "drainage" stores the modeled total upstream drainage (spherical) area of each mesh cell (including its own area) (Figures 9 and 10).

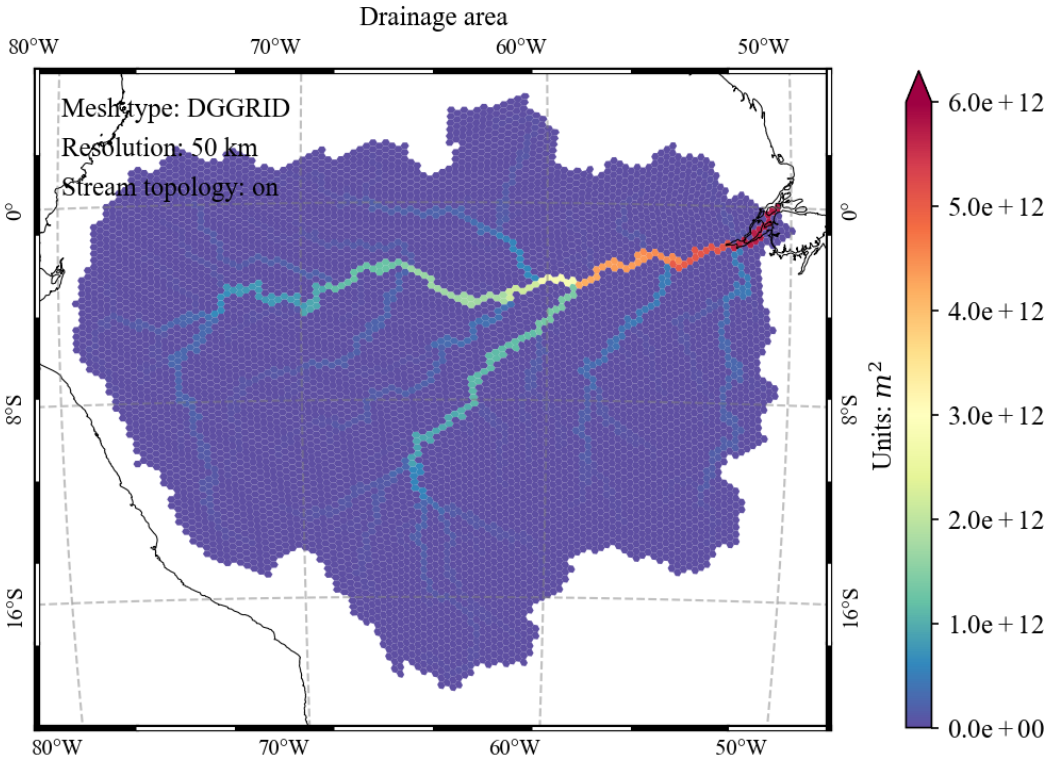

Figure 9. Modeled drainage area at DGGRID ISEA3H level 10 resolution in the Amazon Basin (units: m^2).

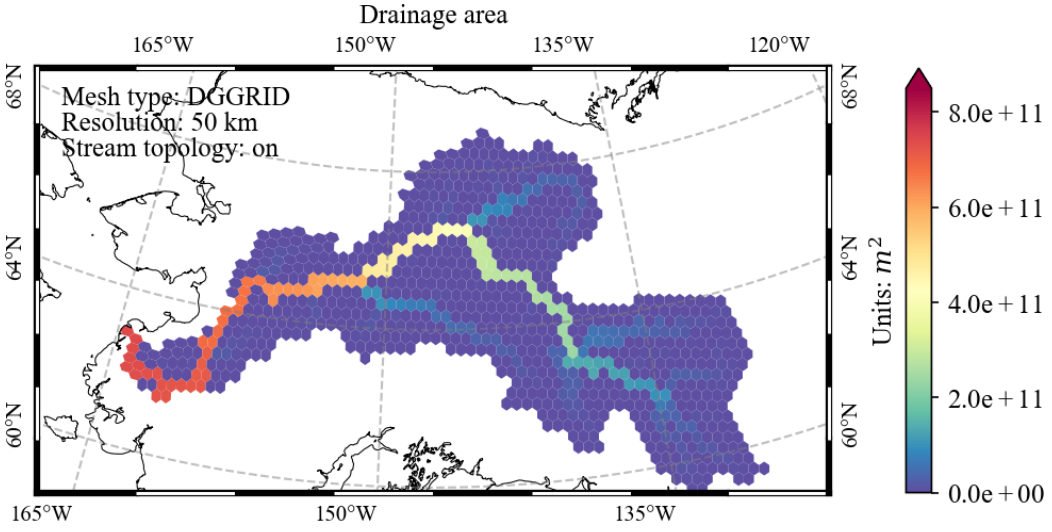

Figure 10. Modeled drainage area at DGGRID ISEA3H level 10 resolution in the Yukon Basin (units: m^2).

3.5 Travel distance

In the **variable_polygon.geojson** data file, the attribute “travel_distance” stores the modeled (great circle) travel distance from each mesh cell to the basin outlets (Figures 11 and 12). This term is also often referred to as downstream flow length.

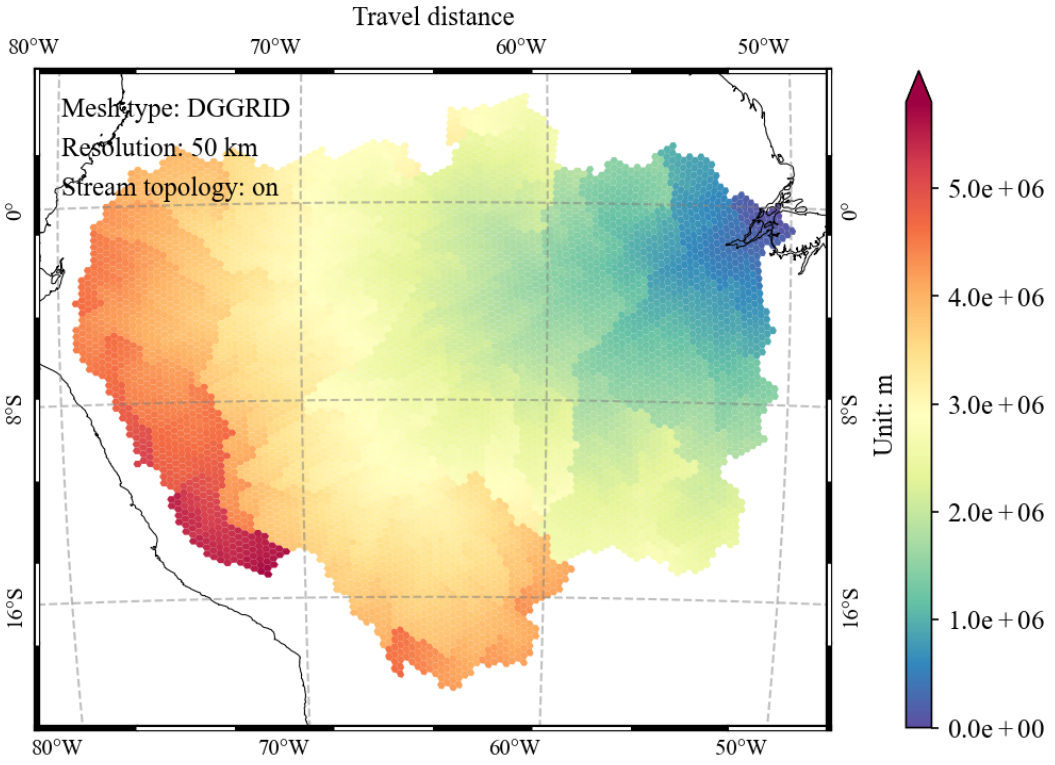

Figure 11. Modeled travel distance to the basin outlet at DGGRID ISEA3H level 10 resolution in the Amazon Basin (unit: m).

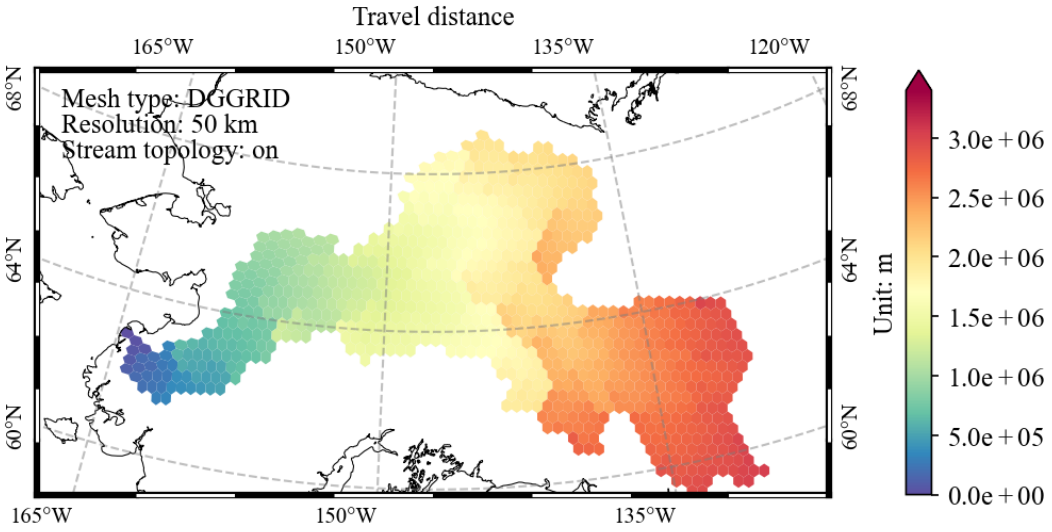

Figure 12. Modeled travel distance to the basin outlet at DGGRID ISEA3H level 10 resolution in the Yukon Basin (unit: m).

4 Technical Validation

Because the fundamental mesh structures differ from most existing datasets, we mainly rely on spatial patterns and geostatistics to evaluate our datasets. Different strategies are used for different data records. We primarily evaluate our datasets using existing flow routing datasets, i.e., HydroSHEDS products, the LBA-ECO CD-06 Amazon River Basin Land and Stream Drainage Direction and DEM datasets (Mayorga et al., 2012; Saatchi, 2013).

4.1 River networks

The river networks are the core intermediate results generated by the HexWatershed model in our workflow (Figure 1). Although these datasets do not cover the entire study domain, they illustrate how closely the modeled river networks match the REACH-generated river networks (Figure 2). In this study, we employed the "area of difference" method to assess their accuracy (Liao et al., 2022, 2023a) (Figure D1 in Section D1). This method uses the area to represent discrepancies in line features, such as river networks. When two or more line features intersect, the intersected segments can create enclosed polygons. We then calculate and compare the area of these polygons to quantify the differences formed by flowline intersections. Figures 13 and 14 illustrate the evaluation of river networks using this method in the Amazon and Yukon Basins, respectively.

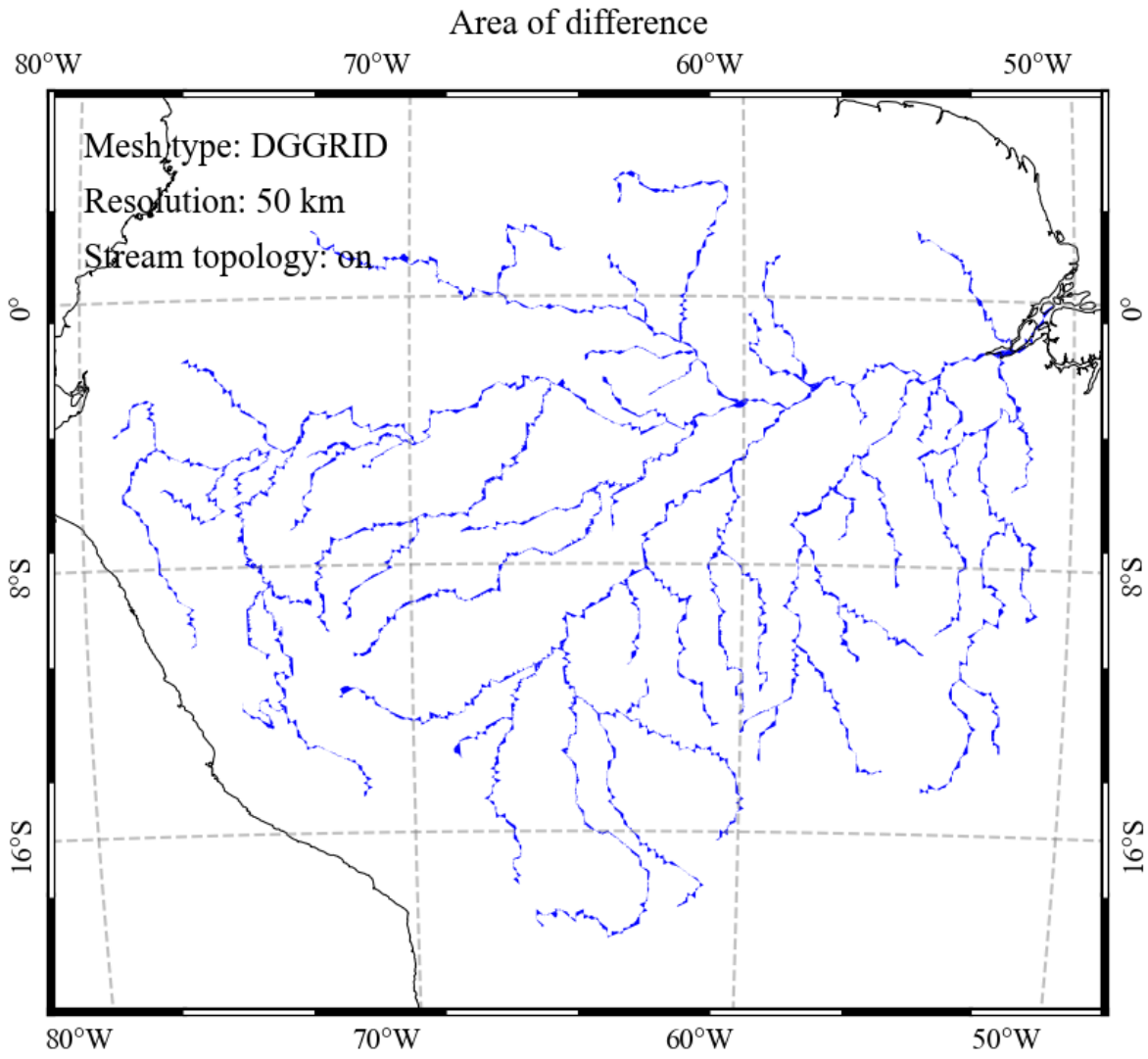

Figure 13. Validation of modeled river networks against HydroSHEDS river networks (REACH simplified) using area of difference at DGGRID ISEA3H level 10 resolution in the Amazon Basin. The total area of difference is 4.5×10^5 km^2.

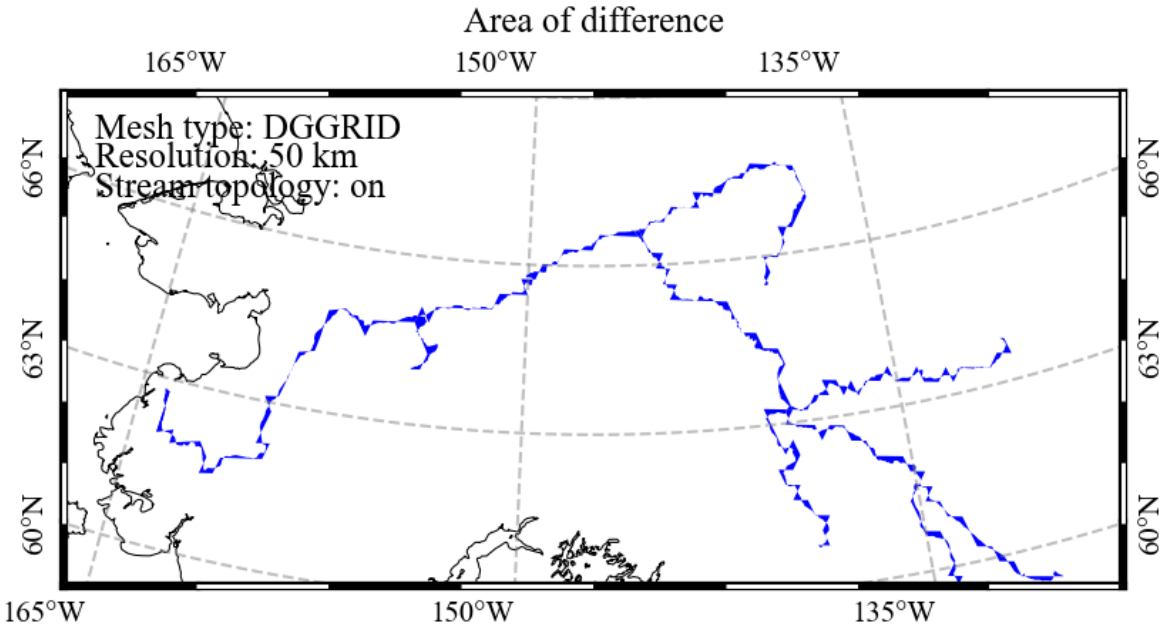

Figure 14. Validation of modeled river networks against HydroSHEDS river networks (REACH simplified) using area of difference at DGGRID ISEA3H level 10 resolution in the Yukon Basin. The total area of difference is $6.0 \times 10^4\,\mathrm{km}^2$.

4.2 Surface elevation

We employed a sphere resampling method to assess the surface elevation data through the following steps: (1) Utilizing the DGGRID ISEA3H level 14 mesh (the highest resolution in the current workflow) as the sampling pool; (2) Randomly selecting N cells as points of interest and recording their center locations; (3) Extracting elevation values from the data records and existing DEM datasets based on the chosen N longitude/latitude pairs. A scatterplot featuring N = 500 sampling points demonstrates that the modeled elevations closely match those of the existing high-resolution raster DEMs (Figure 15).

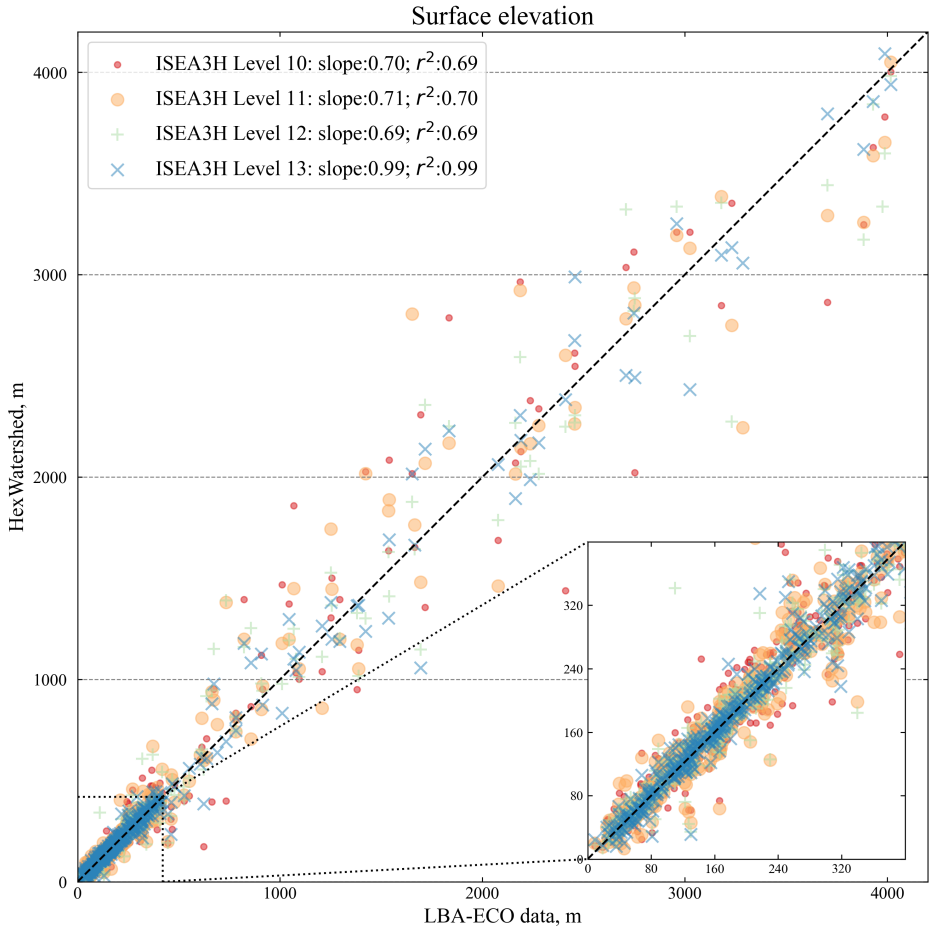

Figure 15. Validation of modeled surface elevation in the Amazon Basin from four DGGRID mesh resolutions. The X-axis is the sampled elevation from the LBA-ECO DEM datesets. The Y-axis is the sampled surface elevation from our records (unit: m). The mini-plot is a zoom-in view of the lower left.

The modeled surface elevation in the Yukon Basin is slightly worse than that in the Amazon Basin (Figure 16). One reason is that the spatial resolution of LBA-ECO DEM (30-second) is twice that of the HydroSHEDS DEM (15-second). Meanwhile, the Amazon Basin has relatively flat terrain compared with the Yukon Basin. This leads to different biases during the zonal mean resampling procedure (Liao et al., 2022).

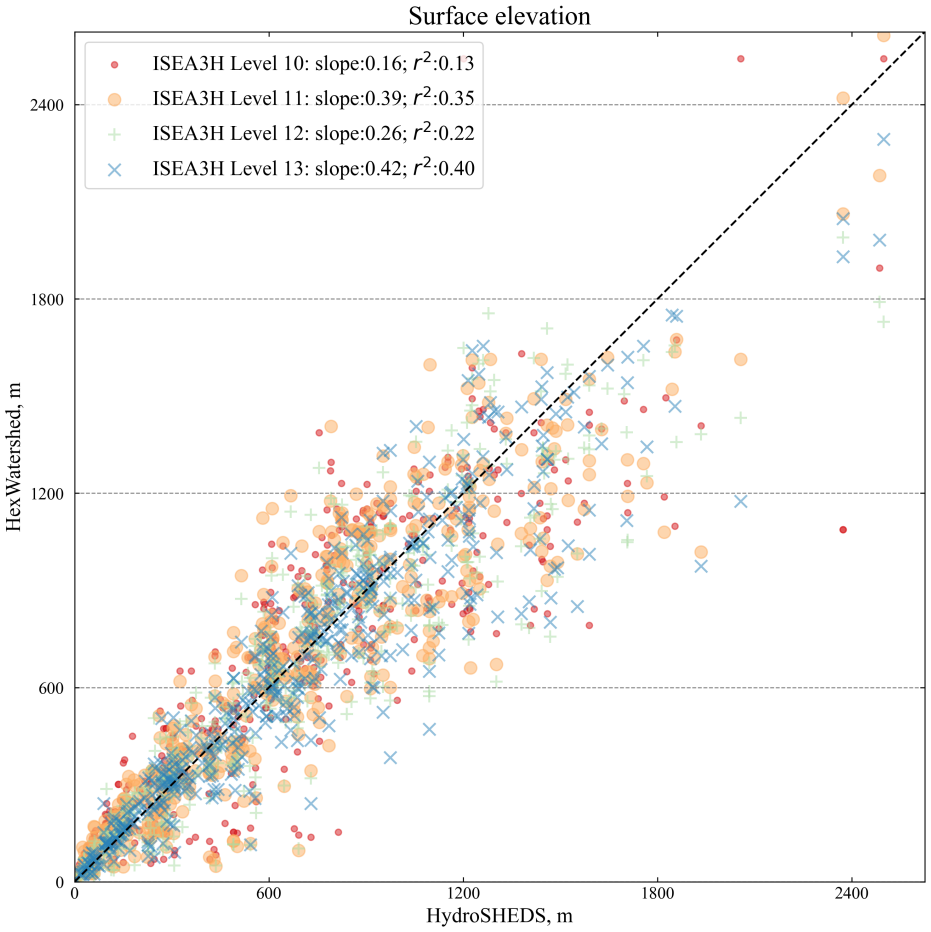

Figure 16. Validation of modeled surface elevation in the Yukon Basin from four DGGRID mesh resolutions. The X-axis is the sampled elevation from the HydroSHEDS DEM datesets. The Y-axis is the sampled surface elevation from our records (unit: m).

4.3 Flow direction

Given that flow direction is a vector field, a direct comparison between the modeled flow directions and existing D4/D8-based flow direction datasets is not feasible. Instead, we conducted a visual examination of the modeled flow directions using the simplified HydroSHEDS river networks. As depicted in Figures 7 and 8, the modeled flow directions consistently align with the simplified HydroSHEDS river networks across all four resolutions, consistent with our previous study (Liao et al., 2023b). Additionally, flow direction can be indirectly validated using the drainage area since they are closely interconnected.

4.4 Drainage area

The sphere resampling method (Section 4.2) could not be directly applied to the drainage area due to issues related to resolution mismatch and spatial dependence. As an alternative, we conducted a comparative analysis using major tributaries along the Amazon River and Yukon River, including their river mouths.

In the Amazon Basin, we selected seven tributary outlets, and their locations are provided in the Supplementary Information section. The scatterplot shows that the modeled drainage areas are consistent with the existing LBA-ECO drainage datasets (Figure 17).

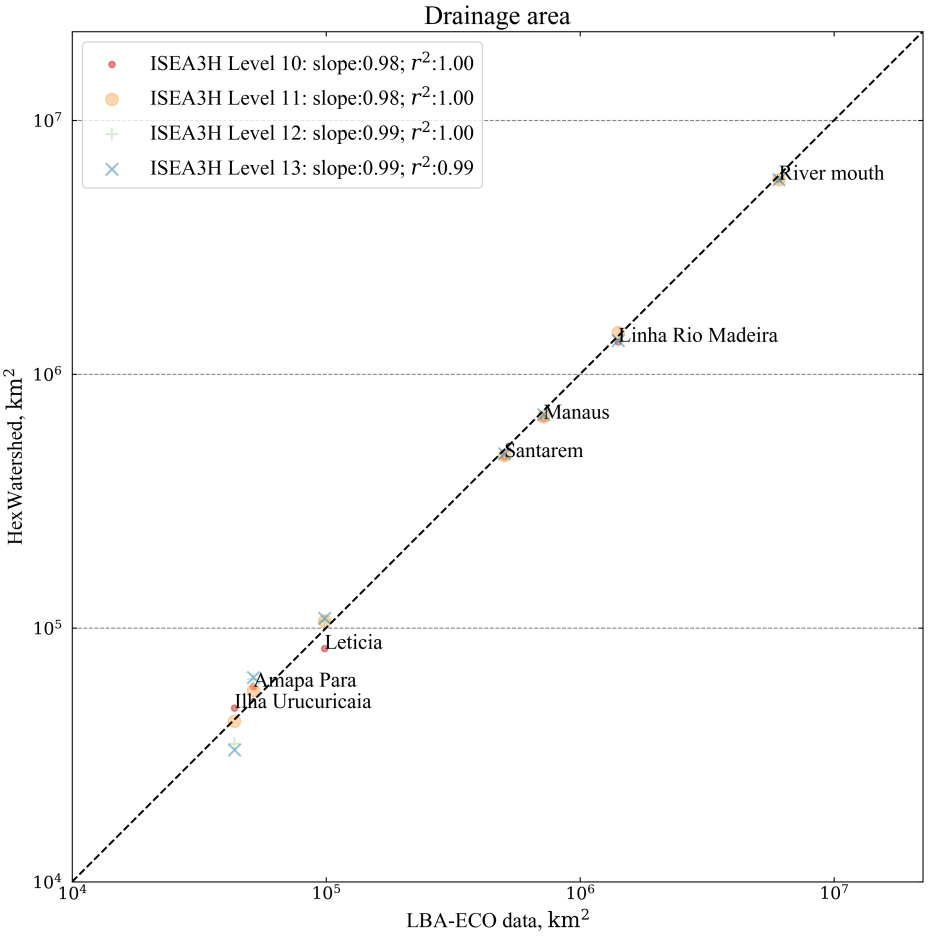

Figure 17. Validation of modeled drainage area of seven tributaries (including the river mouth) along the Amazon River from four DGGRID mesh resolutions. The X-axis is the drainage area from the LBA-ECO CD-06 Amazon River Basin Land and Stream Drainage Direction datesets (converted from flow accumulation). The Y-axis is the modeled drainage area (units: km^2). Both the X and Y axes are in the log scale.

In the Yukon Basin, we selected six tributary outlets, and their locations are provided in the Supplementary Information section. Among these tributaries, only the modeled drainage area at resolution level 10 at the Kooyukuk River is underestimated (Figure 18).

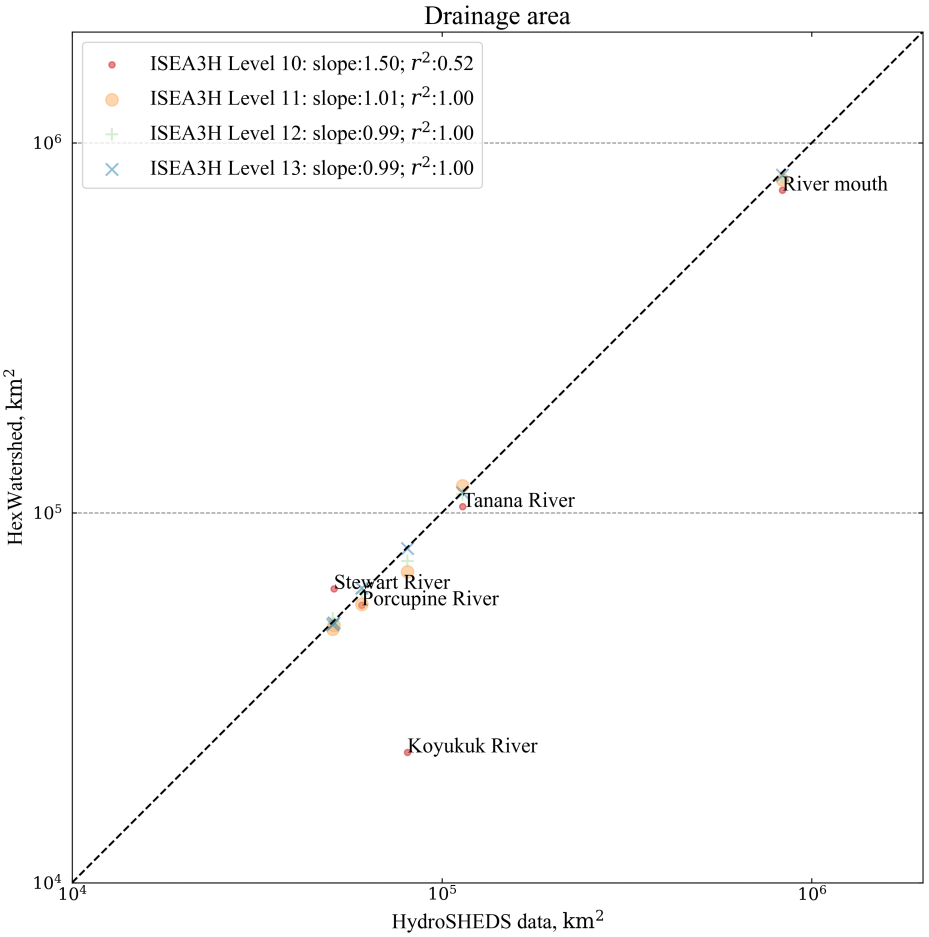

Figure 18. Validation of modeled drainage area of six tributaries along the Yukon River from four DGGRID mesh resolutions. The X-axis is the drainage area from the HydroSHEDS datesets. The Y-axis is the modeled drainage area (units: km^2). Both the X and Y axes are in the log scale.

4.5 Travel distance

Similar to the drainage area, we evaluated the modeled travel distance using the selected tributary outlets, excluding the river mouths. In the Amazon Basin, the scatterplot shows that the modeled travel distances are consistent with the existing LBA-ECO CD-06 flow length datasets (Figure 19). However, the modeled travel distances are slightly higher than the LBA-ECO

datasets. This is possibly caused by the additional length added near the Amazon River delta region. A similar pattern is also observed in the Yukon Basin (Figure 20).

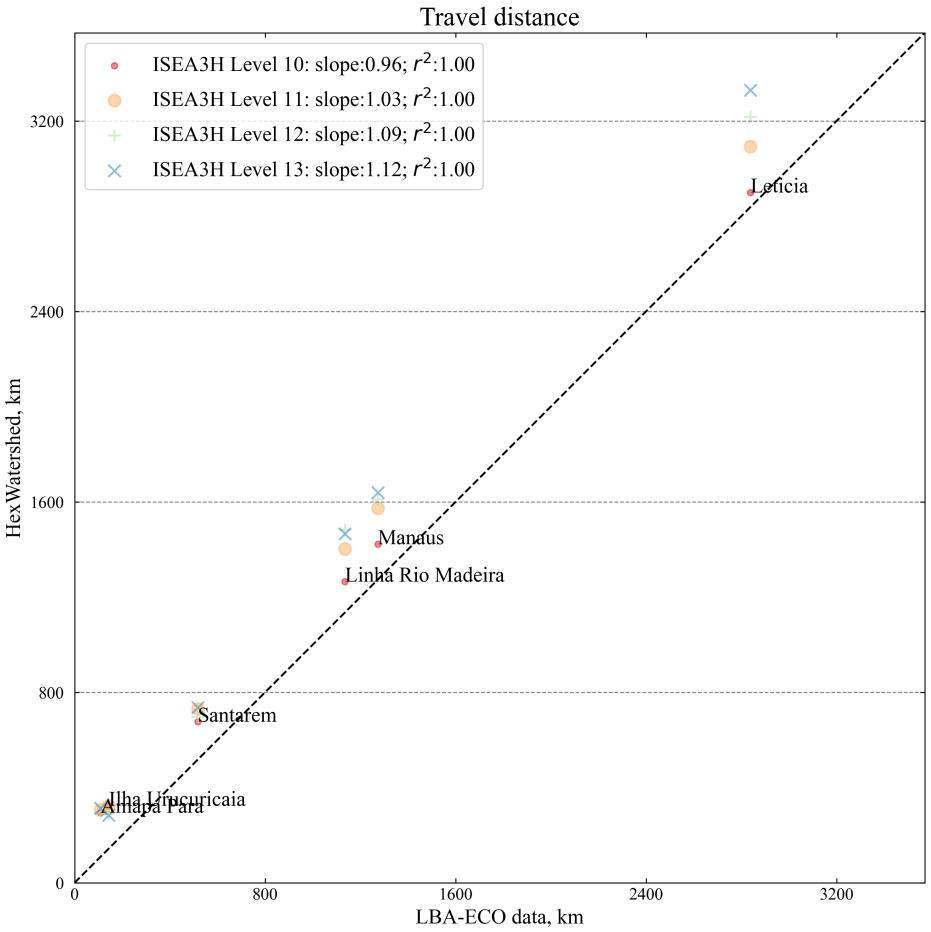

Figure 19. Validation of modeled travel distance of six tributaries along the Amazon River from four DGGRID mesh resolutions. The X-axis is the travel distance from the LBA-ECO CD-06 Amazon River Basin Land and Stream Drainage Direction datesets. The Y-axis is the modeled travel distance (unit: km).

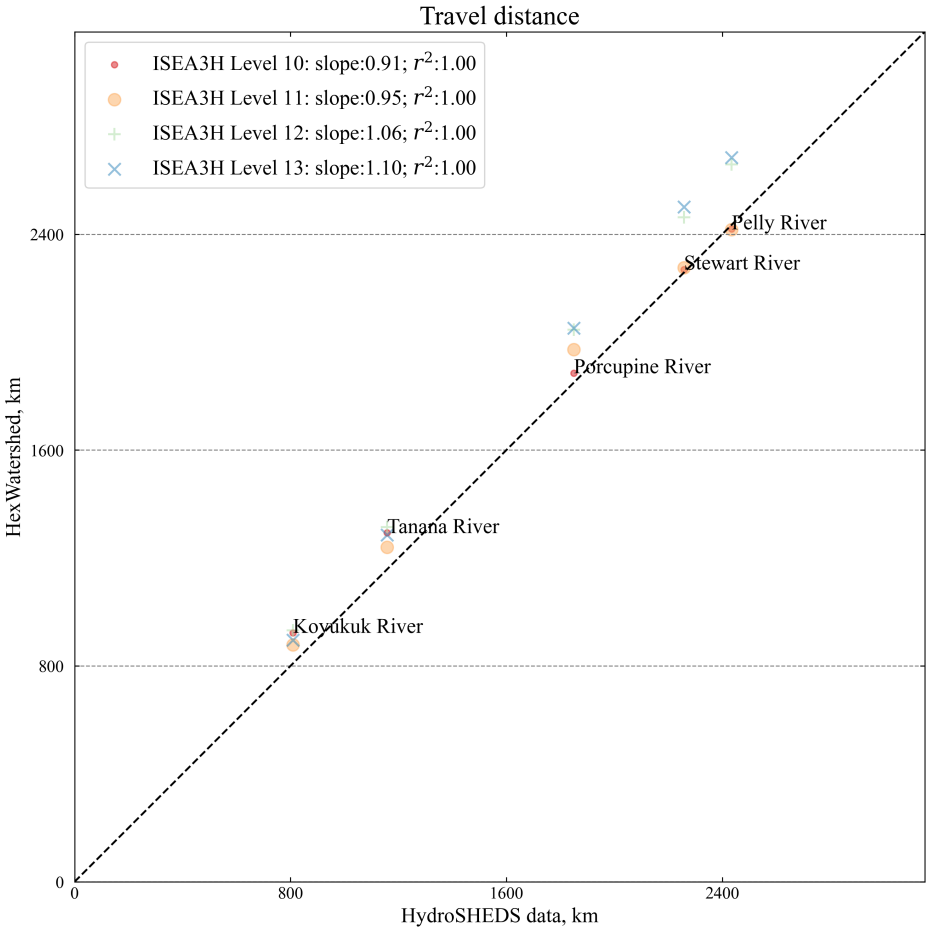

Figure 20. Validation of modeled travel distance of six tributaries along the Yukon River from four DGGRID mesh resolutions. The X-axis is the travel distance from the HydroSHEDS datasets. The Y-axis is the modeled travel distance (unit: km).

5 Discussion

5.1 Limitation

Although our datasets represent an important step forward in flow routing capability, they are not without limitations. Through our practice, we have identified the following limitations and provided corresponding solutions for future improvements:

1. As a pioneering dataset, our dataset records were not produced from a unified data source, especially for the DEM component. This is because nearly all the existing global DEM datasets use the GCS spatial reference, and the quality of DEM gradually decreases from the equator to the high latitudes due to the spatial distortion (Liao et al., 2020). This is

also part of the reason that the modeled surface elevation in the Yukon Basin is not as good as that in the Amazon Basin. However, the workflow we developed is generally robust even though there are potential issues in the DEM datasets.

2. The HexWatershed model does not accommodate braided rivers currently, potentially introducing uncertainty in the flow direction, particularly near river deltas. However, since most hydrologic models also lack support for braided rivers, this limitation is not considered critical.

3. The modeled drainage area and travel distance were slightly higher than observed in the Amazon Basin (Figure 19). We interpret this as due to the inclusion of the complex delta region, which suggests a similar bias could arise in similar deltaic regions worldwide. To address this, it is recommended to use high-resolution meshes (such as the ISEA3H resolution level 13) to mitigate its impact on model performance for large-scale hydrologic and Earth system models. Alternatively, explicit modeling of delta connectivity using HexWatershed can also be an effective solution.

4. Our method does not consider large lake waterbody in the workflow. Therefore, these datasets may not be suitable for hydrologic applications that focus on lake routing. We plan to explicitly consider lakes, especially large lakes, in future developments.

5. Due to computational constraints and input dataset quality, we only generated flow routing datasets at four spatial resolutions in the Amazon and Yukon Basins. To evaluate model performance and suitability at finer spatial resolutions (e.g., $< 5\,\mathrm{km}$), additional simulations are needed. Additionally, a global-scale dataset will be made available once computational efficiency has been enhanced.

5.2 Usage

The datasets are primarily stored using the JavaScript Object Notation (JSON) and GeoJSON formats. Some datasets are also provided in the GeoPackage and (Geo)Parquet file format tailored for high-performance operations and visualizations. Most scientific programming languages, including Python, C++, R, and MATLAB, provide functions or public libraries to read these file formats. The datasets for these two basins are distributed with global coverage, but users can extract portions of the dataset using GIS operations. For example, users can extract sub-basins for regional hydrologic simulations or convert them to other common scientific file formats, including the Network Common Data Format (NetCDF) or Hierarchical Data Format (HDF).

These datasets are suitable for regional and large-scale spatially distributed hydrologic and river routing models, including the Model for Scale Adaptive River Transport (MOSART) (Li et al., 2013). Additionally, users can derive other flow routing parameters like the Topographic Wetness Index (TWI) and Manning's Roughness Coefficient (n) from these datasets.

We also provide an online jupyter notebook tutorial that can be used to reproduce these datasets in a browser at https://github.com/changliao1025/hexwatershed_tutorial. This tutorial can be modified to generate similar datasets using other DGGRID supported mesh types and resolutions or in different river basins.

6 Conclusions

We have produced pioneering ISEA3H DGGS-based hierarchical flow routing datasets in the Amazon Basin and Yukon Basin, available at four spatial resolutions (in length) (29.42 km, 16.99 km, 9.81 km, and 5.66 km). Extensive evaluation confirms their consistency with existing high-resolution terrain data and HydroSHEDS river networks. Because our method is mesh-independent, similar flow routing datasets can also be generated on DGGS with various configurations or even other unstructured grid meshes. Adoption of these datasets by hydrologic models will enhance the performance of spatially distributed hydrological models of these two basins and similar regions worldwide.

7 Code availability

The REACH tool can be accessed from the GitHub repository: https://github.com/dengwirda/reach. The DGGRID model can be accessed from the GitHub repository: https://github.com/sahrk/DGGRID. The HexWatershed model can be installed through the Conda Python platform: https://anaconda.org/conda-forge/hexwatershed (Liao, 2022a; Liao and Cooper, 2022). The source code to reproduce the datasets and figures is stored in the GitHub repository: https://github.com/changliao1025/liao_2023_scidata_dggs.

8 Data availability

The datasets are stored in the Zenodo repository: https://zenodo.org/record/8377765 (Liao, 2023).

Appendix A: Model configurations

A1 DGGRID model configurations

A2 HexWatershed model configurations

Variable name	Data type	Data format	Note
area	float	GEOJSON	Geodesic area
elevation	float	GEOJSON	Mean elevation after the depression removal
slope	float	GEOJSON	Slope between mesh cell in the flow direction
flow direction	Not applicable	JSON/GEOJSON	Dominant flow direction with the steepest slope
drainage area	float	GEOJSON	Geodesic area-based
travel distance	float	GEOJSON	Cell center to cell center distance-based

Table A2. List of data records produced by HexWatershed in the Amazon Basin.

Appendix B: HexWatershed method description

B1 PyFlowline

The PyFlowline model is a core submodule within the HexWatershed model. PyFlowline generates the mesh cell-based conceptual river networks using three steps: (1) flowline simplification, which removes undesired flowlines and builds the topological relationships; (2) mesh generation, which creates customized meshes based on model configuration. For example, it now supports APIs to generate a DGGRID mesh; (3) topological relationship reconstruction. This algorithm uses the intersection between flowlines and mesh cells to reconstruct the cell-to-cell topological relationships.

B2 HexWatershed

HexWatershed is a mesh-independent flow direction model and fully supports all the mesh types generated by PyFlowline. It defines flow direction using a two-step approach. First, it uses a hybrid breaching-filling stream-burning method to define the flow direction for river networks and their riparian zones. Second, it uses a revised priority-flood algorithm to conduct depression filling and defines the flow direction for the remaining mesh cells. A list of other flow routing parameters are generated through this process.

A video describing the hybrid stream burning and depression filling algorithm is provided using the ISEA3H resolution 11 simulation animation at https://www.youtube.com/watch?v=G2uTHlCUMTc.

Appendix C: Full data record visualization

C1 Amazon Basin

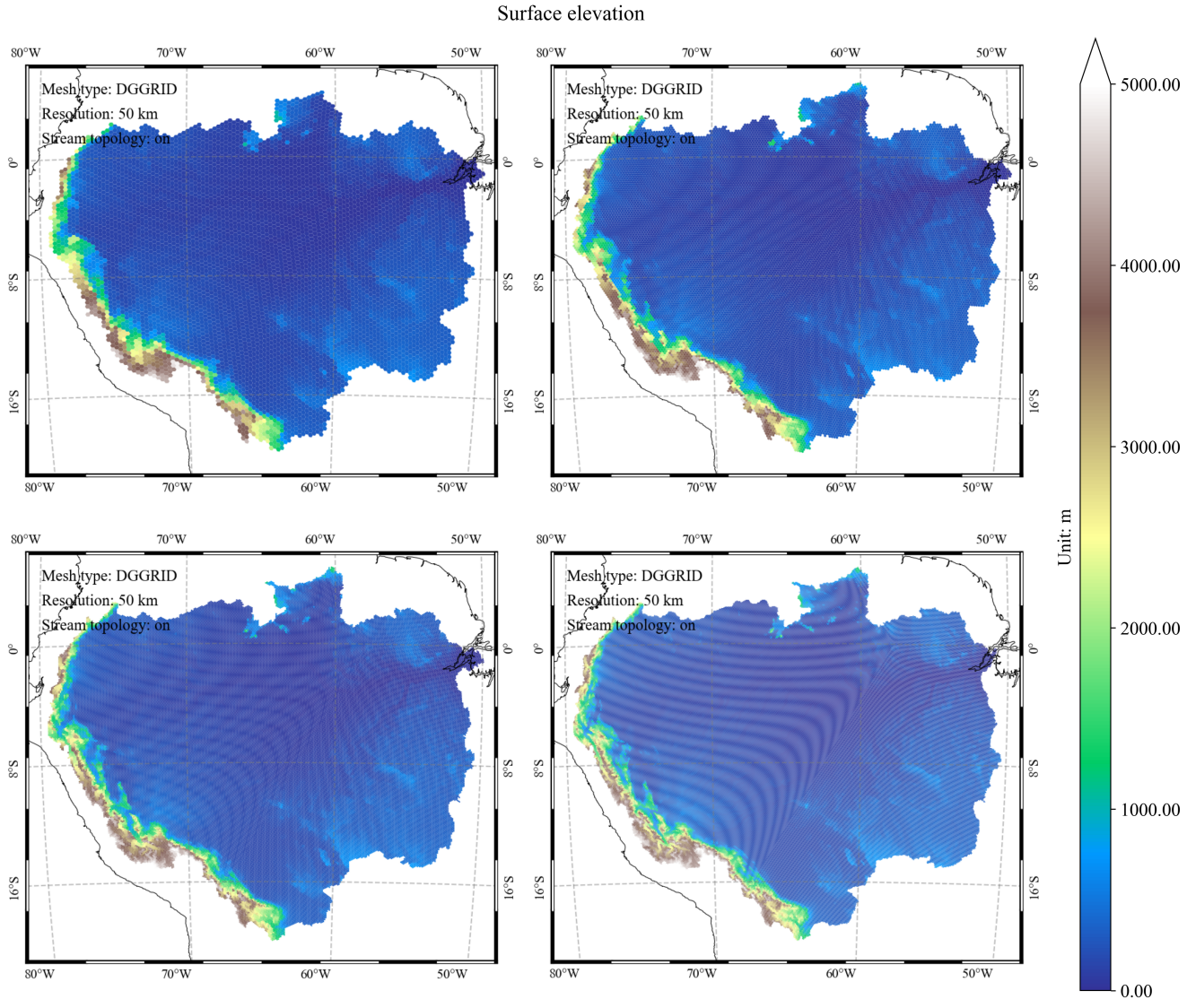

Figure C1. Spatial distribution of modeled surface elevation at DGGRID ISEA3H level 10 to 13 resolutions in the Amazon Basin (unit: m).

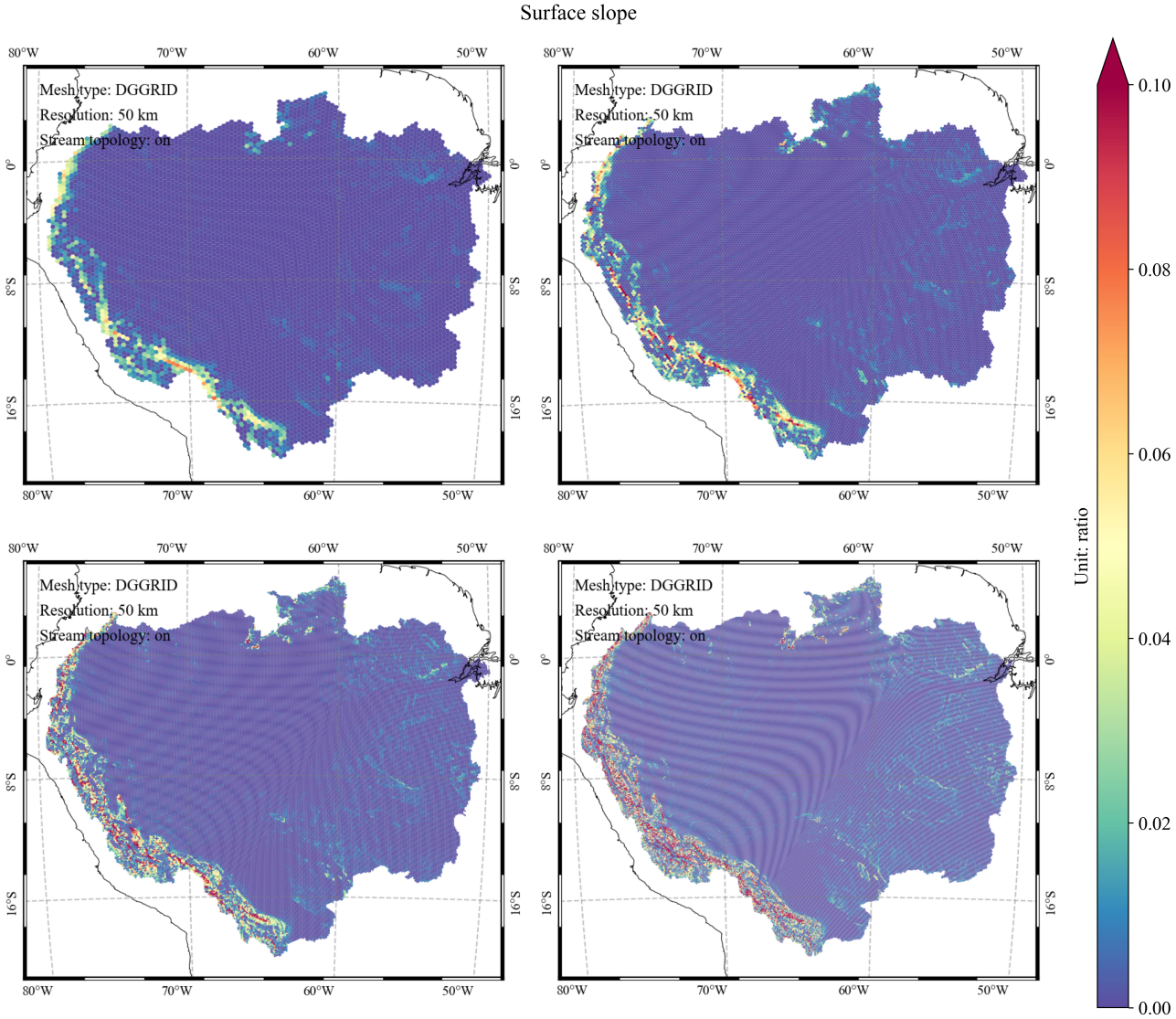

Figure C2. Spatial distribution of modeled mesh cell center to cell center slope at DGGRID ISEA3H level 10 to 13 resolutions in the Amazon Basin (unit: ratio).

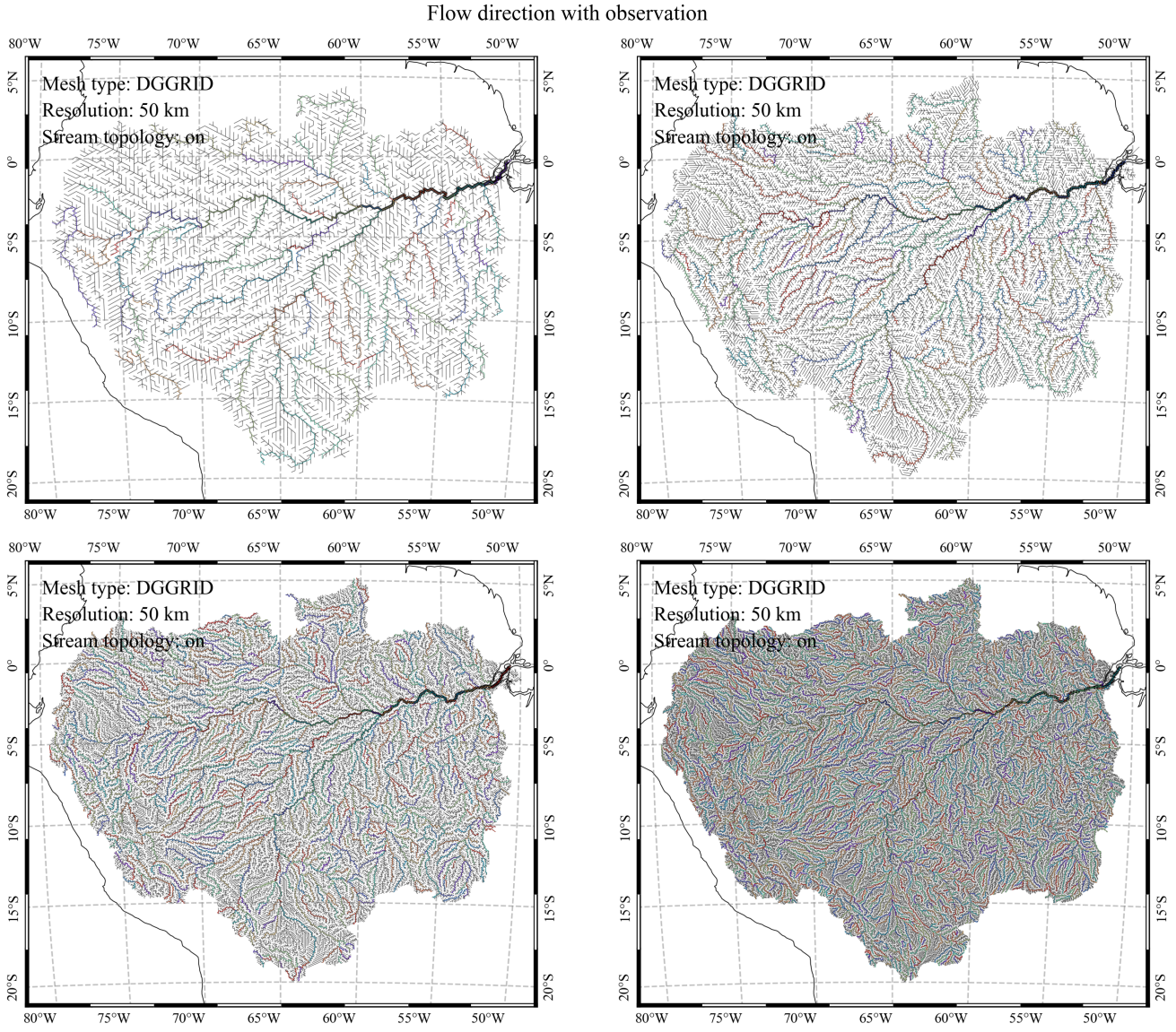

Figure C3. Modeled flow direction at DGGRID ISEA3H level 10 to 13 resolutions in the Amazon Basin. Black lines are cell-to-cell flow direction. Line thickness is scaled with drainage area. Colored and detailed black lines are conceptual and simplified HydroSHEDS river networks.

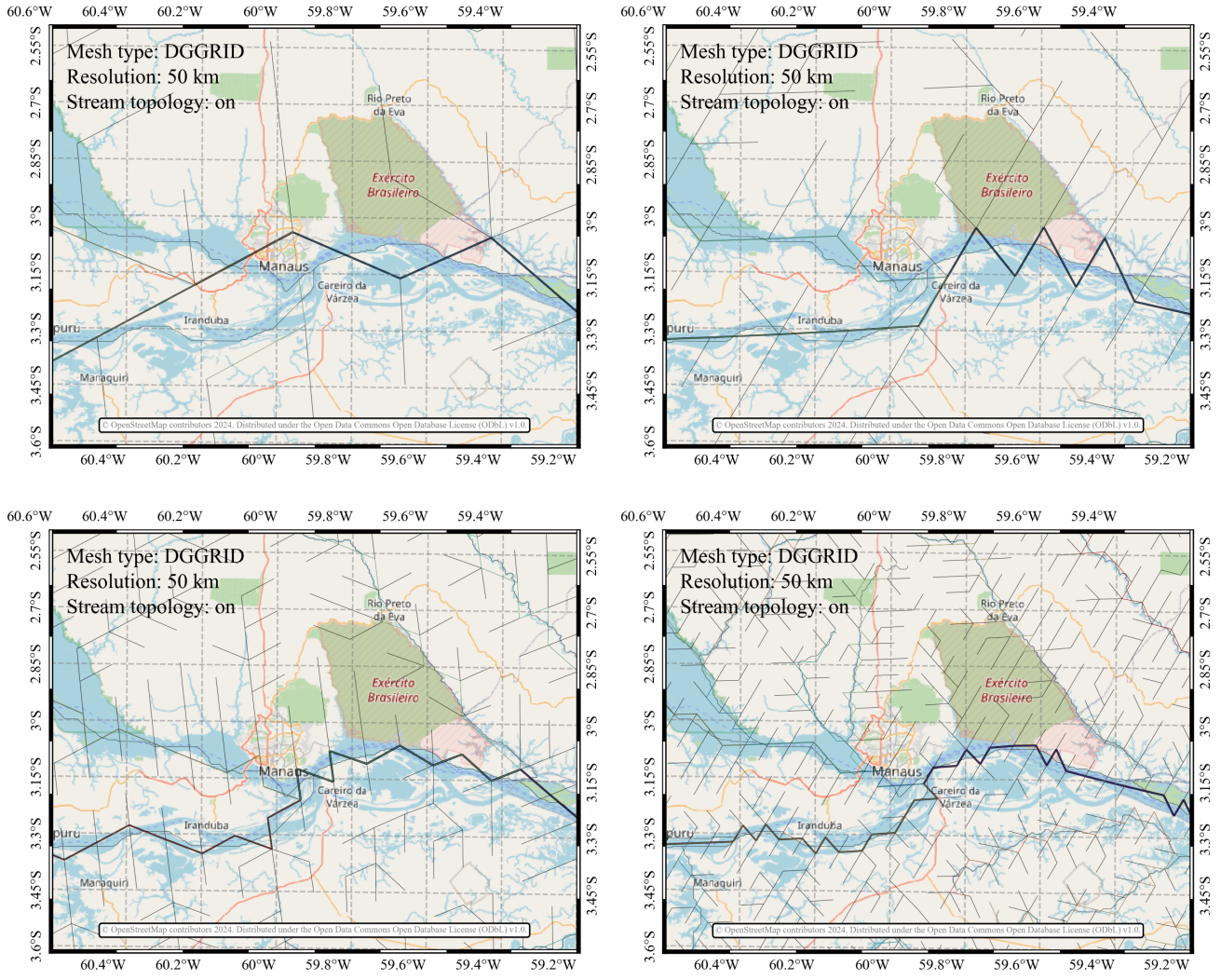

Figure C4. Zoom-in views of modeled flow direction at DGGRID ISEA3H level 10 to 13 resolutions in the Amazon Basin near Manaus. Black lines are cell-to-cell flow direction. Line thickness is scaled with drainage area. Colored and detailed black lines are conceptual and simplified HydroSHEDS river networks. The base images are Openstreet Map contributors 2024. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

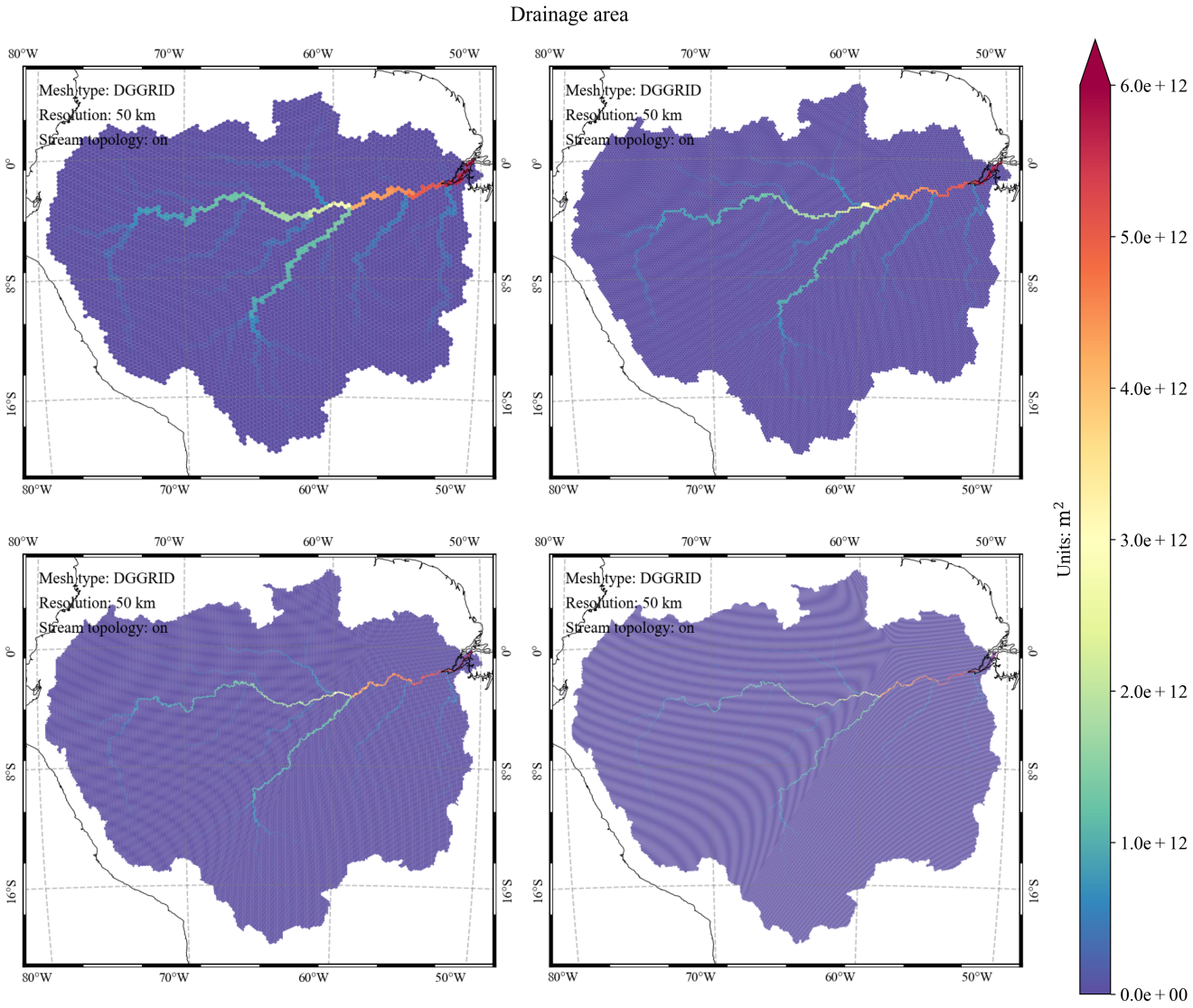

Figure C5. Modeled drainage area at DGGRID ISEA3H level 10 to 13 resolutions in the Amazon Basin (units: m^2).

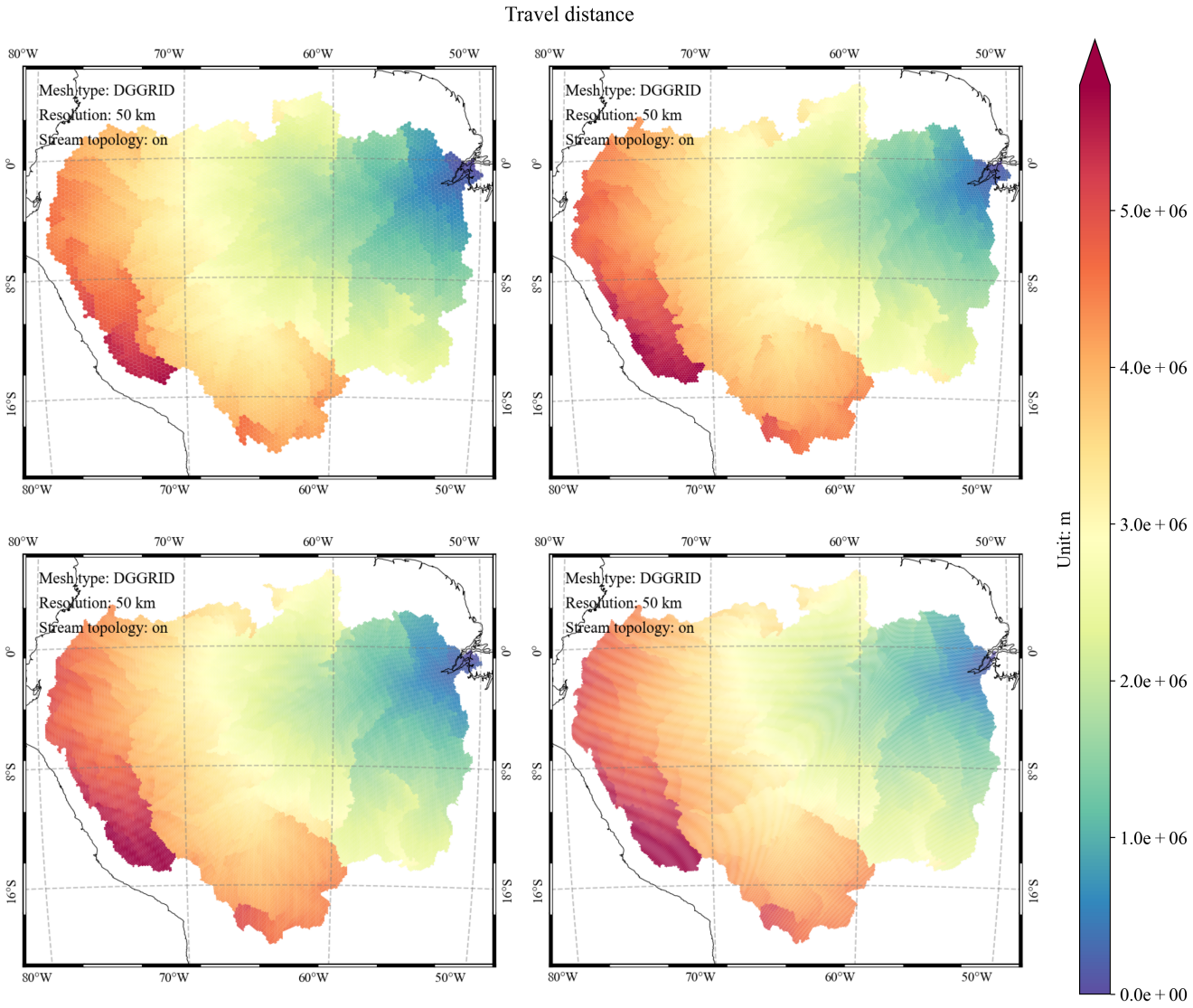

Figure C6. Modeled travel distance to the basin outlet at DGGRID ISEA3H level 10 to 13 resolutions in the Amazon Basin (unit: m).

C2 Yukon Basin

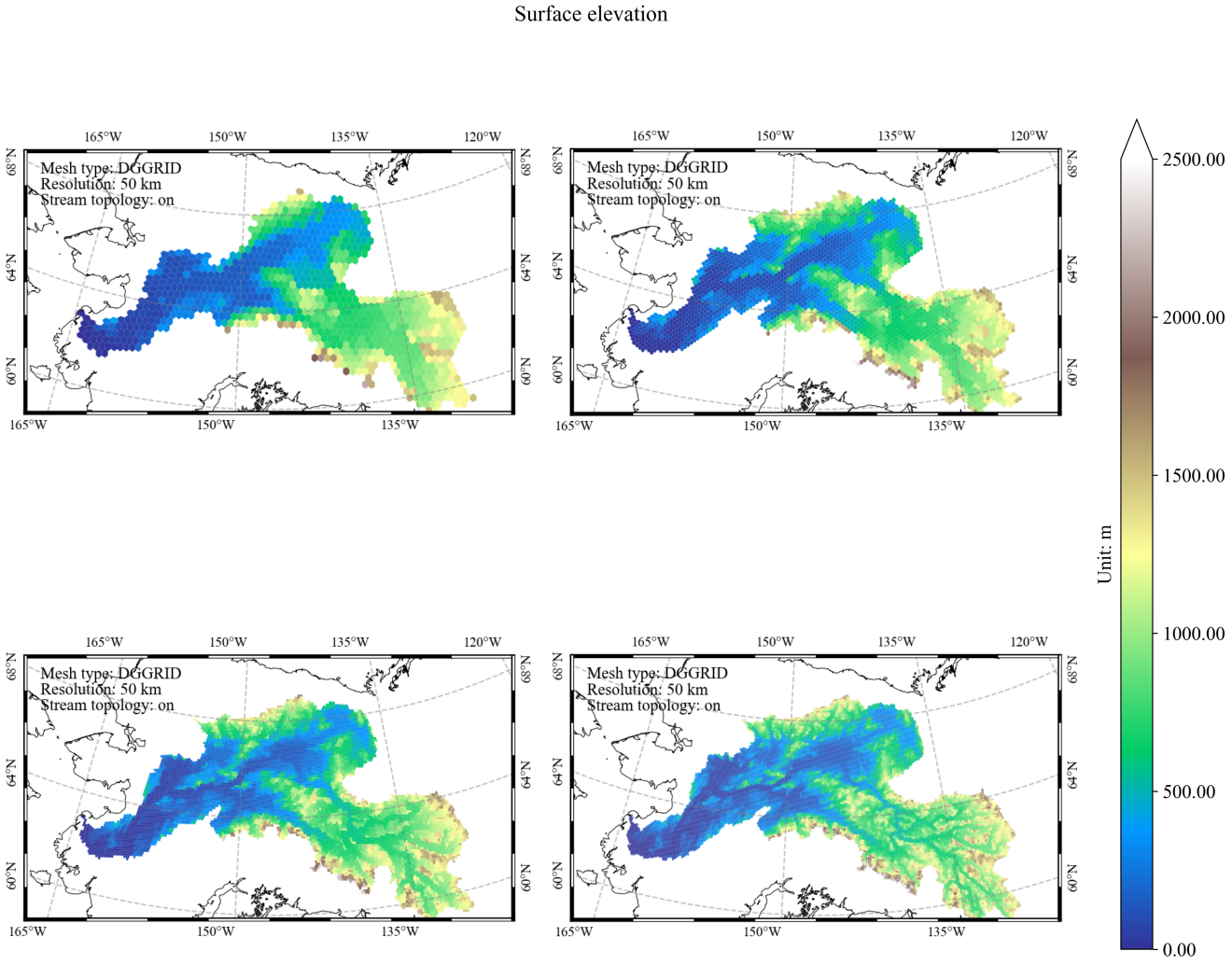

Figure C7. Spatial distribution of modeled surface elevation at DGGRID ISEA3H level 10 to 13 resolutions in the Yukon Basin (unit: m).

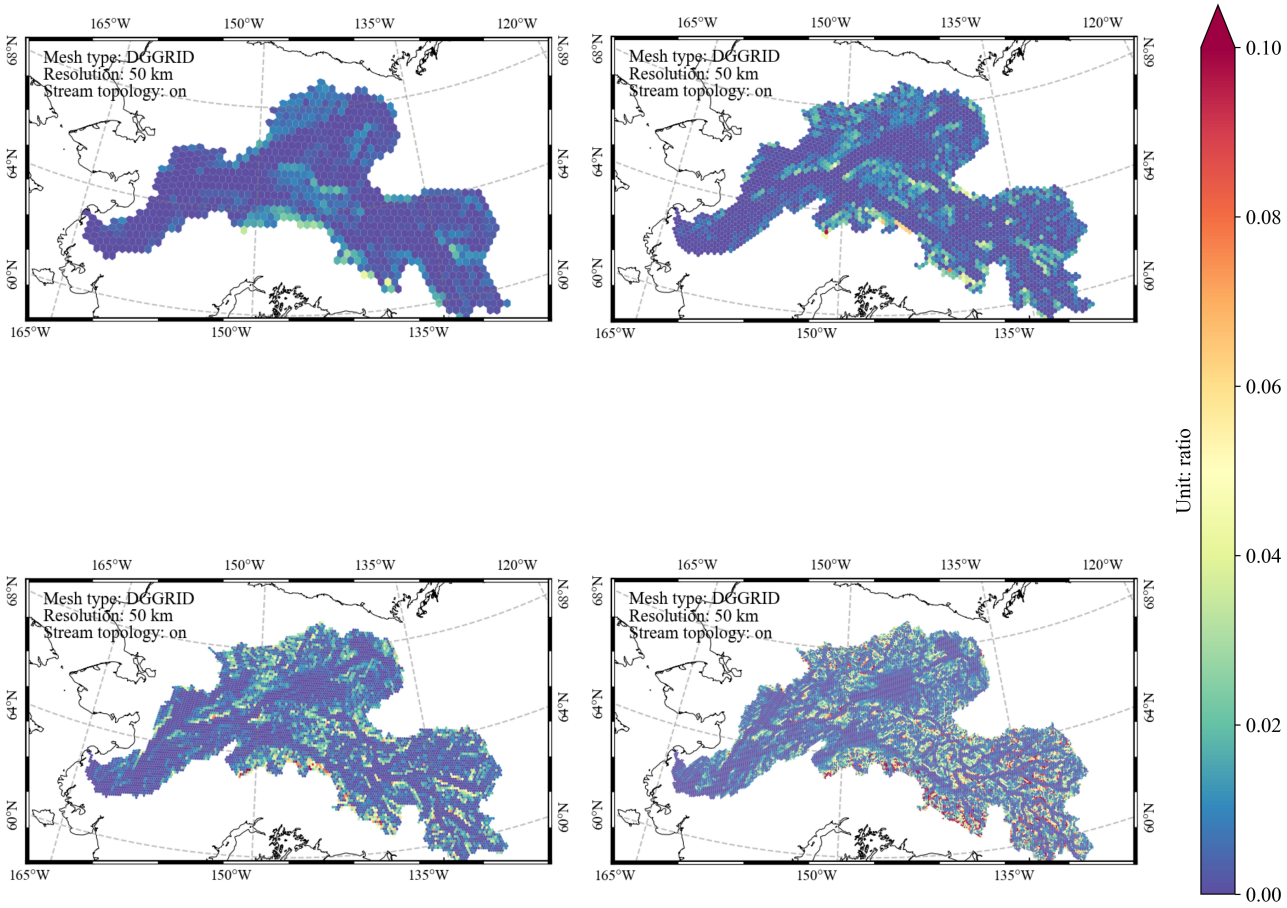

Figure C8. Spatial distribution of modeled mesh cell center to cell center slope at DGGRID ISEA3H level 10 to 13 resolutions in the Yukon Basin (unit: ratio).

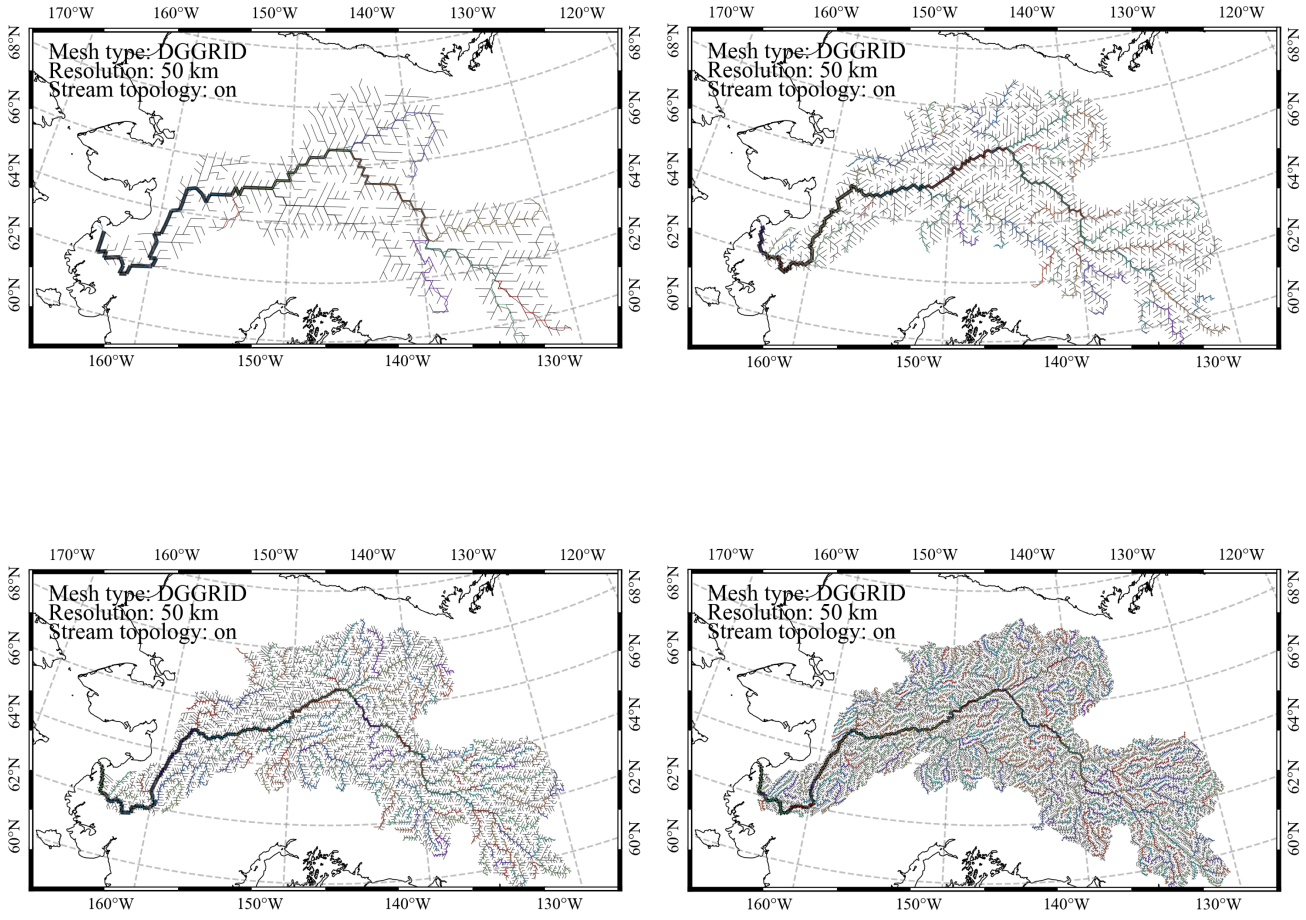

Figure C9. Modeled flow direction at DGGRID ISEA3H level 10 to 13 resolutions in the Yukon Basin. Black lines are cell-to-cell flow direction. Line thickness is scaled with drainage area. Colored and detailed black lines are conceptual and simplified HydroSHEDS river networks.

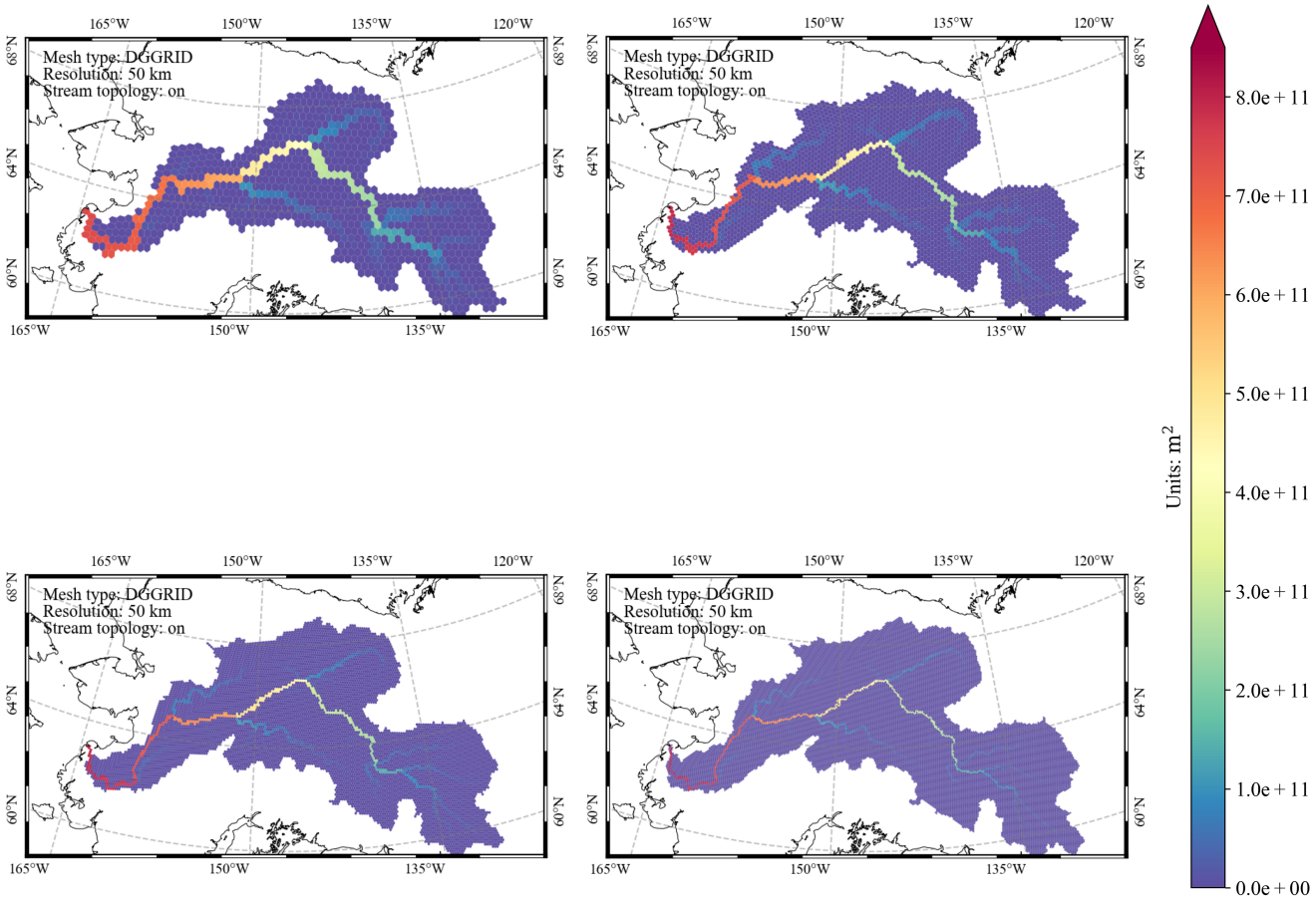

Figure C10. Modeled drainage area at DGGRID ISEA3H level 10 to 13 resolutions in the Yukon Basin (units: m^2).

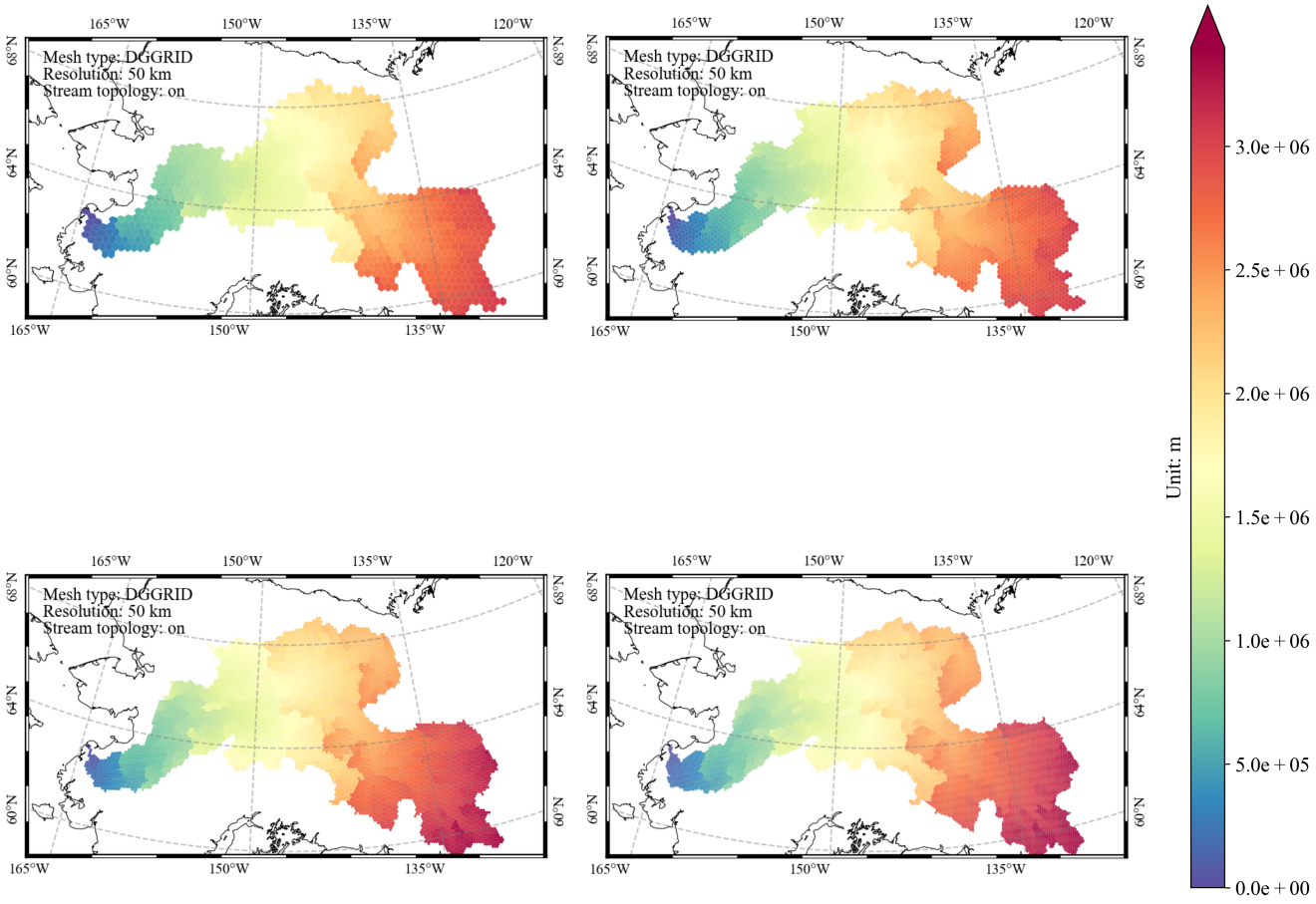

Figure C11. Modeled travel distance to the basin outlet at DGGRID ISEA3H level 10 to 13 resolutions in the Yukon Basin (unit: m).

Appendix D: Data validation

D1 Area of difference method

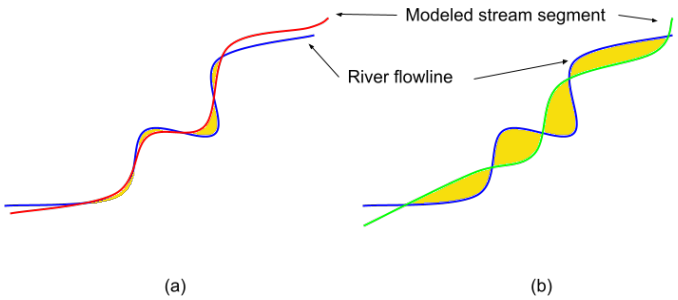

Figure D1. Illustration of the area of differences based on line feature intersections. Blue line features are actual river flowlines. Red and green line features are modeled flowlines from different simulations. Yellow-filled polygons are the product of line intersections from actual and modeled flowlines. The smaller the total area of these polygons, the closer the modeled flowline is to the actual flowline. As a result, the modeled flowline in (a) is closer/better than that in (b).

The area of differences can be calculated using Python scripting in the following steps:

1. Convert modeled conceptual flowlines into edge-based flowlines A;
2. Intersect the edge-based flowlines A with the simplified flowlines B to obtain all the vertices list C;
3. Classify the vertices C into different types of vertices (i.e., shared by the number of flowlines);
4. Split both A and B using C to obtain a list of flowlines D;
5. Build all the polygons enclosed by connected flowlines in D using a cycling algorithm and calculate their areas;
6. Sum up the areas to obtain the total area of difference.

D2 Tributaries along the Amazon River

Tributary	Longitude (°)	Latitude (°)
Amapa Para	-51.20655	-0.03696
Ilha Urucuricaia	-52.23621	-1.54800
Santarem	-54.76348	-2.39293
Linha Rio Madeira	-58.77602	-3.39210
Manaus	-60.02855	-3.14578
Leticia	-70.00767	-4.37262

Table D1. List of tributary outlets along the Amazon River used for drainage area and travel distance validations.

D3 Tributaries along the Yukon River

Tributary	Longitude (°)	Latitude (°)
Tanana River	-151.85660	65.13456
Porcupine River	-141.69291	67.18116
Koyukuk River	-157.55722	64.92744
Stewart River	-139.39908	63.29666
Pelly River	-137.35090	62.80660

Table D2. List of tributary outlets along the Yukon River used for drainage area and travel distance validations.

Author contributions.

Chang Liao prepared the input data and conducted the HexWatershed simulations and analysis. Darren Engwirda developed and tested the REACH library using the HydroSHEDS river network datasets. All the co-authors contributed to the writing and analysis.

Competing interests.

The contact author has declared that none of the authors has any competing interests.

Acknowledgements. This research was funded as part of the multi-program, collaborative Integrated Coastal Modeling (ICoM) project and the Interdisciplinary Research for Arctic Coastal Environments (InteRFACE) project through the Department of Energy, Office of Science,

Biological and Environmental Research program, Earth and Environment Systems Sciences Division, Earth System Model Development (ESMD) program area. This work was also supported by the U.S. Department of Energy Office of Biological and Environmental Research as part of the Terrestrial Ecosystem Systems program through the Next Generation Ecosystem Experiment (NGEE) Tropics project. A portion of this research was performed using PNNL Research Computing at Pacific Northwest National Laboratory. PNNL is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830. We thank Kevin Sahr for his support on the DGGRID software.

References

Amatulli, G., Garcia Marquez, J., Sethi, T., Kiesel, J., Grigoropoulou, A., Üblacker, M. M., Shen, L. Q., and Domisch, S.: Hydrography90m: A new high-resolution global hydrographic dataset, Earth System Science Data, 14, 4525–4550, publisher: Copernicus Publications Göttingen, Germany, 2022.

Chaudhuri, C., Gray, A., and Robertson, C.: InundatEd-v1.0: a height above nearest drainage (HAND)-based flood risk modeling system using a discrete global grid system, Geoscientific Model Development, https://doi.org/10.5194/gmd-14-3295-2021, 2021.

Ellis, E. C., Gauthier, N., Klein Goldewijk, K., Bliege Bird, R., Boivin, N., Díaz, S., Fuller, D. Q., Gill, J. L., Kaplan, J. O., and Kingston, N.: People have shaped most of terrestrial nature for at least 12,000 years, Proceedings of the National Academy of Sciences, 118, e2023483 118, publisher: National Acad Sciences, 2021.

Engwirda, D.: Reach a Python-based river network simplification algorithm, https://doi.org/10.5281/zenodo.8368261, 2023.

Engwirda, D. and Liao, C.: 'Unified' Laguerre-Power Meshes for Coupled Earth System Modelling, Zenodo, https://doi.org/10.5281/ZENODO.5558988, 2021.

Esri Water Resources Team: Arc Hydro Tools - Tutorial [Software], Tech. rep., 2011.

GDAL/OGR contributors: Geospatial Data Abstraction software Library [Software], https://gdal.org, 2019.

Goodchild, M.: Geographical grid models for environmental monitoring and analysis across the globe (panel session), vol. 94, 1994.

Huang, J. and Frimpong, E. A.: Modifying the United States National Hydrography Dataset to improve data quality for ecological models, Ecological informatics, 32, 7–11, publisher: Elsevier, 2016.

Kimerling, J. A., Sahr, K., White, D., and Song, L.: Comparing geometrical properties of global grids, Cartography and Geographic Information Science, 26, 271–288, publisher: Taylor & Francis, 1999.

Lehner, B., Verdin, K., and Jarvis, A.: New global hydrography derived from spaceborne elevation data, Eos, Transactions American Geophysical Union, 89, 93–94, publisher: Wiley Online Library, 2008.

Li, H., Wigmosta, M. S., Wu, H., Huang, M., Ke, Y., Coleman, A. M., and Leung, L. R.: A physically based runoff routing model for land surface and earth system models, Journal of Hydrometeorology, 14, 808–828, https://doi.org/10.1175/JHM-D-12-015.1, 2013.

Li, M., McGrath, H., and Stefanakis, E.: Multi-Scale Flood Mapping under Climate Change Scenarios in Hexagonal Discrete Global Grids, ISPRS International Journal of Geo-Information, 11, 627, publisher: MDPI, 2022.

Liao, C.: HexWatershed: A mesh-independent flow direction model for hydrologic models [Software], https://doi.org/10.5281/zenodo.6425881, 2022a.

Liao, C.: PyEarth: A lightweight Python package for Earth science [Software], https://doi.org/10.5281/ZENODO.6368652, 2022b.

Liao, C.: ISEA3H DGGs based flow routing datasets for the Amazon Basin, https://doi.org/10.5281/zenodo.8377765, 2023.

Liao, C. and Cooper, M. G.: Pyflowline: a mesh-independent river network generator for hydrologic models [Software]., https://doi.org/10.5281/zenodo.6407299, 2022.

Liao, C. and Cooper, M. G.: Pyflowline: a mesh-independent river network generator for hydrologic models, Journal of Open Source Software, 8, https://doi.org/10.21105/joss.05446, 2023.

Liao, C., Tesfa, T., Duan, Z., and Leung, L. R.: Watershed delineation on a hexagonal mesh grid, Environmental Modelling & Software, 128, 104 702, https://doi.org/10.1016/j.envsoft.2020.104702, 2020.

Liao, C., Zhou, T., Xu, D., Barnes, R., Bisht, G., Li, H.-Y., Tan, Z., Tesfa, T., Duan, Z., and Engwirda, D.: Advances in hexagon mesh-based flow direction modeling, Advances in Water Resources, p. 104099, https://doi.org/10.1016/j.advwatres.2021.104099, publisher: Elsevier, 2022.

Liao, C., Zhou, T., Xu, D., Cooper, M. G., Engwirda, D., Li, H.-Y., and Leung, L. R.: Topological relationship-based flow direction modeling: Mesh-independent river networks representation, Journal of Advances in Modeling Earth Systems, n/a, e2022MS003 089, https://doi.org/10.1029/2022MS003089, publisher: John Wiley & Sons, Ltd, 2023a.

Liao, C., Zhou, T., Xu, D., Tan, Z., Bisht, G., Cooper, M. G., Engwirda, D., Li, H.-Y., and Leung, L. R.: Topological Relationship-Based Flow Direction Modeling: Stream Burning and Depression Filling, Journal of Advances in Modeling Earth Systems, 15, e2022MS003 487, https://doi.org/10.1029/2022MS003487, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022MS003487, 2023b.

Lin, P., Pan, M., Wood, E. F., Yamazaki, D., and Allen, G. H.: A new vector-based global river network dataset accounting for variable drainage density, Scientific data, 8, 1–9, https://doi.org/10.1038/s41597-021-00819-9, publisher: Nature Publishing Group, 2021.

Mayorga, E., Logsdon, M. G., Ballester, M. V. R., Richey, J. E., and Richey, J. E.: LBA-ECO CD-06 Amazon River Basin Land and Stream Drainage Direction Maps, https://doi.org/10.3334/ornldaac/1086, 2012.

Mechenich, M. F. and Žliobaitė, I.: Eco-ISEA3H, a machine learning ready spatial database for ecometric and species distribution modeling, Scientific data, 10, 77, publisher: Nature Publishing Group UK London, 2023.

Purss, M. B. J., Gibb, R., Samavati, F., Peterson, P. R., and Ben, J.: The OGC Discrete Global Grid System core standard: A framework for rapid geospatial integration, IEEE International Geoscience and Remote Sensing Symposium, https://doi.org/10.1109/igarss.2016.7729935, 2016.

Randall, D. A., Ringler, T. D., Heikes, R. P., Jones, P., and Baumgardner, J.: Climate modeling with spherical geodesic grids, Computing in Science and Engineering, https://doi.org/10.1109/mcise.2002.1032427, 2002.

Saatchi, S.: LBA-ECO LC-15 SRTM30 Digital Elevation Model Data, Amazon Basin: 2000, https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1181, 2013.

Sahr, K.: Central Place Indexing: Hierarchical Linear Indexing Systems for Mixed-Aperture Hexagonal Discrete Global Grid Systems, Cartographica: The International Journal for Geographic Information and Geovisualization, https://doi.org/10.3138/cart.54.1.2018-0022, 2019.

Sahr, K.: DGGRID version 8.1b: User documentation for discrete global grid software [Software], https://github.com/sahrk/DGGRID, 2024.

Wu, H., Kimball, J. S., Li, H., Huang, M., Leung, L. R., and Adler, R. F.: A new global river network database for macroscale hydrologic modeling, Water resources research, 48, https://doi.org/10.1029/2012WR012313, publisher: Wiley Online Library, 2012.

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., and Pavelsky, T. M.: MERIT Hydro: A high-resolution global hydrography map based on latest topography dataset, Water Resources Research, 55, 5053–5073, publisher: Wiley Online Library, 2019.

Parameter name	Usage	Value	Note
dggrid_operation	Mesh generation purpose	GENERATE_GRID	
dggs_type	DGGS mesh type	ISEA3H	
dggs_res_spec	Resolution level	10, 11, 12, 13, 14	Level 14 is used for validation
clip_region_files	The clip region from the globe	The Amazon Basin boundary file	
update_frequency		10000000	
cell_output_type	Output format	GDAL_COLLECTION	
cell_output_file_name	Output file name	dggrid	
densification	Point on cell edge	0	
max_cells_per_output_file	File size control	0	
neighbor_output_type	File format for neighbor information	GDAL_COLLECTION	Neighbor information stored in the mesh

Table A1. List of the DGGRID mesh generation parameters used in this study.