ICFM 10
Adapting to Global Change:
Innovative Approaches to Flood Management and Resilience
May 20-22, 2026 | Western University, London, Ontario, Canada

Building large-scale flood defence datasets using open data

Podt, M.[1,2], Bron, C.[1], den Heijer[1], F.[1], Rijke, J.S.[1,2]
HAN University of Applied Sciences, Ruitenberglaan 26, 6826 CC Arnhem, The Netherlands[1]
Delft University of Technology, Mekelweg 5, 2628 CD Delft, The Netherlands [2]
E-mail: maarten.podt@han.nl

ABSTRACT

Reliable technical data are essential for flood defence asset management, yet such data are often fragmented across organisations, technologies, and lifecycle stages. While recent advances in data collection technologies have increased data availability, they have also amplified heterogeneity in data formats, ownership, and resolution, shifting the dominant challenge from data acquisition to data integration. Existing national and international flood defence datasets primarily focus on basic asset location and administrative attributes, while detailed technical information required for engineering analysis remains largely absent.

This paper presents a reproducible workflow for building large-scale technical datasets for earthwork levees using open and publicly available data. The workflow is structured around three core components that together form a technical representation of earthwork levees: (1) geometry, (2) construction, and (3) subsurface. Geometry is reconstructed from high-resolution digital elevation models by extracting kink lines that capture slope transitions along the levee body. Construction is derived from cone penetration test data by interpreting soil behaviour types and mapping material composition onto levee geometries. Subsurface characteristics are integrated by converting volumetric geological models into point-based representations suitable for geospatial analysis. Data preparation and processing are automated using GIS and Python-based workflows to support scalability and future updates.

Although demonstrated using national-scale datasets from the Netherlands, the workflow relies on data sources and processing steps that are widely available. By building technical levee components independently from fragmented organisational data structures, the proposed workflow provides a scalable foundation for improving data availability and consistency in flood defence asset management and supports more data-driven assessment, maintenance, and reinforcement decisions.

KEYWORDS: flood defence, asset management, geotechnics, digital transformation, data management

1 INTRODUCTION

Flood defence data are among the most important assets within flood defence asset management. Technical, performance, and condition data underpin decisions on *if*, *when*, and *how* flood defences require assessment, maintenance or reinforcement. This dependency is widely understood, and flood defence asset managers worldwide invest in digital transformation and data-driven decision making.

Flood defences are often located in public space, fulfil multiple societal functions, and are governed and financed by combinations of national, regional and local governments, and in some cases public-private partnerships. This multi-functional, multi-actor, and multi-financed configuration disperses asset management responsibilities across multiple organisations (den Heijer et al., 2023). Although flood safety as the shared goal, these organisations operate under different mandates, incentives, and financing structures. Consequently, flood defence data are produced by different actors for different purposes and are embedded in different organisational contexts, which limits their alignment in a single asset management logic. Organisational diversification therefore is a structural driver of data fragmentation.

In parallel, the flood defence sector has seen rapid advancements in data collection technologies, including satellite observations (Destefanis et al., 2025), unmanned aerial vehicles (Minh, 2025), and in situ sensor networks (Tao et al., 2024). These developments have led to a significant increase in heterogenous data formats, spatial and temporal resolutions, update frequencies, and ownership (Cohen et al., 2025; Towe et al., 2020). As a result, the dominant challenge has been shifting from data acquisition to data integration (Borowicc & Alves-Souza, 2024). Many technological initiatives remain pilot-based, with limited attention paid to integration across the full data lifecycle (Podt & Rijke, 2024). Technological diversification therefore is a structural driver of data fragmentation.

In this paper, data fragmentation is understood as the unintentional misalignment of flood defence data across organisational, technical, and lifecycle dimensions, resulting in incoherent technical asset representations. This fragmentation makes data less accessible, less consistent, and less accurate. Data fragmentation in flood defence management is widely recognised in practice but rarely explicitly described in literature. Empirical studies report fragmentation in different contexts. Vincke et al. (2019) describe how Flemish levee data are stored in separate, disconnected databases with limited GIS integration and a lack of agreed standards for new monitoring campaigns (Vincke et al., 2019). At a governance level, Mohanty et al. (2020) document how intergovernmental distrust and concerns over data reliability hinder the sharing of flood damage data in India (Mohanty et al., 2020). Similarly, Cohen et al. (2025) show how limited access to flood-related data in Colombia is linked to weak institutional data management practices (Cohen et al., 2025). This grasp of studies illustrate that data fragmentation is a recurring and multifaceted challenge in flood defence management.

Several international initiatives have sought to address data fragmentation through data harmonisation. In the AIMS Spatial Flood Defences project, the UK Environmental Agency runs a daily updated open dataset containing the centreline geometries and metadata of flood defence assets (EA, 2025). Comparable initiatives are the Dutch Basisbestand Primaire Waterkeringen, which provides information on asset location, safety standards, and assessment results (IHW, 2025), and the National Asset Database Wales, which harmonises data on levee type, location, ownership, and maintenance responsibility.

Despite these efforts, existing datasets typically focus on asset location and basic administrative attributes, while the technical characteristics remain largely absent. Although large-scale data harmonisation is feasible, many countries still lack a national, uniform, and publicly accessible technical flood defence dataset. This paper therefore presents a workflow to extract data on the geometry, construction, and subsurface characteristics of earthwork levees from open datasets, providing a technical asset representation that is largely independent of fragmented organisational data structures.

2 METHODS

This section presents a reproducible workflow to reconstruct technical characteristics of earthwork levees from open geospatial datasets. The workflow is structured around three core components that together form a technical representation of earthwork levees: geometry, construction, and subsurface. Although the workflow is demonstrated using national-scale datasets from the Netherlands, all processing steps are transferable to other contexts with comparable data availability, and comparable international datasets are indicated where relevant. The methodology combines geospatial processing of elevation data, cone penetration test data, and lithological subsurface data derived from datasets that are often open, public, and FAIR. Widely available spatial analysis toolboxes within GIS software are used to ensure replicability, with key tools and algorithms indicated in italics (e.g. *Polygon to Line*). Data preparation and processing steps are automated using Python scripts to support scalability and for future updates. Throughout the workflow, a distinction is made between vector data, represented as points, lines, and polygons, and raster data, represented as pixel-based surfaces.

2.1 Component 1: Geometry

Component 1 reconstructs the geometry of earthwork levees by extracting slope transitions from digital elevation models and representing these transitions as kink lines along the levee body. The geometry of a levee can be observed using a digital elevation model (DEM), a digital representation of the surface's topography. In geomatics, such surfaces are often represented as a 'raster'. DEMs are useful for querying specific elevation data, such as extracting cross-sections of a levee. DEMs only contain elevation data and lack zone-specific data, such as crest-width and the presence of berms or ramps. Digital identification of levee zones is important for precise geometry extraction and for enriching inspection data with zone context. For example, communicating that a crack is detected on an inner-berm rather than relying solely on coordinates.

A typical earthwork levee consists of seven zones: the foreland (1), outer-berm (2), crest (3), inner-berm (4), hinterland (5), seepage ditch (6), and the slopes connecting them (7). These zones can be categorised into flat and steep surfaces, where each zone boundary is defined by a kink line formed at a slope transition. That same typical earthwork levee subsequently has six unique kink lines: the outer toe (OT), outer berm (OB), outer crest (OC), inner crest (IC), inner berm (IB), and inner toe (IT). The outer and inner toes and crests each occur once, whereas berm kink lines may occur multiple times depending on the number of berms (e.g. OB_1, OB2, …, OBn).

For the identification of zones and kink lines, the highest resolution DEM available should be used. High-resolution DEMs are typically distributed as tiled raster datasets due to their size. For efficient processing, only tiles intersecting the levee footprint are selected and combined into a *mosaic* raster. To minimise storage requirements and computational load in subsequent analyses, each tile is *clipped* to the spatial extent of the levee prior to mosaicking. In the present implementation, this tile selection and clipping procedure is automated using a Python-script that takes a vector feature of the levee footprint as input, retrieves all intersecting DEM tiles from a selected public elevation database, and *clips* them into efficiently sized levee specific rasters. The input vector feature is an approximate levee footprint created from buffered flood defence centrelines, as described in Component 2.

DEMs often miss pixels due to water surfaces or dense features that the sensor could not penetrate. Before performing geometric extraction, these 'voids' should be corrected using an *elevation void fill* algorithm within a GIS environment. Kink lines can then be extracted by converting the DEM into a slope raster. Using a *Slope* algorithm, each elevation cell in the DEM is converted into a slope value that represents the rate of elevation change in degrees within its neighbourhood, typically a 3x3 cell window. The resulting slope raster can be classified into two categories: flat and steep surfaces. Levee slopes generally range between 1:3 to 1:5. While surfaces with slopes less steep than 1:5, or 11.3°, might not appear entirely flat, this threshold effectively identifies gentle transitions, such as ramps (Figure 1). Because optimal slope thresholds vary with local levee morphology, the classification should be locally

calibrated rather then globally defined. In other words, assign slope thresholds to levee trajectories rather than one slope threshold for the whole dataset.

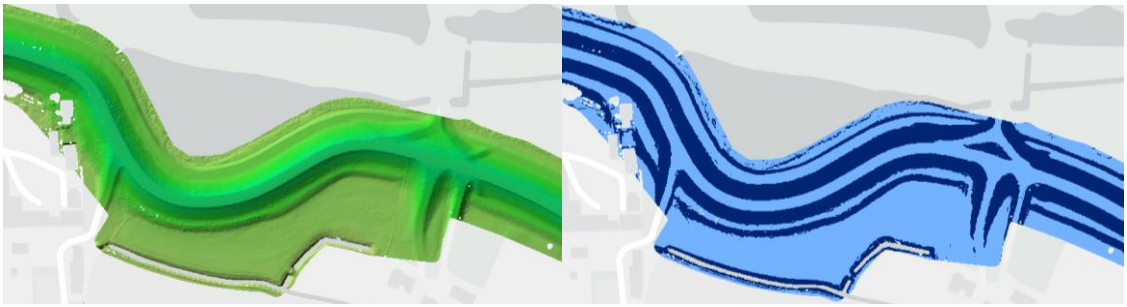


Figure 1 Digital elevation model of an earthwork levee visualised as hillshade (left) and corresponding classified slope raster distinguishing flat and steep surfaces for kink line extraction (right)

After classification of the slope raster, the slope transitions become visible but remain non-interactive. To convert these transitions into useable vector features, the slope raster must be *reclassified* into a binary raster storing integer values. Vector features can only be generated from integer-based raster data within GIS.

Although it might seem logical to extract line features directly from this binary raster, it is advised to first perform a *raster to polygon* conversion with a simplified polygon output. Simplification ensures that the polygons are smoothed rather than constrained by the grid structure of the raster cells, which would otherwise produce blocky edges. Polygon features allow for attribute selection and filtering based on area, unlike line features. To remove noise and minor artifacts that clutter kink lines, polygons smaller than 100m² are selected by attribute and *eliminated* by largest shared area. The results is a smoothed, less-cluttered polygon layer that presents kink-lines already quite well.

The next step in this workflow is to apply a polygon simplification algorithm to reduce feature complexity and eliminate sharp angularities introduced by the raster grid, which do not correspond to real-world geometry. For this purpose, the *bend simplification* (Wang & Müller, 1998) algorithm is particularly well suited, as it preserves geometric characteristics while removing redundant vertices and smoothing non-critical deviations. A simplification tolerance of 8 meters gives clean results. A higher tolerance risks cutting of ramp transitions. During simplification, make sure to resolve topological errors.

The kink lines derived from the DEM now reveal the levee geometry (Figure 3). While these kink lines correspond to features such as crests, berms and toes, the current workflow does not yet assign semantic labels to individual lines. Although humans can reason which line is which, an untrained computer cannot yet make this distinction. Moreover, access ramps cause kink lines associated with different zones to be geometrically connected, for example the crest being connected to the hinterland. This presents challenges for analyses that require computation of distinct features, such as determining crest widths.

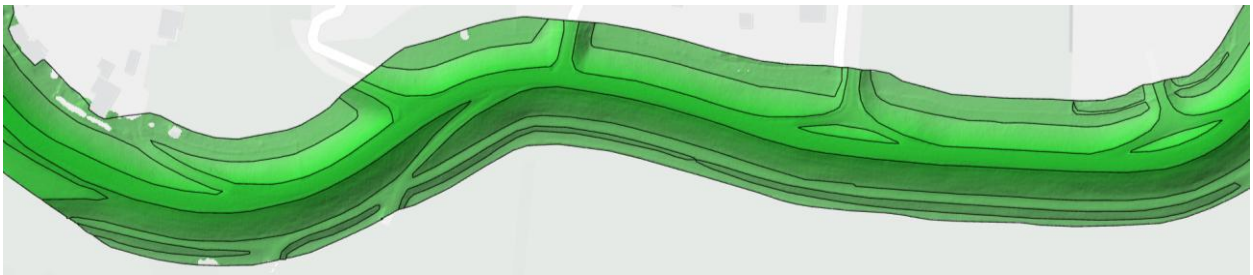
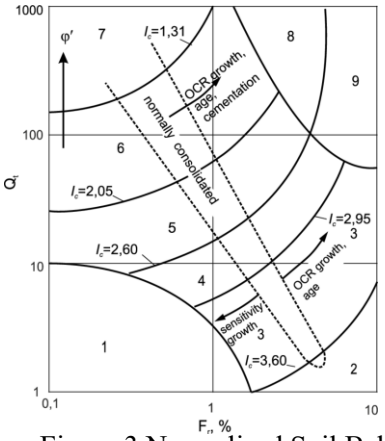


Figure 2 Kink lines extracted from a digital elevation model, representing slope transitions that define the geometry of an earthwork levee

2.2 Component 2: Construction

Component 2 reconstructs the construction of earthwork levees by interpreting material composition from cone penetration tests (CPTs) and mapping interpreted soil types onto levee geometries. The construction of earthwork levees is characterised primarily by its material composition and, where present, additional reinforcement structures such as sheet piles or geotextiles. Data on material composition are generally more extensive and systematically available than data on reinforcement elements, which are often site-specific and inconsistently represented across regional datasets.

Data on material composition are typically obtained from boreholes, which provide lithological descriptions, or CPTs, which record cone tip resistance, sleeve friction, and sometimes pore water pressure. CPTs are generally far more numerous than boreholes due to their practicality and cost-effectiveness and provide high vertical resolution measurements. However, CPTs do not directly indicate soil types and therefore require interpretation. In this workflow, soil types are interpreted using the normalised Soil Behaviour Type (SBTn) chart proposed by Robertson (1990), which provides a widely accepted baseline for translating CPT measurements into material classes. In addition, the authors' soil type interpretation allows for more classification ambiguity in the context of flood defence applications (Figure 3). Distinctions between closely related classes such as 'sand mixtures', 'clean and silty sands', and 'silt mixtures' are of limited practical relevance for large-scale levee characterisation and may introduce unnecessary complexity. It should be noted that the underlying CPT parameters remain unchanged. While artificial neural networks are increasingly applied to capture local soil nuances, these methods typically require extensive training data and site-specific calibration. The SBTn approach therefore offers a robust and reproducible baseline suitable for large-scale and transferable applications.



SBTn code	Robertson interpretation	Authors' interpretation	Hex code
-1	No-data	No-data	#E0E0E0
1	Sensitive fine grained	Sensitive clay	#D3956F
2	Clay organic	Peat	#8B4513
3	Clay & silty clay	Clay	#4B9B5E
4	Silt mixtures	Silt/loam	#A9DB7B
5	Sand mixtures	Clayey sand	#E4D449
6	Sand mixtures	Sand	#FFE600
7	Dense to gravelly sand	Gravelly sand	#D4B36D
8	Stiff sand to clay sand	Compact sand	#9E8E75
9	Stiff fine grained	Compact clay	#9E8E75

Figure 3 Normalised Soil Behaviour Type chart (Robertson, 1990) with adapted soil type interpretations for flood defence characterisation

CPT data are collected from national geological survey databases (e.g. BRO, DOV, NADAG, Jupiter GEUS), where they are stored in SQL-readable format. These databases are developed for interdisciplinary research purposes and therefore extend well beyond levee footprints. To associate CPTs with levee bodies, CPT locations are spatially *clipped* using buffered flood defence centrelines, which are available in the national basic flood defence datasets as mentioned in the Introduction. *Buffer* widths are manually selected based on observed footprint, with difference often between levee types (i.e. coastal, fluvial, urban). In the Dutch implementation, buffer widths ranging from approximately 20 to 100 m are applied, depending on levee type and in a satellite map observed footprint. River flood defences typically use buffer widths of around 40 m, while coastal and estuarine flood defences require wider buffers of approximately 80 m. In the current implementation, buffer distances are manually defined, but this buffer step will be replaced by the derived toe lines once available from Component 1.

CPT measurements are typically recorded at high vertical resolution (often at 2 cm intervals), with cone resistance and sleeve friction stored as depth-indexed records in SQL-readable databases. A Python-

based workflow processes these data by (1) computing derived CPT parameters, (2) assigning soil behaviour types using the normalised Soil Behaviour Type (SBTn) chart, and (3) filtering records that are outdated or lie below current levee geometry and therefore indicate CPT before reinforcement. CPT measurements located below the toe elevation derived from Component 1 can be excluded to distinguish levee construction materials from deeper subsurface stratigraphy. The workflow produces a classified, depth-resolved representation of levee material composition and generates scalable vector graphics (SVG) profiles of interpreted soil types for visual integration with the DEM cross section (Figure 4).

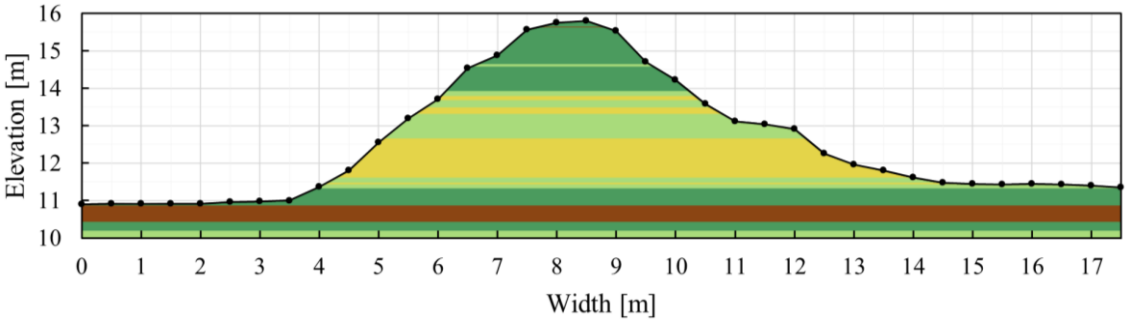


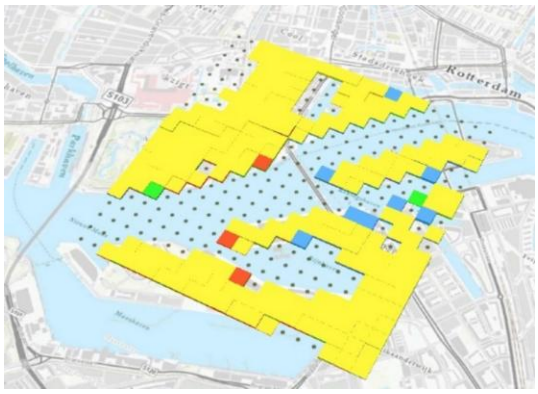
Figure 4 Levee cross-section with soil type interpretation referenced beneath the surface profile extracted from a digital elevation model

2.3 Component 3: Subsurface

Component 3 reconstructs subsurface characteristics beneath earthwork levees by translating volumetric lithological models into point datasets that are suitable for geospatial and geotechnical analysis. Subsurface soil composition is commonly represented in three-dimensional lithological models that integrate borehole and CPT data. These models are often constructed as voxel models (e.g. GeoTOP, DK Model) or volumetric grids (e.g. BGS LithoFrame, USGS Hydrogeological Framework). In many countries, such datasets are provided by national geological surveys and extend to depths of approximately 30 - 50 m below ground level. Although these volumetric datasets store lithological information for each voxel, most GIS environments offer limited support for voxel processing. For spatial integration with levee geometry and construction data from previous components, a volumetric-to-point workflow is therefore applied.

In this workflow, volumetric subsurface models are accessed in NetCDF format via OPeNDAP services are converted into point datasets by extracting the centroid coordinates and lithological attributes of individual voxels. This conversion is performed using a custom Python script (*Compress Voxels*), which compresses the three-dimensional voxel grid into depth-referenced point features. The workflow additionally supports mapping lithological classes from the voxel model to the soil type interpretations used in Component 2, for consistent soil type classes between construction and materialisation. The resulting point datasets (e.g. .csv or .xyz) can be imported into GIS as XY point data, assigned to the appropriate coordinate reference system, and subsequently queried, clipped, and spatially intersected with levee geometries from Component 1.

Figure 5 illustrates a sample of the resulting point dataset, including a visualisation where half of the points are rendered as voxel cells to reflect their volumetric origin (left). The associated attribute table (right) demonstrates the advantage of a point dataset approach: lithological information is stored as depth-referenced records with soil types organised in a single attribute column. This structure enables straightforward identification of aquifer depths, layer thicknesses, and stratigraphic transitions, which can be used in seepage and piping-related risk studies.



Point_ID: 48815 xy: 170100,435000 (centroid of voxel)	
Depth below surface [m]	**Soil type interpretation**
0	Peat
-0.5	Clay
-1.0	Clay
-1.5	Gravelly sand
-2.0	Sand
…	…
-50.0	Sand

Figure 5 Point dataset representation of the GeoTOP voxel model with selected points rendered as volumetric cells to illustrate voxel origin (left), and corresponding attribute table (right) showing depth-referenced soil types

3 CONCLUDING DISCUSSION

This paper addresses data fragmentation in flood defence asset management by demonstrating how technical characteristics of earthwork levees can be reconstructed at scale from open data. The proposed workflow reduces the impact of fragmented and dispersed data ownership and asset management structures. The workflow complements existing flood defence asset management with technical data that is currently absent in many national contexts.

A key strength of the approach lies in its modular structure. By separating geometry, construction, and subsurface into distinct but interoperable components, the workflow allows heterogeneous data sources to be integrated through shared spatial references rather than through harmonised data governance arrangements. Although not shown explicitly in a single figure, the geometry, construction, and subsurface components are spatially consistent and can be directly combined into integrated levee cross-sections where required. This makes the method particularly suited to contexts where flood defence data are distributed across multiple organisations and collected for different purposes. Although developed and demonstrated using Dutch datasets, digital elevation models, cone penetration tests, and geological models are widely available in many contexts, supporting transferability beyond this study.

At the same time, several limitations should be acknowledged. The geometry component extracts kink lines that capture slope transitions but does not yet assign semantic labels such as crest, berm, or toe. Similarly, construction is reconstructed from CPT-based soil interpretations that remain approximations due to material heterogeneity. These limitations reflects deliberate scope choices and point to opportunities for future refinement, including automated zone identification, improved footprint delineation, and advanced material interpretation that incorporates local nuances.

Overall, this study shows that flood defence data availability does not require complete institutional harmonisation. By using open data and reproducible geospatial workflows, it is possible to build large-scale technical flood defence datasets that are available to all. The proposed workflow is currently being further developed as a web-based application in the Netherlands (visit Dijkatlas.com or Dikeatlas.com), and both the methodology and underlying code are being refined to support application in other national contexts.

4 REFERENCES

Borowicc, S., & Alves-Souza, S. (2024). Heterogeneous Data Integration: A Literature Scope Review: *Proceedings of the 26th International Conference on Enterprise Information Systems*, 189–200. https://doi.org/10.5220/0012551000003690

Cohen, J., Mdee, A., Trigg, M., Singhal, S., Cooper, S., Alemu, A., Seifu, E., Lee, C., Bernhofen, M., Bhave, A., Carr, A., C T, D., Haile, A., Pencue-Fierro, L., Sa'adi, Z., Shukla, P., & Solano Correa, Y. (2025). A Politics of Global Datasets and Models in Flood Risk Management. *Water Alternatives*, *18*, 305–329.

den Heijer, F., Rijke, J., Bosch-Rekveldt, M., de Leeuw, A., & Barciela-Rial, M. (2023). Asset management of flood defences as a co-production—An analysis of cooperation in five situations in the Netherlands. *Journal of Flood Risk Management*, *16*(3), e12909. https://doi.org/10.1111/jfr3.12909

Destefanis, T., Guliyeva, S., Boccardo, P., Fissore, V., Destefanis, T., Guliyeva, S., Boccardo, P., & Fissore, V. (2025). Advancing Flood Detection and Mapping: A Review of Earth Observation Services, 3D Data Integration, and AI-Based Techniques. *Remote Sensing*, *17*(17). https://doi.org/10.3390/rs17172943

EA. (2025). *AIMS Spatial Flood Defences*. https://data.europa.eu/data/datasets/aims-spatial-flood-defences-inc-standardised-attributes?locale=en

IHW. (2025). *Nationale Basisbestanden Primaire Waterkeringen*. https://www.nationaalgeoregister.nl/geonetwork/srv/dut/catalog.search#/metadata/9bc22d59-427f-45c9-9f55-7dbe3985a73c

Minh, D. T. (2025). A comprehensive overview on UAV-based applications for flood management. *Measurement Science and Technology*, *36*(8), 086006. https://doi.org/10.1088/1361-6501/adf871

Mohanty, M. P., Mudgil, S., & Karmakar, S. (2020). Flood management in India: A focussed review on the current status and future challenges. *International Journal of Disaster Risk Reduction*, *49*, 101660. https://doi.org/10.1016/j.ijdrr.2020.101660

Robertson, P. K. (1990). Soil classification using the cone penetration test. *Canadian Geotechnical Journal*, *27*(1), 151–158. https://doi.org/10.1139/t90-014

Tao, Y., Tian, B., Adhikari, B. R., Zuo, Q., Luo, X., & Di, B. (2024). A Review of Cutting-Edge Sensor Technologies for Improved Flood Monitoring and Damage Assessment. *Sensors*, *24*(21), 7090. https://doi.org/10.3390/s24217090

Towe, R., Dean, G., Edwards, L., Nundloll, V., Blair, G., Lamb, R., Hankin, B., & Manson, S. (2020). Rethinking data-driven decision support in flood risk management for a big data age. *Journal of Flood Risk Management*, *13*. https://doi.org/10.1111/jfr3.12652

Vincke, L., Visser, K. P., Peeters, P., & Leus, B. (2019). Dike Data Management in Flanders. *Proceedings of the XVII European Conference on Soil Mechanics and Geotechnical Engineering*, *Geotechnical Engineering, foundation of the future*, 1387–1393. https://doi.org/10.32075/17ECSMGE-2019-0261

Wang, Z., & Müller, J.-C. (1998). Line Generalization Based on Analysis of Shape Characteristics. *Cartography and Geographic Information Systems*, *25*(1), 3–15. https://doi.org/10.1559/152304098782441750