A global open-source database of flood-protection levees on river deltas (openDELvE)

Joey O'Dell[1], Jaap H. Nienhuis[1], Jana R. Cox[1], Douglas A. Edmonds[2], and Paolo Scussolini[3]

[1] Department of Physical Geography, Universiteit Utrecht, Postbus 80.115, 3508 TC Utrecht, Netherlands
[2] Department of Earth and Atmospheric Sciences, Indiana University Bloomington, 1001 East 10th Street, Bloomington, IN 47405-1405, United States of America
[3] Institute for Environmental Studies, Vrije Universiteit Amsterdam, De Boelelaan 1111, 1081 HV Amsterdam, Netherlands

Correspondence to: Joey O'Dell (data@opendelve.eu)

Abstract. Flood-protection levees have been built along rivers and coastlines globally. Current datasets, however, are generally confined to territorial boundaries (national datasets) and are not always easily accessible, posing limitations for hydrologic models and assessments of flood hazard. Here we present our work to develop a single, **open**-source global river **de**lta **lev**ee data **e**nvironment (openDELvE) which aims to bridge a data deficiency by collecting and standardising global flood-protection levee data for river deltas. In openDELvE we have aggregated data from national databases as well as data stored in reports, maps, and satellite imagery. The database identifies the river delta land areas that the levees have been designed to protect, and where additional data is available, we record the extent and design specifications of the levees themselves (e.g., levee height, crest width, construction material) in a harmonised format. openDELvE currently contains 5,089 km of levees on deltas, and 44,733.505 km^2 of leveed area in 1,601 polygons. For the 152 deltas included in openDELvE, on average 19% of their habitable land area is confined by verifiable flood-protection levees. Globally, we estimate that between 5% and 54% of all delta land is confined by flood-protection levees. The data is aligned to the recent standards of Findability, Accessibility, Interoperability and Reuse of scientific data (FAIR) and is open-source. openDELvE is made public on an interactive platform (www.opendelve.eu), which includes a community-driven revision tool to encourage inclusion of new levee data and continuous improvement and refinement of open-source levee data.

1 Introduction

1.1 What are levees and what do they do?

Levees are banks of sediment or artificial material that prevent water from entering areas where it is not desirable. They are common in deltas and protect their populations and urban areas, including floodplains, from water level fluctuations of rivers and the sea. Levees have been constructed to mitigate flood risk and direct water flows throughout human civilisation. Recorded building of levees along the River Nile (Egypt) began around 4600 BP (Westermann, 1919) which indicates the innate link between the settlement of coastal populations and the development of levees. Modern materials and engineering concepts have altered the overall appearance and effectiveness of levees, but the basic principle remains the same.

Levees can also have negative environmental consequences. They alter sediment transport and sedimentation patterns, as sediment deposition behind levees is usually reduced. Areas protected by levees can subside relative to the surrounding (Middelkoop et al., 2010) resulting in increased risk of coastal and river flooding in the longer term (Criss and Shock, 2001; Munoz et al., 2018; Pinter, 2005; Pinter et al., 2008). Deltas in particular are at risk to be locked-in (Santos and Dekker, 2020), as areas become sediment starved and cease to keep up with sea level rise (Pinter et al., 2016). Another example of the negative effect of levees is in Australia, where undocumented private levees intending to protect land, resulted in degradation of the floodplain ecosystem, and contributed to flash flood risk by disconnecting the floodplain and channel (Steinfeld et al., 2013). Because of the negative consequences, contemporary fluvial and flood management scenarios

40 therefore often prioritise Nature-based Solutions that limit the need for levees (Cohen-Shacham et al., 2016; Esteves, 2014; Van Wesenbeeck et al., 2014), but this is not always possible.

1.2 Why (data of) levees matter

Data on levees is important, especially for river deltas. Modern urbanized deltas tend to be heavily embanked by levees because high population densities have demanded protection against river and coastal flooding. Despite these levees, people

45 living in coastal deltas face mounting threats; they are disproportionately affected by coastal flooding and relative sea level rise (Edmonds et al., 2020) and rely on ever diminishing river sediment supply (Dunn et al., 2019). Data on levees can help to assess these threats.

Mapping levee presence is useful for hydrologic and hydrodynamic modelling. Such models often aim to predict inundation

50 during high discharge events and help active management of risk and hazard to life, but large-scale modelling without accurate levee locations results in modelling inaccuracies (Fleischmann et al., 2019). Opposingly, models are also used to design levees by matching a specific return period or flood scenario, and so modelling data accuracy has a direct impact on the suitability of the height of the levee.

55 Data on levees can help to better understand human-landscape interactions (Werner and McNamara, 2007). Levees are deployed, (re-)engineered, and altered in response to threats – both present and future – so their dynamics are challenging to simulate, but remain of interest to the scientific and public communities (Di Baldassarre et al., 2018).

Levee data can also help studies on levee failures, which are a globally significant source of flood risk. Özer et al. (2019)

60 have developed the International Levee Performance Database (ILPD), focussed on assimilating data on levee testing and failure events whilst also producing an interactive and queryable interface. Levee data for hazard assessment purposes is additionally useful outside of the realm of pure geophysical modelling, and is core to civil engineering and emergency response management for levee performance, such as during the safety and risk calculation of hurricanes (Mitchell et al., 2013). The insurance industry, local residents and homeowners are additional users of levee data and modelling outputs

65 (National Research Council, 2013, p.68 Box 5-1) for hazard and risk assessments, where this data is usually commercially sensitive and not shared, whereas in the case of the US National Levee Database, open data being made available to all prevents this from being a hidden factor.

1.3 A (data) gap in levees

Despite the potential use of levee data, locations and characteristics of levees are often poorly documented (Özer et al., 2019;

70 Scussolini et al., 2016), resulting in inaccuracies and challenges for flood risk modelling (Sampson et al., 2015; Trigg et al., 2016; Winsemius et al., 2016), hazard modelling (Di Baldassarre et al., 2009), and projections of delta land loss from sea-level rise (Nienhuis and Wal, 2021). Accurate models require data input about levees including their spatial extent, protected area, and basic attributes, which currently does not exist in a coherent and harmonised single geospatial data format.

75 It is typically governments and municipal organisations who plan and construct levees, as safe and dry places to live are key to maintaining public health and wellbeing. These institutions also maintain them as part of their daily operations and produce maps and datasets about their design, operation, and failure. This gives a plethora of data such as reports and design specifications, which allows for accurate data gathering and collection processes without the need for in-person observation. Generally, this results in good quality central national databases, sometimes with higher resolution localised variants (e.g.,

80   New South Wales' Distinctive Land Surface Dataset, Australia) but that do not extend past administrative borders. Data
      availability can also be restricted.

      The lack of global registration, and therefore data, of these levees further complicates flood management efforts, however
      just registration of their existence cannot alone be considered a solution to the problem, and indeed we understand that

85   communication and awareness can be just as important in the whole-system oversight and evaluation of change. We
      recognise that attempts exist to document and aggregate specific aspects of levee failure (i.e., Özer et al., 2019) but we are
      not aware of any competing open-source approaches that collect and harmonise levee extent and attribute data.

1.4 Levees in hydrologic and hydrodynamic models

      Levees, by design, affect the flow of water. They can be incorporated in detailed models (e.g. HEC-RAS [US Army Corps of

90   Engineers, 2020], or Delft3D [https://oss.deltares.nl/web/delft3d]) as a geometric feature against initial surface topography.
      For models on larger scales, levees are too small to be included directly and are sometimes presented as a sub-grid feature or
      through a flood-attenuation proxy (Sampson et al., 2015). In both cases, poor data on levee existence and levee properties
      have made it such that their presence is often disregarded in global flood modelling (Trigg et al., 2016) and global delta
      modelling (Nienhuis et al., 2020). The lack of levee data (which change and control water and sediment discharge) results in

95   suboptimal   modelling   scenarios,   such   as   the   WRI   AQUEDUCT   flood-risk   tool
      (https://www.wri.org/applicaitons/aqueduct/floods/) which provides exceptional global-level data but does not include levees
      and results in abstract scenarios for heavily leveed areas such as the Netherlands.

      As an alternative to global levee data, FLOPROS (Scussolini et al., 2016) presents a global dataset on existing and policy-

100   level flood protection standards. FLOPROS provides uniform, global coverage, however individual feature level data is
      omitted. Other approaches exist that use (semi-)automated algorithms to locate and specify levees from LIDAR data (e.g.
      Steinfeld et al., 2013; Wing et al., 2019) but these are generally focussed on specific problem definitions and lack global
      applicability. A global levee database can help inform those algorithms and provide validation and calibration data.

1.5 Objective

105   The objective of openDElvE is to provide an attestable source of delta levee protection delta, for both primary use in flood
      and hazard modelling, as well as secondary community use through increased data availability by publishing the data on a
      public website (http://www.opendelve.eu) following standard data types, and a user-led amendment reporting function.

2 Methods

2.1 Overview

110   openDELvE is a collection of existing data on levees and protection features on deltas. We have collected data from vector,
      raster, and documentary sources. This results in two geospatial layers – one for leveed areas, and one for leveed lines – and a
      supporting index dataset, linked to the respective delta by a unique identifier and cross-mapped to the river delta dataset of
      Edmonds et al., 2020. Our methods allow for replicable tracing, processing, assimilation, and display of the data. By storing
      individual level references and data quality, we aim to provide data that is open and transparent. Our work is underpinned by

115   the principles of FAIR science to support reuse by producing data that is Findable, Accessible, Interoperable, and Reusable
      (Wilkinson et al., 2016). openDELvE development followed these steps: data definition (Sect. 2.2), data collection (2.3),
      data processing (Sect. 2.4), data attribution (Sect. 2.5), data management (Sect. 2.6), and data assurance (Sect. 2.7).

2.2 Data definition

We followed our definition of levees from Sect. 1.1. Levees exist along coasts and rivers globally, but the scope of
120   openDELvE is limited to river deltas (Sect. 2.4.1). We made use of a database of deltaic locations and deltaic area extent by
Caldwell et al. (2019) and Edmonds et al. (2020). We further limited ourselves to only storing information on defences that
are permanent features, and not temporary/reactive measures. Temporary measures, such as sandbags and hoardings
deployed for flash flooding or imminent but irregular flood issues are not temporally constant, and so are usually not
mapped, nor were considered for inclusion in this database.

125

openDELvE is designed to represent levees as geospatially explicit vector data: lines and polygons. For source data that
exists in reports on maps and technical drawings levee presence is often reduced to a raster map element, and so needed to be
sufficiently georeferenced and assessed for quality. However, this is still a valid data source and is included in our process.
We consider the age, source document, and data quality as we recognise that data may be reworked and requoted a number
130   of times in its lifespan.

openDELvE consists of three data elements: an index table and two vector layers (Table 1), each with a set of standardised
attributes (Table 2).

Data Entity	Type	Exported Elements	Purpose
Delta Index	Table	CSV	Contains data decision logs and linking characteristics at delta level
Leveed Area	Polygon	SHP, KML	A vector layer containing polygons of the areas protected by levees
Levee Lines	Line	SHP, KML	A vector layer containing lines of the levees and including standardised attributes

135   **Table 1: Data entities in the live viewing environment and their exported file types as in the research data store**

Data Entity	Attribute	Purpose
Delta Index	FriendlyName	Name of the delta, if known
	Status	Processed, No Result, Pending, or Not Processed (as per Sect. 2.3)
	PolygonID	Delta ID following Edmonds et al., 2020
	ISO_2	2-digit code identifying the country where the majority of the delta lies, following ISO 3166-1:2020 alpha-2
	Journal	A timestamped text log of activity at a delta level
	MainRefAPA7	Literature reference for the overall source material for the delta, formatted in APA 7th Edition, if available
	MainRefDOI	Digital Object Identifier for the source material, if available
	NeedsReview	Boolean indicator of requirement for later review of delta
	LastChkDate	Date field signalling last check date of the delta
	LastChkBy	Two-character identifier of the last user who updated the dataset
Leveed Area	NAME	Name of the leveed area feature from the source dataset, if available
	REFERENCE	The identifier for the feature from the source dataset, if available
	DOI	Digital Object Identifier for the source material, if available
	URL	Uniform Resource Locator (web link) for the source material, if available
	LITREF	Literature reference for the source material, formatted in APA 7th Edition

	PolygonID	Delta ID following Edmonds et al., 2020
	DataQuality	Data quality classification (following Table 3)
Levee Lines	NAME	The name for the feature from the source dataset, if available
	REFERENCE	The identifier for the feature from the source dataset, if available
	DOI	Digital Object Identifier for the source material, if available
	URL	Uniform Resource Locator (web link) for the source material, if available
	LITREF	Literature reference for the source material, formatted in APA 7th Edition
	DefenceLength	The length of the levee feature, as provided in the source dataset, if available (metres)
	DefenceHeight	The height of the levee feature, as provided in the source dataset, if available (metres)
	DefenceWidth	The width of the levee feature, as provided in the source dataset, if available (metres)
	FoundationWidth	The width of the levee foundation, as provided in the source dataset, if available (metres)
	Construction	The primary material that the levee is composed of
	ClassType	Construction or formation type of the feature
	CutoffMaterial	The material that the levee cutoff is composed of
	DesignStandard	Design storm rating of the feature (1/n, decimal)
	DataQuality	Data quality classification (following Table 3)
	PolygonID	Delta ID following Edmonds et al., 2020

Table 2: openDELvE attributes for the three data elements (as per Table 1). Conversion factors and mapping of fields is given in Supplementary Table S2.

140   Levee data in openDELvE include a data quality class and a direct link to the source dataset. We devised the data quality criteria included in Table 3:

Class	Criteria
A (Excellent)	Vector data
	First-order data source (i.e., scientific papers, governmental geospatial data, original publication)
	Spatially complete[a] (with respect to geopolitical boundaries)
	Existence verifiable with satellite imagery
B (Good)	Raster data (suitably georeferenced, little to no variance)
	First-order or re-cited/modified (original accessible) but published within a scientific or government publication
	Existence verifiable with satellite imagery
C (Acceptable)	Raster data (loosely georeferenced, variance due to old base map or similar)
	Conjectural or non-scientific source (ex: newspaper)
	Source >20 years of age, regardless of type
	Existence (partially[b]) verifiable with satellite imagery
X[c] (Invalid)	Data inaccessible (blocked, hidden, unpublished)
	Irrecoverable issues with data quality
	Could not confirm existence of data from other sources using satellite imagery with resolution ≤25m
	Temporary or reactive measures only (ex: sandbags)

Table 3: Data quality definition for levees based upon data provenance, both for use in initial data classification and ongoing maintenance. Criteria are exclusively applied, and all categories must be met to meet that classification.

145   [a]*Data that were attributable to class X have not been included in the published dataset but are documented in the delta index*

[b]*We included 'partially verifiable' due to incident patchy local coverage of openly accessible satellite data, as there are instances where sufficient high-resolution imagery was not accessible, but standard-resolution imagery indicated the presence of the feature that was elsewhere published.*

[c]*Spatially complete was defined as being of the entire levee run, which may be comprised of several subsection maps.*

2.3 Data collection

We conducted extensive literature searches using a variety of web searching platforms (i.e., Clarivate Web of Science, Google Search, Google Scholar, OCLC WorldCat) as well as data aggregation platforms (e.g., re3data.org, DataCite, data.gov.uk, data.gov, data.gov.au). Data was collected in a search process that is documented as a log with diary-style entries in the Delta Index table (see Table 1) and recorded at a delta level. Sources for each individual levee are stored at the feature level. This allowed us to record rationale and decision-making process so that both viewers and onward developers of the dataset are aware of the steps taken and explanations for decisions taken in data hand.

With an international scope, searching often required country or location-specific terms (e.g., '*tanggul*' meaning levee or embankment in Indonesian) to aid data discovery, and these were regionally supplemented along with a vocabulary of common delta and levee terms when using academic paper and internet indexing services.

Funding reports from the World Bank projects on flood defence activities has also contributed to the database. Financing documents often contain maps and so we include data from the World Bank where it was discovered in our searches, released publicly, had been reviewed, and contained levee feature level data.

When it was not possible to find data in areas where levees were expected, the place was identified by name using the address search (*gazetteer*) function in ArcGIS and then basic internet searching was performed to find reports of floods or sea level rise related damage. Finally, we made use of the world satellite imagery layer within ArcGIS to review areas where levee source data was inaccessible, and assess by visual means whether it was likely levees were present. We verified areas that we believe may be uninhabited areas using this imagery and classified them accordingly, where satellite imagery confirmed no visible levees, the delta was set to No Result. If levees were visible but we could not verify them with alternative data sources, we set the delta to 'Pending' where external enquiries were taking place and the relevant note was entered in the Journal. We identify deltas as 'Not Processed' if we have yet to manually review available sources, and no national vector dataset was discoverable for processing via our automated tool.

Many deltas in the delta dataset may be small and uninhabited (Edmonds et al., 2020), have inaccessible data, or have data that we were unable to convert into a format that we could add to the database. We collectively group these deltas as having "No result" in terms of data collection. Note that this does not always mean there is no data. For example, data from the Database nazionale della AgriNature in TErra (DANTE, *formerly known as: ItaliaN LEvee Database [INLED]*) (Barbetta et al., 2015) was not suitable for processing because it only contains a levee start and end point coordinate. We classify these deltas under "No result" because it requires access to a detailed regional-level watercourses database and high-resolution DEM so that an interpretational algorithm could be trained to infer the levee course.

Where available, we include levee attributes (e.g., design storm, wall height, levee material, Table 2). This can inform modelling and therefore work as a stand-alone spatial tool for investigating river delta dynamics. Additionally, the data layers can be used for verification of deductive models for the detection of levees by other means, including LIDAR and remotely sensed data as well as corroborating other data sources, such as OpenStreetMap. As we intend for the database to be globally comparable, we set up a cross matching list (Supp. Table S2) within the project documentation to ensure that the

attributes of the levee lines layer were consistent between sources and languages. This was then used for both manual and
190   automated input so that different units of measure, classifications of levee and construction type, and key engineering data
were harmonious.

2.4 Data processing

2.4.1 Vector data processing

Where data was sourced in vector format, we defined a data processing algorithm in the ArcGIS® Model Builder (Supp. Fig.
195   S1) to clip the imported data to the extent of river deltas from Edmonds et al. (2020) with a 100 km 'buffer zone'. This
buffer zone is included to maximize OpenDELvE data usability but it does not affect reported statistics on delta coverage.
All reported data statistics in this paper are for levees strictly within delta boundaries (Fig. 1). The buffer zone is included to
allow extended use of the dataset for upstream fluvial and sediment transport modelling and additionally, should dataset of
Edmonds et al. (2020) be updated, reduces the likelihood that levees are missed from the layer.
200

The ArcGIS® Model Builder automated import process created is distributed with the dataset so that data can be repeatedly
processed and added to the database both now and in the future. We supplemented this by the creation of conversion tables
(Supplementary Table S2) so that levee attributes, where available, are comparable at a global scale.

2.4.2 Non-vector data processing

205   We performed georeferencing of map/documentary data where the location was visible using a contemporary map and the
map could be referenced in less than 5 reference points. This ensured that we were not extensively distorting the source map
and therefore it was possible for us to trace in the features as accurately as possible. Where no georeferencing within 5
reference points was possible, or where the map had too few defining features to be georeferenced at all (e.g. map created
with too few topographical features, substantial engineered or geological change resulted in difference between map and
210   modern day situation) then the appropriate data quality class was assigned, and where the map was impossible to suitably
georeference, the data source was set aside and documented in the log. Furthermore, where aerial photography was analysed,
we defined a set protocol for the inference of leveed area (Supp. Fig. S3).

Data in the "Levee Lines" layer is currently limited to vector levee data sources and does not exist for raster data sources.
215   Ongoing work includes manual review and development of (semi)automated processing steps to retrieve levee lines from
raster sources.

2.5 Data attribution

Every task performed in the journal is recorded for audit purposes, and each entry into the layers is attributed to the data
source, including a full literature ref, the source URL, and a DOI (where available). This ensures that we can display this
220   data interactively and that the original source remains permanently available. We also included any digital identifiers from
vector datasets so that the individual feature can be tracked and mapped over subsequent data revisions.

We timestamp each entry into the delta index and additionally flag deltas that need manual review in the future. This has no
effect on data quality, however it ensures that there is a robust process in the future to signal amendments needed or entries
225   where it is apparent that there are undocumented or inaccessible data sources available. This not only supports local
maintenance, but also prevents repetition of previous search activities.

2.6 Data management

The resulting data layers for levee area and levee line feature were created in ArcGIS Pro and hosted on an ArcGIS Online data hub (http://www.opendelve.eu). Additionally, we maintain ongoing research data exports in the DataverseNL

230   environment as the database develops. Data is stored in three defined entities as per Table 1, and we store each layer within their own container in the public ArcGIS Online® environment. These layers are then publicly published to be used as part of the ArcGIS Online Directory and through modern GIS clients via Web Feature Service (WFS).

The openDELvE platform facilitates an interactive and community driven maintenance of the dataset through an amendment

235   form and additional messages in all metadata files. The project remains actively maintained by the authors at Utrecht University, and by assigning permanent identifiers (DOIs) to the research dataset, as well as developing the website alongside, there project remains actively maintained.
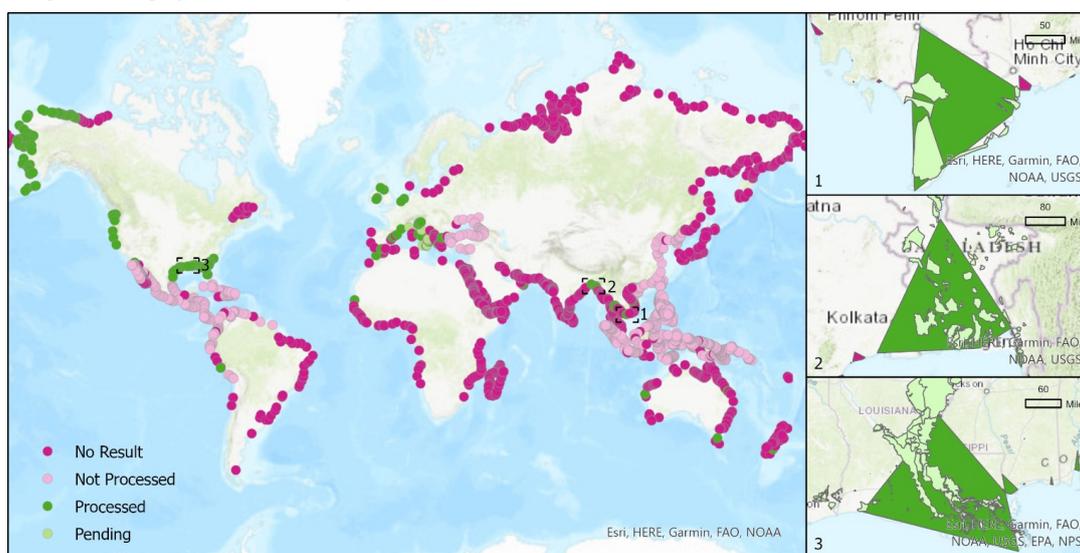
Figure 1: Distribution of delta levee dataset completeness and data availability in release of openDELvE (v1.0). Polygons
240   **encompass the four-point deltaic extent as defined by Edmonds et al. (2020).**

2.7 Data assurance

Before releasing the dataset, we performed several checks on the data and metadata (Table 4). We then generated metadata compliant with the EU INSPIRE geospatial metadata standard (European Parliament, 2007) using the built in ArcGIS® Pro wizard for each data element (Table 2), and for the dataset in its entirety. This included interactive help-text for the model

245   builder GUI. We self-validated the Metadata files using the metadata wizard in the ArcGIS® Pro system.

Type	Criteria
Duplicate Check	There are no duplicate delta polygon IDs (*PolygonID)* in the index
Orphan Check	All linked delta polygon IDs matched a delta polygon in the dataset
	There were no unsuccessful joins between the data layers
Null Check	Where there was no match to a delta polygon, this returned -1
	Where it was not (yet) possible to match the polygon to a delta, this returned null
Visual Check	Visually verify data appears as should be reasonable to expect (i.e., within 100 km of delta
	polygon border, within proximity of water feature, of a shape that is coincident to fluvial

	morphology)
Metadata Check	All fields in the ArcGIS® Pro metadata wizard completed

Table 4: Categories and criteria for the data validation performed on the dataset

3. Results

Our leveed area data layer contains 44,734 km[2] of identified delta area protected by levees (following definition in Sect. 2.4)

250 (Table 5). Levee line data that we have processed contains 5,089 km of levee. We have processed levee information for 152 deltas of the 2,174 deltas identified by Caldwell et al. (2019), representing 28% of delta area (239,044 km[2] of 847,936 km[2]). We find 1,098 deltas that are pristine and/or small, where levees are unlikely; or, where no data on levee presence could be found, and we could not identify levees visually either (No Result category) (59% of delta area). A further 924 deltas remain unprocessed or await data input from external sources, however this represent only 12% of global geomorphic delta area, and

255 as such we have processed the largest deltas and the remaining deltas to process are small in surface area.

Comparing flood-protected delta area against delta extent, we find that 19% (44,734 km[2] of 239,044 km[2]) of the processed delta area is contained within levees. Total global delta area is 847,936 km[2] (Edmonds et al., 2020), which means that, at least, 5% of global delta area is within verifiable levees. This number should be considered a minimum as many deltas

260 remain unprocessed and we suspect that many levees exist that are not (yet) in the openDELvE dataset as even when discovering data, there existed data sources that were incompatible with the licensing of the dataset, for which we document in the delta index.

Percentage coverage between continental zones (using the UN Region from Edmonds et. al. (2020) ranges from 13%

265 (Africa) to 54% (Europe and C. Asia) and this broadly reflects the different data publishing policies in these regions. As can be seen in Fig. 1, we recognise that the global distribution of data is sub-optimal, and we investigate the imbalance further in the discussion (Sect. 4.2). As discussed in Sect. 2.4.1, the data from Edmonds et al. (2020) consists of polygons drawn from four maximal extent points to create a four sided polygon which represents maxima and not the absolute extent of the delta, so we have calculated these statistics based upon the geomorphic area data provided by Edmonds et al. (2020).

270

Of the leveed area data in openDELvE, 90% of the levee area dataset by area (1,641 unique features) is considered of excellent or good quality (Fig. 3). This indicates that most of our sources are vector and high-quality raster data, which we believe supports high-quality onward data propagation to, and consumption by, the hydrological and risk modelling communities.

275

Continental Zone (UN Region)	Number of deltas with levee data present in openDELvE[a]	Total number of unique leveed areas represented in dataset	Total geomorphic deltaic area[a] [km2]	Area protected by levees within the delta[a] [km2]	Coverage of delta area by levees as % of deltaic area (computed)
Africa	3	9	4,358	570	13 %
Americas	100	301	99,262	15,282	15 %
Asia-Pacific	19	83	128,970	25,396	20 %
Europe & C. Asia	30	68	6,454	3,486	54 %
Processed Total	**152**	**461**	**239,044**	**44,734**	**19 %**

No Result	1,098	-	502,928	-	-
Unprocessed (Pending & Not Processed)	924	-	105,964	-	-
Global Total	**2,174**	**461ᵃ**	**847,936ᵇ**	**44,734**	**5 %**

Table 5: Summary of processed features and deltaic area at openDELvE release v1.0 per geographic region and area totalled, figures are rounded to nearest whole integer.

ᵃopenDELvE contains 1,601 leveed area polygons mainly due to the structure of administrative units in the USACE NLD. Overlapping sections are therefore only counted once and for the purpose of this article we 'dissolved' the layer to ensure that area
280 **was not double counted.**

ᵇCalculated by dissolving leveed area dataset to a single layer and compared to supplementary data from Edmonds et al. (2020).
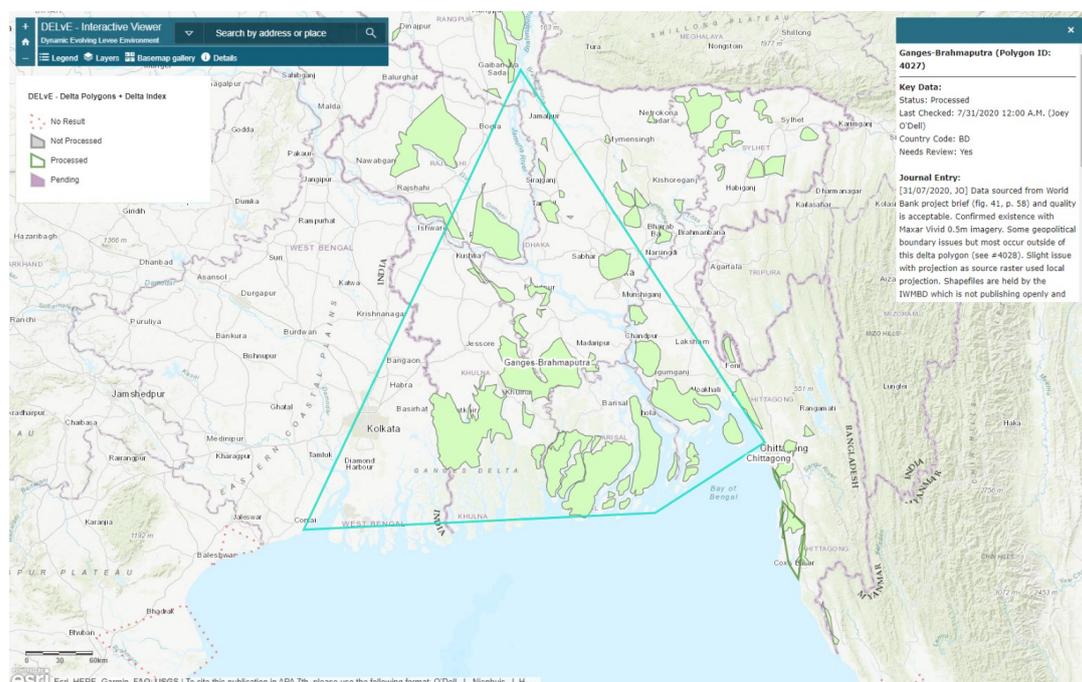
Figure 2: Interactive browsing interface to openDELvE built using the ArcGIS Online platform. Area of focus is the Ganges-Brahmaputra delta, Bangladesh. Data content as per openDELvE version 1.0. Available publicly at: http://opendelve.eu

285
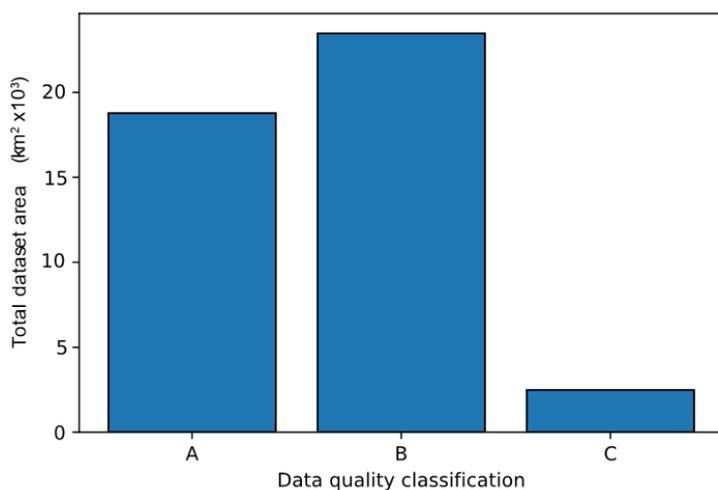
Figure 3: Distribution of data quality classification in openDELvE (v1.0) given for each individual leveed area feature, classified according to the data quality matrix (Table 3)
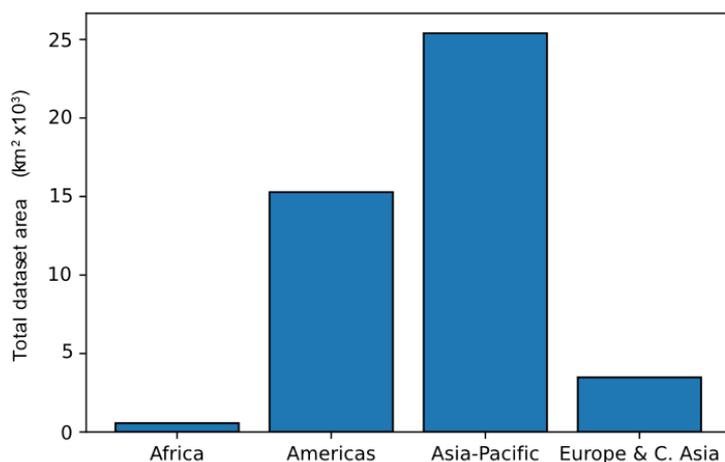
Figure 4: Distribution of leveed area data in openDELvE (v1.0) by UN Region using allocation of deltas to region by Edmonds et al. (2020)

4. Discussion

4.1 How representative is openDELvE?

As summarised in Table 5, we found that 19% of the geomorphic delta area (which can include the shallow marine portions of the delta front) processed in openDELvE is protected by a levee. This should be considered a rough estimate. For deltas covered by nationally maintained databases (e.g. Mississippi, Rhine-Meuse) the data quality is good. There is rich metadata and there is little chance of false negatives (no levee in openDELvE but levees present nevertheless). Data quality and coverage in other deltas (e.g. Ganges-Brahmaputra, Mekong) is poorer, and this appears to be linked to the lack of a

nationally or regionally coordinated platform for levee data sharing. While individual levees or leveed areas are represented,
300   there is a high chance of false negatives and undercounting for the delta as a whole.

Extrapolating our findings to a global level (Table 5) by including the unprocessed and no result categories, we might
conclude that global leveed delta area is likely lower than 19%. Most deltas where we could not find information on levees
(no result) are small and uninhabited, those represent 59% of global delta area. We expect those to have fewer levees
305   compared to deltas within openDELvE. Global delta levee area is probably higher than 5%, given that this would mean
openDELvE currently includes all levees on deltas. 54% is a likely maximum because this is what we found for Europe and
Central Asia, where many deltas are included and data quality is generally good.

4.2 Global barriers to data availability

Data sovereignty is an emerging topic within global modelling that revolves around the value, sharing, and ownership of data
310   in a global context. Whilst we acknowledge that breakthroughs have been made in the academic world of data sharing,
through the formation of data initiatives (i.e., FAIR) and for standardised data sharing (e.g., INSPIRE, European Parliament,
2007), data in the private and governmental sectors can still be considered as an internal asset. Tang et al. (2020) define the
term 'data sovereign' to identify someone with the capabilities, skill set, and hierarchical position to facilitate data sharing
across global borders, in turn facilitating the share of knowledge and contributing overall to the global economy.

315

We identified that countries and governmental organisations which have core values supporting open (governmental) data
tend to treat levee information as a 'product' and therefore appoint a central data repository or facilitated ordering process to
act as a 'data sovereign'. Examples of these are the US data.gov platform (which holds record locators for the US Army
Corps of Engineers *National Levee Database*), the UK data.gov.uk Open Data platform (which holds record locators for the
320   UK Environment Agency *Asset Information Management System*), the Dutch data.overheid.nl (which holds record locators
for the Rijkswaterstaat *Dataregister*), and Australian data.gov.au (which holds record locators for the various state-led
systems in place across the country). These repositories may not themselves hold the actual data but act as centrally
maintained indices of 'open government' data.

325   Data ownership and 'sovereignty' can act as a roadblock to progress towards a harmonised global database. We have
attempted to counter this by using extensive, disparate sources, but we understand that it is beneficial when the creators and
maintainers of such levees additionally release information on their locations and engineered properties. We recognise that
data is often locked away in local, national, or offline archives that we were unable to access, and as such this is only a
partially complete dataset. Acknowledging this, we have attempted to conduct research using a best effort multilingual
330   approach but recognise that local knowledge will ultimately prevail in the provision of high-quality data.

4.3 Bias in data availability

There is a clear global divide in data availability, with countries such as the UK, USA, Australia, and bloc systems such as
the EU gathering vector flood defence data on a centralised platform. We identify that these data environments have defined
a clear 'data sovereign' and this is enshrined both in law and local process. In contrast, levee data availability in deltas in
335   Africa, South-East Asia, the Southern and Central Americas - as well as those in the Russian Federation and late-accession
members to the EU either simply do not have this data, or it is locked away in archives and not available in globally
published formats. This creates a bias in our database, as there is an unfair co-occurrence of data inavailability and the
requirement for data to enhance modelling and management in these areas.

340   We hope that by publishing our data as openly as possible (following FAIR princples) we can encourage not only external inspection but also suggestion of changes and further data additions. We explicitly have made our map and submission system as lightweight as possible with the minimal number of questions and responses. We also encourage the furthest and widest possible spectrum for reuse. Additional crowdsourced or "*volunteer geographic information" (VGI,* Young et al., 2020) projects such as 510 - an initiative of the Netherlands Red Cross - and OpenStreetMap.org (amongst other platforms)

345   may be able to further expand data on levees. We acknowledge that for greater global adoption a local presence is needed, and recognise the work of Young et al. (2020) in documenting the deployment of, and challenges associated with, a globally diverse data collection project, however including crowed-sourced data was out of the scope of our research.

4.4 Future outlook

Our dataset is incomplete, mainly due to the varying nature of data publishing. By publishing data as openly and FAIR as

350   possible, and incorporating a feedback system into our web application, we aim to set a solid foundation for the development of a global levee database. Publishing openDELvE provides a foundation for the development not only of this dataset further, but also for the inclusion of higher-quality levee data in hydrodynamic models in general. Aside from direct use in global models, there is also the possibility of openDELvE to function as a training dataset for statistical (machine learning) models for levee and flood detection (Wing et al., 2019). By publishing our data with an open licence (Creative Commons

355   Attribution) we encourage onward consumption and reworking of the database.

The data availability landscape is continually improving, and further hope that by showing the encouraging work we have achieved, further data sources will become available, and this will further foster the culture of open data and data sharing in the earth sciences and beyond. Data accessibility remains a factor in the limitation and applicability of any such system.

360   Publishing levee data could be an important step not only in supporting flood-risk modelling, but additionally in the increased visibility and ownership of their existence. We greatly encourage further data submissions and amendments to our database as the global geoscience community identify local sources of data, but furthermore encourage authorities and public works bodies to make their data openly accessible.

5. Conclusion

365   OpenDELvE is a global delta levee database of 461 unique leveed areas and 5,089 km of levee extent. We have standardised levee attributes and features to allow global comparability, finding that, for the deltas we processed, 44,733.505 km^2 is contained within levees. This represents 19% of their area and 5% of global delta area.

The database is FAIR, openly available and we encourage contributions from other researchers or levee experts. Our

370   database is biased due to data availability, with UN regions of Asia-Pacific and Americas having a higher data availability than Africa and Europe & Central Asia.

Additionally, we acknowledge that the lack of levee data for countries whose political, financial, or administrative structure prevents open data publication, and indeed we identify the great work of many academics such as Wang et al. (2021) in

375   creating calculated attempts as flood protection layers to facilitate disaster management and hydrological research despite this lack of data. We recommend that governments, where possible, seek to appoint a 'data sovereign' to take on this task.

Code availability

The ArcGIS® Model Builder template used to process vector data is published within the research dataset.

Data availability

380 The research dataset is publicly available on DataverseNL at https://doi.org/10.34894/2WZ0S9.

The layers and viewing interface are publicly consultable at http://www.opendelve.eu and are additionally hosted in the ArcGIS Online Portal for use with ArcGIS® and other OGC-compatible GIS packages.

Author contribution

JO curated and maintained the data, performed the investigation, and prepared the draft manuscript with contributions from
385 the co-authors. JHN conceptualised, validated, and supervised the project, and edited the manuscript. JRC supervised the project, visualised the results and edited the manuscript. DAE and PS provided key digital resources and reviewed the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Disclaimer

390

Figure 1, 2 & Supplementary Figure 1 were created using ArcGIS® software by Esri (ArcGIS® Pro [ver. 2.6.2] and ArcGIS® Online). Cartographic data displayed in the images is supplied by Esri under licence from HERE, Garmin, FAO, NOAA, EPA, NPS, and the USGS.

Acknowledgements

395 The authors wish to thank the many organisations who have actively chosen to embrace open data standards and share such rich datasets as this is a crucial part of the database. We are grateful to Ece Özer (TNO) and Alex Curran (formerly of TU Delft) whose work on the SAFElevee project and the ILPD, and their detailed publications as a result, were instrumental in the discovery of closed datasets. We further thank Silvia Barbetta, Albert Kettner, Yoshiki Saito, and Dhruvesh Patel for their help in gaining access to international data. We would additionally like to thank Fergus Miller Kerins for his help in
400 testing and exploring deployment possibilities for the dataset. Finally, we are very appreciative of the support from Maarten Zeylmans van Emmichoven that allowed the data processing to continue remotely during the pandemic.

References

Di Baldassarre, G., Castellarin, A., Montanari, A. and Brath, A.: Probability-weighted hazard maps for comparing different flood risk management strategies: a case study, Nat. Hazards, 50(3), 479–496, doi:10.1007/s11069-009-9355-6, 2009.

405 Di Baldassarre, G., Kreibich, H., Vorogushyn, S., Aerts, J., Arnbjerg-Nielsen, K., Barendrecht, M., Bates, P., Borga, M., Botzen, W., Bubeck, P., De Marchi, B., Llasat, C., Mazzoleni, M., Molinari, D., Mondino, E., Mård, J., Petrucci, O., Scolobig, A., Viglione, A. and Ward, P. J.: Hess Opinions: An interdisciplinary research agenda to explore the unintended consequences of structural flood protection, Hydrol. Earth Syst. Sci., 22(11), 5629–5637, doi:10.5194/hess-22-5629-2018, 2018.

410 Barbetta, S., Camici, S., Maccioni, P. and Moramarco, T.: National Levee Database: monitoring, vulnerability assessment and management in Italy, in Geophysical Research Abstracts, vol. 17, pp. EGU2015-10170., 2015.

Caldwell, R. L., Edmonds, D. A., Baumgardner, S., Paola, C., Roy, S. and Nienhuis, J. H.: A global delta dataset and the

environmental variables that predict delta formation on marine coastlines, Earth Surf. Dyn., 7(3), 773–787, doi:10.5194/esurf-7-773-2019, 2019.

415   Cohen-Shacham, E., Walters, G., Janzen, C. and Maginnis, S.: Nature-based solutions to address global societal challenges, edited by E. Cohen-Shacham, G. Walters, C. Janzen, and S. Maginnis, IUCN International Union for Conservation of Nature., 2016.

Criss, R. E. and Shock, E. L.: Flood enhancement through flood control, Geology, 29(10), 875, doi:10.1130/0091-7613(2001)029<0875:FETFC>2.0.CO;2, 2001.

420   Dunn, F. E., Darby, S. E., Nicholls, R. J., Cohen, S., Zarfl, C. and Fekete, B. M.: Projections of declining fluvial sediment delivery to major deltas worldwide in response to climate change and anthropogenic stress, Environ. Res. Lett., 14(8), 084034, doi:10.1088/1748-9326/ab304e, 2019.

Edmonds, D. A., Caldwell, R. L., Brondizio, E. S. and Siani, S. M. O.: Coastal flooding will disproportionately impact people on river deltas, Nat. Commun., 11(1), 4741, doi:10.1038/s41467-020-18531-4, 2020.

425   Esteves, L. S.: What is Managed Realignment?, pp. 19–31, Springer, Dordrecht., 2014.

European Parliament: Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), European Parliament., 2007.

Fleischmann, A., Paiva, R. and Collischonn, W.: Can regional to continental river hydrodynamic models be locally relevant? A cross-scale comparison, J. Hydrol. X, 3, 100027, doi:10.1016/j.hydroa.2019.100027, 2019.

430   Middelkoop, H., Erkens, G. and van der Perk, M.: The Rhine delta-a record of sediment trapping over time scales from millennia to decades, J. Soils Sediments, doi:10.1007/s11368-010-0237-z, 2010.

Mitchell, T., Kluskens, R., Woldringh, B., Van Der Meer, M. T., Kamp, R. G., De Gooijer, C. and Hillen, M. M.: Integrating Levee Performance Assessments into Complex Flood Protection Systems, in US Society of Dams Annual Conference., 2013.

435   Munoz, S. E., Giosan, L., Therrell, M. D., Remo, J. W. F., Shen, Z., Sullivan, R. M., Wiman, C., O'Donnell, M. and Donnelly, J. P.: Climatic control of Mississippi River flood hazard amplified by river engineering, Nature, 556(7699), 95–98, doi:10.1038/nature26145, 2018.

National Research Council: Levees and the National Flood Insurance Program, National Academies Press, Washington, D.C., doi: 10.17226/18309, 2013.

440   Nienhuis, J. H. and Wal, R. S. W. van de: Projections of Global Delta Land Loss From Sea-Level Rise in the 21st Century, Geophys. Res. Lett., 48(14), e2021GL093368, doi:10.1029/2021GL093368, 2021.

Nienhuis, J. H., Ashton, A. D., Edmonds, D. A., Hoitink, A. J. F., Kettner, A. J., Rowland, J. C. and Törnqvist, T. E.: Global-scale human impact on delta morphology has led to net land area gain, Nature, 577(7791), 514–518, doi:10.1038/s41586-019-1905-9, 2020.

445   Özer, I. E., van Damme, M. and Jonkman, S. N.: Towards an International Levee Performance Database (ILPD) and Its Use for Macro-Scale Analysis of Levee Breaches and Failures, Water, 12(1), 119, doi:10.3390/w12010119, 2019.

Pinter, N.: One Step Forward, Two Steps Back on U.S. Floodplains, Science (80-.)., 308(5719), 207–208, doi:10.1126/science.1108411, 2005.

Pinter, N., Jemberie, A. A., Remo, J. W. F., Heine, R. A. and Ickes, B. S.: Flood trends and river engineering on the
450   Mississippi River system, Geophys. Res. Lett, 35, 23404, doi:10.1029/2008GL035987, 2008.

Pinter, N., Huthoff, F., Dierauer, J., Remo, J. W. F. and Damptz, A.: Modeling residual flood risk behind levees, Upper Mississippi River, USA, Environ. Sci. Policy, 58, 131–140, doi:10.1016/j.envsci.2016.01.003, 2016.

Sampson, C. C., Smith, A. M., Bates, P. D., Neal, J. C., Alfieri, L. and Freer, J. E.: A high-resolution global flood hazard model, Water Resour. Res., 51(9), 7358–7381, doi:10.1002/2015WR016954, 2015.

455   Santos, M. J. and Dekker, S. C.: Locked-in and living delta pathways in the Anthropocene, Sci. Rep., 10(1), 1–10,

doi:10.1038/s41598-020-76304-x, 2020.

Scussolini, P., Aerts, J. C. J. H., Jongman, B., Bouwer, L. M., Winsemius, H. C., De Moel, H. and Ward, P. J.: FLOPROS: an evolving global database of flood protection standards, Nat. Hazards Earth Syst. Sci., doi:10.5194/nhess-16-1049-2016, 2016.

460   Steinfeld, C. M. M. M., Kingsford, R. T. and Laffan, S. W.: Semi-automated GIS techniques for detecting floodplain earthworks, Hydrol. Process., 27(4), 579–591, doi:10.1002/hyp.9244, 2013.

Tang, C., Plasek, J. M., Zhu, Y. and Huang, Y.: Data sovereigns for the world economy, Humanit. Soc. Sci. Commun., 7(1), 184, doi:10.1057/s41599-020-00664-y, 2020.

Trigg, M. A., Birch, C. E., Neal, J. C., Bates, P. D., Smith, A., Sampson, C. C., Yamazaki, D., Hirabayashi, Y.,
465   Pappenberger, F., Dutra, E., Ward, P. J., Winsemius, H. C., Salamon, P., Dottori, F., Rudari, R., Kappes, M. S., Simpson, A. L., Hadzilacos, G. and Fewtrell, T. J.: The credibility challenge for global fluvial flood risk analysis, Environ. Res. Lett., 11(9), 094014, doi:10.1088/1748-9326/11/9/094014, 2016.

US Army Corps of Engineers: River Analysis System (HEC-RAS), 2020.

Wang, D., Scussolini, P. and Du, S.: Assessing Chinese flood protection and its social divergence, Nat. Hazards Earth Syst.
470   Sci., 21(2), 743–755, doi:10.5194/nhess-21-743-2021, 2021.

Werner, B. T. and McNamara, D. E.: Dynamics of coupled human-landscape systems, Geomorphology, 91(3–4), 393–407, doi:10.1016/j.geomorph.2007.04.020, 2007.

Van Wesenbeeck, B. K., Mulder, J. P. M., Marchand, M., Reed, D. J., De Vries, M. B., De Vriend, H. J. and Herman, P. M. J.: Damming deltas: A practice of the past? Towards nature-based flood defenses, Estuar. Coast. Shelf Sci., 140, 1–6,
475   doi:10.1016/j.ecss.2013.12.031, 2014.

Westermann, W. L.: The Development of the Irrigation System of Egypt, Class. Philol., 14(2), 158–164, doi:10.1086/360222, 1919.

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,
480   S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., t Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S. A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., Van Der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B.: Comment: The FAIR Guiding Principles for scientific data management and
485   stewardship, Sci. Data, 3(1), 1–9, doi:10.1038/sdata.2016.18, 2016.

Wing, O. E. J., Bates, P. D., Neal, J. C., Sampson, C. C., Smith, A. M., Quinn, N., Shustikova, I., Domeneghetti, A., Gilles, D. W., Goska, R. and Krajewski, W. F.: A New Automated Method for Improved Flood Defense Representation in Large-Scale Hydraulic Models, Water Resour. Res., 55(12), 11007–11034, doi:10.1029/2019WR025957, 2019.

Winsemius, H. C., Aerts, J. C. J. H., Van Beek, L. P. H., Bierkens, M. F. P., Bouwman, A., Jongman, B., Kwadijk, J. C. J.,
490   Ligtvoet, W., Lucas, P. L., Van Vuuren, D. P. and Ward, P. J.: Global drivers of future river flood risk, Nat. Clim. Chang., 6(4), 381–385, doi:10.1038/nclimate2893, 2016.

Young, J. C., Lynch, R., Boakye-Achampong, S., Jowaisas, C., Sam, J. and Norlander, B.: Volunteer geographic information in the Global South: barriers to local implementation of mapping projects across Africa, GeoJournal, 1–17, doi:10.1007/s10708-020-10184-6, 2020.

495