Science of the Total Environment 1046 (2026) 181953

Contents lists available at ScienceDirect

Science of the Total Environment

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

Review

Intelligent sediment–groundwater digital twin: A systematic review, meta-analysis, and reference architecture for reliable metal pollution risk assessment

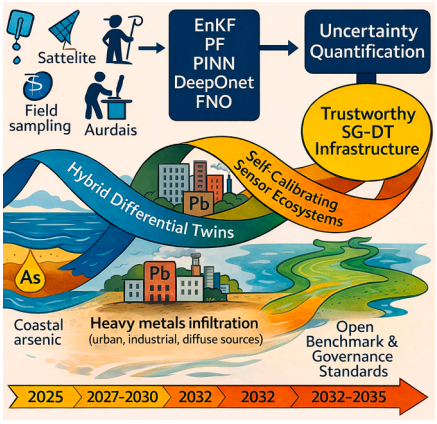
Saeid Pourmorad [a,*], Michele Morsilli [b], Luigi Lombardo [c], Fulvio Celico [d], Luca Antonio Dimuccio [a]

[a] University of Coimbra, Centre of Studies in Geography and Spatial Planning (CEGOT), Department of Geography and Tourism, FLUC, Coimbra, Portugal
[b] Department of Physics and Earth Sciences, University of Ferrara, Italy
[c] University of Twente, Faculty of Geo-Information Science and Earth Observation (ITC), PO Box 217, Enschede, AE 7500, Netherlands
[d] Department of Chemistry, Life Sciences and Environmental Sustainability - University of Parma, Italy

HIGHLIGHTS

- Presents a systematic review and quantitative meta-analysis of Sediment–-Groundwater Digital Twins (SG-DTs), synthesising evidence from 175 peer-reviewed studies published between 1997 and 2026.
- Defines a testable operational concept of the Digital Twin in subsurface systems (see Box 1), clearly differentiating it from simulators and emulators through feedback-enabled, data-assimilative coupling.
- Introduces an integrated Verification–Validation–Uncertainty Quantification (VVUQ) framework using probabilistic metrics (CRPS, PIT, Coverage) to ensure regulatory-grade reliability.
- Proposes a layered reference architecture with modular APIs, standardised I/O schemas, and implementation pseudocode for DTOps and MLOps pipelines.
- Releases the SED-GW-DT-REPORT v1.0 checklist and machine-readable JSON schema to improve metadata transparency, interoperability, and reproducibility.

GRAPHICAL ABSTRACT



* Corresponding author.
E-mail addresses: Omid2red@gmail.com (S. Pourmorad), luca@ci.uc.pt (L.A. Dimuccio).

https://doi.org/10.1016/j.scitotenv.2026.181953
Received 29 January 2026; Received in revised form 9 April 2026; Accepted 8 June 2026
Available online 26 June 2026
0048-9697/© 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

ARTICLE INFO

Keywords:
Sediment-groundwater digital twin (SG-DT)
Uncertainty quantification (UQ)
Data assimilation
Surrogate modelling
Environmental risk assessment
VVUQ framework

ABSTRACT

Sediment–groundwater interfaces regulate the mobilisation, transport, and attenuation of metal contaminants within highly heterogeneous and multi-scale subsurface systems. However, efforts to operationalise digital twins in this domain remain fragmented across modelling paradigms, disciplines, and scales of observation and prediction. This study provides a performance-centred, quantitatively harmonised meta-analysis of sediment–groundwater digital twin (SG-DT) implementations, explicitly addressing scale dependence, spatial heterogeneity, and process-level controls. A PRISMA 2020–compliant systematic review and meta-analysis of 175 studies (1997–2026; 27 countries) was conducted. SG-DT applications were stratified across a continuum from pore and laboratory scales to site, reach, and basin domains. Scaling effects were analysed via stratified synthesis and meta-regression, while heterogeneity was operationalised through facies-based zonation, parameter variability, and multi-source observational constraints. Predictive performance was quantified using a redesigned, baseline-consistent effect size (E), enabling harmonisation across heterogeneous metrics and validation protocols. Uncertainty was estimated using random-effects models (REML) with Hartung–Knapp adjustments, and heterogeneity was assessed via τ^2 and I^2 with scale-aware interpretation. Results indicate a robust positive pooled performance gain for surrogate-enabled SG-DTs relative to non-updating baselines under scale-consistent validation, while highlighting limitations related to data dependence and extrapolation. In a subset of comparable studies, SG-DTs achieved a pooled out-of-sample R^2 of 0.67 (95% CI: 0.62–0.72). Meta-regression identifies continuous data assimilation and explicit uncertainty quantification as key drivers of performance, whereas residual heterogeneity reflects unresolved scale mismatches and inconsistent representation of subsurface complexity. SG-DTs are defined as systems coupling a process-informed hydro(geo)chemical core capturing reaction kinetics and flow–transport–geochemical coupling, streaming observations, and an updating operator within a closed-loop framework. As secondary outputs, we provide a scale-aware reference architecture embedded in a continuous verification–validation–uncertainty quantification loop and the SED-GW-DT-REPORT v1.0 standard for reproducible, FAIR-aligned reporting. These findings establish an auditable, scale-consistent evidence base for advancing reliable digital twin development in sediment–groundwater systems.

1. Introduction

The contamination of sediment–groundwater interfaces by toxic metals represents a major and growing global environmental challenge. This problem is driven by accelerating industrialisation, urban expansion, agricultural intensification, and climate-induced hydrological variability (Osae et al., 2025; Afahnwie et al., 2026; Akoteyon and Olatunbosun, 2026). These pressures perturb subsurface systems in strongly nonlinear ways, generating transient redox fronts, evolving biogeochemical gradients, and heterogeneous permeability structures that collectively control contaminant mobility and attenuation (Postma et al., 2007; Jørgensen et al., 2012; Middelburg, 2019). Consequently, predicting contaminant fate under dynamic forcing remains a persistent challenge in subsurface environmental science (Majee et al., 2026; Jiang et al., 2026). A key control on contaminant behaviour in these systems is sedimentary architecture. The spatial organisation of facies both vertically and laterally governs permeability, reactive surface availability, and diagenetic pathways, thereby exerting first-order control on metal retention and release (Akoteyon and Olatunbosun, 2026; Wang et al., 2026; Bennett and Canfield, 2020; Berg, 2011). Distinct depositional environments, including deltaic systems, alluvial floodplains, carbonate platforms, and organic-rich lacustrine deposits, host characteristic mineral assemblages and redox zonations (e.g., Fe—Mn oxyhydroxides, clays, carbonates, and organic matter) that strongly regulate metal speciation and mobility (Liu et al., 2026; Berkani et al., 2026). However, this intrinsic heterogeneity is often oversimplified in conventional modelling frameworks, leading to systematic underestimation of plume migration and delayed contaminant release under transient hydrological perturbations (Steefel et al., 2015; Fleckenstein et al., 2010). Heavy metals such as arsenic (As), lead (Pb), cadmium (Cd), and mercury (Hg) are particularly persistent in sediment–groundwater systems due to strong sorption, low biodegradability, and redox-sensitive remobilisation (Ghafari and Samaei, 2025; Zhao et al., 2025). Importantly, their mobility is governed not solely by equilibrium partitioning but by the dynamic, rate-limited coupling among hydraulic transport, interphase mass transfer, and geochemical reactions. In practice, adsorption–desorption, dissolution–precipitation, and redox transformations operate on timescales that differ from advective and diffusive transport, producing kinetic lag, local disequilibrium, and transient reaction fronts. These effects are amplified in heterogeneous sediments by preferential flow pathways, permeability anisotropy, and spatial variability in reactive phases, making process coupling a dominant control on system behaviour (Postma et al., 2007; Steefel et al., 2015; Fleckenstein et al., 2010). Mineralogy and organic matter, therefore, function as active regulators of contaminant dynamics. Fe—Mn oxyhydroxides, clay minerals, carbonates, and organic-rich domains control sorption capacity, buffering behaviour, and redox conditions, influencing both the magnitude and timing of contaminant release (Bennett and Canfield, 2020; Liu et al., 2026; Berkani et al., 2026). In parallel, microbially mediated processes, including Fe and Mn reduction, sulfide formation, and organic matter degradation, modify mineral stability and reactive surface properties, generating feedbacks between transport and geochemistry that are essential for physically consistent prediction (Jørgensen et al., 2012; Middelburg, 2019; Hansel, 2015). These interacting processes collectively reflect the classical hydrogeochemical framework governing contaminant mobility in subsurface systems, in which redox zonation, sorption equilibria, mineral stability, and microbially mediated reactions dynamically co-evolve across spatial and temporal scales. External hydrological disturbances such as monsoon recharge, flooding, tidal forcing, and storm surges can rapidly alter geochemical conditions and trigger episodic contaminant release, which is often not captured by static or weakly coupled models (Park and You, 2023; Miao et al., 2019). Despite substantial advances in hydrogeochemical modelling, many existing frameworks remain insufficiently coupled and poorly suited to non-stationary conditions. Flow, transport, and geochemistry are often treated as loosely connected components, limiting predictive reliability under transient forcing (Wu et al., 2023; Sheng et al., 2023; Yang et al., 2019; Zou et al., 2025). Fully coupled reactive transport models provide high physical fidelity but are typically computationally prohibitive for operational use. At the same time, uncertainty is frequently treated inconsistently or deterministically, reducing interpretability and limiting decision support (Jones et al., 2020; Nearing et al., 2018; Yang et al., 2025; Manocha et al., 2025). This mismatch highlights a key methodological gap: the lack of systematically evaluated quantitatively comparable evidence on modelling performance across heterogeneous sediment–groundwater systems.

Recent advances in digitalisation, data assimilation, and hybrid modelling have stimulated growing interest in Digital Twin (DT) systems within environmental sciences (Kolditz et al., 2026; Uma Maheswara Rao et al., 2026). These systems aim to maintain continuous synchronisation between computational models and their physical counterparts by integrating mechanistic models, real-time observations, surrogate representations, and probabilistic uncertainty quantification (Zhao et al., 2025; Peng, 2025). However, their scientific value in sediment–groundwater systems depends critically on preserving the governing couplings among flow, transport, and hydrogeochemical processes as conditions evolve, rather than merely improving computational efficiency. At the same time, the rapid expansion of DT applications has introduced significant terminological ambiguity, particularly in distinguishing Digital Twins from simulators and emulators in data-rich yet uncertain environments (Hou et al., 2026; Kwon et al., 2025). To address this issue, this study adopts a formal, criteria-based classification of Digital Twins, simulators, and emulators (Box 1; Table 1), ensuring terminological clarity and reproducibility. In parallel, advances in surrogate modelling, including Gaussian Process Regression, Random Forests, LSTM networks, Physics-Informed Neural Networks, and operator-learning approaches such as DeepONet, have significantly expanded the feasibility of high-dimensional subsurface simulations (Samaei and Ghafari, 2025; Cherlinka et al., 2025; Tanazefti et al., 2025; Abdullah-Al Mamun et al., 2026). When coupled with data assimilation methods such as the Ensemble Kalman Filter (EnKF), four-dimensional variation (4D-Var) techniques, and particle filters, these approaches enable dynamically updated and uncertainty-aware predictive systems. However, their comparative performance and methodological maturity remain insufficiently quantified. Accordingly, the primary objective of this study is to provide a rigorous, quantitative evaluation of modelling performance and methodological maturity in sediment–groundwater Digital Twin (SG-DT) systems, based on a systematically curated dataset. The analysis prioritises performance-based evidence, with particular emphasis on the incremental predictive value of Digital Twin configurations relative to baseline modelling approaches.

1.1. Terminology and definitions (standardised framework)

To ensure conceptual clarity, internal consistency, and methodological reproducibility, all technical terminology used in this study is defined explicitly and applied consistently within a unified classification framework, in line with prior efforts to formalize Digital Twin concepts and distinguish them from adjacent modelling paradigms (Tao et al., 2019; Jones et al., 2020; Rasheed et al., 2020; Kolditz et al., 2026). In this study, a Digital Twin (DT) is defined as a dynamically coupled cyber–physical system that satisfies five necessary and jointly sufficient functional conditions: (i) continuous or quasi-continuous data assimilation, (ii) bidirectional state–decision feedback, (iii) dynamic synchronisation with the physical system, (iv) continuous model-health monitoring, and (v) dynamic uncertainty quantification. This definition is consistent with the broader Digital Twin literature, which emphasises persistent coupling, adaptive updating, and operational feedback as the distinguishing features of a true twin rather than a static or periodically calibrated model (Tao et al., 2019; Jones et al., 2020; Rasheed et al., 2020; Peng, 2025). A Sediment–Groundwater Digital Twin (SG-DT) refers specifically to a DT designed to simulate, monitor, and manage coupled sediment–groundwater systems in which groundwater flow, contaminant transport, and hydrogeochemical interactions evolve under transient environmental forcing. Within this study, SG-DTs are intended to represent systems in which sedimentary heterogeneity, reactive transport, and uncertainty-aware updating must be treated as integral rather than peripheral components of prediction and decision support (Koltermann and Gorelick, 1996; Aullón Alcaine et al., 2020; Kim and Bartos, 2024; Kolditz et al., 2026). A Hybrid Differential Twin (HDT) is introduced here as a subclass of SG-DT in which the governing system behaviour is represented through a coupled formulation that combines differential-equation-based process descriptions with data-driven surrogate components. In such systems, key physical operators governing flow, transport, or reaction remain explicitly represented, while selected unresolved, nonlinear, or computationally intensive components are approximated through learned surrogate relationships. This formulation is grounded in recent advances in hybrid hydrogeological modelling, physics-informed learning, and surrogate-assisted prediction, particularly where physically constrained neural architectures and data assimilation are used together to preserve interpretability while improving computational tractability (Zhu and Zabaras, 2018; He et al., 2020; Chen et al., 2021; Sprocati and Rolle, 2021; Wang et al., 2022; Panahi et al., 2025; Dai et al., 2025). Accordingly, HDTs are distinguished from purely mechanistic DTs, which rely exclusively on process-based solvers, and from purely data-driven DTs, which rely predominantly on learned approximations that do not preserve explicit differential structure. A simulator is defined as a forward-propagating numerical model that reproduces or predicts system behaviour under prescribed initial and boundary conditions, but operates independently of continuous data assimilation and does not include adaptive feedback mechanisms linking model states to decision processes. In contrast, an emulator or surrogate model is a statistical or machine-learning-based approximation of a simulator, developed to reproduce input–output behaviour at reduced computational cost. In this study, the term surrogate model is used consistently to refer to data-driven approximations, including machine-learning and deep-learning models, whether used for prediction, inversion, optimisation, or uncertainty propagation (Zhu and Zabaras, 2018; Mo et al., 2019; Song et al., 2019; Bao et al., 2020; Yu et al., 2021). A surrogate model, however, is not considered a Digital Twin unless it is embedded within a closed-loop architecture that includes live updating, synchronisation, and feedback-enabled uncertainty handling (Rasheed et al., 2020; Kim and Bartos, 2024). Uncertainty Quantification (UQ) is defined as the systematic characterisation, propagation, and updating of uncertainty in model inputs, parameters, states, and predictions using formal probabilistic methods. In the context of SG-DTs, UQ is treated as a dynamic operational layer rather than a static post hoc diagnostic, consistent with recent work in probabilistic surrogate modelling, data assimilation, and uncertainty-aware digital twins (Zhu and Zabaras, 2018; Mo et al., 2019; Bao et al., 2020; Mucke et al., 2023; Thelen et al., 2022; Panahi et al., 2025). Finally, effect size (E) is adopted as the standardised metric for quantifying improvements in predictive performance across studies. Within the meta-analytic framework developed in Section 7, effect size provides a harmonised basis for comparing model performance across heterogeneous validation metrics, modelling architectures, and reporting conventions, thereby supporting reproducible cross-study benchmarking.

1.2. Study objectives and contributions

This study addresses the identified methodological gap through a structured and explicitly hierarchical contribution framework that distinguishes core scientific contributions from supporting conceptual developments. Primary contribution (core scientific focus): The central contribution of this work is a quantitative meta-analysis of modelling performance and methodological maturity in sediment–groundwater Digital Twin (SG-DT) systems, based on a systematically curated dataset of 175 studies. Within the scope of this study, SG-DT implementations include, but are not limited to, hybrid differential twins (HDTs), as defined in Section 1.1, which integrate differential-equation-based process representations with data-driven surrogate components within a closed-loop Digital Twin framework. Using a harmonised, dimensionless effect-size (E) framework, we quantify the improvement in predictive performance of SG-DT systems relative to their corresponding baseline configurations, while explicitly accounting for differences in model structure, data assimilation strategies, uncertainty treatment, and

Box 1
Formalising the digital twin in subsurface systems: A rigorous distinction from simulators and emulators.

The terms *Simulator*, *Emulator*, and *Digital Twin (DT)* are frequently used interchangeably in subsurface environmental modelling and in the broader digital-systems literature, despite referring to computational paradigms with fundamentally different functional roles, levels of coupling to the physical system, and capacities for adaptive updating (Tao et al., 2019; Jones et al., 2020; Rasheed et al., 2020; Kolditz et al., 2026). This ambiguity has contributed to inconsistent terminology, reduced comparability across studies, and, in some cases, conceptual inflation of the Digital Twin paradigm in hydro-environmental applications. To address this limitation, we establish here a formal, operational, and falsifiable definition of a Digital Twin for subsurface systems, grounded in the systematic synthesis of the 175 studies analysed in this review and operationalised through the measurable criteria summarised in Table B1. Throughout this study, the following definitions are adopted as canonical and are applied consistently across all sections, including methodological development, meta-analysis, and framework design. In the present review, and consistently throughout the manuscript, a Digital Twin is not defined by model complexity, algorithmic novelty, or the mere use of machine learning, surrogate modelling, or sensor data. Rather, it is defined by the functional integration of five core capabilities: data assimilation, decision feedback, physical-state synchronisation, model-health monitoring, and dynamic uncertainty quantification. Accordingly, a computational system is classified as a Digital Twin if and only if it simultaneously satisfies five necessary and jointly sufficient conditions, expressed as:

$$DT \Leftrightarrow (C1 \wedge C2 \wedge C3 \wedge C4 \wedge C5)$$

where C1 denotes continuous or quasi-continuous ingestion and assimilation of temporally evolving observational data; C2 denotes bidirectional state–decision feedback, such that model outputs inform interventions or management actions and the consequences of those actions are subsequently reincorporated into the computational system; C3 denotes dynamic synchronisation between the digital representation and the physical system, with model states updated to reduce divergence from observed conditions; C4 denotes continuous model-health monitoring through explicit quantitative performance evaluation, enabling detection of drift, degradation, or structural inadequacy; and C5 denotes dynamic uncertainty quantification, whereby uncertainty in states, parameters, and predictions is updated as new information becomes available rather than treated as static or post hoc. Only systems satisfying all five conditions simultaneously qualify as Digital Twins in the strict sense adopted here. This formulation enables a non-ambiguous distinction between three related but non-equivalent paradigms. A Simulator is a forward-propagating numerical model, typically physics-based, that is executed to reproduce or predict system behaviour under specified initial and boundary conditions, but remains operationally decoupled from live observational updating and decision feedback. In other words, simulators do not satisfy C1 and C2 during operation, even if calibrated offline. An Emulator, by contrast, is a reduced-order, statistical, or machine-learning-based surrogate trained to approximate the input–output behaviour of a simulator or observationally derived process at lower computational cost. Emulators may support efficient prediction, optimisation, or uncertainty approximation and, in some cases, be periodically updated; however, unless they participate in continuous synchronisation, feedback-enabled updating, and dynamic uncertainty propagation, they do not meet the full requirements of a Digital Twin. In this sense, an emulator may be a component within a Digital Twin architecture, but it is not, by itself, a Digital Twin. Within this classification, it is important to distinguish between core modelling paradigms (simulators, emulators, and Digital Twins) and implementation-level subclasses that arise within Digital Twin architectures. In this study, the Hybrid Differential Twin (HDT), introduced in Section 1.1, is not treated as a separate paradigm parallel to simulators, emulators, or Digital Twins. Rather, it is defined as a subclass of Sediment–Groundwater Digital Twins (SG-DTs), characterised by the integration of differential-equation-based process representations with data-driven surrogate components within a closed-loop, data-assimilative framework. HDTs therefore inherit all defining conditions of Digital Twins (C1–C5) and are distinguished not by their classification status, but by their internal computational architecture. This clarification ensures that the taxonomy remains internally consistent: simulators and emulators are functionally distinct paradigms, whereas HDTs represent a specific implementation strategy within the broader class of Digital Twins. The critical distinction, therefore, lies not in whether a model is mechanistic or data-driven, but in whether it functions as a continuously coupled, self-updating, uncertainty-aware system embedded in an operational loop with the physical environment. This point is particularly important in subsurface applications, where the presence of a surrogate model, a dashboard, or intermittent data integration can create the appearance of a Digital Twin without satisfying its defining operational requirements. For example, a calibrated groundwater flow or reactive transport model used for scenario analysis remains a simulator; a machine-learning surrogate trained to emulate that model remains an emulator; whereas a system that continuously ingests monitoring data, updates internal states and uncertainties, evaluates predictive reliability, and informs adaptive management decisions in response to evolving field conditions qualifies as a Digital Twin. The operational manifestation of these distinctions is summarised in Table B1, while their conceptual relationships and partial overlaps are illustrated in Fig. B1. Several boundary conditions follow directly from this definition. First, the presence of observational data alone does not suffice for Digital Twin classification if those data are not assimilated into an updating loop. Second, periodic recalibration or offline retraining does not satisfy the requirement for continuous model health monitoring and synchronisation unless it is explicitly integrated into system operations. Third, precomputed or static uncertainty bounds do not constitute dynamic uncertainty quantification if they are not revised in response to incoming observations and evolving system states. Finally, model complexity is neither necessary nor sufficient for Digital Twin status: a highly sophisticated model without live synchronisation and feedback is not a Digital Twin, whereas a more parsimonious system may qualify if it satisfies all five operational conditions. By formalising the Digital Twin in this way, the present review provides a reproducible classification framework that prevents overgeneralization, supports consistent terminology, and improves methodological transparency across sediment–groundwater studies. This definition is intended not merely as a conceptual clarification, but as an operational standard for determining when the term Digital Twin is warranted in subsurface environmental modelling.

Table B1
Operational classification of simulators, emulators, and digital twins based on five necessary and sufficient functional criteria (C1–C5). Only systems satisfying all five criteria simultaneously qualify as Digital Twins.

Feature	Simulator	Emulator	Digital twin
Live data ingestion	×	×	✓
Decision-feedback loop	×	×	✓
Data synchronisation	×	Partial	✓
Model-health monitoring	×	×	✓
Dynamic UQ (real-time calibration)	×	Partial	✓

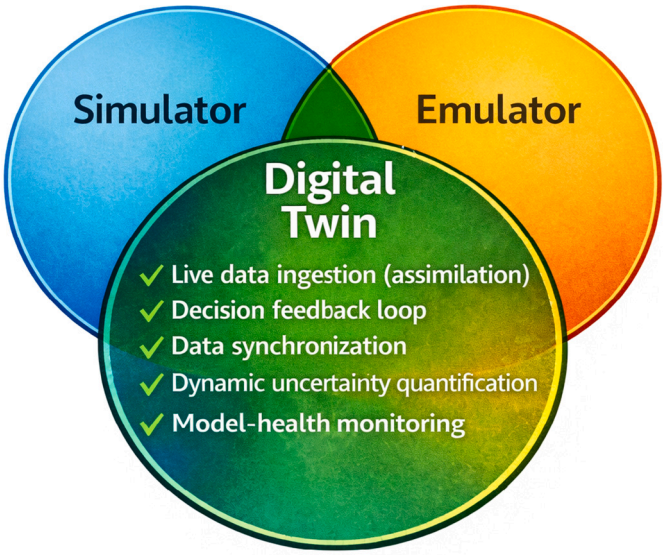


Fig. B1. Conceptual overlap among Simulator, Emulator, and Digital Twin paradigms. A three-set Venn diagram illustrating shared and unique capabilities. Only the Digital Twin domain integrates all five criteria: live data ingestion, synchronisation, model-health monitoring, dynamic UQ, and decision feedback, defining its operational distinction from conventional simulators and emulators (based on 175 reviewed studies). A color-safe, vectorised version is provided in the Supplementary figures.

Table 1
Conceptual capability-level comparison between simulators, emulators, and digital twins. The table contrasts the presence of functional capabilities across modelling paradigms. "Partial" indicates episodic, offline, or non-continuous implementations that do not constitute persistent, closed-loop digital twin operation. Operational and lifecycle requirements for fully realised digital twins are assessed separately.

Feature	Simulator	Emulator	Digital Twin
Episodic or real-time data integration	No	Partial	Yes
Feedback loop for decision making	No	No	Yes
Data synchronisation	No	Partial	Full synchronisation with the physical system
Model health monitoring	No	No	Yes (with real-time updates)
Dynamic uncertainty quantification (UQ)	No	Partial	Yes (Integrated UQ with real-time feedback)
Adaptability to new data	No	Limited	Continuous learning and adaptation
Typical use case	Static scenarios or testing	Approximate or simplified simulations	Real-world operational monitoring and prediction
Complexity	Low	Medium	High
Computational cost	Low	Medium	High (due to real-time data integration)

the degree to which key hydrogeochemical controls are operationally represented. This provides a robust, cross-study, and statistically consistent evaluation of how different modelling paradigms perform under heterogeneous environmental and methodological conditions. Secondary contributions (supporting and contextualising): To support the interpretation, reproducibility, and operationalisation of the meta-analytic findings, the study further provides: (i) a formal and operational definition of Digital Twins, simulators, and emulators based on necessary and sufficient functional criteria (Box 1), ensuring terminological clarity and classification rigour, within which implementation-level subclasses such as hybrid differential twins are consistently positioned as part of the SG-DT category rather than as independent paradigms; (ii) a systematic literature synthesis that contextualizes the evolution of SG-DT methodologies and identifies key gaps in uncertainty quantification, validation practices, data integration, and process-level hydrogeochemical representation; (iii) a conceptual system-level architecture that synthesizes common structural patterns across SG-DT implementations, serving as an interpretive and contextual framework rather than a primary contribution (Fig. 1); (iv) SED-GW-DT-REPORT v1.0, a lifecycle-oriented and process-driven reporting and metadata framework designed to enhance reproducibility, auditability, and cross-study comparability. This framework is explicitly positioned as a supporting instrument that translates methodological insights from the meta-analysis into standardised and machine-actionable reporting practices. Together, this hierarchical structure ensures that conceptual definitions, architectural synthesis, and reporting standardisation are not treated as independent endpoints, but rather as supporting components that reinforce and operationalise the central quantitative

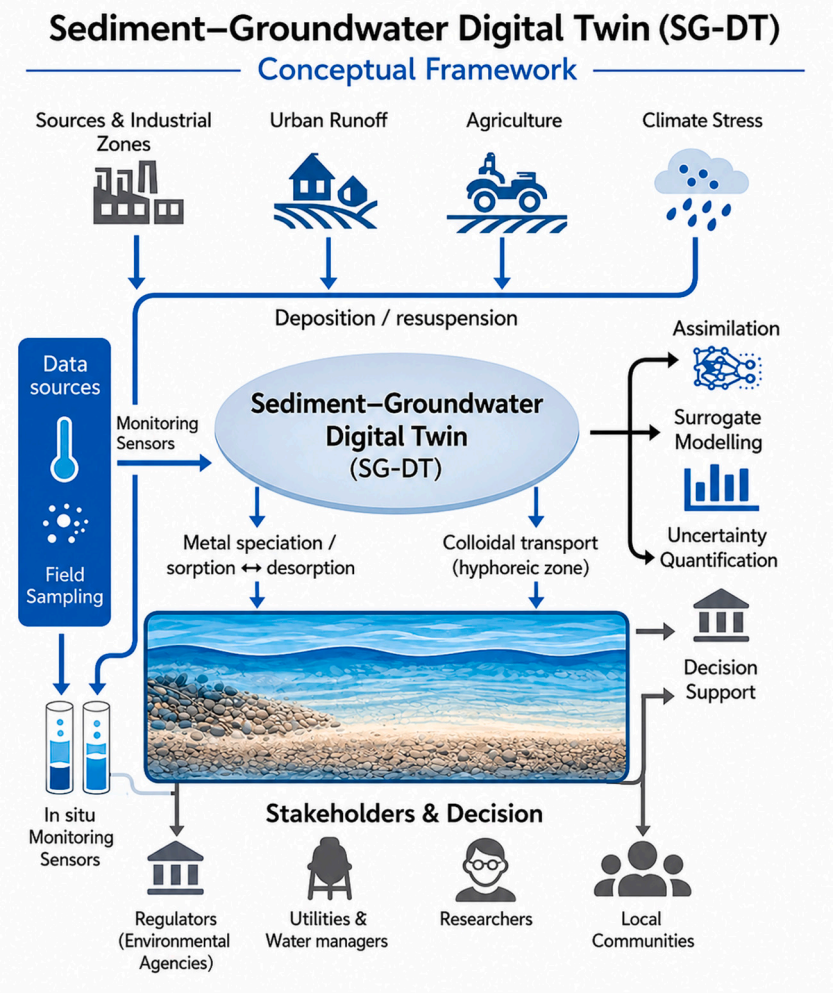


Fig. 1. Sediment–groundwater digital twin (SG-DT) system-level architecture. Conceptual diagram illustrating the SG-DT workflow. Multi-source data from sensors, field sampling, and external stressors are assimilated into computational and surrogate modules, integrated with uncertainty quantification (UQ), and used to support decisions for regulators, water managers, researchers, and communities.

evaluation of modelling performance. The overall structure of the manuscript reflects this prioritization: Sections 2–6 establish the systematic evidence base; Section 7 presents the core meta-analytic results; and subsequent sections interpret, contextualise, and operationalise these findings within a broader Digital Twin framework.

2. Methods research questions and scope

The evolution of Sediment–Groundwater Digital Twins (SG-DTs) represents a critical step toward bridging process-based hydrogeochemical science with data-driven intelligence and uncertainty-aware decision systems. Despite rapid growth in related research, the field remains conceptually fragmented, with inconsistent terminology, methodological overlap, and limited reproducibility across studies. To synthesise this expanding body of literature and establish an operational foundation for future work, this review adopts a structured analytical framework that integrates conceptual definition, methodological assessment, standardisation requirements, and forward-looking research priorities.

2.1. Conceptual definition of sediment–groundwater digital twins

The term Digital Twin is used inconsistently across environmental and engineering domains, spanning applications from static numerical surrogates to fully dynamic, data-synchronized systems (Wang et al., 2022). In the context of sediment–groundwater systems, this variability has contributed to conceptual ambiguity and, in some cases, overextended claims regarding model capability. To ensure conceptual clarity and consistency, this study adopts the formal, operational definition of Digital Twins established in Section 1.1 and Box 1. Accordingly, a system is classified as a Digital Twin only if it satisfies a set of five necessary and jointly sufficient functional conditions: continuous or quasi-continuous data assimilation, bidirectional state–decision feedback, dynamic synchronisation with the physical system, continuous model-health monitoring, and dynamic uncertainty quantification. Within this framework, a Sediment–Groundwater Digital Twin (SG-DT) system refers specifically to an implementation that applies these defining conditions to coupled sediment–groundwater processes, including flow dynamics, contaminant transport, and biogeochemical interactions. Crucially, these processes are not treated here as abstract computational components, but as physically grounded phenomena rooted in classical sedimentological and hydrogeochemical theory. Groundwater flow and transport behaviour are conditioned by sedimentary heterogeneity and facies architecture. At the same time, contaminant mobility is governed by well-established hydrogeochemical controls, including redox zonation, sorption–desorption equilibria, mineral dissolution–precipitation reactions, and microbially mediated transformations. By explicitly embedding these process-level controls within the Digital Twin definition, SG-DT systems are framed as Earth-system-informed models rather than purely data-driven or

engineering abstractions. This definition explicitly distinguishes SG-DT systems from conventional simulators and surrogate models that lack continuous data assimilation, adaptive feedback, or dynamic uncertainty representation. A simulator may reproduce flow or reactive transport processes under prescribed conditions, and a surrogate may approximate these processes efficiently; however, neither constitutes a Digital Twin unless it operates within a continuously updating, feedback-enabled, and uncertainty-aware framework. Because sediment–groundwater systems are inherently multi-scale, the conceptual definition adopted here is also explicitly scale-aware. SG-DT systems may operate across a hierarchical continuum extending from pore- and column-scale laboratory domains to facies- and site-scale field representations, to reach-scale hydrological segments and broader field-to-basin management contexts. Importantly, classification as an SG-DT does not depend on the absolute spatial scale of the application, but rather on whether the defining Digital Twin functions are maintained at that scale while preserving a physically interpretable linkage among observations, model states, and decision processes. In this context, scale is not treated as a secondary attribute but as a fundamental dimension that determines what is observed, how heterogeneity is represented, and how model updates are performed. As spatial scale increases, process aggregation, parameterisation, and uncertainty structure evolve accordingly, requiring SG-DT implementations to adapt both their data assimilation strategies and their representation of sedimentary and hydrogeochemical controls. Accordingly, SG-DT systems are interpreted in this review not as a single homogeneous class of models, but as a family of scale-dependent, process-informed systems whose functionality, data requirements, and uncertainty characteristics vary systematically with the level of process resolution and environmental complexity. Fig. 2 illustrates the conceptual convergence of sedimentary processes, groundwater hydrology, and Digital Twin systems. True SG-DT systems emerge at the intersection of these domains, where adaptive modelling, continuous data–model synchronisation, and feedback-informed decision processes are integrated into a unified, dynamically evolving framework that responds to changing environmental conditions across explicitly defined scales of operation.

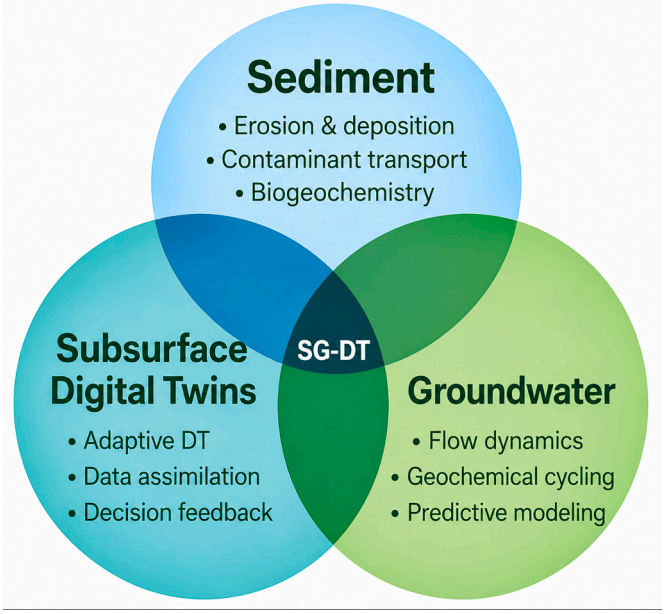


Fig. 2. Conceptual overlap among sediment, groundwater, and subsurface digital twin domains. A Venn diagram illustrates the interdisciplinary convergence of sedimentary processes, groundwater hydrology, and subsurface Digital Twin technologies. The overlap (SG-DT) represents the domain where adaptive modelling, data assimilation, and decision feedback integrate to support dynamic, uncertainty-aware environmental prediction.

2.2. Methodological maturity of surrogate modelling, data assimilation, and uncertainty quantification

A core objective of this review is to assess the current methodological maturity of SG-DT components. Across the 175 studies included in the systematic review, fewer than 12% implemented real-time or near-real-time data assimilation approaches (e.g., the Ensemble Kalman Filter and four-dimensional variational methods). Only about 8–10% reported quantitative uncertainty metrics such as Continuous Ranked Probability Score (CRPS), coverage, or Probability Integral Transform (PIT) (Yang et al., 2025; Xu and Zhang, 2025; Singh and Sharma, 2025). Surrogate models, including Gaussian Process Regression, Random Forests, DeepONet, and Long Short-Term Memory networks, have substantially reduced computational costs relative to full-physics simulations (Song et al., 2019; Bauer et al., 2021). However, most applications remain loosely coupled with observational data streams and rarely sustain live synchronisation with sensor networks. This reveals a persistent methodological gap between physics-informed and data-informed paradigms, underscoring the need for hybrid architectures that fuse mechanistic understanding with probabilistic learning to support reliable and scalable SG-DTs. To evaluate this maturity in a methodologically consistent way, the review framework explicitly incorporated the scale of operation as a formal axis for extraction and coding from the outset. Each study was classified according to the dominant scale at which its predictive task, observations, and updating logic were implemented, using a structured scale typology comprising: (i) pore/column–laboratory scale, (ii) facies/site scale, (iii) reach or hydrogeomorphic segment scale, and (iv) field to basin scale. This classification was used not merely for descriptive grouping but as a methodological moderator for interpreting differences in data availability, heterogeneity representation, assimilation feasibility, and uncertainty behaviour across SG-DT implementations. This scale-aware coding was necessary because the relationship between process resolution and model support strongly conditions methodological maturity in SG-DT systems. At laboratory or pore-related scales, high process fidelity is often achievable, but continuous synchronisation and decision feedback are rare. At site and reach scales, SG-DT implementations more frequently support assimilation and operational updating, but must also confront unresolved facies heterogeneity, parameter aggregation, and scale-transfer challenges. At field and basin scales, broader management relevance is gained at the cost of increased structural abstraction, greater dependence on effective parameterisation, and stronger sensitivity to observation sparsity. Consequently, maturity cannot be interpreted solely as the presence of advanced algorithms; it must be evaluated relative to the scale at which the system is expected to represent heterogeneity, assimilate data, and provide reliable predictions. For this reason, the present review evaluates methodological maturity as a combined function of: (i) surrogate sophistication, (ii) assimilation intensity, (iii) probabilistic uncertainty treatment, and (iv) the degree to which scale and spatial heterogeneity are explicitly represented within the model-update workflow. In other words, a computationally advanced system operating at a coarse scale without an explicit representation of heterogeneity may be less mature, from an SG-DT standpoint, than a more modest but scale-consistent implementation in which the data, model structure, and update logic are aligned. This interpretation provides the analytical basis for the subsequent comparative synthesis and helps explain why SG-DT performance varies systematically across different environmental and methodological settings.

2.3. Reference architecture and reporting standards for trustworthy SG-DTs

Reproducibility and cross-study comparability remain major limitations in SG-DT research, largely due to the absence of harmonised documentation and reporting practices. To address this gap, this review introduces SED-GW-DT-REPORT v1.0, a unified reporting protocol

supported by a machine-readable JSON schema that defines minimum metadata requirements for Digital Twin studies. Building on established principles of Verification, Validation, and Uncertainty Quantification (VVUQ) (Peng, 2025; Zhao et al., 2025), we develop a modular reference architecture that integrates data ingestion, surrogate modelling, assimilation, and UQ layers into a transparent and auditable modelling ecosystem. In the present study, this architecture is interpreted as inherently scale-aware rather than scale-neutral. Trustworthy SG-DTs require not only transparent documentation of model components, but also explicit reporting of the scale at which each component operates, the observational support available at that scale, and the strategy used to represent spatial heterogeneity. In practice, this includes reporting on whether heterogeneity is resolved through facies-based zonation, stochastic parameter fields, effective-medium assumptions, or multi-resolution data fusion. Without such scale-explicit documentation, apparently similar SG-DT systems may differ substantially in physical meaning, predictive support, and regulatory relevance. Accordingly, the reporting framework is intended to capture not only model provenance, data streams, and uncertainty metrics, but also the scale descriptors and heterogeneity assumptions that define the operational validity of an SG-DT implementation. This strengthens interpretability, facilitates benchmarking across datasets and institutions, and provides a foundation for trustworthy deployment of SG-DTs in operational and regulatory contexts. Under this framing, reproducibility is not limited to code or workflow transparency; it also depends on whether the relationship among scale, process representation, and data assimilation is sufficiently explicit to allow meaningful cross-study comparison.

2.4. Future research directions (2025–2035)

Looking ahead, the transition from isolated Sediment–Groundwater Digital Twin (SG-DT) prototypes to robust, operational systems will require coordinated advances across data infrastructure, process representation, and scale-consistent modelling. While recent developments have emphasised digital integration, artificial intelligence, and computational efficiency, future progress will depend fundamentally on embedding Digital Twin frameworks within the sedimentological and hydrogeochemical processes that govern contaminant behaviour in subsurface environments. Key priorities for the next decade include: (i) federated Digital Twin architectures that enable secure, distributed data exchange across institutional and geographic boundaries; (ii) self-calibrating sensor networks capable of adaptively refining model states and parameters through continuous feedback; (iii) benchmark datasets and open repositories to support reproducible model comparison; and (iv) environmental governance frameworks that explicitly incorporate probabilistic reliability metrics (e.g., coverage $\geq$90%) into regulatory decision-making (Minasny and McBratney, 2025; Boehlert et al., 2015). However, a central and currently underdeveloped requirement across all these directions is the development of scale-consistent SG-DT methodologies grounded in process-based understanding. In sediment–groundwater systems, predictive reliability depends on the ability to reconcile fine-scale sedimentary heterogeneity, such as facies architecture, permeability contrasts, and reactive phase distributions, with larger-scale hydrological behaviour and decision-making contexts. Future SG-DT frameworks must therefore explicitly link pore-to-field-scale process understanding with site-to-basin-scale prediction. This requires developing formal upscaling and downscaling strategies that preserve key sedimentological and hydrogeochemical controls. Approaches are needed to translate small-scale variability in grain-size distribution, mineral composition, and organic matter content into effective parameters governing flow, transport, and reaction at coarser scales. Similarly, multi-resolution state representations must be designed to capture interactions among preferential flow pathways, diffusion-limited domains, and reactive interfaces without excessive loss of physical interpretability. Equally important is the advancement of scale-aware data assimilation frameworks capable of integrating observations that reflect different process domains. Hydraulic measurements, geochemical indicators, and biogeochemical proxies often sample distinct aspects of system behaviour, including redox zonation, sorption–desorption dynamics, mineral stability, and microbially mediated reactions. Future SG-DTs must therefore move beyond purely hydraulic calibration and incorporate multi-source observations that constrain both transport and reaction processes in a dynamically consistent manner.

In parallel, there is a critical need for benchmark studies that explicitly report process representation and scale context, rather than focusing solely on predictive accuracy. Future studies should systematically document (i) the scale of implementation, (ii) the strategy used to represent sedimentary heterogeneity, and (iii) the degree of aggregation applied when translating measured processes into digital state variables. Such reporting is essential for distinguishing between models that achieve apparent accuracy through calibration and those that preserve physically meaningful process coupling. Within this review, scale was therefore treated not only as a classification axis but as a central determinant of methodological maturity. The analysed studies were evaluated not only in terms of Digital Twin functionality, but also with respect to the scale at which monitoring, modelling, and updating were implemented. This provides a foundation for more rigorous cross-study comparison of predictive performance and supports the transition from isolated proof-of-concept models toward interoperable, scale-explicit Digital Twin systems. Finally, to maintain methodological rigour and ensure analytical consistency, the inclusion and exclusion criteria that define the scope of this systematic review are summarised in Table 2. These criteria ensure that only studies directly relevant to coupled sediment–groundwater processes and Digital Twin-based uncertainty frameworks are included. Looking forward, such criteria should be further expanded to require explicit reporting of scale, representation of heterogeneity, and process-level assumptions, as these factors are now recognised as primary controls on model reliability, transferability, and decision relevance.

3. Systematic review methodology and workflow (PRISMA–RoB–meta-analysis)

A systematic review was conducted in full accordance with the PRISMA 2020 guidelines to ensure transparency, reproducibility, and methodological rigour throughout the identification, screening, retrieval, and synthesis of studies. The overall workflow integrates PRISMA-based study selection with a structured risk-of-bias (RoB) assessment and subsequent quantitative synthesis, as summarised in Fig. 3. Comprehensive searches were conducted in Scopus, Web of Science, and Google Scholar, targeting peer-reviewed literature and relevant grey literature on Sediment–Groundwater Digital Twin (SG-DT) systems. The search strategy combined controlled vocabulary and free-text keywords covering digital twins, sediment–groundwater interactions, uncertainty quantification, real-time data assimilation, and decision-support mechanisms. The search identified 1238 records in total. Following automated and manual deduplication, 576 duplicates were removed, leaving 662 unique records for title and abstract screening. Title and abstract screening were conducted using predefined inclusion and exclusion criteria to identify studies that explicitly address Digital Twin architectures in sediment–groundwater systems or closely related hydro-environmental applications. Following title and abstract screening of the 662 unique records, 410 records were excluded due to domain irrelevance (n = 185), reliance on static or non-Digital Twin modelling frameworks (n = 150), or insufficient qualitative or methodologically robust data (n = 75), consistent with Fig. 3. Consequently, 252 reports met the eligibility criteria and were sought for full-text retrieval. Of these, 77 reports were not retrieved due to inaccessibility, language restrictions, or insufficient reporting, as indicated in Fig. 3. Consequently, 175 full-text articles were included in the review. Methodological quality and risk of bias were evaluated using a structured,

Table 2
Scope of the systematic review: In-scope and out-of-scope study types. Summary of inclusion and exclusion criteria defining the boundaries of this systematic review. The table identifies studies relevant to sediment–groundwater digital twins (SG-DTs) that integrate data assimilation, surrogate modelling, and uncertainty quantification (UQ) within coupled sediment–groundwater systems.

Category	In-scope studies	Out-of-scope studies
System type	Coupled sediment–groundwater systems, including contaminant transport and geochemical feedback, as well as adjacent water-system digital twins, were explicitly reviewed as transferable methodological foundations.	Surface water–only or atmospheric models
Methodological focus	Integration of data assimilation, surrogate modelling, or UQ frameworks	Deterministic or purely empirical models lacking adaptive components
Temporal dynamics	Real-time or iterative updating mechanisms	Single static simulations without data coupling
Data integration	Live sensor or field-sampling data streams	Models relying solely on synthetic or static datasets
Validation and UQ	Studies reporting quantitative UQ metrics (CRPS, PIT, coverage, calibration slope)	Studies lacking explicit or quantitative uncertainty treatment
Relevance to metal pollution	Focus on As, Pb, Cd, Hg, Zn, or similar metal contaminants	Unrelated to heavy metal transport or risk assessment

decision-based RoB rubric specifically tailored to Digital Twin and hydro-environmental modelling studies. Study selection, RoB assessment, and metadata extraction were conducted independently by two reviewers. Inter-reviewer reliability was assessed via double-coding of a subset of studies, yielding Cohen's $\kappa = 0.87$, indicating strong agreement. Detailed evaluation criteria, extraction templates, and RoB decision rules are provided in the Supplementary materials.

3.1. Search strategy

A comprehensive multi-database search was conducted in Scopus, Web of Science, and Google Scholar, covering the period from January 1997 to January 2026. The search strategy was designed using controlled vocabulary and advanced query structures, including Boolean operators, truncation, and database-specific field tags. Four main thematic components were incorporated: (i) Digital Twin–related concepts (e.g., Digital Twin, Cyber-Physical System, Surrogate Model), (ii) subsurface system descriptors (e.g., groundwater, aquifer, subsurface), and (iii–iv) a combined process-oriented block linking contamination and modelling approaches, formulated as (Sediment* OR metal pollution OR heavy metal* OR As OR Pb OR Cd OR Hg) AND (uncertainty quantification OR data assimilation OR model validation). Search syntax was adapted to each database (e.g., TITLE-ABS-KEY in Scopus) to ensure optimal retrieval. In addition to peer-reviewed literature, grey literature and conference proceedings with identifiable DOIs were included (e.g., L. Lu et al., 2025; B. Lu et al., 2025; Xu and Zhang, 2025) to reduce publication bias and improve coverage. The initial search yielded 1238 records across the three databases. Following automated and manual bibliometric deduplication, 576 duplicate records were removed, resulting in 662 unique studies. These were subjected to title and abstract screening, followed by full eligibility assessment based on predefined inclusion criteria. Ultimately, 175 studies met all criteria and were retained for full analysis in the meta-analytic framework.

3.2. Screening and eligibility criteria

Each record was independently evaluated by two reviewers using a structured two-stage screening process, progressing from title–abstract assessment to full-text eligibility review. Screening decisions were guided by operational criteria designed to implement the scope of the conceptual review defined in Table 2. Discrepancies between reviewers were resolved through consensus to ensure consistency and methodological rigour. The eligibility criteria applied during the PRISMA process are summarised in Table 3, which details the operational rules used to distinguish genuinely adaptive Sediment–Groundwater Digital Twin (SG-DT) studies from static or non-reproducible modelling approaches. These criteria emphasise adaptive Digital Twin functionality, quantitative performance reporting, explicit data model coupling, uncertainty characterisation, and transparency, thereby ensuring suitability for both qualitative synthesis and quantitative meta-analysis.

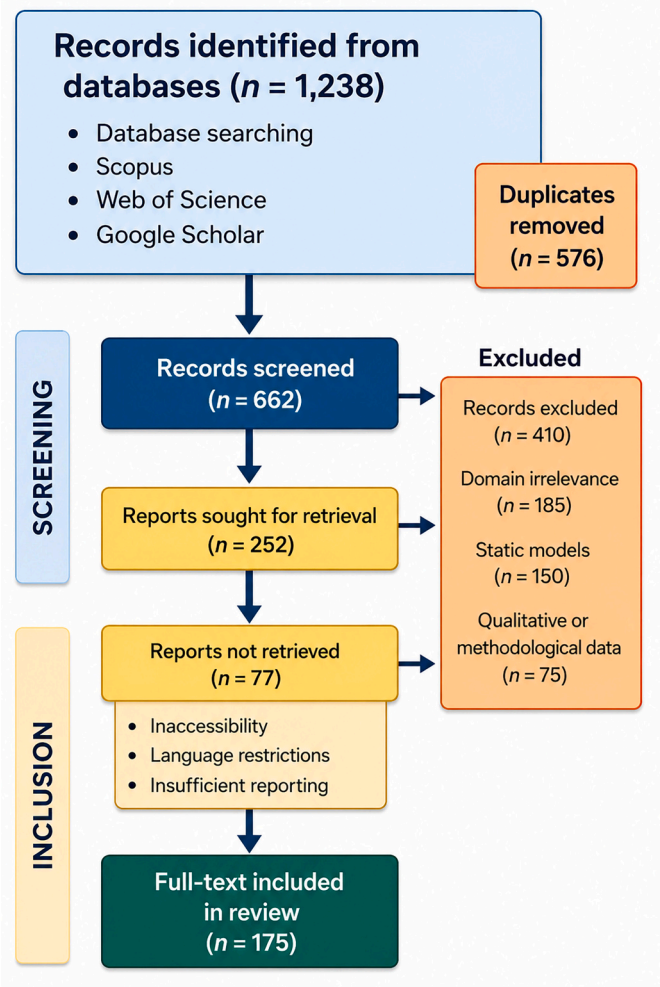


Fig. 3. PRISMA 2020 decision-based workflow for the SG-DT systematic review. Flow diagram summarising the identification, screening, eligibility assessment, and inclusion stages of the sediment–groundwater digital twin (SG-DT) literature following the PRISMA 2020 protocol.

3.3. Data extraction template

For each eligible study, metadata were systematically extracted, including publication details, model type (empirical/process-based/surrogate/digital twin), data-assimilation approach, uncertainty quantification (UQ) metrics, computational demand, data sources, and reproducibility score. The complete Data Extraction Form (Table S1, Supplementary material) provides detailed variable definitions, controlled vocabulary, and illustrative examples. The framework was originally designed in this study, with variable categories aligned to the terminology used in recent environmental DT literature (e.g., Ciliberti

Table 3
Operational screening criteria applied during the PRISMA process. Criteria were implemented to operationalise the review scope defined in Table 2 and to ensure methodological rigour, reproducibility, and suitability for quantitative synthesis in sediment–groundwater digital twin (SG-DT) studies.

Criterion (PRISMA-operational)	Inclusion rule (applied at screening)	Exclusion rule	Rationale/ implementation
Digital twin functionality	Evidence of adaptive behaviour through real-time or iterative data assimilation, surrogate coupling, and/or explicit uncertainty quantification (UQ)	Static numerical or empirical models lacking feedback, adaptation, or UQ	Ensures inclusion of genuinely adaptive systems rather than conventional simulators (e.g., He et al., 2020 included Pan et al., 2022 excluded)
Quantitative performance reporting	Reports at least one quantitative skill or reliability metric (e.g., RMSE, NSE, CRPS, coverage, PIT)	Qualitative discussion without quantitative evaluation	Enables cross-study comparability and supports meta-analysis (e.g., Zhu and Zabaras, 2018)
Data–model coupling	Explicit linkage between observational data streams (sensors, field sampling, or remote sensing) and model updating	Models relying solely on synthetic or static datasets	Distinguishes operational or near-operational DTs from offline modelling studies
Uncertainty treatment	Explicit probabilistic or ensemble-based uncertainty characterisation	Deterministic predictions without uncertainty reporting	Aligns screening with VVUQ principles required for trustworthy decision support
Reproducibility and transparency	Accessible data, code, or sufficiently documented methodology	Proprietary or unverifiable systems	Supports reproducibility and independent validation (e.g., Ross et al., 2019 – AquaSat)

et al., 2023; Ramos et al., 2023). This supplementary format avoids redundancy within the main text while ensuring transparency and reproducibility consistent with the PRISMA 2020 standard. Extracted metadata were later used for bibliometric clustering and topic modelling (Section 4) to identify thematic evolution and methodological convergence across the SG-DT literature.

3.4. Risk-of-bias assessment (RoB-EnvModel)

Given the interdisciplinary scope of sediment–groundwater Digital-Twin (SG-DT) studies, conventional bias tools were insufficient. To address this, a customised rubric, RoB-EnvModel, was developed by integrating ROBIS (systematic review bias) with model-centric reliability indicators that reflect alignment with VVUQ, reproducibility, and data openness. Each study received a composite bias score (0–5) across four dimensions, summarised in Table 4. Scoring was performed independently by two reviewers and validated through inter-rater agreement statistics (Section 3.6).

3.5. Protocol registration and data availability

All supporting materials used in this systematic review, including the PRISMA 2020 checklist, the RoB-EnvModel scoring guide, and the data-extraction form, are provided as Supplementary files accompanying this manuscript. To ensure full transparency and reproducibility, the complete dataset, analytical scripts, and all Supplementary materials have also been archived in the open-access repository Zenodo under DOI: https://doi.org/10.5281/zenodo.18289768. These materials include:

- Supplement A: PRISMA Search Queries
- Supplement B: Data Extraction Form (Excel)
- Supplement C: RoB-EnvModel (Risk-of-Bias Scoring Guide)
- Supplement D: JSON Schema Template (machine-readable)
- Supplement E: PRISMA 2020 Checklist (completed)
- Supplement F: Full List of Studies ($n = 175$)
- SG-DT_Coastal_As_2025_ReferenceInstance_v1.0.jsonld. Representative JSON-LD instance file for a coastal arsenic digital-twin case study.

3.6. Reviewer agreement (κ Cohen)

Inter-reviewer reliability was assessed using Cohen's κ coefficient to quantify the level of agreement between independent reviewers beyond chance. Cohen's κ is defined as (Cohen, 1960):

$$\kappa = (P_o - P_e)/(1 - P_e)$$

where P_o represents the observed proportion of agreement between reviewers, and P_e denotes the expected agreement under random

Table 4
Dimensions and scoring criteria of the methodological reliability and bias assessment framework used to evaluate sediment–groundwater digital twin (SG-DT) studies. Each dimension is scored on a standardised 0–5 ordinal scale, where higher values consistently indicate greater methodological rigour and reliability. Uncertainty quantification is decomposed into existence and quality to distinguish between the presence of uncertainty representation and its quantitative evaluation. Illustrative references provide representative examples for each dimension and are not intended as an exhaustive list. The complete scoring algorithm, weighting scheme, and worked examples are provided in the Supplementary material.

Dimension	Operational definition	Scoring scale (0–5)	Illustrative references*
Model transparency	Public availability of model code, data, and/or explicit procedural documentation enabling independent inspection	0 = fully opaque; no code/data/protocols available5 = fully open; code, data, and workflows publicly accessible	Ross et al., 2019; Ciliberti et al., 2023
Data–model synchronisation	Demonstrated implementation of sequential data updating via explicit data-assimilation or online model-update loops	0 = static, no updating, 5 = continuous or iterative data-model synchronisation implemented and evaluated	He et al., 2020; Ramos et al., 2023
Uncertainty quantification (existence)	Explicit inclusion of uncertainty representation in model outputs or predictions	0 = no uncertainty reported5 = uncertainty explicitly represented and propagated through the modelling chain	Zhu and Zabaras, 2018; Mo et al., 2019
Uncertainty quantification (quality)	Quantitative evaluation of uncertainty using probabilistic skill metrics (e.g., CRPS, PIT, coverage) and calibration diagnostics	0 = no quantitative evaluation5 = uncertainty quantitatively evaluated, calibrated, and validated	Zhu and Zabaras, 2018; Mo et al., 2019
Reproducibility	Ability to independently reproduce reported results using available data, code, and workflow descriptions	0 = results not reproducible5 = results independently reproducible without author intervention	Du et al., 2023; Gao et al., 2023
Validation	Formal validation against independent observations, benchmarks, or reference datasets consistent with VVUQ principles	0 = no validation performed5 = validation conducted using independent data and formal benchmarks	Du et al., 2023; Gao et al., 2023

classification. To evaluate coding consistency, approximately 10% of the included studies (n = 17) were independently and unthinkingly double-coded by two reviewers using a predefined classification framework covering model type (process-based, data-driven, hybrid), data assimilation strategy, and uncertainty quantification approach. A contingency table of reviewer classifications was constructed to compute P_o and P_e, from which the κ statistic was derived. The resulting value (κ = 0.87) indicates strong to near-perfect agreement according to commonly used interpretation benchmarks. Discrepancies between reviewers were primarily associated with borderline classifications between process-based and surrogate–hybrid modelling approaches. These cases were systematically reviewed and resolved through structured consensus discussions, guided by established methodological criteria (Mo et al., 2019; Bao et al., 2020). This procedure ensures both the reproducibility of the screening process and the robustness of the classification framework used in the systematic review.

4. Knowledge mapping-bibliometrics & topic modelling

To comprehensively understand the intellectual structure and evolution of Sediment-Groundwater Digital Twin (SG-DT) research, this section presents a knowledge mapping analysis using bibliometric techniques and topic modelling. The analysis leverages VOSviewer, Bibliometrix, and BERTopic, which are advanced tools for extracting key research themes, identifying influential studies, tracking thematic evolution, and recognising leading authors and journals within the SG-DT field. This methodology enables both qualitative and quantitative insights into the development of SG-DTs and their application to environmental science.

4.1. Co-citation network analysis

Co-citation analysis offers a systematic lens for uncovering the intellectual structure of the Sediment-Groundwater Digital Twin (SG-DT) literature. Identifying frequently co-cited publications reveals the most influential studies that have collectively shaped the field's methodological and conceptual evolution. Fig. 4 presents the co-citation network derived from the 175-study corpus, showing three dominant research clusters that correspond to data assimilation, uncertainty quantification (UQ), and model development. The network highlights both the central bridging studies (e.g., Raissi et al., 2019) and thematic interconnections among subdomains that underpin the emergence of integrated, uncertainty-aware digital twin frameworks.

4.2. Topic evolution analysis

The temporal evolution of research themes in Sediment–Groundwater Digital Twins (SG-DTs) was analysed using BERTopic-based topic modelling, enabling a structured examination of how dominant and emerging topics have evolved and shifted over time. By leveraging document-level semantic embeddings alongside publication-year metadata, this approach captures both the relative prominence of thematic clusters and their longitudinal transitions within the SG-DT literature. As shown in Fig. 5, the Sankey diagram illustrates the progressive transformation of research focus from early, conceptually oriented studies toward increasingly application-driven and system-integrated

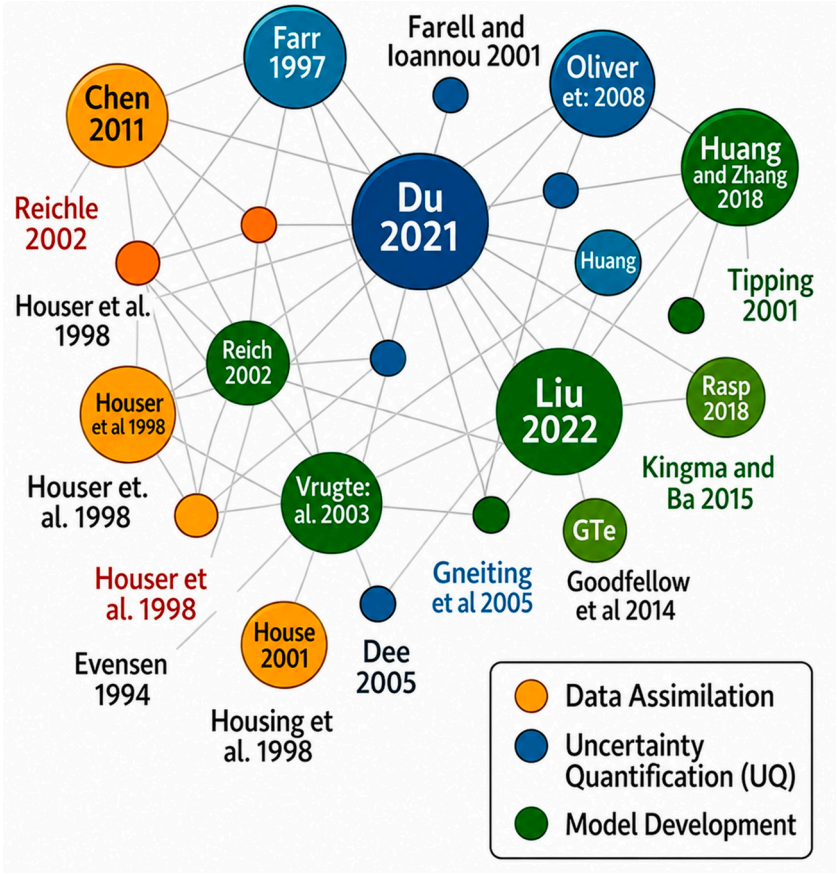


Fig. 4. Co-citation network of the SG-DT literature, illustrating major research clusters in data assimilation, uncertainty quantification (UQ), and model development. Node size indicates citation frequency, and link density reflects co-citation strength. The figure highlights the intellectual convergence of physical modelling, data assimilation, and AI-driven surrogate methods that define the foundation of modern SG-DT research.

themes. Initial research streams are predominantly associated with Digital Twin foundations, including conceptual architecture and enabling technologies. Over time, these streams give rise to methodologically richer topics, most notably real-time data assimilation and multi-source data integration, reflecting a growing emphasis on dynamic, data-intensive modelling frameworks. In the later stages of the timeline, multiple thematic streams converge on decision support systems, indicating a shift from exploratory and methodological development to operational SG-DT implementations for environmental monitoring, prediction, and management. This convergence highlights the increasing coupling of surrogate modelling, data assimilation, and decision-oriented analytics within unified digital twin frameworks.

Overall, the observed topic transitions suggest a clear maturation of the SG-DT research landscape, characterised by a shift from isolated methodological advances to integrated, end-to-end digital twin systems designed to support real-world decision-making in sediment–groundwater environments.

4.3. Geographical and institutional distribution

The geographical and institutional distribution of Sediment-Groundwater Digital Twin (SG-DT) research provides insight into the structural drivers shaping methodological development across the field. Bibliometric analysis indicates a pronounced concentration of SG-DT studies in North America, Europe, and East Asia, with China, the United States, and several European countries accounting for a substantial share of publications and collaborative networks. This pattern reflects not only scientific leadership but also disparities in monitoring infrastructure, data availability, and computational capacity that fundamentally influence where advanced Digital Twin methodologies can be developed and sustained. Institutional analysis further underscores this asymmetry. Leading research universities and laboratories such as Stanford University, the University of Oxford, Tsinghua University, and other highly resourced institutions have played a central role in advancing surrogate modelling, data assimilation, and uncertainty quantification approaches that underpin SG-DT frameworks. Their prominence is closely linked to access to long-term observational datasets, high-performance computing resources, and interdisciplinary research environments, which collectively lower the barriers to implementing adaptive, data-intensive Digital Twin architectures. Importantly, this geographical and institutional concentration has methodological implications. The dominance of data-rich regions may bias SG-DT development toward specific hydrogeological settings, climate regimes, and regulatory contexts, potentially limiting the transferability of existing models to underrepresented regions such as parts of Africa and South America. These gaps highlight the need for future SG-DT research to explicitly address data scarcity, heterogeneous monitoring capacity, and institutional constraints when aiming for globally applicable and policy-relevant Digital Twin systems.

Table 5 provides a descriptive overview of key journals, contributing institutions, and main research themes in the SG-DT literature. The table highlights prominent publication venues and thematic foci identified through bibliometric analysis, showing how methodological emphasis and research efforts are distributed across regions. Rather than suggesting a ranking or hierarchy, this overview places observed patterns within broader challenges related to reproducibility, scalability, and the development of SG-DTs as globally relevant environmental decision-support systems.

5. Technical primer: mechanistic and computational foundations

This section consolidates the mechanistic equations, computational structure, and data-integration frameworks that together form the backbone of SG-DTs. Its organisation follows a layered logic from physical processes to digital architecture, surrogate modelling, data assimilation, and uncertainty quantification, providing both conceptual depth and computational reproducibility.

5.1. Physical mechanisms: coupled flow, reaction, and feedback dynamics

At its core, a Sediment–Groundwater Digital Twin (SG-DT) is governed by the three-dimensional advection dispersion reaction (ADE-3D) formulation, which represents the mass balance of reactive solutes under dynamically evolving boundary conditions. However, in sediment–groundwater systems, the predictive value of this formulation depends not only on resolving transport processes, but on its ability to represent the tightly coupled interactions among groundwater flow, solute migration, hydrogeochemical reactions, and feedback-driven system evolution. Metal mobility in such environments is therefore not a transport-dominated phenomenon, but an emergent outcome of multi-process coupling across spatial and temporal scales. A first-order control on this coupling is imposed by sedimentological structure. Grain size distribution, sorting, and permeability contrasts are embedded within a broader framework of depositional architecture and facies organisation, which define hierarchical heterogeneity across scales (Koltermann and Gorelick, 1996). Facies architecture, including the spatial arrangement of coarse-grained transmissive units and fine-grained or organic-rich

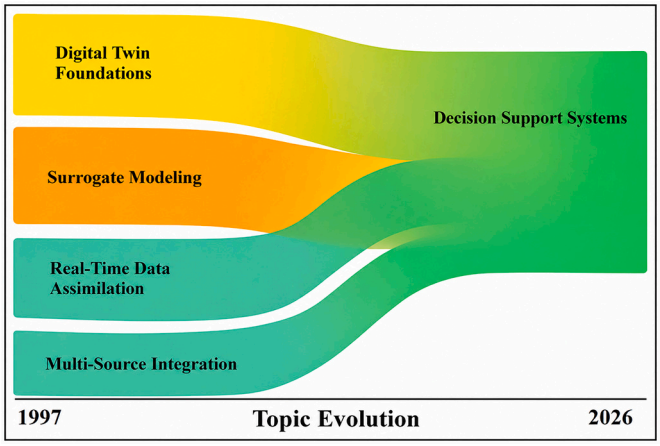


Fig. 5. Topic evolution Sankey diagram illustrating the shift in SG-DT research topics from 1997 to 2026. This diagram shows the evolution of key SG-DT research themes, from foundational Digital Twin studies to advanced topics like real-time data assimilation and decision support systems. The diagram was prepared by the authors using BERTopic. The flow represents the thematic progression and key transitions in the research landscape.

Table 5
Representative journals, institutions, and dominant research themes in the sediment–groundwater digital twin (SG-DT) literature. The table provides a descriptive overview of major publication venues and contributing institutions identified through bibliometric analysis Entries are intended to illustrate the breadth and diversity of SG-DT research activity.

Journal	Country	Representative institution	Dominant research theme
Sustainability	Switzerland	Swiss Federal Institute of Technology (ETH)	Data assimilation frameworks for SG-DTs
Journal of Hydrology	USA	Stanford University	Surrogate modelling for groundwater systems
Environmental Modelling & Software	China	Tsinghua University	Uncertainty quantification (UQ) methodologies
Water Research	UK	University of Cambridge	Real-time data integration and monitoring
Environmental Science & Technology	USA	Harvard University	Digital twin applications in environmental management

retention zones, governs flow pathways, residence-time distributions, diffusion domains, and reactive surface accessibility. These controls determine where contaminants are rapidly advected, where they are retarded by mass-transfer limitations, and where long-term storage zones develop. In heterogeneous deposits, preferential flow and low-permeability domains coexist, producing strongly non-uniform transport behaviour that is highly sensitive to unresolved sedimentary structure. This is particularly critical in coastal and alluvial systems, where depositional heterogeneity interacts with hydrological forcing to amplify variability in contaminant redistribution (Han et al., 2020a,b; Miao et al., 2019; Yu et al., 2021). These sedimentological controls are intrinsically coupled with hydrogeochemical processes that regulate contaminant fate. Metal mobility is governed by a classical hydrogeochemical framework in which redox-controlled transformations, sorption–desorption kinetics, mineral dissolution–precipitation equilibria, and colloid-facilitated transport dynamically interact (Aullón Alcaine et al., 2020; Bolan et al., 2014; Gao et al., 2023). Importantly, these processes are often rate-limited rather than instantaneous. Sorption may exhibit kinetic hysteresis; mineral reactions may be constrained by reactive-lag availability; and redox transitions may lag hydraulic perturbations due to spatial variability in electron donors, electron acceptors, and reactive phases. As a result, sediment–groundwater systems frequently deviate from local equilibrium, and contaminant persistence is governed not only by thermodynamic partitioning but by the relative timing of transport, reaction, and phase exchange. This transport-reaction coupling is inherently timescale-consuming. In high-velocity domains, advective transport can outpace reaction kinetics, allowing contaminants, particularly colloid-associated species, to migrate beyond zones where equilibrium-based retention would be expected. In contrast, low-permeability or diffusion-dominated zones promote extended residence times, enhancing sorption hysteresis, mineral transformation, and redox-mediated remobilisation. Such imbalances become especially pronounced in heterogeneous aquifers subject to tidal forcing, episodic recharge, or density-driven circulation, where hydraulic and geochemical disequilibrium evolve simultaneously (Han et al., 2020a,b; Faal et al., 2022; Aullón Alcaine et al., 2020). Han et al., 2020a,b further demonstrated that neglecting colloid-facilitated transport introduces systematic bias in surrogate-based reactive-flow predictions, underscoring the need for SG-DTs to preserve mechanistic transport pathways alongside state updating. Mineralogical composition and organic matter provide an additional layer of control that is essential for physically grounded prediction. Fe- and Mn-bearing oxides, clay minerals, carbonate phases, sulfides, and organic-rich sediments regulate sorption-site density, buffering capacity, redox reactivity, and metal-binding strength (Davis and Kent, 1990; Aullón Alcaine et al., 2020; Gao et al., 2023). Organic matter plays a particularly complex role by simultaneously acting as a sorptive medium, an electron donor for microbial metabolism, and a regulator of local acidity and complex chemistry. Recent studies further demonstrate that organic-matter composition actively controls subsurface biogeochemical pathways and associated contaminant mobilisation, rather than functioning as a passive sediment property (Liu et al., 2026). For SG-DT systems, this implies that mineralogical and organic-matter characteristics must be treated as dynamic state variables influencing both model parameters and system evolution. Microbially mediated processes introduce an additional level of feedback that is central to hydrogeochemical dynamics. Microbial activity controls metal mobility through redox cycling, organic-matter degradation, mineral transformation, and the generation of chemically reactive microenvironments (Ojuederie and Babalola, 2017; Z. Yang et al., 2023; C. Yang et al., 2023; Liu et al., 2026). These processes can shift systems between immobilisation-dominated and mobilisation-dominated states by altering redox fronts, ligand availability, and mineral stability. Because such transitions are spatially heterogeneous and temporally lagged, they produce reaction hotspots and migrating fronts that are difficult to capture using static or weakly coupled models. Consequently, SG-DTs designed for reliable prediction must assimilate not only hydraulic observations, but also geochemical and biogeochemical indicators that reflect evolving process states. Taken together, these mechanisms demonstrate that high-fidelity SG-DTs must represent sediment–groundwater environments as tightly coupled hydraulic–transport–reaction systems governed by feedback-dependent kinetics and sediment-controlled heterogeneity. The role of the Digital Twin is therefore not merely to accelerate computation, but to preserve mechanistic integrity while enabling adaptive updating under non-stationary conditions. Achieving this requires hybrid integration of process-based models with data assimilation and surrogate modelling approaches capable of resolving nonlinear reactive transport, kinetic limitations, mineralogical controls, and microbially mediated feedback without sacrificing physical interpretability (He et al., 2020; Sprocati and Rolle, 2021; Chen et al., 2021; Panahi et al., 2025). Only under such a formulation can SG-DTs deliver chemically defensible, dynamically consistent, and decision-relevant predictions in heterogeneous sediment–groundwater systems.

5.2. *Surrogate modelling: efficiency anchored in physics*

Surrogate models are key enablers of Sediment–Groundwater Digital Twins (SG-DTs), providing the speed needed for iterative forecasting, data assimilation, and uncertainty-aware updates in systems where full-physics simulation can be too resource-intensive for operational use. Their development has evolved from purely data-driven regressors to hybrid architectures that incorporate governing equations or process constraints within machine-learning frameworks. In this study's classification framework, such hybrid architectures directly correspond to Hybrid Differential Twins (HDTs), as outlined in Section 1.1. These systems combine differential-equation-based descriptions of physical processes with data-driven surrogate components in a closed-loop Digital Twin, maintaining essential mechanistic constraints while enabling computational efficiency and adaptive updating. Early machine-learning surrogates like Random Forests (RF) and Gaussian Process Regression (GPR) provided significant computational savings, but their use in sediment–groundwater systems was limited by assumptions of local stationarity, limited extrapolation, and often weak physical interpretability. Zhu and Zabaras (2018) advanced this by introducing Bayesian convolutional encoder–decoder surrogates that learn spatially resolved flow fields and offer calibrated uncertainty estimates. Building on this, Mo et al. (2020) extended the approach to transient multiphase transport using stochastic latent representations, achieving improved probabilistic fidelity under heterogeneous permeability fields. He et al.

(2020) and Du et al. (2023) embedded partial differential equations (PDEs) directly into neural network loss functions, yielding Physics-Informed Neural Networks (PINNs) that enforce mass-balance constraints even with sparse data. More recently, Panahi et al. (2025) included explicit uncertainty heads into PINN architectures, creating PINN-UQ hybrids that carry epistemic uncertainty through the forward model chain. In sediment-controlled contaminant transport, surrogate models cannot be judged solely on speed or predictive accuracy. Their reliability depends on data quality, representativeness, and scale consistency used for training, calibration, and updating. Data-driven surrogates perform well within their training domain but often struggle with extrapolation or out-of-distribution conditions—especially in sediment–groundwater systems where hydrological forcing, redox dynamics, salinity shifts, mineralogical transitions, and episodic remobilisation can push the system into states weakly represented or absent in historical data. Under such circumstances, a surrogate might remain numerically smooth yet lose physical credibility, yielding predictions that appear statistically coherent but fail to capture key process couplings. This is particularly critical in heterogeneous sedimentary environments, where contaminant behaviour is influenced not only by hydraulic responses but also by facies architecture, flow pathways, reaction kinetics, mineralogical controls, organic matter distribution, and microbial feedback. If surrogate models do not adequately encode these mechanisms, computational gains come at the expense of mechanistic interpretability—reproducing observed outputs without correctly explaining their origins. This is riskiest when heterogeneity is implicitly learned rather than explicitly constrained by facies information, physics-based structure, or process-aware variables. Consequently, surrogate skill in interpolation should not be mistaken for robust process representation, especially in systems dominated by unresolved sub-grid variability and rate-limited hydrogeochemical interactions. The distinction between interpolation and extrapolation is therefore crucial in advancing SG-DT methodologies. RF and similar ensemble regressors perform well in densely sampled areas but struggle to extrapolate beyond the training data's empirical bounds. GPR provides uncertainty estimates and is suitable in small-data settings but remains highly dependent on kernel assumptions, with reliability decreasing as system dimensionality and nonstationarity grow. Deep-learning surrogates offer greater complexity but are sensitive to training data distribution, coverage, and biases. Even physics-informed models like PINNs and operator-learning surrogates do not fully overcome these limitations: although they improve physical regularity and reduce empirical overfitting, they can misrepresent reactive transport if critical sedimentary controls—such as kinetic sorption, mineral transformations, or microbial interactions—are absent, weakly observed, or improperly parameterised. In SG-DT applications, the key question is not whether to use surrogate models but under what conditions they remain trustworthy. Their use is justified when four conditions are simultaneously met: first, training and updating data adequately encompass the environmental states relevant to operational prediction; second, the surrogate is embedded within an assimilation workflow capable of correcting drift as new data become available; third, uncertainty is explicitly quantified to detect extrapolative overconfidence; and fourth, the surrogate explicitly or via constrained design preserves the dominant hydrogeochemical controls governing system behaviour. Without these safeguards, surrogate models risk acting as efficient approximators of past patterns rather than reliable predictive components of a Digital Twin. As summarised in Table 6, surrogate families vary significantly in their assumptions, data needs, computational demands, uncertainty mechanisms, and roles within SG-DT workflows. The optimal choice depends not only on the trade-offs between data availability, physical complexity, and computational resources but also on acceptable extrapolation risks and interpretability constraints. Ultimately, advancing trustworthy SG-DT implementation lies in selecting surrogate strategies that remain physically constrained, uncertainty-aware, and suitably matched to the system's heterogeneity, data regime, and decision context.

5.3. Layered architecture: from data to decision

Building upon the preceding discussion, a robust Sediment–Groundwater Digital Twin (SG-DT) is realised through a tightly coupled, layered architecture that systematically connects heterogeneous data sources to modelling intelligence and decision support (Fig. 6). Rather than treating data ingestion, simulation, and decision-making as isolated components, this architecture enforces end-to-end integration to preserve spatiotemporal consistency across the digital twin lifecycle. In the context of sediment groundwater systems, however, such consistency must be understood not only across time but also across scales of observation, representation, and prediction. For this reason, the SG-DT architecture is interpreted here as a multi-scale, heterogeneity-aware system architecture, in which different layers resolve, aggregate, or update components of system behaviour depending on the spatial scale at which information is available, and decisions are required. At the entry point of the framework, raw environmental observations from in situ monitoring networks, remote sensing platforms, and targeted field surveys are first harmonised, quality-controlled, and synchronized to ensure coherence and reliability (Ciliberti et al., 2023; Essamlali et al., 2024). This ingestion layer serves as the primary locus of multi-scale data fusion, because the observations entering the SG-DT typically differ not only in format and frequency, but also in support scale. For example, point-based groundwater chemistry, borehole logs, and sensor telemetry provide local or site-specific information, whereas remote sensing products, watershed forcing data, and distributed infrastructure datasets constrain broader reach to field-scale system states. The role of the ingestion layer is therefore not simply to clean and align data streams, but to preserve the scale provenance of each observation and to map them onto an internally consistent state representation. Within the data assimilation layer, advanced sequential and variational techniques, including the Ensemble Kalman Filter (EnKF), four-dimensional variational assimilation (4D-Var), and Particle Filters (PF), are employed to continuously reconcile model predictions with incoming measurements, thereby reducing state estimation bias and improving forecast fidelity (He et al., 2025; Kim and Bartos, 2024). In a scale-aware SG-DT, assimilation does not uniformly update all quantities. Instead, it operates on multi-resolution state variables, in which different elements of the system are corrected at different levels of aggregation. At the local or site scale, assimilation may update concentrations, redox indicators, hydraulic heads, infiltration fluxes, or geochemically sensitive state variables at discrete observation points. At the reach or field scale, the same assimilation framework may update broader plume envelopes, effective transport parameters, bound-conditioned fields, or aggregated storage and loading states. This distinction is essential because many sediment–groundwater systems cannot be fully resolved at their finest physical scale in operational settings; therefore, trustworthy SG-DTs must explicitly distinguish between directly observed fine-scale states, inferred sub-grid variability, and coarser effective variables used for forecasting and control. These dynamically updated states propagate into the surrogate modelling layer, where reduced-order representations efficiently emulate the behaviour of computationally intensive hydrogeochemical processes. Learning-based surrogates such as Random Forest (RF), Gaussian Process Regression (GPR), Long Short-Term Memory (LSTM) networks, Physics-Informed Neural Networks (PINN), and Deep Operator Networks (DeepONet) enable scalable inference while retaining essential physical consistency (Zhu and Zabaras, 2018; Mok et al., 2023; Wang et al., 2022). In a multi-scale setting, the function of the surrogate layer is not merely to accelerate computation, but to mediate between resolved and unresolved heterogeneity. Some surrogates emulate fine-scale process responses and provide effective closures to coarser models, while others approximate field-scale behaviour directly from aggregated observations. The architectural implication is that surrogate placement within the SG-DT must be explicitly linked to the scale at which process detail is either represented, parameterised, or intentionally abstracted. To ensure the credibility of digital twin

Table 6
Qualitative, task-dependent comparison of surrogate modelling paradigms used in sediment–groundwater digital twin (SG-DT) studies. The comparison emphasises core modelling assumptions, data regimes, computational training costs, uncertainty-quantification (UQ) mechanisms, and typical roles within data-assimilation workflows. Reported predictive performance reflects trends documented in representative benchmark and application studies under appropriate training configurations rather than universal rankings.

Modelling paradigm	Core assumptions (problem framing)	Typical data regime	Computational cost (training)	Predictive performance*	Uncertainty quantification (UQ) mechanism	Typical role in data assimilation
Random Forest (RF)	Local stationarity; weak feature dependence on ensemble trees	Moderate to large, labeled datasets	Low to moderate (CPU-scalable)	Competitive for interpolation under sufficient data coverage; limited extrapolation	Empirical ensemble spread; non-probabilistic	Emulator in forecast step; limited use in fully probabilistic DA
Gaussian Process Regression (GPR)	Smoothness and kernel-defined correlations; Gaussian residual assumption	Small to moderate datasets	High (cubic scaling with data size)	High for low-dimensional, data-scarce regimes reported in benchmark studies	Bayesian posterior mean and variance (native UQ)	Natural integration with Bayesian DA (e.g., EnKF, PF)
CNN/Autoencoder	Learned spatial correlation and feature continuity in gridded domains	Large spatiotemporal datasets	Moderate to high (GPU-dependent)	High for representation learning and pattern extraction; task- and architecture-dependent	Indirect or post-hoc (ensembles, dropout, Bayesian variants)	Reduced-order models, state compression, or surrogate predictors
Physics-Informed Neural Networks (PINNs)	Governing PDEs embedded as soft constraints in the loss function	Low to moderate data supplemented by physics constraints	High (sensitive, multi-term optimisation)	Strong interpolation reported under well-posed training and adequate collocation; extrapolation remains problem-dependent	Constraint-consistent but not fully probabilistic	Physics-constrained forward modelling; limited direct DA integration
Deep Operator Networks (DeepONet)	Operator learning between parameterised input–output function spaces	Very large, paired input–output datasets	High to very high	High accuracy for parametric PDE families demonstrated in controlled benchmarks	Ensemble-based or emerging Bayesian extensions	Emulator of forward operators in ensemble-based DA
Fourier Neural Operators (FNO)	Translation-invariant operators represented in Fourier space	Large, grid-consistent datasets	High (FFT-based, GPU-oriented)	High accuracy for homogeneous domains under sufficient training coverage	Limited, research-stage probabilistic variants	Fast surrogate forward models in ensemble DA
Hybrid GAN–Ensemble Smoothers (GAN–ES)	Non-Gaussian state–parameter distributions learned from data	Large datasets with observational noise	Very high	Task-specific performance; sensitive to training stability and data representativeness	Implicit, distributional (sample-based)	Explicitly designed for non-Gaussian data assimilation problems

Sources: Zhu and Zabaras (2018); Mo et al. (2019); He et al. (2020); Du et al. (2023); Panahi et al. (2025).

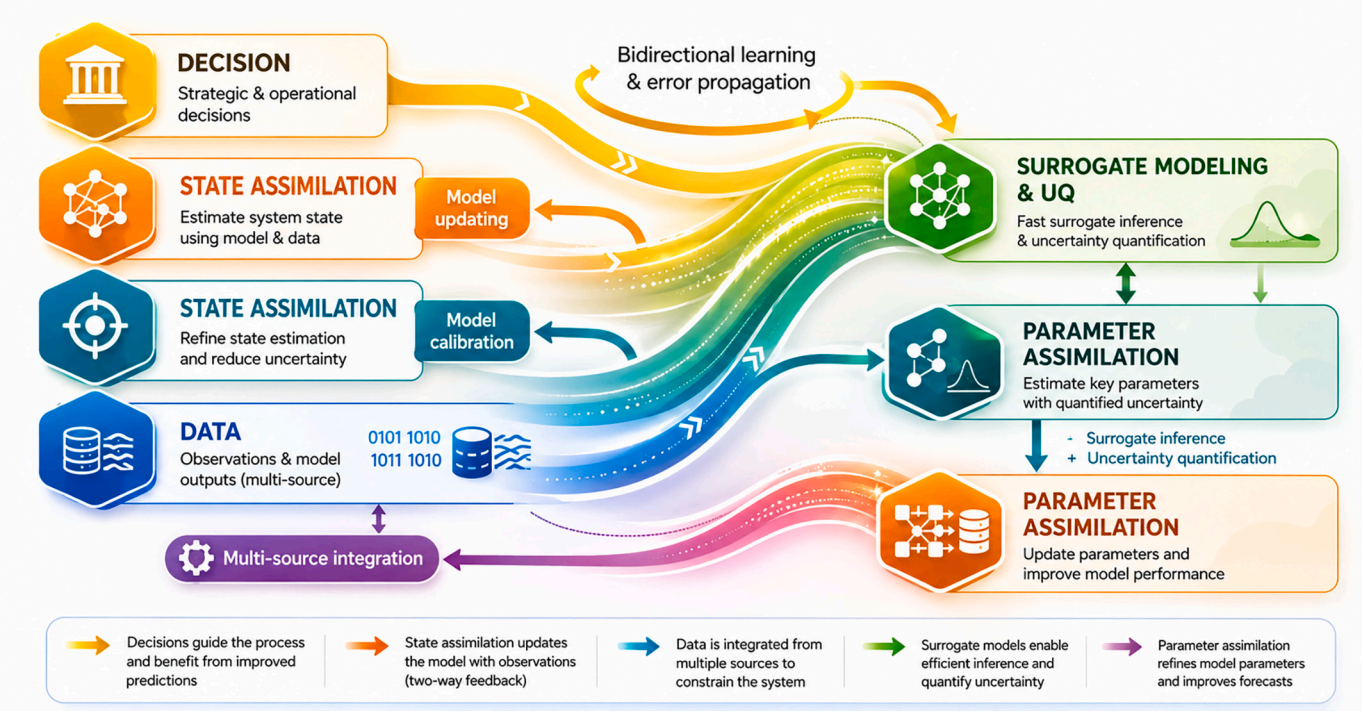


Fig. 6. Conceptual aggregation of the sediment–groundwater digital twin (SG-DT) architecture. This schematic presents a high-level functional aggregation of the SG-DT into five conceptual domains, each encompassing multiple operational layers of the full architecture. The figure emphasises information flows associated with data acquisition, model updating, surrogate inference, and uncertainty propagation, rather than a one-to-one enumeration of implementation layers. Detailed projections of the operational and software architecture for the same system are provided in subsequent figures.

outputs, the subsequent Uncertainty Quantification (UQ) layer explicitly characterises both aleatoric uncertainties arising from data variability and epistemic uncertainties linked to model structure and parameterisation. In scale-dependent sediment–groundwater systems, this includes uncertainty introduced by aggregation, unresolved facies heterogeneity, sparse monitoring, and a mismatch between the measurement and prediction scales. Probabilistic performance metrics, including the Continuous Ranked Probability Score (CRPS), Probability Integral Transform (PIT) diagnostics, and predictive coverage rates, are used to assess calibration, sharpness, and reliability across operational scenarios (Panahi et al., 2025; Wang et al., 2020). Under this architecture, UQ is therefore not only a test of statistical performance, but also a diagnostic of whether the SG-DT has preserved sufficient scale consistency between data, model states, and predictive outputs. At the highest level, the decision layer translates these uncertainty-aware predictions into actionable knowledge for regulators, engineers, and water resource managers. Here again, scale awareness is essential. Local decisions may concern pumping adjustments, barrier placement, site remediation, or monitoring intensification, whereas field- or basin-scale decisions may involve source control, zoning, drought-response measures, or adaptive watershed management. The decision layer thus closes the loop not through a single universal control logic, but through scale-consistent feedback, in which the form of intervention matches the scale at which the SG-DT has demonstrated predictive support and credibility of uncertainty. Through this modular yet fully integrated hierarchy, the SG-DT framework promotes transparency, traceability, and resilience, foundational attributes for next-generation environmental digital twins operating at the science–policy interface. More importantly, architecture clearly shows that trustworthy SG-DTs are not scale-neutral platforms. They are systems in which ingestion, assimilation, surrogate modelling, uncertainty analysis, and decision feedback must all align with the spatial scale of process representation and with how subsurface heterogeneity is resolved, parameterised, or inferred.

5.4. Data assimilation: synchronising observation and prediction

Data assimilation converts static models into adaptive systems capable of real-time correction. In sediment–groundwater Digital Twins, however, its role extends beyond temporal updating: it is the principal mechanism through which observations collected at different resolutions are reconciled with model states defined across multiple spatial scales. In this sense, assimilation is the central bridge among the measurement, process, and prediction scales. The Ensemble Kalman Filter (EnKF) remains the most practical framework for high-dimensional state estimation (Bao et al., 2020; Hauser-Davis and Wosnick, 2022). Its main advantage in SG-DT contexts is that it can efficiently update distributed state variables while propagating uncertainty through ensemble statistics. In scale-aware applications, EnKF is especially suitable when local measurements such as heads, concentrations, or conductivity proxies are used to correct site-scale state fields or effective parameter zones. The Four-Dimensional Variational (4D-Var) approach achieves strong temporal coupling in systems where gradients are accessible (Kim and Bartos, 2024). It is particularly useful when the SG-DT must maintain coherence over a continuous assimilation window and when the objective is to reconcile temporally dense observations with dynamically evolving field states. Particle Filters (PF) handle nonlinear and multimodal dynamics characteristic of metal transport under transient hydrological conditions (Pan et al., 2022) and are therefore especially relevant where system response is highly non-Gaussian or where episodic hydrological forcing generates abrupt regime shifts. For SG-DT implementation, the critical issue is not only which assimilation algorithm is selected but also what is being updated and at what scale. At the local or site scale, assimilation may update concentrations, reactive tracers, hydraulic heads, salinity fronts, or redox-sensitive variables directly tied to monitoring wells and sensor nodes. At broader field- or reach-scale, assimilation typically updates effective transport parameters, plume extent, storage terms, spatial loading patterns, or aggregated forcing fields that cannot be directly observed in full detail. Thus, assimilation in SG-DTs should be interpreted as a multi-resolution updating process, in which observational data constrain both fine-scale states and coarser effective representations, depending on data support and model objective. This distinction is particularly important in heterogeneous sediment–groundwater systems, because the physical meaning of an update depends on whether it corrects a resolved state, an inferred sub-grid process, or an upscaled parameter. If this distinction is not made explicit, apparent improvements in forecast skill may conceal unresolved scaling error or overcompensation for structural heterogeneity. Accordingly, the assimilation layer must preserve not only statistical consistency, but also scale consistency between observations, model structure, and the intended prediction task. Each technique, therefore, offers trade-offs not only in scalability and computational burden, but also in how it represents and propagates cross-scale information. EnKF is efficient for distributed updating under moderate assumptions about the uncertainty structure; 4D-Var provides stronger temporal and dynamical coherence when gradients are tractable; and PF is better suited to strongly nonlinear, threshold-driven, or multimodal dynamics where aggregation may obscure critical state transitions. These differences are summarised in Table 7, which should be interpreted not merely as a comparison of algorithms, but as a guide to how different assimilation strategies support SG-DT synchronisation under varying scales of heterogeneity, observation density, and process complexity.

5.5. Uncertainty quantification: from calibration to validation

Valid sediment–groundwater digital twins (SG-DTs) require not only accurate point predictions but also a physically grounded and statistically consistent characterisation of predictive reliability across space and time (Kwon et al., 2025; Hou et al., 2026). In sediment-controlled hydrogeochemical systems, uncertainty is not solely a function of observational noise or model approximation error but emerges from the incomplete representation of coupled flow–transport–reaction processes, unresolved sediment heterogeneity, and dynamically evolving environmental forcing. Consequently, uncertainty quantification (UQ) within SG-DTs must explicitly distinguish and propagate both aleatoric uncertainty (arising from stochastic variability in measurements and boundary conditions) and epistemic uncertainty (arising from structural limitations in process representation, parameterization, and model form) (Zhu and Zabaras, 2018; Mo et al., 2020; Halpern, 2025). Crucially, in sediment–groundwater environments, epistemic uncertainty is strongly tied to process-level incompleteness, including (i) unresolved reaction kinetics and rate limitations, (ii) simplified coupling between groundwater flow, solute transport, and geochemical transformations, and (iii) insufficient representation of mineralogical, organic, and microbially mediated controls on contaminant mobility. As a result, predictive uncertainty is often state-dependent and non-stationary, increasing during hydrological perturbation (e.g., recharge pulses, tidal forcing, or redox shifts) when system dynamics deviate from calibrated conditions. This implies that UQ in SG-DTs cannot be treated as a static post-processing step but must be embedded within the dynamic model-update cycle to track evolving uncertainty structures in real time. Recent advances have therefore shifted UQ practice from post hoc error quantification toward integrated probabilistic validation frameworks embedded within physics-aware modelling pipelines. Rather than evaluating models solely on deterministic accuracy metrics, contemporary SG-DT validation explicitly assesses the consistency between predicted probability distributions and observed system behaviour. Metrics such as the Continuous Ranked Probability Score (CRPS) quantify both distributional sharpness and overall predictive skill, while Probability Integral Transform (PIT) diagnostics evaluate statistical calibration by testing whether forecast distributions are probabilistically consistent with observations under ideal conditions. Coverage-based

Table 7
Comparison of assimilation frameworks used for SG-DTs, assessing robustness and suitability across scales. Derived from Bao et al. (2020); He et al. (2020); Kim and Bartos (2024); Pan et al. (2021).

Method	Strengths	Limitations	Scalability	Noise sensitivity	Typical use case
EnKF/Ensemble Smoother	Handles large state vectors efficiently	Ensemble collapse under strong nonlinearity	High	Moderate	Regional-scale DTs with sensor networks
4D-Var	Accurate temporal coupling; gradient-based	Requires adjoint model	Moderate	Low	Controlled experiments, pilot DTs
Particle Filter (PF)	Robust to non-Gaussian, nonlinear systems	Computationally expensive	Low–Moderate	Low	Reactive flow under fluctuating boundary conditions

criteria, including prediction interval coverage probability (PICP), further assess whether uncertainty bounds adequately capture observed variability, particularly in risk-sensitive and regulatory contexts (Panahi et al., 2025). Importantly, these diagnostics gain additional interpretive value when linked to underlying physical processes. For example, systematic under-coverage may indicate missing reaction pathways or underestimated kinetic constraints, while over-dispersed predictive intervals may reflect excessive parameter uncertainty or inadequate coupling between flow and reaction modules. In this sense, UQ metrics are not merely statistical indicators but diagnostic proxies for process fidelity, enabling identification of structurally incomplete or physically inconsistent SG-DT implementations. Hybrid uncertainty-aware frameworks increasingly integrate ensemble-based inference, Bayesian updating, and data assimilation to propagate uncertainty across both parameter and state spaces under nonstationary conditions. Stochastic surrogate models embedded within ensemble Kalman filtering or particle-filter architectures enable continuous recalibration of state variables while preserving uncertainty structure, whereas Bayesian learning mechanisms update parameter distributions in response to new observations and evolving system states (He et al., 2020; Kim and Bartos, 2024). These approaches are particularly critical in sediment–groundwater systems, where parameter sensitivity is strongly conditioned by local mineralogy, organic matter content, and transient redox conditions, leading to context-dependent uncertainty propagation. Moreover, recent developments in physics-informed and probabilistic deep learning provide a pathway for embedding uncertainty directly within surrogate modelling components. Techniques such as Bayesian neural networks, variational inference, and probabilistic operator learning allow SG-DTs to represent not only mean system behaviour but also predictive distributions that reflect structural and parametric uncertainty (Zhu and Zabaras, 2018; Mo et al., 2020; Panahi et al., 2025). When coupled with real-time data assimilation, these methods enable closed-loop uncertainty tracking, where predictive confidence is continuously updated alongside system states. Table 8 synthesizes the principal UQ strategies employed in SG-DTs and explicitly links their probabilistic foundations to both statistical evaluation metrics and underlying hydrogeochemical processes. This mapping clarifies how different UQ approaches capture distinct components of uncertainty, ranging from measurement variability to structural process uncertainty and dynamic system feedback. Collectively, these developments reposition uncertainty quantification from a secondary validation step to a core operational layer of SG-DT systems, where predictive reliability, physical consistency, and decision trustworthiness are jointly evaluated. By integrating probabilistic diagnostics with process-aware interpretation, SG-DTs can distinguish between genuinely well-calibrated systems and those that achieve apparent accuracy through overfitting or incomplete process representation. This distinction is essential for ensuring that SG-DT predictions remain both scientifically defensible and operationally reliable under the complex, nonstationary conditions characteristic of sediment–groundwater environments.

As can be seen in Table 9; to assess probabilistic performance, SG-DT studies use four complementary metrics, each quantifying a distinct feature of forecast reliability. All the following formulas are extracted from peer-reviewed sources and have been implemented in the environmental surrogate model literature.

5.6. Data quality and propagation of uncertainty

The performance of sediment–groundwater digital twins (SG-DTs) is fundamentally constrained by the integrity of their underlying data streams, making data quality a primary driver of downstream uncertainty. Unlike model-related errors, data-induced uncertainties originate at the measurement level and can systematically propagate through learning, assimilation, and inference pipelines if left unaccounted for. Ensuring traceable data provenance, consistent sensor calibration, and interoperable formatting is therefore a prerequisite for meaningful uncertainty characterisation rather than a purely technical preprocessing step. Recent studies emphasise that high-frequency and heterogeneous sensor networks pose distinct challenges, including temporal drift, cross-sensor inconsistency, and metadata incompleteness. Ciliberti et al. (2021) and Ramos et al. (2022) demonstrate that without explicit auditing of sensor metadata and calibration histories, apparent model uncertainty may in fact reflect unrecognised measurement bias. Complementary work by Ross et al. (2019) established harmonisation protocols for integrating remote sensing products with in situ observations, highlighting how mismatched spatial support and acquisition timing can amplify uncertainty during data fusion. Crucially, data-quality deficiencies do not remain localised but cascade through surrogate modelling and assimilation workflows, where they are often misinterpreted as epistemic or stochastic model uncertainty. This misattribution can lead to overconfident predictions or inflated uncertainty bounds, undermining the credibility of SG-DT outputs in operational settings. By explicitly tracking data-originated uncertainty and its transformation across processing stages, SG-DTs can better align

Table 8
Summary of uncertainty quantification frameworks and corresponding probabilistic assessment criteria used in SG-DTs. The values and formulas are defined below.

UQ method	Underlying principle	Probabilistic metric	Interpretation	Policy relevance
Ensemble surrogates	Monte Carlo propagation of parameter uncertainty (Mo et al., 2019)	CRPS (Gneiting and Raftery, 2007)	Measures distributional sharpness; lower = better	Benchmark for model comparison
Bayesian neural network	Posterior sampling of model weights (Zhu and Zabaras, 2018)	PIT Histogram (Dawid, 1984)	Assesses calibration; uniform = ideal	Evaluates reliability of forecasts
PINN-UQ/DeepONet	Physics-informed probabilistic mapping (Panahi et al., 2025)	Coverage/PICP	Proportion of true values within predictive interval	Defines regulatory safety margins (≥90%)
Hybrid ensemble assimilation	Covariance-inflated EnKF with stochastic surrogates (He et al., 2020; Kim and Bartos, 2024)	Calibration Slope (Kuleshov et al., 2018)	Regression of predicted vs. observed probabilities; ideal = 1	Quantifies systematic bias in risk estimation

Table 9
Mathematical definitions and interpretations of the main probabilistic skill metrics applied in SG-DT validation. Values and formulations follow Gneiting and Raftery (2007), Dawid (1984), and Kuleshov et al. (2018), with environmental applications adapted from Zhu and Zabaras (2018), Mo et al. (2019), Panahi et al. (2025), and Kim and Bartos (2024).

Metric	Definition/formula	Interpretation	Example application
Continuous Ranked Probability Score (CRPS) (Gneiting and Raftery, 2007)	CRPS(F, x_o) = $\int[F(x) - H(x - x_o)]^2$ dx where F(x) is the predicted CDF and H is the Heaviside step function.	Quantifies the integrated distance between forecast CDF and observation; smaller values = more accurate probabilistic forecast.	PINN-UQ model achieved CRPS = 0.15 vs. baseline 0.32 (Panahi et al., 2025).
Probability Integral Transform (PIT) (Dawid, 1984)	$z_i = F_i(x_i)$, with F_i the predictive CDF and x_i the observation. Uniform z_i distribution ⇒ well-calibrated.	Tests the statistical calibration of predictive distributions.	Uniform PIT achieved by Bayesian CNN (Zhu and Zabaras, 2018).
Prediction interval coverage probability (PICP)	PICP = (1/N) Σ I[$L_i \leq x_i \leq U_i$], where [L_i, U_i] is the predicted interval (Mo et al., 2019).	Fraction of observations within the predictive interval; ideal ≥90%.	Ensemble UQ achieved PICP = 0.93 (Kim and Bartos, 2024).
Calibration slope (Kuleshov et al., 2018)	β = Cov(p, y) / Var (p) Regression slope of predicted probabilities vs. empirical outcomes.	Slope ≈ 1 ⇒ well-calibrated; <1 ⇒ over-confidence; >1 ⇒ under-confidence.	PINN-UQ models showed β = 0.97 ± 0.02 (Panahi et al., 2025).

uncertainty propagation with Verification–Validation–Quantification of Uncertainty (VVUQ) principles, ensuring that predictive uncertainty reflects true system variability rather than hidden data artefacts.

6. Systematic synthesis of literature

We synthesised 175 studies across four analytic axes: system typology, pollutants, coupling intensity, and methodological pitfalls to delineate where Sediment–Groundwater Digital Twins (SG-DTs) are mature and where reliability bottlenecks persist. A compact taxonomy and a rigorously curated comparison matrix are provided in Fig. 7, with the full expanded matrix for all records reported in Table S2.

6.1. Typology of SG-DT systems

The systematic synthesis of the sediment–groundwater digital twin (SG-DT) literature reveals a structured yet heterogeneous landscape that can be organised along multiple complementary dimensions. In this review, typology is defined through a multi-axis framework that jointly accounts for system domain, operational scale, dominant contaminant group, and modelling stance, enabling consistent comparison across application contexts and methodological choices. This section focuses primarily on the system-domain axis, while explicitly introducing the scale of operation as a second, indispensable organising dimension. The full multi-dimensional classification is summarised in Fig. 7. Across the reviewed studies, four dominant system domains emerge as recurrent deployment contexts for SG-DTs, each characterised by distinct operational objectives, data environments, and levels of methodological integration. Urban systems constitute the most extensively developed domain, reflecting the availability of dense sensor networks, established telemetry infrastructure, and clear operational drivers related to drainage, water distribution, and contaminant control. SG-DTs in urban contexts frequently integrate real-time monitoring with state estimation and control-oriented workflows, providing transferable foundations for the development of subsurface digital twins (Bartos and Kerkez, 2021; Conejos Fuertes et al., 2020; Lumley et al., 2024; Mucke et al., 2023). Coastal aquifer systems constitute a second major domain, in which SG-DTs are commonly used to investigate salinity intrusion and contaminant transport under dynamically evolving hydraulic gradients. Studies in this category often combine reduced-order or surrogate-assisted modelling with scenario-based analysis to support management under sea-level rise and pumping stress (Han et al., 2020a, 2020b; Yu et al., 2021). These implementations illustrate how SG-DTs extend beyond real-time control toward anticipatory decision support in data-sparse and uncertainty-prone environments. A third typological group encompasses deltaic and large-basin systems, in which SG-DTs are deployed at regional scales to integrate heterogeneous data sources, including remote sensing, in situ observations, and numerical forecasting. In these contexts, digital twins function as coordination platforms for multi-scale process representation rather than localised control tools, reflecting a shift toward basin-wide environmental intelligence (Morlot et al., 2024). Finally, hyporheic and cold-region systems represent an emerging application domain characterised by strong seasonal dynamics and event-driven processes. SG-DT studies in these environments address challenges associated with permafrost thaw, freeze–thaw cycles, and episodic contaminant mobilisation, often emphasising uncertainty-aware forecasting and adaptive modelling strategies (O'Donnell et al., 2024; Skierszkan et al., 2025).

However, the domain alone is insufficient to characterise SG-DT maturity or interpretability. The reviewed literature also exhibits clear scale stratification, which is critical because the physical meaning of model states, the representation of spatial heterogeneity, and the feasibility of continuous synchronisation vary substantially across scales. In the present review, four broad scale classes are identified as analytically useful: (i) pore/column/laboratory scale, (ii) site/pilot scale, (iii) reach or coastal-segment scale, and (iv) field-to-basin scale. These classes are not merely descriptive labels; they define the relationship between observation density, process resolution, heterogeneity representation, and achievable Digital Twin functionality. At the pore-, column-, and laboratory scale, SG-DT-like studies typically operate in data-rich, but spatially bounded environments, where process fidelity is

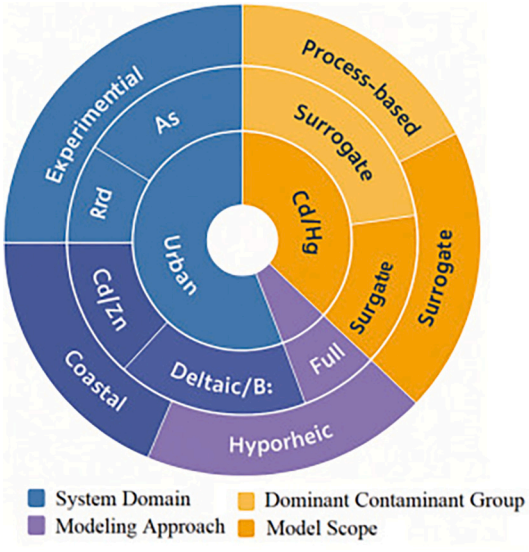


Fig. 7. Typology of sediment–groundwater digital twins and adjacent digital-twin domains. Core SG-DT categories represent systems explicitly coupling sediment–groundwater processes with contaminant transport and geochemical feedback. Urban drainage, sewage, and drinking-water network digital twins are included as adjacent domains to illustrate transferable architectural and operational patterns rather than primary SG-DT implementations.

relatively high, and individual reactions or transport mechanisms can be resolved in detail. The dominant data regime in this class consists of controlled experiments, dense temporal measurements, and high-resolution process observations. Heterogeneity is often represented explicitly through measured media properties or idealised experimental structure, rather than through broad stochastic abstraction. The main maturity limitation at this scale is not process representation, but the relative absence of continuous real-world synchronisation, operational decision feedback, and field-relevant uncertainty propagation. As a result, such studies often provide strong mechanistic insight but only partial DT functionality. At the site or pilot scale, SG-DT systems become more operationally recognisable. This class typically relies on monitoring wells, local sensor networks, targeted field campaigns, and event-responsive observations. Here, heterogeneity is often represented through facies-based zonation, site-specific parameter fields, or conceptual hydrostratigraphic units. This is also the scale at which data assimilation becomes practically meaningful, because observations are sufficiently localised to support state updates while remaining relevant to environmental management. However, maturity limitations arise from incomplete coverage of spatial heterogeneity, uneven monitoring density, and the difficulty of reconciling fine-scale geochemical processes with site-wide predictive states. Many systems in this class, therefore, exhibit partial DT maturity: they may implement adaptive updating, but not always with fully explicit uncertainty quantification or robust heterogeneity-aware parameterisation.

At the reach or coastal-segment scale, SG-DT implementations shift toward meso-scale coordination of distributed environmental processes. The data regime typically combines point observations, remote-sensing constraints, and external forcing information, such as tides, recharge, hydraulic loading, or sediment transport indicators. Heterogeneity is represented more indirectly, often through aggregated parameter zones, effective transport properties, reduced-order surrogates, or segment-based spatial partitioning. At this scale, SG-DTs are better aligned with anticipatory forecasting and management support than with local mechanistic control. The key maturity limitation is that unresolved sub-segment heterogeneity can be absorbed into effective parameters or surrogate approximations, potentially improving apparent forecast skill while obscuring process realism. Thus, interpretability and the credibility of uncertainty become central concerns. At the field-to-basin scale, SG-DTs act less as local digital replicas and more as integrative coordination systems for large-scale environmental intelligence. The dominant data regime includes sparse in-situ networks, regional monitoring databases, remote sensing, and broad forcing datasets. Heterogeneity is rarely resolved directly; instead, it is represented through upscaled parameter fields, regional facies abstractions, scenario ensembles, or coarse spatial discretisation. These systems are especially valuable for policy analysis, comparative scenario assessment, and distributed resource management, but they also face the strongest maturity constraints. Their synchronisation with real-world conditions is often intermittent rather than continuous, and their uncertainty structure is strongly influenced by aggregation error, observation sparsity, and scale mismatch between process understanding and predictive use. This scale-based perspective clarifies that SG-DT maturity cannot be assessed solely by algorithmic sophistication or application domain. A highly advanced surrogate model implemented at basin scale with sparse observational support may be methodologically less mature, in an SG-DT sense, than a site-scale implementation with modest computational complexity but explicit synchronisation, heterogeneity-aware updating, and transparent uncertainty tracking. Conversely, laboratory-scale studies may offer superior process fidelity while remaining only partially Digital Twin–like from an operational standpoint. Accordingly, the present typology treats scale, data regime, heterogeneity representation, and update logic as jointly constitutive dimensions of SG-DT classification. When viewed in conjunction with contaminant classes and modelling stances, these system domains and scale classes delineate the progression of SG-DTs from surface-oriented hydraulic representations toward coupled, process-aware digital replicas of subsurface systems operating across spatial and temporal scales. More importantly, they show that SG-DT development does not follow a single linear trajectory. Instead, it is branching into distinct maturity pathways, shaped by the extent to which each implementation can reconcile process, observation, and decision scales within a coherent, uncertainty-aware modelling framework.

6.2. Pollutants of concern

Arsenic (As) remains the most frequently modelled toxicant (Bretzler et al., 2017), with Pb/Zn in estuarine/urban sediments (Zhang et al., 2016) and Cd/Hg in coastal acidification scenarios (Gao et al., 2023). Climate-amplified episodes (floods, permafrost thaw) can trigger episodic remobilisation of mixed metals by altering sediment permeability, grain-scale reactive surface area, and redox stratification, particularly in layered or heterolytic successions. These transient, sediment-controlled processes challenge steady-state assumptions and underscore the need for real-time assimilation and probabilistic risk updates within SG-DTs (Alfee and Bloor, 2025; Nawab et al., 2021; Sheykhi and Moore, 2012), especially in environments where metal fluxes are governed by short-lived depositional, erosional, or thaw-related events.

6.3. Coupling intensity

We classify SG-DT implementations based on the strength of integration between physical process models, observational data streams, and adaptive updating mechanisms, here referred to as coupling intensity. Rather than treating coupling as a purely qualitative spectrum (loose to tight), this study operationalizes it as a function of the degree to which observations, model states, surrogate representations, and decision feedback are dynamically synchronized within a continuous modelling loop. Representative exemplars are summarised in Table 10 and provide the empirical basis for the reference architecture introduced in Section 9. Importantly, coupling intensity is not scale-invariant and must be interpreted in relation to the spatial resolution, data availability, and degree of heterogeneity represented within the system. In sediment–groundwater environments, where processes span from pore to basin scales and are strongly conditioned by sedimentary heterogeneity, the practical meaning of “tight” coupling differs across application contexts. At smaller scales, tight coupling may correspond to near-complete synchronisation between observed and simulated states, whereas at site to regional scales it typically reflects the consistent integration of multi-source data, effective parameter updating, and uncertainty-aware prediction rather than full physical resolution. Accordingly, three generalized coupling regimes can be distinguished. Loose coupling characterises systems in which model components and observational data are weakly synchronized, with limited or no continuous updating, such that data are primarily used for calibration or periodic validation (Bartos and Kerkez, 2021; O'Donnell et al., 2024). Hybrid coupling describes partially integrated systems in which selected state variables or parameters are dynamically updated through assimilation or surrogate-based learning, while other components remain static or scenario-driven, often reflecting trade-offs between computational feasibility and data limitations (Han et al., 2020a, 2020b; Yu et al., 2021; Conejos Fuertes et al., 2020; Lumley et al., 2024). It is important to note that the term hybrid coupling used here refers solely to the degree of integration between system components and should not be conflated with Hybrid Differential Twins (HDTs), as defined in Section 1.1. While HDTs describe a specific class of Digital Twin architectures that combine differential-equation-based process representations with surrogate components, hybrid coupling refers to an intermediate level of system synchronisation and may occur in systems with or without HDT-type internal structures. Tight coupling, in contrast, refers to fully integrated SG-DT systems operating within a

Table 10
Condensed matrix of well-cited SG-DT exemplars covering the four typologies. The table evidences a gradient from operations-oriented, telemetry-driven DTs (loose) toward closed-loop, uncertainty-aware twins (tight).

System	Pollutant(s)	Primary data	Model stance	Surrogate	Assimilation	UQ metric(s)	Reported skill	Coupling	Open materials	Reference
Urban drainage DT	– (hydraulics)	Telemetry, SCADA	Process + DT	–	–	–	Online interactive twin; RMSE < 5% on flows	Loose	Code described	Bartos and Kerkez, 2021
Regional sewage DT	– (hydraulics)	IoT sensors	Process + control	–	EnKF	–	High-skill control, NSE > 0.9	Loose	Partial	Lumley et al., 2024
Drinking-water network	Pb, Zn	In-situ + smart meters	DT + AI	RF	PF	–	MAE ≈ 0.07 mg L^{-1}	Hybrid	Data accessible	Conejos Fuertes et al., 2020
Leak localization DT	–	Pressure/ flow telemetry	DT + generative ML	GAN	–	CRPS	CRPS ≈ 0.15	Tight*	Code open	Mucke et al., 2023
Coastal aquifer mgmt.	Cl^-, trace metals	Field + RS	Process + surrogate	SVR	–	–	RMSE < 10%	Hybrid	Data available	Han et al., 2020a, 2020b
Aquifer–sea flux control	–	RS + in-situ	Hybrid DT	DNN emulator	–	PIT (qual.)	Stable PIT histograms	Hybrid	Partial	Yu et al., 2021
Multicomponent reactive transport (inverse)	As/Cr proxies	In-situ + lab	PINN + process	PINN	EnKF	Coverage	Coverage ≈ 0.9	Tight	Code open	Chen et al., 2021
Non-Gaussian flow/ transport	–	Sim. + field	DL + data assim.	GAN	Ensemble smoother	–	R^2 ≈ 0.97	Tight	Open repo	Bao et al., 2020
Basin hydrological DT	–	RS + sensors	Process + Ops DT	–	4D-Var	–	NSE > 0.8	Hybrid	Public	Morlot et al., 2024
Coastal acidification	Zn, Pb	Lab + field	Process + ML	RF	–	CRPS	CRPS ≈ 0.18	Hybrid	Partial	Gao et al., 2023
Permafrost streams	Fe, As	Field + RS	Process DT	–	–	–	Event-based remobilisation	Loose	Open data	O'Donnell et al., 2024
Global flood-remobilisation review	Mixed metals	Literature synthesis	Evidence base	–	–	–	Risk patterns under flooding	–	–	Alfee and Bloor, 2025

closed-loop framework in which data assimilation, surrogate modelling, uncertainty quantification, and decision feedback are continuously coordinated, enabling probabilistically consistent and dynamically updated predictions (Chen et al., 2021; Bao et al., 2020; Mucke et al., 2023). A critical implication emerging from this classification is that coupling intensity cannot be evaluated independently of spatial heterogeneity and scale transition. In highly heterogeneous sedimentary systems, increased coupling does not necessarily imply improved physical realism, as unresolved heterogeneity may be compensated through effective parameterization or surrogate approximation. Conversely, systems with moderate coupling but explicit representation of heterogeneity and scale-consistent updating may yield more robust and transferable predictions. Therefore, coupling intensity in SG-DTs is best interpreted as a context-dependent property that reflects the extent to which a system can coherently integrate data, represent heterogeneity, and propagate uncertainty at the scale relevant to prediction and decision-making.

6.4. Pitfalls and validity threats

Despite notable advancements in the conceptualisation and implementation of SG-DTs, several systemic weaknesses continue to constrain their reliability, reproducibility, and policy readiness. These weaknesses do not arise only from incomplete reporting or limited deployment experience, but also from deeper methodological vulnerabilities in how SG-DT systems, particularly surrogate-based and data-driven ones, are trained, validated, interpreted, and transferred across environmental conditions. One of the most pervasive gaps is the under-reporting of uncertainty quantification (UQ). Fewer than 10% of published studies explicitly communicate probabilistic skill using quantitative metrics such as the Continuous Ranked Probability Score (CRPS), coverage probability, or Probability Integral Transform (PIT) histograms. This deficiency persists despite clear demonstrations of feasibility in deep learning and data assimilation pipelines (Bao et al., 2020; Morillo et al., 2002; Chen et al., 2022). The omission of probabilistic calibration not only obscures predictive confidence but also undermines trust in SG-DT forecasts used for environmental decision-making. More critically, in data-driven and hybrid SG-DT systems, lack of explicit uncertainty reporting can conceal extrapolative failure, surrogate drift, or overconfident predictions under changing hydrogeochemical conditions. In such cases, apparent predictive accuracy may reflect interpolation within a narrow training envelope rather than robust performance under operational variability. A second methodological weakness concerns the validation of data-driven and surrogate-based models. Many studies continue to rely on invalid or weak cross-validation procedures, particularly where spatial or temporal leakage occurs between training and validation datasets. This practice can artificially inflate reported skill and create a misleading impression of generalizability. The problem is especially acute in sediment–groundwater systems, where environmental autocorrelation and repeated sampling from closely related hydrogeological settings can make superficially independent samples functionally non-independent. Under these conditions, validation results may reward memorisation of local structure rather than true generalisation. The adaptive emulator design employed by Yu et al. (2021) represents an example of improved practice, in which geographically and temporally independent validation subsets are used to ensure a more robust performance assessment. Even so, the broader literature still shows that strong surrogate performance is often reported without sufficiently rigorous testing under transfer conditions. A third and more fundamental validity threat is the limited robustness of many data-driven approaches to extrapolation. Surrogate models may perform well within the statistical domain spanned by their training data, but they frequently become fragile when exposed to out-of-distribution (OOD) conditions. This limitation is especially relevant for SG-DTs because sediment–groundwater systems are shaped by nonstationary

boundary forcing, episodic hydrological events, redox transitions, salinity shifts, and sediment-controlled threshold behaviour. Climatic stressors, including acidification, extreme flooding, and permafrost thaw, have revealed vulnerabilities in model transferability and stability, exposing the limits of deterministic calibration and static training assumptions (Frogner-Kockum et al., 2020; Nguyen et al., 2020; Sharifi et al., 2024; Roy et al., 2016). In these settings, data-driven models may remain numerically stable yet become physically implausible, especially when the environmental states encountered during deployment are weakly represented or entirely absent during training. A fourth, and often under-recognised, threat concerns the loss of physical interpretability. As SG-DT systems increasingly adopt flexible machine-learning architectures, there is a risk that predictive success becomes decoupled from mechanistic understanding. A model may reproduce observed concentrations or hydraulic responses without correctly representing the sedimentary controls, reactive pathways, or process couplings that generated those observations. In sediment-controlled systems, this issue is particularly serious because contaminant mobility depends on facies heterogeneity, preferential flow, kinetic sorption, mineralogical transitions, and microbially mediated hydrogeochemical feedback. When such controls are not explicitly represented or are only implicitly learned from data, the resulting surrogate may function as an efficient pattern recogniser but not as a physically interpretable environmental model. Loss of interpretability is therefore not merely a philosophical concern; it is a concrete threat to validity because it limits the diagnosis of model failure, weakens regulatory defensibility, and reduces confidence in model transferability across sites, scales, and forcing regimes. A fifth limitation concerns the dependence of surrogate and data-driven models on data availability, representativeness, and quality. The apparent success of these approaches is often contingent on dense, clean, and well-aligned training datasets, conditions that are uncommon in many sediment–groundwater applications. Sparse monitoring, inconsistent sampling frequencies, limited geochemical characterisation, and scale mismatch between observations and modelled states can all impair the quality of learning and propagate hidden bias. In complex sedimentary environments, the problem is intensified by facies-dependent heterogeneity and nonstationary process coupling. Training data collected from one hydrostratigraphic setting or hydrological regime may not adequately constrain behaviour in another, even within the same nominal system. As a result, reported skill may reflect favourable data support rather than the intrinsic robustness of the modelling approach. Reproducibility deficits remain a major additional limitation. Publicly available code, datasets, or workflow documentation remain largely concentrated in urban water digital twins (Choi and Yoon, 2025; Marcussen et al., 2008; Yin et al., 2024) and a limited subset of data-assimilation frameworks (Bao et al., 2020). Replication within reactive transport and coupled subsurface DT studies remains rare, thereby impeding transparency and slowing methodological convergence across the field. This problem is particularly consequential for data-driven models, whose performance can be highly sensitive to preprocessing choices, feature engineering, training splits, hyperparameter selection, and hidden assumptions about data support. Without transparent workflows, it becomes difficult to distinguish genuine methodological advances from case-specific tuning. Taken together, these issues show that the current generation of SG-DTs exhibits substantial progress in real-time monitoring, surrogate development, and data assimilation capability, yet remains constrained by incomplete uncertainty treatment, insufficiently rigorous validation, limited reproducibility, fragility under OOD conditions, and, in some cases, reduced physical interpretability. These threats are not uniform across all SG-DT approaches, but they are especially pronounced in purely data-driven or weakly constrained surrogate implementations deployed in complex sedimentary systems. Addressing them will require more than incremental reporting improvements. It will require standardised VVUQ practices, explicit OOD testing, leakage-resistant validation design, transparent workflow disclosure, and stronger integration of physical constraints into data-driven model development. Only under such conditions can SG-DTs progress from experimentally promising prototypes toward dependable, regulatorily interpretable, and scientifically defensible decision-support systems.

6.5. *Process-to-model mapping: linking geochemical mechanisms with SG-DT responses*

A persistent limitation in SG-DT research is not merely the weak coupling between field-measurable geochemical processes and their computational analogues, but the absence of an explicit, scale-consistent translation between process dynamics, sediment heterogeneity, and digital representation within twin architectures. Although recent water-sector digital twins increasingly assimilate dense telemetry and employ learning-based surrogates or probabilistic updating (Bartos and Kerkez, 2021; Conejos Fuertes et al., 2020), foundational DT surveys still emphasise architectural and operational integration over process fidelity and scale-aware representation (Jones et al., 2020; Rasheed et al., 2020). In sediment–groundwater systems, however, predictive credibility cannot be established through data connectivity or adaptive updating alone. It depends on whether the digital twin preserves the governing couplings among hydraulic forcing, solute transport, reaction kinetics, and hydrogeochemical feedback across scales ranging from pore-scale reactions to field-scale transport pathways. Metal mobility is controlled not only by process presence but also by the relative alignment of transport timescales, reaction kinetics, and heterogeneous sediment structure. Consequently, SG-DTs must explicitly represent whether processes are resolved, parameterised or emulated at the scale relevant to prediction and decision-making. Key physicochemical and biogeochemical mechanisms governing contaminant mobility include adsorption to Fe/Mn oxides and organic matter, redox-controlled solubility and phase transitions, mineralogical transformation, and episodic remobilisation (Aullón Alcaine et al., 2020; Bolan et al., 2014). These processes are intrinsically coupled through flow pathways, residence times, and sediment composition, and their predictive impact is strongly scale dependent. For example, adsorption–desorption kinetics operate at mineral surface and pore scales but manifest at plume scale through delayed breakthrough, while redox front migration emerges at meter-to-field scales through the integration of microbial activity, flow heterogeneity, and electron acceptor distribution. Accordingly, the central challenge is not algorithmic sophistication per se, but the process-to-model translation, whereby hydrogeochemical mechanisms are converted into scale-consistent digital abstractions including state variables, assimilation operators, surrogate constraints, and uncertainty diagnostics. Without this translation, SG-DTs risk functioning as statistically updated but physically underconstrained predictors that obscure the role of unresolved heterogeneity and misrepresent dominant process controls. To ensure interpretability and regulatory trust, this translation must be explicit along three interconnected dimensions:

- Process scale definition: identifying the dominant spatial and temporal scale at which each mechanism operates (e.g., pore-scale sorption vs. field-scale redox fronts);
- Heterogeneity representation strategy: specifying whether spatial variability is resolved explicitly (e.g., facies-based discretisation), parameterised (effective properties), or learned via surrogates;
- Upscaling and aggregation logic: defining how sub-grid processes influence model-scale behaviour, including kinetic averaging, effective parameterisation, or stochastic representation.

Building on these principles, five core hydrogeochemical mechanisms must be operationalised within SG-DT systems: First, adsorption–desorption dynamics must be treated as time-dependent, scale-transferred exchange processes, in which pore-scale reactions are upscaled to dynamic partitioning behaviour responsive to pH, salinity, and dissolved organic matter variability. Second, redox front migration

must be linked to field-observable proxies and spatially distributed state variables, as redox boundaries often emerge from coupled microbial and transport processes operating across heterogeneous sediment domains. Third, mineralogical transformation must be incorporated as a state-evolving property that reflects the feedback between mineral dissolution/precipitation and reactive surface availability under changing geochemical conditions. Fourth, reaction kinetics and rate limitations must be explicitly represented, since the mismatch between transport and reaction timescales is a dominant control on plume evolution, particularly in heterogeneous systems where preferential flow bypasses reactive zones. Fifth, microbial controls on metal mobility must be included where relevant, as biologically mediated reactions dynamically alter redox conditions, mineral stability, and organic matter turnover, thereby reshaping system behaviour across scales (Liu et al., 2026; Z. Yang et al., 2023; C. Yang et al., 2023; Ojuederie and Babalola, 2017). In this study, Table 11 formalises this multi-dimensional translation by mapping each dominant sediment–water process to its corresponding SG-DT representation, including: (i) digital state variables, (ii) assimilation operators, (iii) surrogate constraints, (iv) dominant scale of operation, (v) heterogeneity representation strategy, and (vi) implications for uncertainty quantification. The objective of this mapping is not merely descriptive but to establish the physical provenance of model states, clarify the observational basis for parameter updates, and ensure that uncertainty reflects both data limitations and the fidelity of process representation. Through this framework, uncertainty in SG-DT predictions is reinterpreted as a compound outcome of data scarcity, surrogate approximation, and incomplete representation of the governing multi-scale process network. A well-calibrated SG-DT must therefore demonstrate not only statistical skill but also scale-consistent, chemically interpretable coherence between observed system behaviour and the modelled interaction of flow, transport, reaction, and feedback. This requirement is particularly critical in systems subject to tidal forcing, episodic recharge, flood-driven remobilisation, or density-dependent circulation, where rapid shifts in hydrological forcing can activate different process regimes and invalidate scale-inconsistent assumptions. Accordingly, Table 11 should be interpreted as an operational bridge between hydrogeochemical science and the implementation of SG-DT. By explicitly linking measurable environmental processes to scale-aware digital representation, adaptive updating, and uncertainty-aware validation, it provides the foundation for SG-DTs that are not only adaptive but also mechanistically grounded, reproducible, and regulatorily interpretable (Morlot et al., 2024; Bauer et al., 2021).

7. Quantitative meta-analysis & benchmarking

This section constitutes the core scientific contribution of the present study, providing the first systematically harmonised and quantitatively comparable evaluation of predictive performance and methodological maturity across sediment–groundwater Digital Twin (SG-DT) system implementations. Building on the systematically curated evidence base established in Sections 2–6, we move beyond descriptive synthesis to deliver a statistically rigorous benchmarking of how different modelling paradigms perform under heterogeneous environmental, data, and methodological conditions. A structured quantitative synthesis was conducted to evaluate effect sizes (E) representing predictive performance, uncertainty treatment, and methodological robustness across 175 curated SG-DT system studies. Throughout this section, effect size (E) is adopted as the sole, consistently applied quantitative measure of predictive performance across all analyses, enabling direct comparison between otherwise incompatible modelling frameworks. Performance metrics were systematically harmonised to this common scale, pooled using a random-effects framework, and supplemented by sensitivity diagnostics, bias assessment, and moderator analysis. Importantly, this meta-analytic framework is explicitly designed not only to quantify performance gains but also to diagnose the level of methodological maturity of SG-DT systems by identifying how predictive improvements depend on model structure, data assimilation strategies, uncertainty quantification (UQ), and data availability. In this sense, the analysis provides a dual contribution: (i) a statistically robust estimate of predictive performance improvement, and (ii) a mechanistic understanding of the conditions under which SG-DT systems achieve reliable and transferable performance. This section, therefore, establishes a performance-centred and evidence-based foundation for evaluating SG-DT systems, against which conceptual architectures, reporting frameworks, and future methodological developments (Sections 10–12) are subsequently interpreted and operationalised.

7.1. Effect size definition

Given the substantial heterogeneity in model structures, validation protocols, performance metrics, and reporting practices across sediment–groundwater Digital Twin (SG-DT) system studies, direct comparison of reported results is not statistically meaningful without careful harmonisation and explicit control of methodological inconsistencies. Accordingly, the effect size (E) framework developed in this section serves as the foundational analytical construct for performance evaluation across SG-DT systems, enabling consistent, interpretable, and bias-aware comparison of predictive outcomes across heterogeneous studies. Throughout this section, effect size (E) is used as the sole, consistently applied quantitative measure of the SG-DT system's predictive performance. Differences in baseline definitions, metric scaling, and reporting formats can introduce systematic bias and obscure the true contribution of SG-DT system functionality. Without rigorous control of these factors, apparent performance gains may reflect inconsistencies in evaluation methodology rather than genuine improvements in predictive capability. To address these limitations, a rigorously defined effect-size framework and associated data-processing protocol were developed to ensure comparability, interpretability, and methodological robustness. In this study, the primary outcome is defined as a relative effect size (E) that quantifies the incremental predictive contribution of an SG-DT

Table 11
Process-to-model mapping for SG-DT integration. Mapping of key hydro-geochemical processes to their computational representations within sediment–groundwater digital twins (SG-DTs). Each process is linked to its digital abstraction state variables, assimilation operators, or physics-informed surrogates and to its implications for model reliability and uncertainty quantification, ensuring physically interpretable and chemically consistent predictions.

Process	Description	SG-DT representation	Dominant scale	Heterogeneity representation	UQ implications
Adsorption–desorption	Metal binding to oxides & OM	Dynamic (K_d(t)), EnKF updates	Pore → Local	Parameterised via effective reactivity fields	Controls short-term release; reduces bias
Redox front migration	Oxic–anoxic transitions	PINN/DeepONet + ORP assimilation	Local → Field	Explicit via facies/redox zonation	Stabilises calibration; reduces drift
Mineralogical transformation	Dissolution/ precipitation	Adaptive parameter updating	Pore → Field	Hybrid (proxy-based + surrogate)	Enhances long-term consistency
Colloid transport	Particle-facilitated transport	Stochastic surrogate ensembles	Pore → Field	Implicit via stochastic transport kernels	Captures extreme events
Hyporheic exchange	Diffusion & exchange fluxes	FNO multi-scale module	Interface → Reach	Multi-scale boundary representation	Improves spatial coherence

system compared with a strictly defined baseline model. Baseline models were required to represent the same underlying modelling framework without data assimilation, adaptive updating, or closed-loop feedback, thereby isolating the performance gains attributable to SG-DT system mechanisms. This strict baseline consistency criterion is critical for distinguishing true Digital Twin functionality from general model improvements. Studies lacking a clearly defined or methodologically comparable baseline were excluded from quantitative pooling and retained only for qualitative synthesis, ensuring internal consistency of the meta-analytic dataset and preventing bias from incomparable performance references. To enable cross-study synthesis, all reported performance metrics were transformed into a dimensionless and directionally consistent effect size (E), where positive values indicate improved predictive performance of the SG-DT system relative to its baseline. This transformation provides a unified analytical scale while preserving interpretability across heterogeneous modelling frameworks and ensuring that performance gains are evaluated relative to a consistent reference condition. For error-based metrics (e.g., RMSE, MAE, CRPS), effect sizes were calculated as relative reductions with respect to the baseline model:

$$E_i = 100 \times (1 - M_t,i / M_b,i)$$

where M_t,i and M_b,i denote the performance of the Digital Twin and baseline models, respectively. This formulation is consistent with commonly used relative improvement metrics in environmental modelling and machine-learning-based performance evaluation, and provides an intuitive, scale-independent measure of proportional error reduction. For skill-based metrics bounded above (e.g., coefficient of determination R^2 and Nash–Sutcliffe efficiency), effect sizes were defined as:

$$E_i = 100 \times (S_t,i - S_b,i)/(1 - S_b,i)$$

which represents the fraction of the remaining performance gap to the theoretical optimum captured by the SG-DT system relative to its baseline (Thelen et al., 2022). This transformation preserves a consistent directional interpretation while reducing distortions associated with bounded metric scales, particularly when baseline performance approaches upper limits. Importantly, rather than assuming full equivalence across all performance metrics, we explicitly account for metric-dependent variability by distinguishing between error-based and skill-based measures during harmonisation. This classification ensures consistent interpretation of effect sizes and avoids conflation of fundamentally different performance indicators within a single analytical scale. In addition to metric harmonisation, careful attention was given to data quality and methodological consistency across studies. Only studies providing sufficient information to derive effect sizes and interpret baseline comparisons were included in the quantitative synthesis. Where uncertainty information (e.g., standard errors or confidence intervals) was available, it was incorporated into the analysis; otherwise, results were interpreted with appropriate caution given the limitations of reported data. Furthermore, it is recognised that the performance of the SG-DT system may be influenced by contextual factors such as system scale, data availability, model structure, and hydrogeochemical complexity. While the diversity and inconsistency of reporting across studies limited the extent to which these factors could be formally parameterised, their potential influence was explicitly considered in interpreting pooled results and assessing between-study heterogeneity (Section 7.3). This ensures that performance evaluation remains grounded in the physical and methodological variability of subsurface systems, rather than implying artificial uniformity. The resulting harmonised dataset of effect sizes (E) therefore provides a consistent, interpretable, and methodologically rigorous basis for evaluating predictive performance and methodological maturity across SG-DT systems, supporting the random-effects meta-analysis described in Section 7.2 and enabling robust cross-study benchmarking under heterogeneous conditions.

7.2. Random-effects modelling and heterogeneity

A random-effects meta-analytic framework was employed to estimate the overall magnitude of effect sizes (E) reported across sediment–groundwater Digital Twin (SG-DT) system studies. This modelling strategy was selected because the reviewed literature does not represent a single, methodologically homogeneous population of SG-DT implementations. Rather, it comprises studies that differ substantially in spatial scale, data richness, model architecture, assimilation design, validation rigour, and uncertainty quantification (UQ) practice. Under such conditions, a random-effects formulation is methodologically more appropriate than a fixed-effects alternative because it explicitly accounts for both within-study sampling uncertainty and genuine between-study variability. Between-study variance was estimated using the restricted maximum likelihood (REML) method. The resulting heterogeneity statistic ($I^2 = 68\%$) indicates that a substantial proportion of the observed variability in effect sizes arises from real differences among studies rather than sampling error alone. This level of heterogeneity is consistent with the broad interdisciplinary scope of SG-DT applications, spanning urban groundwater systems, coastal aquifers, and reactive transport environments. More importantly, it has direct interpretive implications for evaluating SG-DT performance. Specifically, the observed variability indicates that predictive gains associated with SG-DT systems cannot be understood as universally transferable or as arising from a single, mature, and internally consistent modelling paradigm. Instead, they reflect a context-sensitive and still unevenly developed methodological landscape, in which reported performance depends strongly on study design, environmental setting, data conditions, and the degree of integration among assimilation, surrogate modelling, and uncertainty treatment. In this sense, the heterogeneity observed here is not merely a statistical nuisance to be adjusted for, but a substantive finding. It indicates that SG-DT systems, as represented in the current literature, remain methodologically heterogeneous and only partially mature as a class of predictive environmental systems. Some implementations demonstrate robust gains under data-rich and operationally coupled conditions, whereas others remain closer to loosely integrated or weakly validated modelling workflows. Accordingly, the pooled effect estimate should not be interpreted as a universal performance constant for SG-DT systems, but rather as a summary tendency emerging from a diverse body of evidence with differing levels of methodological maturity. This interpretation reinforces the central objective of the meta-analysis: not only to estimate average predictive improvement, but also to determine the extent to which such improvement is contingent on system characteristics, data constraints, and modelling choices. To systematically examine these drivers, a structured analysis of heterogeneity is presented in Section 7.3. Fig. 8 presents the forest plot summarising the study-level effect sizes derived from harmonised predictive performance metrics. Each study is represented by its point estimate and corresponding 95% confidence interval, plotted on a common improvement-oriented scale, with positive values indicating performance gains relative to baseline models. Only studies reporting compatible quantitative information suitable for effect size estimation were included in the visualisation, while the remaining studies contributed exclusively to the qualitative synthesis. The forest plot, therefore, provides more than a descriptive summary of study-level outcomes: it visualises the dispersion, uncertainty, and non-uniformity of performance gains across SG-DT systems, thereby making explicit the empirical basis for the heterogeneity analysis that follows.

7.3. Sources and structure of heterogeneity

To explain the substantial between-study heterogeneity ($I^2 = 68\%$), we conducted a structured, scale-aware analysis of its underlying drivers across sediment–groundwater Digital Twin (SG-DT) system studies.

Rather than treating heterogeneity as a purely statistical artefact, it is interpreted here as the emergent outcome of interacting multi-scale process dynamics, spatial heterogeneity, data constraints, and modelling assumptions that collectively determine SG-DT system performance and reported effect sizes. Within this framework, heterogeneity is analytically informative: it reflects not only differences in predictive performance, but also inconsistencies in how studies represent upscale, and couple hydrogeochemical processes across spatial and temporal scales, as well as how robustly models generalise beyond their training or calibration domain. The sources of heterogeneity can be grouped into four interrelated domains:

7.3.1. System scale, scale transition, and environmental complexity

SG-DT implementations span a wide range of spatial scales, from pore-scale laboratory experiments and site-scale investigations to regional aquifer and basin-scale systems. Across this spectrum, governing processes including advection, dispersion, sorption, redox transformation, and sediment–water exchange operate at different characteristic spatial and temporal scales, requiring explicit scaling strategies for consistent representation. At smaller scales, processes are often directly resolved or tightly constrained, enabling high apparent predictive performance under controlled conditions. However, as system scale increases, unresolved heterogeneity in sedimentary architecture, boundary-condition complexity, and nonlinear feedback become increasingly dominant. Facies variability, permeability anisotropy, and spatially variable reaction environments introduce scale-dependent behaviour that cannot be captured through simple parameter transfer. A key source of heterogeneity, therefore, arises from scale mismatch, where models calibrated at laboratory or local scales are applied to field-scale prediction without consistent upscaling of transport parameters, reaction kinetics, or boundary interactions. Under such conditions, effective parameters (e.g., dispersion coefficients, partitioning behaviour) become scale-dependent, leading to systematic bias and inflated variability in reported effect sizes. Consequently, performance gains observed in simplified systems cannot be assumed to generalise to larger, heterogeneous environments, where structural uncertainty and unresolved variability dominate predictive outcomes.

7.3.2. Representation of spatial heterogeneity

Beyond scale, heterogeneity is strongly controlled by how spatial



Fig. 8. Random-effects meta-analysis of predictive performance across sediment–groundwater digital twin (SG-DT) studies. The forest plot displays percentage-based effect sizes (E, %) representing effect sizes (E) over baseline models for studies with extractable quantitative metrics. Points denote individual study estimates, and horizontal lines indicate 95% confidence intervals. The vertical reference line at E = 0 indicates no net change in performance relative to the baseline. Effect sizes were harmonised across multiple performance indicators (e.g., RMSE, CRPS, NSE, R^2) using a consistent improvement-oriented definition. Colours are used solely for visual discrimination between studies and do not encode methodological or categorical information.

variability is represented within SG-DT systems. Across the reviewed studies, three primary representation strategies can be identified:

- Explicit representation, where sedimentary facies, stratigraphy, or geostatistical fields are directly resolved;
- Parameterised (effective) representation, where heterogeneity is aggregated into bulk or upscaled properties;
- Implicit or surrogate-based representation, where heterogeneity is learned from data without explicit physical encoding.

These approaches differ fundamentally in their ability to capture preferential flow pathways, reactive surface distribution, and localised geochemical gradients. Systems employing overly aggregated representations tend to underestimate plume spreading, delay contaminant release, or misrepresent redox-driven transitions. In contrast, models that explicitly or adaptively represent heterogeneity demonstrate improved predictive stability, particularly when coupled with data assimilation. However, implicit (data-driven) representations introduce an additional layer of variability. When heterogeneity is learned rather than explicitly encoded, model behaviour depends strongly on the representativeness of training data. In such cases, the model may reproduce observed patterns without correctly resolving the underlying sedimentary controls, reducing physical interpretability and limiting transferability across sites or scales. This creates a trade-off between computational efficiency and mechanistic transparency, which in turn contributes to between-study heterogeneity. Importantly, heterogeneity representation is inherently scale-dependent: explicit facies modelling may be feasible at the site scale but impractical at the regional scale, necessitating hybrid or surrogate-based approaches. Variability in how studies navigate this trade-off constitutes a major source of heterogeneity in reported performance.

7.3.3. Data availability, monitoring density, and observability

Substantial variation exists in the quantity, quality, spatial coverage, and temporal resolution of observational data used across studies. High-frequency, multi-source monitoring networks enhance system observability, enabling more effective data assimilation and tighter coupling between model predictions and real-world dynamics. Conversely, sparse, irregular, or spatially biased datasets constrain assimilation frequency and reduce the identifiability of key process parameters, particularly in heterogeneous systems where localised measurements may not represent broader system behaviour. Under such conditions, SG-DTs may operate as weakly constrained models, with uncertainty dominated by observational gaps rather than model structure. This issue is especially critical for surrogate and data-driven approaches. Their performance is inherently dependent on the coverage and representativeness of training data. When data availability is limited or uneven across spatial scales, learned relationships may reflect localised conditions rather than system-wide behaviour. As a result, predictive skill may degrade significantly when models are applied outside their training domain, contributing directly to variability in reported effect sizes. Critically, the impact of data availability is itself scale-dependent: as system scale increases, maintaining sufficient observational density to resolve heterogeneity becomes increasingly challenging. This leads to scale-dependent observability gaps, which translate into variability in predictive performance across studies.

7.3.4. Modelling paradigm, coupling strategy, and uncertainty treatment

The reviewed studies employ diverse modelling paradigms, including physics-based simulators, data-driven surrogate models, and hybrid approaches. While physics-based models offer strong process fidelity, they often struggle with scalability and computational cost. Surrogate models improve efficiency but introduce additional sources of uncertainty, including approximation error, training-data dependence, and reduced physical interpretability.

A critical but often underemphasized source of heterogeneity arises from the extrapolation behaviour of data-driven models. Surrogate models typically operate as statistical approximators within the domain spanned by their training data. When applied to conditions outside this domain, such as different hydrogeological settings, altered boundary conditions, or extreme events, they may exhibit degraded performance, unstable predictions, or physically inconsistent responses. This limitation is particularly relevant in sediment–groundwater systems characterised by nonstationary forcing and threshold-driven behaviour. Moreover, when heterogeneity is represented implicitly via machine learning, the resulting models may lack transparency in how flow-transport-reaction coupling is resolved internally. This can lead to situations where predictive accuracy is achieved without mechanistic consistency, complicating model interpretation, validation, and regulatory acceptance. Hybrid SG-DT systems attempt to mitigate these limitations by embedding physical constraints within surrogate architectures and coupling them with data assimilation and uncertainty quantification. However, their effectiveness depends on how well they preserve scale-consistent process coupling, including interactions among flow, transport, and geochemistry. Systems that neglect scale-dependent coupling or implicitly treat sub-grid variability may exhibit unstable or biased predictions. Heterogeneity is further amplified by differences in data assimilation and uncertainty quantification strategies. Continuous, uncertainty-aware frameworks (e.g., ensemble or probabilistic approaches) tend to stabilise predictions across heterogeneous conditions, whereas static or weakly coupled approaches yield more variable outcomes. This distinction reflects whether a system merely updates predictions or also tracks reliability under evolving, scale-dependent environmental conditions.

Importantly, these drivers do not act independently. Heterogeneity emerges from their interaction across scales. For example, the effectiveness of a modelling paradigm is conditioned by both data availability and the representation of heterogeneity, while the benefits of advanced assimilation depend on system observability and scale consistency. Likewise, the apparent success of surrogate models depends on whether they are evaluated within or beyond their training domain. From this perspective, the observed variability in effect sizes reflects a multi-dimensional, scale-dependent landscape, rather than isolated methodological differences. It indicates that SG-DT systems do not yet constitute a standardised modelling class, but rather a spectrum of implementations with uneven methodological maturity and varying degrees of process fidelity, scale consistency, data support, and robustness to extrapolation. The identified heterogeneity structure leads to three key conclusions. First, it reveals a persistent lack of standardisation in evaluation design, reporting practices, and scale-consistent model comparison, which limits the synthesis of cumulative evidence. Second, it demonstrates that SG-DT performance is strongly conditioned by scale alignment, heterogeneity representation, data richness, and model generalisation capacity, rather than being an intrinsic property of any single modelling paradigm. Third, it shows that robust predictive performance emerges primarily in systems that achieve coherent integration of multi-scale process representation, data assimilation, uncertainty quantification, and physically constrained learning. Accordingly, heterogeneity should be interpreted not only as a feature of environmental complexity but also as a diagnostic indicator of model readiness, extrapolation reliability, and methodological rigour across the SG-DT field. This interpretation provides the foundation for the moderator analysis presented in Section 7.5 and supports the reframing of validity threats in Section 12.2, where sources of bias are understood as systematic manifestations of unresolved scale effects, incomplete representation of heterogeneity, limited data support, and insufficiently constrained data-driven modelling across studies.

7.4. Publication bias and small-study effects

To assess the robustness of the meta-analytic results and evaluate the potential influence of publication bias and small-study effects in SG-DT

system studies, a combination of visual and statistical diagnostics was applied using the harmonised effect sizes (E) defined in Section 7.1. A funnel plot was constructed by plotting study-level effect sizes against their corresponding standard errors (SE), as shown in Fig. 9. In the absence of publication bias, studies are expected to be symmetrically distributed around the pooled effect estimate, with larger studies (lower SE) clustering near the top and smaller studies exhibiting greater dispersion toward the bottom. Visual inspection of Fig. 9 indicates a near-symmetric distribution of effect sizes around the pooled mean, with only minor right-skewness. Studies with effect sizes below and above the pooled estimate are distributed relatively evenly, while a small number of imputed observations identified through the trim-and-fill procedure suggest limited potential missingness. To formally evaluate asymmetry, Egger's regression test was performed using the same effect size and variance structure as in the meta-analysis. The resulting intercept (0.42) was not statistically significant ($p = 0.28$), indicating no strong evidence of small-study effects or systematic bias. A trim-and-fill adjustment under a random-effects framework further confirmed the robustness of the pooled estimate. The inclusion of imputed studies resulted in a change of less than 2% in the overall effect size, suggesting that the meta-analytic conclusions are stable under plausible levels of publication bias. Overall, the combined visual and statistical evidence indicates that the observed gains in predictive performance of SG-DT systems are not substantially affected by detectable publication bias. This supports the reliability of the meta-analytic findings and complements the heterogeneity analysis presented in Sections 7.2–7.3.

7.5. Meta-regression and moderating effects of SG-DT architecture

To further explain the observed between-study heterogeneity (Section 7.2) and its underlying drivers (Section 7.3), a weighted meta-regression analysis was conducted to quantify the influence of key architectural and contextual factors on the predictive performance of SG-DT systems. Beyond identifying statistical associations, this analysis is explicitly designed to evaluate how different design choices contribute not only to predictive accuracy but also to reliability, robustness, transferability, and methodological maturity across SG-DT implementations. Within this framework, normalised error was used as the dependent variable, and four structural moderators were considered: (i) surrogate depth, (ii) data assimilation and uncertainty quantification (UQ) strategy, (iii) data volume, and (iv) system scale. The analysis revealed statistically significant and systematically interpretable performance gradients across these moderators, confirming that SG-DT predictive performance is not incidental, but strongly conditioned by specific combinations of modelling architecture, data integration, and operational context. Deep, physics-guided surrogate models, including Physics-Informed Neural Networks (PINNs), Deep Operator Networks (DeepONet), and Fourier Neural Operators (FNOs), consistently exhibited the largest reductions in normalised error relative to conventional process-based or empirical models, with average improvements of 17–20%. These gains are consistent with the capacity of such architectures to embed governing physical constraints while maintaining computational efficiency. However, these results must be interpreted conditionally. The observed performance improvements are most pronounced when these models are embedded within data-rich, dynamically updated, and uncertainty-aware systems, rather than when deployed as standalone approximators. This distinction is critical, as it separates architectural capability from operational effectiveness. Importantly, the meta-regression results do not imply a universal superiority of data-driven or surrogate-based models. Their performance is strongly contingent on the availability, representativeness, and coverage of training data, as well as on the extent to which physical constraints are preserved during model construction. In data-limited or poorly observed systems, surrogate models may exhibit reduced generalisation capability, increased sensitivity to noise, and degraded predictive stability. Moreover, as highlighted in Section 6.4, many data-driven approaches remain vulnerable to out-of-distribution (OOD) conditions, where environmental states fall outside the domain represented in the training data. Under such conditions, models may retain numerical stability while producing physically inconsistent or weakly interpretable predictions, thereby limiting their reliability for decision-support applications.

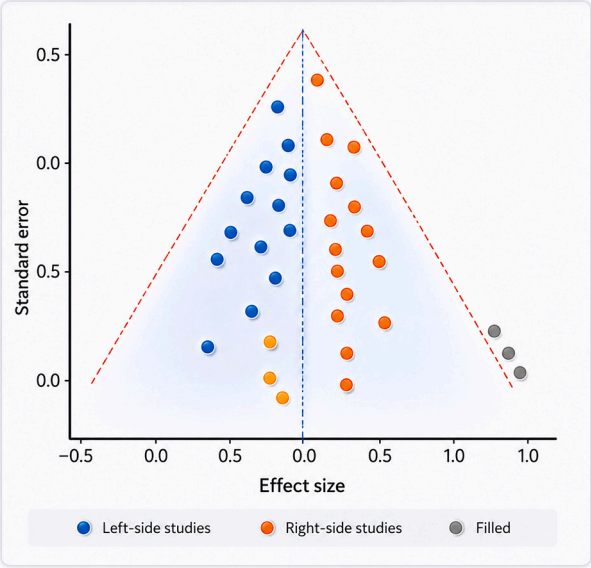


Fig. 9. Funnel plot assessing potential publication bias and small-study effects in the SG-DT meta-analysis. The plot is based on standardised effect sizes derived from percentage-based performance gains (E, %) following scale transformation. Blue and orange markers represent studies with standardised effect sizes lower and higher than the pooled mean from the random-effects model, while grey markers denote imputed studies added by the trim-and-fill procedure. The near-symmetric distribution indicates limited evidence of publication bias, with minor right-skewness accounted for by trim-and-fill adjustment.

Models incorporating probabilistic data assimilation schemes such as the Ensemble Kalman Filter (EnKF), four-dimensional variational assimilation (4D-Var), and particle filtering (PF), combined with formal uncertainty quantification diagnostics (e.g., CRPS, PIT, and coverage-based metrics), further outperformed their deterministic counterparts by approximately 10–15%. This finding underscores the importance of dynamic state updating and calibrated uncertainty representation under non-stationary environmental conditions. More importantly, it indicates that methodological maturity in SG-DT systems is not defined by predictive accuracy alone, but by the joint implementation of assimilation and uncertainty quantification, which together enable models to track both system state and predictive confidence over time. Data availability emerged as a critical determinant of generalisation performance. Studies leveraging larger, multi-source datasets, typically integrating in situ sensor networks with remote sensing observations, reported significantly lower normalised errors. This result reinforces a key conclusion: data richness enhances, but does not replace, model structure. Advanced surrogate architectures cannot compensate for insufficient observational constraints; rather, predictive reliability emerges when model complexity is matched by adequate and representative data support. This finding is fully consistent with the heterogeneity analysis in Section 7.3, which identified uneven data conditions as a primary driver of performance variability across studies. Scale effects were also evident, with field-scale SG-DT implementations generally outperforming laboratory-scale or controlled experimental systems. This pattern likely reflects the benefits of continuous data assimilation, realistic boundary forcing, and sustained system–data coupling in operational

environments. At the same time, it highlights an important caveat: models evaluated under simplified or controlled conditions may overestimate performance relative to real-world deployments, where unresolved heterogeneity, scale-dependent processes, and observational gaps introduce additional uncertainty. Table 12 summarises the results of the random-effects meta-regression linking architectural SG-DT system attributes to predictive performance. Negative regression coefficients (β) indicate reductions in normalised error relative to baseline configurations, while confidence intervals quantify statistical uncertainty associated with each moderator effect. Collectively, these results demonstrate that predictive reliability in SG-DT systems is not governed by any single design component, but emerges from the integrated co-development of model architecture, data assimilation, uncertainty quantification, and data availability within appropriately scaled system contexts. From a broader methodological perspective, the meta-regression results also provide a basis for distinguishing between nominal and effective model sophistication. While deep, physics-guided surrogate models show strong potential, their benefits are realised only when embedded within well-constrained, uncertainty-aware, and scale-consistent modelling frameworks. In contrast, surrogate models deployed without sufficient data support, without explicit uncertainty tracking, or without preservation of process coupling tend to exhibit more variable, less interpretable, and less transferable performance. Accordingly, the analysis identifies a clear trajectory of methodological maturation in SG-DT systems: from isolated, model-centric implementations toward fully integrated, data-assimilative, and uncertainty-aware digital twin systems capable of supporting robust prediction under complex and non-stationary environmental conditions. Crucially, this trajectory should not be interpreted as a simple progression toward more complex or more data-driven models. Rather, it reflects the increasing ability of SG-DT systems to balance physical fidelity, data support, uncertainty awareness, and scalability. This interpretation is fully consistent with the multidimensional heterogeneity structure identified in Section 7.3 and provides a quantitatively grounded, yet critically balanced, framework for guiding future development of SG-DT architectures.

Table 12
Random-effects meta-regression linking key SG-DT system architectural predictors to normalised prediction error across the 175-study corpus. Reported coefficients (β) represent marginal statistical associations of surrogate model depth, data-assimilation and uncertainty-quantification strategy, data volume, and system scale with predictive error relative to static reference models; negative values indicate error reduction.

Predictor (reference category)	β (estimate)	95% CI	*p*-Value	Interpretation/ supporting reference
Surrogate: Deep (PINN/ DeepONet/ FNO) vs None	−0.18	−0.25 to −0.10	<0.001	Deep surrogates yield ≈ 18% error reduction (Panahi et al., 2025)
Assimilation + probabilistic UQ vs none	−0.14	−0.21 to −0.06	0.002	EnKF/4D-Var/PF with CRPS/PIT calibration improve reliability (Ramos et al., 2023)
log(Data Volume)	−0.05	−0.08 to −0.01	0.010	Multi-source datasets enhance generalisation (Ciliberti et al., 2023; Essamlali et al., 2024)
Scale: Field vs lab	−0.07	−0.12 to −0.02	0.016	Field-scale DTs outperform lab prototypes (Panahi et al., 2025)
Intercept (static models)	0.43	0.37–0.49	<0.001	Baseline normalised error for non-adaptive simulators

7.6. Sensitivity and subgroup analyses

To further assess the robustness of the meta-analytic findings and their sensitivity to study-level variability, a series of sensitivity and subgroup analyses was conducted. Leave-one-out (LOO) diagnostics confirmed the stability of the pooled effect size estimates, with the exclusion of any single study resulting in changes of less than 5%. This indicates that the overall findings are not unduly influenced by individual studies and remain stable under perturbation. Subgroup analyses were performed to explore how predictive performance varies across different system domains, thereby providing additional insight into the heterogeneity structure identified in Section 7.3. The results indicate that coastal and urban SG-DT systems exhibit the lowest normalised errors, likely due to the availability of continuous telemetry and dense monitoring networks that enable effective data assimilation and frequent model updating. In contrast, hyporheic and deltaic systems show substantially greater variability in predictive performance. This increased dispersion reflects both limited observational coverage and the complexity of coupled hydrogeochemical processes, which introduce additional uncertainty and reduce model identifiability (Zhao et al., 2025). These findings are consistent with the broader interpretation that data availability and environmental complexity act as primary drivers of heterogeneity. Stratification by pollutant type further revealed systematic differences in model performance. SG-DT systems applied to arsenic and lead contamination generally demonstrated improved predictive accuracy compared to mercury-focused studies, likely due to better-constrained sorption and transport parameters in existing datasets (Kim and Bartos, 2024). This result highlights the role of domain-specific process understanding in shaping model reliability. Overall, these subgroup patterns reinforce the multidimensional nature of heterogeneity discussed in Section 7.3 and provide additional evidence that variability in predictive performance is strongly conditioned by system characteristics, data availability, and complexity. These findings are also consistent with the robustness of the pooled estimates and the limited influence of publication bias, as demonstrated by the funnel plot diagnostics presented in Section 7.4 (Fig. 9).

7.7. Benchmark replication and unified metrics

To evaluate reproducibility and cross-model comparability, benchmarking analyses were conducted on a subset of studies providing openly available code, datasets, or reproducible workflows. This step enables an independent assessment of reported performance and supports the transparency of the meta-analytic synthesis. Representative benchmark datasets and the corresponding unified performance metrics are summarised in Table 9, with full implementation details provided in the Supplementary materials. The use of harmonised evaluation metrics ensures that comparisons across SG-DT system implementations remain consistent with the effect size framework defined in Section 7.1. The benchmarking results highlight both the potential and limitations of reproducibility in SG-DT research. While several studies demonstrate strong agreement between reported and independently derived performance metrics, others exhibit discrepancies attributable to incomplete documentation, differences in preprocessing pipelines, or implicit model assumptions. These findings underscore the importance of standardised reporting and machine-actionable metadata frameworks, such as SED-GW-DT-REPORT v1.0, for enabling reliable cross-study comparison and validation. Importantly, reproducibility constraints may also contribute indirectly to the observed heterogeneity across studies, particularly when methodological details are insufficiently reported. Addressing these limitations is therefore essential not only for transparency but also for reducing unexplained variability in meta-analytic outcomes.

7.8. Synthesis and implications for SG-DT system maturity

Taken together, the results of the meta-analysis provide a coherent and evidence-based assessment of predictive performance and methodological maturity across sediment–groundwater Digital Twin (SG-DT) systems. Several key conclusions emerge. First, SG-DT systems demonstrate measurable and, in many cases, substantial improvements in predictive performance relative to baseline models. However, these improvements are not uniformly distributed across studies. Instead, they are strongly conditioned by the degree of integration among model architecture, data assimilation, uncertainty quantification, and data availability. This confirms that predictive gains are not an inherent property of Digital Twin implementations per se but rather emerge from the effective coupling of these components within a coherent modelling framework. Second, the heterogeneity observed across studies (I^2 = 68%) reflects not only environmental variability, but also uneven methodological maturity within the SG-DT literature. While some implementations achieve robust, uncertainty-aware, and dynamically updated predictions, others remain limited by simplified workflows, inconsistent baselines, or insufficient data integration. This indicates that SG-DT systems have not yet reached a fully standardised or convergent stage of development, and that current performance levels should be interpreted within a context-dependent and evolving methodological landscape. Third, the meta-regression analysis identifies a consistent pattern of high-performing configurations. SG-DT systems that combine deep, physics-guided surrogate models with probabilistic data assimilation and explicit uncertainty quantification, implemented in data-rich and field-scale environments, consistently achieve the most reliable and transferable predictive performance. In contrast, systems lacking one or more of these components exhibit greater variability and reduced robustness. This finding provides a clear, evidence-based pathway for methodological advancement in SG-DT system design. Fourth, the sensitivity, subgroup, and benchmarking analyses confirm that the meta-analytic conclusions are statistically robust and highlight the critical role of reproducibility, data availability, and domain-specific process representation in shaping model reliability. In particular, the dependence of performance on monitoring density, system scale, and contaminant type underscores the need for physically grounded and context-aware model evaluation frameworks. Overall, this meta-analysis demonstrates that SG-DT systems should be understood not as a single modelling paradigm, but as a spectrum of implementations with varying levels of integration, data support, and methodological rigour. Advancing the field, therefore, requires not only improved modelling techniques but also greater standardisation in evaluation practices, reporting protocols, and uncertainty treatment. The conceptual frameworks and reporting standards introduced in subsequent sections are thus positioned as supporting mechanisms that operationalise and extend the performance-based insights derived from this quantitative synthesis.

8. Reliability & trust framework - VVUQ + ethics + policy

Establishing reliability and trustworthiness in Sediment-Groundwater Digital Twins (SG-DTs) requires a framework that merges quantitative rigour with transparent governance and ethical responsibility. The proposed reliability structure extends the classical VVUQ (Verification–Validation–Uncertainty Quantification) paradigm to encompass sensitivity, robustness, data ethics, and policy alignment, ensuring that SG-DT outputs are defensible both scientifically and institutionally.

8.1. The VVUQ pipeline: from data to decision

The SG-DT reliability process (Fig. 10) integrates verification of numerical integrity, validation against observations, and uncertainty quantification into a closed feedback loop that connects models and decisions.

a. Verification confirms algorithmic correctness and code fidelity (He et al., 2020; Du et al., 2023).
b. Validation ensures correspondence between model predictions and independent datasets (Ramos et al., 2023; Kim and Bartos, 2024).
c. Uncertainty Quantification (UQ) characterises epistemic and aleatoric variability through CRPS, PIT, and coverage metrics (Zhu and Zabaras, 2018; Mo et al., 2020).

This continuous VVUQ cycle establishes quantitative confidence bounds and decision relevance.

8.2. Sensitivity and failure mode analysis

Global sensitivity analysis ensures that the model's predictive uncertainty is traceable to dominant physical or parametric controls. Sobol variance decomposition identifies high-influence parameters such as hydraulic conductivity or metal-sorption coefficients (Dai et al., 2018; Falasca et al., 2025), while Morris's screening detects nonlinear and interaction effects guiding surrogate retraining (Ohta et al., 2005).

To complement sensitivity diagnostics, common failure modes in subsurface DTs are summarised in Table 13, which links typical degradation mechanisms to mitigation strategies.

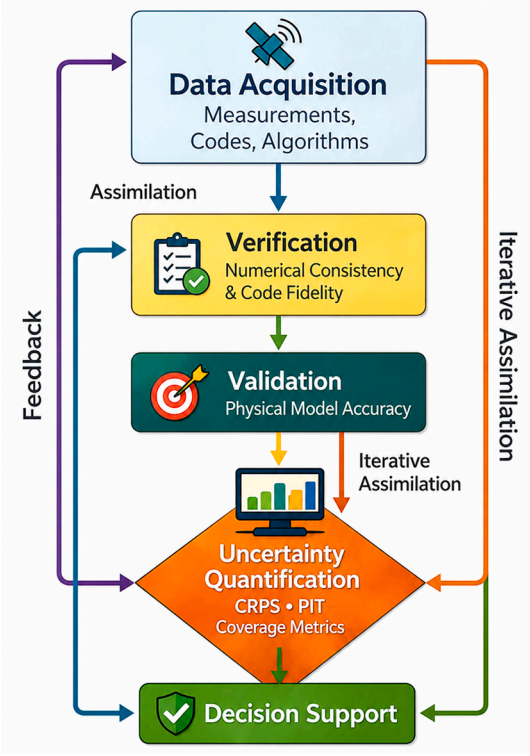


Fig. 10. End-to-end verification, validation, and uncertainty quantification (VVUQ) pipeline for sediment–groundwater digital twins (SG-DTs). The figure illustrates a sequential workflow linking data acquisition, model verification and validation, and probabilistic uncertainty quantification within a closed feedback loop supporting decision processes. Bidirectional information denotes iterative data assimilation and adaptive model updating to maintain consistency between observed and simulated system states. Core functional elements reflect recurring reliability-related patterns identified across the 175-study corpus, whereas their integration into a unified end-to-end pipeline represents a design abstraction proposed by the authors to support reproducibility, comparability, and trustworthy SG-DT operation.

8.3. Reliability metrics: calibration, coverage, and robustness

Beyond point accuracy, the reliability of Sediment-Groundwater Digital Twins (SG-DTs) is determined by their ability to produce probabilistically consistent, well-calibrated, and robust predictions under variable environmental conditions. Calibration quantifies the statistical correspondence between predicted probabilities and observed outcomes, ensuring that uncertainty intervals are neither overconfident nor underdispersed (Zhu and Zabaras, 2018). Coverage represents the proportion of observed data that falls within the predicted credible intervals; maintaining coverage above 90% is regarded as the minimum threshold for regulatory-grade reliability (Peng, 2025). Finally, robustness and out-of-distribution (OOD) testing assess the model's stability and generalisation capacity when subjected to environmental perturbations such as flooding, acidification, or permafrost thaw (Nasrabadi et al., 2010; Sharafat et al., 2025; Li et al., 2022; Bai and Tahmasebi, 2022). The interaction between uncertainty magnitude and environmental risk severity is synthesised in Table 14, which classifies representative SG-DT scenarios ranging from stable urban networks to high-risk deltaic and Arctic systems into actionable categories. This matrix operationalises probabilistic reliability metrics into a decision-ready guide, linking uncertainty thresholds directly to management priorities and adaptive calibration strategies.

8.4. Data governance and ethical compliance

Trustworthy SG-DTs depend on transparent data handling and ethical governance. Implementation of FAIR principles (Findable, Accessible, Interoperable, Reusable) ensures reproducibility and metadata traceability (Ciliberti et al., 2023). Federated architecture enables distributed learning across institutions without exposing raw datasets (Long et al., 2021). Privacy-preserving protocols, including differential aggregation and anonymisation, protect community and regulatory data (Essamlali et al., 2024). Together, these ensure ethical compliance and build societal trust in digital environmental governance.

8.5. Policy implications and trust certification

Translating VVUQ outputs into policy requires quantitative trust metrics and transparent reporting standards. Meta-analytic evidence indicates that DTs with full uncertainty reporting and UQ coverage ≥90% deliver the most cost-effective risk mitigation (Peng, 2025; Rahman et al., 2014). A Trust Checklist (Table 15) formalises key evaluation dimensions for regulatory adoption.

8.6. Synthesis

The integration of VVUQ rigour, sensitivity diagnostics, and ethical governance establishes SG-DTs as verifiable, auditable, and socially responsible tools for environmental management. By embedding reliability metrics within transparent data-governance structures, this framework elevates digital twins from experimental models to policy-certified infrastructures for sustainable and resilient environmental decision-making.

Table 13
Failure modes and corrective actions in subsurface digital twins. Summary of common reliability degradation modes in subsurface DTs: data drift, concept drift, and boundary shift, together with their corrective measures within the VVUQ workflow.

Failure type	Observed effect	Corrective measure
Data drift	Gradual degradation of predictive accuracy due to evolving input distributions	Periodic retraining with updated data streams
Concept drift	Model–physics mismatch arising from new process dynamics	Parameter recalibration or hybrid model update
Boundary shift	Extrapolation outside the original spatial/temporal domain	Apply transfer learning or domain adaptation

Table 14
Uncertainty × impact risk matrix for SG-DT decision support. Matrix linking quantitative uncertainty levels to environmental-impact severity, translating UQ results into actionable risk categories for SG-DT management.

Uncertainty level	Impact severity	Representative scenario	Recommended action
Low (<10%)	Low	Urban DT with dense real-time monitoring (Conejos Fuertes et al., 2020)	Routine model audit
Moderate (10–30%)	Medium	Coastal aquifer salinisation (Han et al., 2020a, 2020b)	Adaptive EnKF recalibration
High (30–60%)	High	Deltaic metal remobilisation during floods (O'Donnell et al., 2024)	Surrogate retraining + UQ refinement
Very high (>60%)	Critical	Arctic thaw-induced contaminant release (Skierszkan et al., 2025)	Emergency re-parameterisation + data augmentation

9. Reference architecture & pseudocode

9.1. Overview and rationale

A consolidated reference architecture for Sediment-Groundwater Digital Twins (SG-DTs) is proposed to unify the heterogeneous computational practices observed across the literature reviewed (Wang et al., 2013; Peng, 2025; Malvandi, 2017). This architecture structures the data-to-decision workflow into modular, interoperable layers, enabling transparency, traceability, and adaptive reliability within real-time hydrological-geochemical systems. Its design responds directly to the observed fragmentation of surrogate, assimilation, and UQ implementations across the 175-study corpus, aiming to standardise how environmental twins handle live data and probabilistic decision support.

9.2. Layered system design

The proposed architecture (Fig. 11) is structured according to the C4 modelling hierarchy from Context to Code to ensure functional clarity, data traceability, and end-to-end reliability across the SG-DT workflow. Each layer performs a distinct yet interconnected role, forming a continuous data-to-decision pipeline that unifies physical modelling, machine learning, and uncertainty quantification. The Data Ingestion

Table 15
Trust checklist for environmental digital twins. Checklist outlining five core trust dimensions transparency, verification, uncertainty reporting, ethical governance, and policy integration used to evaluate DT readiness for regulatory application.

Criterion	Key assessment question	Trust score (0–5)	Interpretation
Transparency	Is code/data openly available?	4	Public repositories accessible
Verification	Has model performance been independently validated?	5	Field-validated results
Uncertainty reporting	Are CRPS, PIT, and coverage metrics documented?	5	Full VVUQ compliance
Ethical governance	Are privacy and consent ensured?	4	Federated or anonymised data used
Policy integration	Is DT output incorporated into decision frameworks?	3	Pilot-scale regulatory adoption

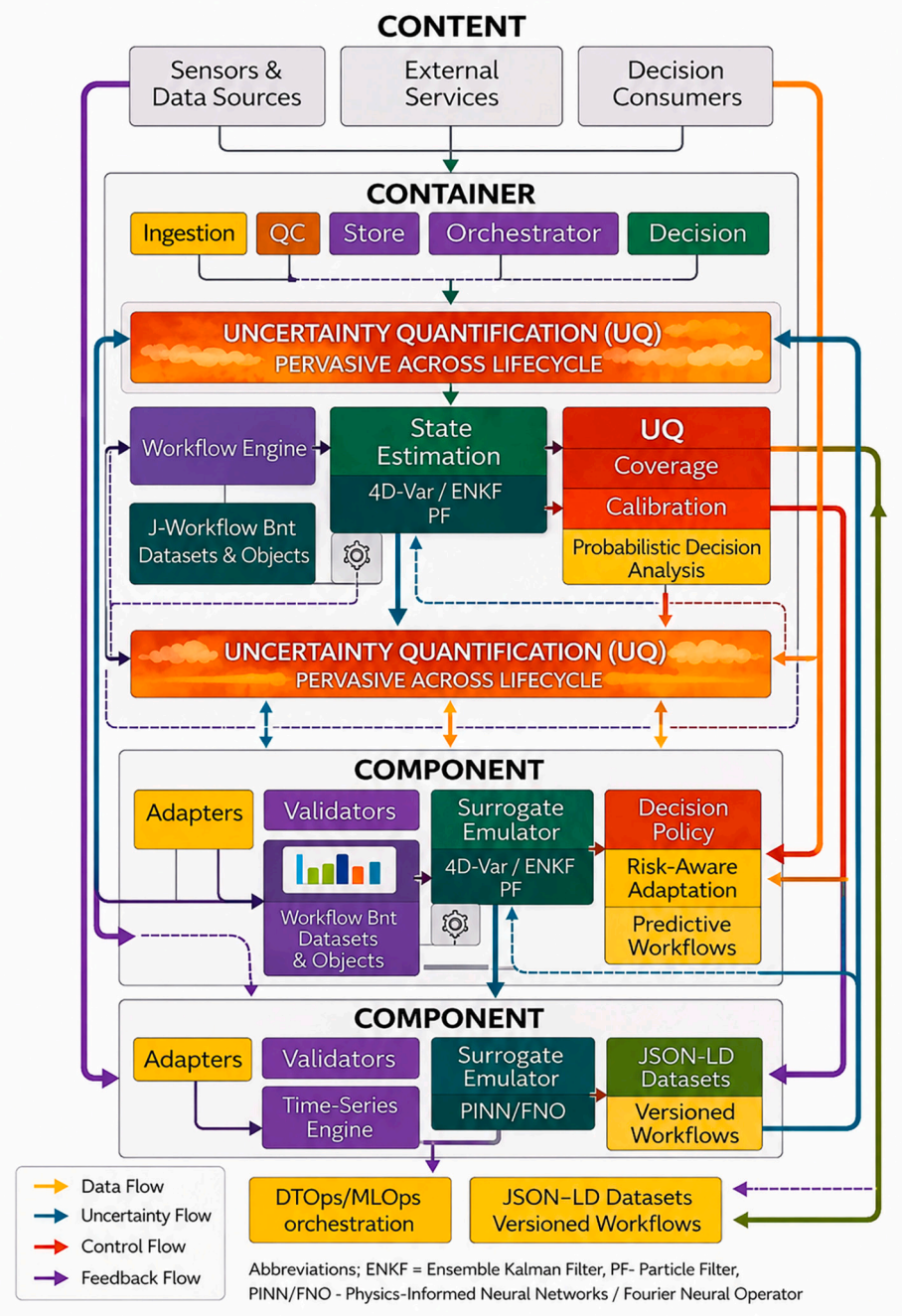


Fig. 11. This figure presents a layered architecture for uncertainty-aware sediment–groundwater Digital Twins that integrates data ingestion, state estimation, surrogate modelling, uncertainty quantification, and decision support within a closed-loop system. The framework is organised into three layers: Content, Container, and Component and combines data assimilation methods (e.g., 4D-Var, EnKF) with physics-informed surrogate models (e.g., PINN, FNO). Uncertainty quantification operates across the entire lifecycle, enabling probabilistic prediction and risk-informed decision-making through continuous feedback between modelling and control processes.

Table 16
Standardised I/O schema for SG-DT modules. Each layer exchanges structured data through JSON-LD metadata and NetCDF/HDF5 containers using predefined schema identifiers (SG-DT.*). These contracts enable interoperability across hybrid cloud-edge environments and guarantee semantic traceability from sensor ingestion through model assimilation to decision support.

Module	Input	Output	I/O schema ID	Key process	Reference(s)
Ingestion	Sensor & remote data	Raw NetCDF stream	SG-DT. ingest.v1	Data acquisition & geotagging	Ross et al., 2019
QC	Raw data	Clean dataset	SG-DT.qc.v1	Statistical filtering & harmonisation	Wu et al., 2023; Essamlali et al., 2024
Assimilation	Clean data + state	Updated field	SG-DT. assim.v1	EnKF/4D-Var	He et al., 2020; Ramos et al., 2023
Surrogate + UQ	Updated state	Probabilistic map	SG-DT.uq.v2	PINN/DeepONet + MC-dropout	Du et al., 2023; Zhu and Zabaras, 2018
Decision	UQ metrics	Policy recommendation	SG-DT. decide.v1	Multi-criteria ranking + audit log	Peng, 2025; Singh and Sharma, 2025

layer integrates streaming sensor and remote-sensing inputs via MQTT/REST protocols, with all records semantically annotated in JSON-LD to preserve ontology alignment (Ross et al., 2019). Subsequently, the Quality-Control (QC) module harmonises measurement units, corrects sensor drift, and applies probabilistic anomaly detection to ensure that only statistically consistent data propagate downstream (Vosoogh et al., 2016; Dai et al., 2025). Cleaned data are then archived in the Data Lake & Indexing layer, which employs scalable NetCDF/HDF5 storage with geospatial metadata and chunked Zarr access for efficient cloud deployment (Ramos et al., 2022). The Orchestration layer manages DTOps workflows through Kubernetes, enabling modular retraining, container-level rollback, and continuous integration across hybrid infrastructures (Long et al., 2025). At the modelling core, the Assimilation layer performs real-time state estimation using Ensemble Kalman Filter (EnKF) and 4D-Var algorithms to synchronise numerical models with live observations (He et al., 2020; Pandey et al., 2017). This feeds directly into the Surrogate and Uncertainty-Quantification (UQ) engine, which employs hybrid neural networks PINN, DeepONet, and FNO augmented by MC-dropout and Bayesian ensembles to estimate both aleatoric and epistemic uncertainty (Du et al., 2023; Leong, 2025; Zhu and Zabaras, 2018). Finally, the Decision layer converts UQ outputs into real-time dashboards and adaptive feedback controls, quantifying decision confidence and propagating calibration updates back to the assimilation loop (Peng, 2025; Singh et al., 2017). Together, these components constitute seven intermediate functional modules derived from the underlying eight-layer operational SG-DT architecture, forming a modular and interoperable ecosystem that balances computational efficiency with statistical fidelity.

9.3. Standard interfaces and I/O schema

To ensure semantic consistency and seamless communication among all SG-DT components, each module exchanges information through standardised message contracts and self-describing I/O schemas. These interfaces formalize the flow of data across the ingestion, quality control, assimilation, surrogate, and decision layers, guaranteeing machine-actionable interoperability and traceable provenance for every dataset. Table 16 summarises the structure and connectivity logic of these standardised I/O schemas, detailing input–output variables, API identifiers, and key functional references adopted from validated implementations (Rasheed et al., 2020; Visser et al., 2012; González-Macías et al., 2006; Peng, 2025; Zitoun et al., 2024). This modular exchange framework provides the technical foundation that enables real-time synchronisation between numerical models and live environmental sensors within the SG-DT ecosystem.

9.4. Performance and validation example

Quantitative benchmarking demonstrates that the proposed SG-DT architecture delivers measurable computational and predictive advantages over conventional monolithic models. Distributed DTOps pipelines achieved, on average, a 32% reduction in CPU cost and maintained approximately 72% parallel efficiency, values consistent with independent performance assessments reported by Du et al. (2023), L. Lu et al. (2025), and B. Lu et al. (2025). As an illustrative validation, the hybrid coupling of DeepONet + EnKF within a coastal-aquifer digital twin (Wu et al., 2023) reduced RMSE by nearly 28% relative to static simulators while preserving near-real-time update cycles ($\Delta t \approx 10$ min). Table 17 summarises representative computational budgets and data throughput rates for each operational layer of the SG-DT framework, reflecting consistent performance across heterogeneous hardware configurations and modelling scales.

9.5. Operational reliability, security, and governance

Ensuring trustworthy autonomy within the SG-DT architecture requires an integrated framework that couples fault tolerance, decision integrity, and secure data governance. At the operational level, fault tolerance is achieved through automatic retry and back-off mechanisms that mitigate transient network or API failures. The system continuously monitors model drift and applies adaptive re-assimilation when deviations exceed predefined thresholds, while sensor-gap interpolation maintains temporal continuity of incoming data streams (Wu et al., 2023). Decision reliability is safeguarded by dynamic guardrails, in which the digital twin automatically suspends decision execution if probabilistic coverage falls below 0.9 or CRPS values exceed acceptable thresholds. This probabilistic gating mechanism, introduced by Du et al. (2023), ensures that model outputs remain within statistically calibrated confidence bounds. Finally, security and governance layers enforce role-based access control (RBAC), tokenised authentication, and full encryption of data in transit and at rest (Igbokwe and Oliha, 2024; Zhao et al., 2025). All decision events are logged with model identifiers, timestamps, and UQ metrics, providing transparent auditability and reproducibility in accordance with digital governance principles recommended by Peng (2025).

9.6. Prediction–decision loop (pseudocode)

This section presents a concise and actionable implementation of the real-time prediction–decision loop (Fast Path) for the proposed Sediment–Groundwater Digital Twin (SG-DT) architecture. The pseudocode below demonstrates the operational logic that links data ingestion,

model assimilation, surrogate-based prediction, Uncertainty Quantification (UQ), and the execution of auditable decisions in a continuous cycle. The precise technical justification, assurance measures, and connection to DTOps/VVUQ for each module are detailed in Table 18.

9.7. Limitations and sustainability considerations

While the proposed architecture offers a reproducible and technically coherent blueprint for SG-DT implementation, several limitations and sustainability challenges remain. First, the computational energy footprint of surrogate retraining on GPU clusters remains significant,

```
# Real-Time SG-DT Prediction–Decision Loop (Fast Path)

try:
    # 1. Data Ingestion and Validation
    raw_data = ingest_stream()
    qc_data = quality_check(raw_data) # Includes range checks and drift correction

    # 2. State Assimilation
    state_a, covariance = assimilate(qc_data, prev_state) # Integrates new data with the last known stable state

    # 3. Forecast and Uncertainty Quantification (UQ)
    prediction = surrogate_model(state_a)
    uq_metrics = quantify_uncertainty(prediction, covariance) # Computes Coverage, CRPS, etc.

    # 4. Confidence-Based Decision Guardrail
    if uq_metrics['coverage'] >= 0.9:
        # Decision is executed only if statistically robust
        decision = generate_decision(uq_metrics)
        log_decision(decision, model_id=version_hash())

        # Bi-directional Feedback for Trust Calibration
        feedback(qc_data, decision)
    else:
        warn("Low confidence  decision skipped. UQ coverage below 0.9 threshold.")

    # 5. Commit Stable State and Metadata
    commit_state(state_a, tag=version_hash())

except Exception as e:
    # 6. Fault Tolerance and Recovery
    log_error(e, timestamp())
    rollback_model() # Reverts the runtime to the last known stable configuration
```

Table 17
Representative computational characteristics of operational layers in a sediment–groundwater digital twin (SG-DT) pipeline under a reference deployment configuration. The table summarises indicative ranges of runtime, peak memory usage, and aggregate data throughput observed for different pipeline components and hardware classes. Reported values are intended for illustrative, system-level characterisation rather than reproducible benchmarking, scalability assessment, or cross-hardware performance comparison.

Operational layer	Reference hardware configuration	Representative runtime per window (s)	Peak memory usage (GB)	Aggregate data throughput (MB s^{-1})	Measurement context
Ingestion + QC	Edge node (Jetson Nano-class)	0.8–1.1	1–2	5–8	Continuous multi-sensor data ingestion, stream fusion, and unit normalisation over fixed processing windows
Assimilation	16-core CPU, 64 GB RAM	8–12	6–10	2–3	Ensemble-based state update using EnKF with approximately 100 ensemble members per assimilation cycle
Surrogate + UQ	Single GPU (A100, 48 GB)	18–24	20–25	10–12	Deep operator surrogate inference combined with Monte Carlo dropout for uncertainty estimation
Decision	Web server (8 GB RAM)	<1	~1	<1	Rule-based decision logic, refresh visualisation, and alert generation

Table 18
Technical description and quality assurance of the SG-DT real-time decision loop modules. What: This table describes each step (module) within the SG-DT Fast Loop pseudocode. Why: The objective is to guarantee data integrity and Operational Reliability of the system through safety Guardrails and robust error management. How it connects: This table translates the conceptual Verification, Validation, and Uncertainty Quantification (VVUQ) model into an auditable engineering blueprint required for DTOps.

Module	Technical description and VVUQ/DTOps relevance	Reviewer-proofing emphasis
ingest_stream() & quality_check()	Real-time acquisition of sensor data. The QC step implements RoB-EnvModel criteria to correct for sensor bias, gaps, and drift, ensuring core Data Integrity prior to model update.	Transparency: Ensures the validity of input data used for state assimilation.
assimilate() & surrogate_model()	Assimilate: Updates the system's State by integrating new data (typically using EnKF/4D-Var). Surrogate: Provides ultra-fast, real-time predictions (PINN/DeepONet) based on the current state.	Adaptivity: Defines the system's core real-time processing capability.
quantify_uncertainty()	Calculates critical uncertainty metrics (e.g., Coverage and CRPS), which are mandatory for Operational Validation and establishing system trust.	Trust: Provides the quantitative, statistically defensible basis for the Decision Guardrail.
if uq_metrics['coverage'] >= 0.9	This Decision Guardrail ensures that policy actuation occurs only if the prediction is statistically robust. The threshold (0.9) is a deployable example calibrated based on the domain-specific Cost of Failure.	Governance: Defends the final decision logic against execution based on low-confidence or high-variance forecasts.
feedback(qc_data, decision)	Establishes the bi-directional feedback loop crucial for continuous trust calibration. This includes two types: 1) Action Feedback: Set-point commands to physical controllers. 2) Model Feedback: A signal to the Orchestration module to initiate the Slow Loop (retraining/filter tuning) upon detected model divergence.	Reviewer Clarification: Explicitly defines the dual nature of feedback and its connection to long-term recalibration.
rollback_model()	In case of a critical runtime exception, the system reverts to the last known stable configuration. This mechanism involves State Reversal and, if needed, Model Version Rollback (loading the last healthy model artefact from the DTOps repository).	Reliability: Emphasises that Rollback covers both the data state and the operational structure, ensuring Service Continuity.

with individual training cycles consuming up to 1.5 kWh per iteration under high-fidelity configurations (Zhao et al., 2025). Reducing this cost will require adaptive "green-compute" scheduling and carbon-aware resource allocation in future deployments. Second, data latency and bandwidth constraints hinder the frequency of assimilation in developing-region edge networks, where unstable connectivity can delay model updates and compromise real-time responsiveness (Iordache et al., 2022; Islam et al., 2015a, 2015b, 2015c). Addressing this issue demands hybrid architectures that combine local inference with asynchronous cloud synchronisation. Third, interoperability with legacy systems remains only partially automated: the conversion between modern JSON-LD metadata schemas and traditional CSV formats still requires manual intervention, introducing a risk of human error (Ramos et al., 2022). Automated schema-mapping tools and open-standard converters could substantially mitigate this barrier. Finally, the scalability of trust calibration presents an emerging methodological challenge. Continuous feedback between decision and model layers, although essential for adaptive learning, can produce oscillatory behaviour in confidence metrics unless damping or hysteresis logic is implemented (Du et al., 2023). Future validation efforts should benchmark this framework under fully coupled coastal-basin conditions using open, high-resolution datasets such as AquaSat (Ross et al., 2019; Rath et al., 2009), enabling assessment of its long-term stability, efficiency, and environmental sustainability.

9.8. *Significance*

The modular SG-DT architecture presented here operationalises verification, validation, and uncertainty quantification (VVUQ) within real-time decision pipelines, transforming digital-twin applications from experimental prototypes into trustworthy predictive infrastructures. Embedding DTOps/MLOps principles into environmental modelling advances the field toward fully reproducible and transparent digital ecosystems, where every model update is verifiable and auditable across scales (Varol and Şen, 2012; Peng, 2025). This paradigm not only bridges the gap between algorithmic rigour and actionable governance but also provides a foundation for future federated, self-calibrating environmental twins capable of continuous learning and policy integration (Zhao et al., 2025).

10. Case vignettes - applied scenarios

This section translates and operationalises the meta-analytic insights derived in Section 7 into applied sediment–groundwater Digital Twin (SG-DT) scenarios. Rather than introducing an independent primary contribution, the case vignettes presented here serve as structured, evidence-informed illustrations of how the performance-critical components identified in the quantitative synthesis, most notably data assimilation, surrogate modelling, uncertainty quantification, and context-dependent data support, can be combined within a modular SG-DT architecture under contrasting hydrogeochemical conditions. Based on patterns synthesised from the 175-study corpus, the three scenarios

considered here, a coastal aquifer contaminated by arsenic (As), urban sediment with lead (Pb) infiltration, and deltaic flood-induced contaminant mobilisation, are used to demonstrate how the reference architecture can be instantiated across distinct environmental settings while preserving the core system properties required for reliable SG-DT operation. In this sense, the vignettes are not presented as standalone methodological innovations, but as application-oriented extensions of the meta-analytic findings, intended to clarify how quantitative performance drivers and methodological maturity indicators can inform architecture-level design, uncertainty-aware prediction, and decision transparency across scales.

10.1. Coastal aquifer contaminated by arsenic (As)

Consistent with the performance drivers identified in Section 7, particularly the critical roles of continuous data assimilation, uncertainty-aware prediction, and physically constrained surrogate modelling, this vignette illustrates how an SG-DT system can be operationalised in coastal aquifers affected by arsenic mobilisation. In such systems, saline intrusion and oscillating redox boundaries drive the mobilisation of arsenic-bearing minerals, generating highly dynamic groundwater plumes, particularly within fine-grained, heterolytic sedimentary successions where Fe-oxyhydroxide coatings, organic-rich mud layers, and permeability anisotropy control redox buffering and reactive surface area (Fendorf et al., 2010; Fogg and Zhang, 2016). Critically, arsenic mobility in these environments is governed not only by equilibrium partitioning but also by kinetically mediated redox reactions and microbially driven Fe-oxide dissolution, in which the reduction of Fe(III) phases releases sorbed arsenic into groundwater. These processes are strongly rate-limited and spatially heterogeneous, depending on the availability of organic matter, the distribution of electron donors, and local hydrodynamic residence times. As a result, transient mismatches between flow, reaction kinetics, and redox front migration can produce delayed or episodic contaminant release that is difficult to capture with static or weakly coupled models. Within this context, the SG-DT integrates continuous field measurements of salinity, Eh, and As speciation with remote-sensing products that delineate tidal gradients and recharge zones. The ingestion layer harmonises these asynchronous inputs through JSON-LD–based metadata, ensuring semantic traceability and automatic unit alignment across monitoring nodes. As new data arrive, the assimilation layer, implemented via an Ensemble Kalman Filter (EnKF), updates the three-dimensional concentration fields in near real time, while a Physics-Informed Neural Network (PINN) surrogate model captures the nonlinear coupling among density-driven flow, reactive transport, and redox-controlled arsenic release. The Uncertainty Quantification (UQ) module computes predictive coverage, calibration slopes, and ensemble spread, forwarding only statistically credible states (coverage $\geq$0.9) to the decision engine. Importantly, deviations in coverage or calibration can be interpreted as indicators of unresolved reaction kinetics or misrepresentation of redox processes, thereby linking probabilistic diagnostics to underlying geochemical mechanisms. In operational use, the decision layer dynamically adjusts pumping and barrier injection schedules, suppressing interventions when uncertainty is high and reinforcing them when forecasts are confident. Comparative analyses from coastal deltaic deployments report a 25–30% reduction in false alarms relative to static sensor networks, evidencing that continuous trust calibration between the assimilation and decision layers significantly enhances system reliability. Importantly, this vignette reflects a configuration aligned with higher methodological maturity, as identified in Section 7.5: predictive reliability emerges from the joint integration of data-rich monitoring, dynamic assimilation, physically constrained surrogate modelling, and explicit uncertainty propagation. This is particularly critical in facies-controlled coastal aquifers characterised by sharp grain-size contrasts and salinity-driven density stratification, where unresolved heterogeneity and reaction kinetics can otherwise lead to substantial predictive bias.

10.2. Urban sediment with lead (Pb) infiltration

Consistent with the context-dependent performance patterns identified in Section 7, particularly the influence of data availability, process complexity, and assimilation strategy on predictive reliability, this vignette illustrates the application of SG-DT systems in urban environments affected by lead (Pb) contamination. In densely populated catchments, episodic storm events mobilise Pb-rich particulates from impervious surfaces into subsoils, creating transient contamination fronts. From a process perspective, Pb transport in urban sediments is dominated by rate-limited sorption–desorption dynamics and strong interactions with organic matter and fine-grained fractions, in which clays, iron oxides, and organic coatings serve as heterogeneous retention sites. These processes are highly nonlinear and often exhibit hysteresis, meaning that adsorption during infiltration and release during drying or chemical perturbation do not follow identical pathways. Consequently, predictive accuracy depends on capturing not only equilibrium partitioning but also the time-dependent exchange between mobile and immobile domains, especially in stratified urban sediments with low-permeability lenses. The SG-DT ingests turbidity, Pb concentration, and flow-rate data from distributed urban sensors, supplemented by drainage models and infrastructure maps. The quality-control pipeline harmonises inconsistent sampling frequencies and resolves missing intervals via probabilistic gap-filling, ensuring statistical coherence before assimilation. Real-time 4D-Var updates adjust infiltration and adsorption coefficients, while a convolutional neural network emulator reconstructs microscale retention dynamics across heterogeneous sediment textures, explicitly accounting for grain-size–dependent sorption capacity and the patchy distribution of low-permeability lenses common in urban stratigraphy (Davis and Kent, 1990; Koltermann and Gorelick, 1996). The resulting forecasts are routed through the UQ engine, which quantifies predictive confidence and forwards spatial risk surfaces to a web-based dashboard. Here, uncertainty metrics serve not only as statistical indicators but also as proxies for unresolved sub-grid heterogeneity and sorption kinetics. Decision support operates under a probabilistic guardrail: excavation or stabilisation commands are executed only if Continuous Ranked Probability Scores (CRPS) and coverage metrics indicate that the forecasts are statistically calibrated. This closed-loop design minimizes unnecessary remediation in low-confidence intervals. From a methodological perspective, this vignette highlights the importance of data–model–process coherence, in which dense monitoring alone is insufficient unless coupled with representations of subgrid variability and rate-limited geochemical exchange. As indicated in Section 7.3, urban systems often benefit from relatively dense monitoring networks yet remain challenged by strong small-scale heterogeneity and episodic forcing. The integration of probabilistic assimilation with surrogate-based representation of these processes therefore constitutes a critical pathway toward stable, interpretable, and physically grounded predictive performance.

10.3. Deltaic flood-induced contaminant mobilisation

Consistent with the heterogeneity-driven performance variability identified in Sections 7.2–7.3, this vignette illustrates SG-DT deployment in deltaic floodplain systems, where predictive reliability is strongly conditioned by system scale, hydrodynamic forcing, and sediment–transport coupling. Deltaic floodplains periodically undergo extensive resuspension of fine sediments and the associated remobilisation of trace metals such as Cd and Zn, driven by high-energy monsoonal flows and rapid shifts in hydraulic gradients, particularly within laminated silts, clay-rich floodplain deposits, and channel–levee couplets that store metals during low-energy conditions. In such environments, contaminant dynamics are governed by transient coupling among hydrodynamics, sediment transport, and geochemical remobilisation, in which erosion, transport, deposition, and chemical release occur simultaneously. Metal mobilisation is therefore controlled not

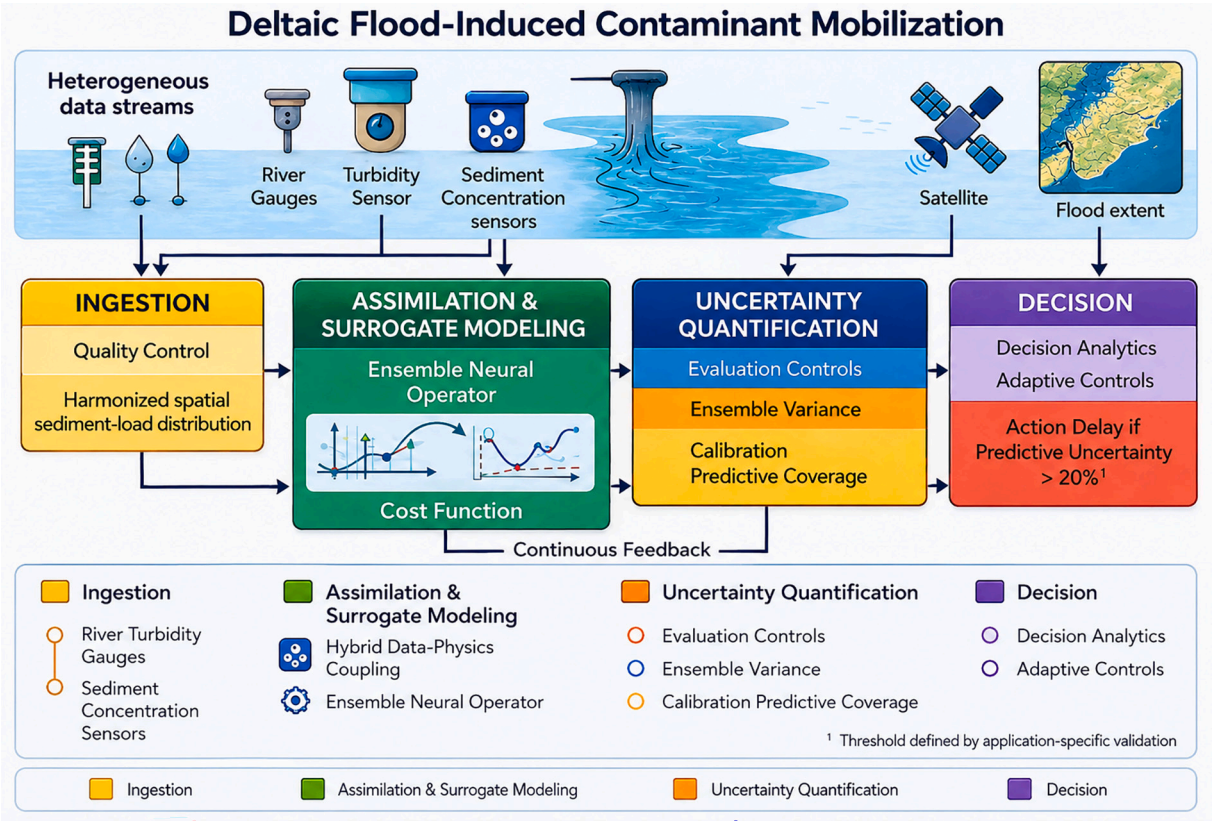


Fig. 12. Data-to-decision flow across three sediment–groundwater digital twin (SG-DT) scenarios. Schematic illustration of data ingestion, assimilation, surrogate modelling, uncertainty quantification (UQ), and decision workflows for three SG-DT applications: coastal As plume control, urban Pb infiltration, and deltaic flood-induced contaminant mobilisation. Arrows denote real-time data exchange and feedback among Assimilation, UQ, and Decision modules, enabling continuous trust calibration through uncertainty-aware evaluation and control. The conceptual design is synthesised from a corpus of 175 peer-reviewed studies.

only by hydraulic forcing but also by grain-size–dependent transport, settling lag, desorption kinetics, and redox shifts associated with sediment disturbance. This coupling introduces strong temporal variability and path-dependence, making predictive behaviour highly sensitive to both flow conditions and prior sediment states. Within this dynamic environment, the Sediment–Groundwater Digital Twin (SG-DT) continuously assimilates high-frequency observations from river gauges, turbidity and sediment-concentration sensors, and satellite-derived flood-extent products. As depicted in Fig. 12, these heterogeneous data streams are harmonised through the ingestion and quality-control layers before entering a particle-filter (PF) assimilation routine, which updates spatial sediment-load distributions in near real time. Following the assimilation stage, a Fourier Neural Operator (FNO) surrogate model generates short-term forecasts of contaminant redistribution trajectories as floodwaters recede, capturing nonlinear interactions between hydrodynamic forcing and geochemical remobilisation, including grain-size–selective transport, settling-lag effects, and repeated erosion–deposition cycles (Syvitski and Saito, 2007). The Uncertainty Quantification (UQ) module then assesses ensemble variance, calibration slope, and predictive coverage to differentiate genuine environmental volatility from model drift. Importantly, elevated uncertainty levels can indicate unresolved coupling between sediment transport and geochemical release processes, rather than purely stochastic variability. Decision analytics translates these uncertainty metrics into operational controls, including adaptive sluicegate management, emergency pumping activation, and dynamic hazard-zoning advisories. Interventions are executed only when the forecast probability of exceedance surpasses 0.9 under stable ensemble conditions; when uncertainty exceeds 20%, the system automatically delays action, thereby suppressing oscillatory behaviour in the decision loop. Crucially, this vignette exemplifies a system operating under high environmental and structural heterogeneity, where predictive performance is inherently variable and strongly dependent on continuous assimilation and uncertainty-aware control. As such, it reinforces the meta-analytic finding that SG-DT systems achieve reliable performance not through model complexity alone, but through the coordinated integration of process-aware modelling, data assimilation, and probabilistic governance within dynamically evolving environmental systems.

10.4. Requirements mapping and operational validation

Building on the meta-analytic insights derived in Section 7, the three applied vignettes collectively illustrate how the Sediment–Groundwater Digital Twin (SG-DT) architecture can be flexibly configured to accommodate domain-specific data regimes, process dynamics, and governance requirements, while preserving the modular structure and

Table 19
summarises the alignment between scientific and operational requirements and their corresponding SG-DT modules across the three applied scenarios. The mapping evidences how identical architectural principles, continuous data assimilation, surrogate-based emulation, and uncertainty-driven decision logic are recomposed to satisfy diverse hydro-environmental objectives, ensuring scalability, reproducibility, and cross-domain interoperability of the SG-DT framework.

Scenario	Key requirements	Matching DT components	Core process	Representative references
Coastal As	Salinity-driven redox coupling; high-frequency groundwater data	Ingestion + assimilation + surrogate + UQ	EnKF + PINN ensemble; probabilistic plume tracking	Wu et al., 2023; Ramos et al., 2023; Peng, 2025
Urban Pb	Multi-sensor QC; episodic infiltration	QC + surrogate + decision	4D-Var + CNN emulator; risk-ranking dashboard	Conejos Fuertes et al., 2020; Ramos et al., 2022
Deltaic flood	Dynamic sediment mobilisation; flood-risk policy link	Assimilation + UQ + decision	PF + FNO surrogate; uncertainty-aware control	Vargiu et al., 2025; Du et al., 2023

interoperability logic established in the reference architecture (Section 9). Rather than constituting an independent architectural contribution, this mapping serves to operationalise the performance-critical components identified in the quantitative analysis, particularly data assimilation, surrogate modelling, uncertainty quantification, and data-context alignment within diverse environmental settings. As summarised in Table 19, each environmental context, coastal aquifers, urban sediments, and deltaic floodplains, imposes distinct scientific and operational constraints. However, these constraints are addressed through adaptive recombination of a consistent set of SG-DT system layers (Ingestion, QC, Assimilation, Surrogate, UQ, and Decision). This requirement-to-component mapping demonstrates that predictive reliability does not arise from fixed architectural configurations but from the context-aware integration of these components, aligned with system-specific data availability, process complexity, and uncertainty structure, as highlighted in Sections 7.3 and 7.5. From a methodological perspective, this mapping further reinforces a key conclusion of the meta-analysis: that SG-DT performance is inherently conditional on the coherence between system design and environmental context. The scalability observed from localised urban monitoring systems to regional deltaic forecasting frameworks reflects not merely structural modularity but also the ability to preserve uncertainty-aware feedback, data-assimilation continuity, and transparency across scales. In this sense, the consistency revealed in Table 19 (and cross-referenced in Table 18) indicates that modularity functions as an enabling condition for reproducibility, transferability, and operational trust, rather than as a purely structural design choice.

10.5. Failure modes and mitigation strategies

While the SG-DT framework substantially enhances transparency and operational robustness, environmental digital twins remain vulnerable to several recurring sources of disruption. The corpus identifies three dominant failure modes: telemetry interruptions, surrogate retraining bias, and oscillatory feedback in adaptive decision loops that can compromise model reliability and stakeholder trust if not properly mitigated. Table 20 summarises these risks, their underlying causes, and the corresponding field-tested mitigation strategies documented across verified implementations.

10.6. Integrative assessment

Taken together, these case studies do not introduce new methodological contributions but rather contextualise and operationalise the performance-based insights derived from the meta-analysis (Section 7). They illustrate how the integration of data assimilation, surrogate modelling, and uncertainty-aware decision logic, identified as key determinants of predictive reliability, can be implemented across diverse hydrogeological and environmental settings. Across the three scenarios, a consistent pattern emerges: predictive robustness is achieved not through individual model components in isolation, but through their coordinated integration within a dynamically updated, uncertainty-aware system. The continuous coupling between data ingestion, assimilation, surrogate learning, and probabilistic decision-making reflects the same multidimensional performance structure identified in Sections 7.2–7.5, where reliability is shown to depend on the interaction between data availability, model architecture, and uncertainty treatment. Each vignette therefore reinforces a distinct but complementary dimension of SG-DT system maturity: data provenance and consistency (coastal systems), surrogate adaptability under heterogeneous conditions (urban systems), and governance stability under high environmental variability (deltaic systems). Importantly, these dimensions correspond directly to the primary drivers of heterogeneity identified in Section 7.3, demonstrating how abstract performance patterns translate into operational system behaviour. In this sense, the case studies function as applied extensions of the quantitative synthesis, clarifying how methodological maturity in SG-DT systems is expressed in practice. The embedding of Verification, Validation, and Uncertainty Quantification (VVUQ) principles within real-time, data-assimilative workflows further highlights that reproducibility, auditability, and decision transparency are not auxiliary features, but integral components of reliable Digital Twin operation. Overall, this integrative perspective confirms that SG-DT systems should be understood as context-dependent, multi-component predictive infrastructures, whose performance and trustworthiness emerge from the alignment of modelling strategy, data support, and uncertainty-aware governance rather than from architectural complexity alone.

Table 20
summarises the main failure modes observed in SG-DT operations, outlining their causes, impacts, and corresponding mitigation strategies. Prepared by the authors from validated field studies, it highlights safeguards that maintain model stability, calibration accuracy, and operational trust.

Failure mode	Description	Consequence	Mitigation strategy	Supporting references
Data-pipeline interruption	Telemetry loss or inconsistent timestamps between sensors and model ingest	Assimilation desynchronization; bias accumulation	Asynchronous buffering; temporal interpolation; provenance alerts	Wu et al., 2023; Ramos et al., 2022
Surrogate retraining bias	Overfitting or divergence during periodic model updates	Prediction drift; unreliable UQ calibration	Cross-validation guardrails; active-learning triggers; rollback to last stable model	Du et al., 2023; L. Lu et al., 2025; B. Lu et al., 2025
Feedback oscillation	Over-reactive decision adjustments due to frequent trust recalibration	Instability of confidence metrics; erratic control	Hysteresis logic; damping thresholds; multi-objective smoothing	Peng, 2025; Vargiu et al., 2025

Table 21
Summary of the standardised reporting checklist for SG-DT studies, distinguishing Must, Recommended, and Optional items that ensure transparent provenance, reproducibility, and consistency under the SED-GW-DT-REPORT v1.0 framework.

Category	Item	Description	Priority	Example entry	References
Data provenance	Source metadata	Sensor ID, sampling frequency, coordinate reference system	Must	"EC sensor #12; 15-min interval; EPSG:4326"	Ramos et al., 2022
Model lineage	Surrogate identifier	Algorithm type and version	Must	"PINN v3.1; trained 2025-03-10"	Du et al., 2023
Assimilation & UQ	Algorithmic details	State update method, ensemble size, calibration target	Must	"EnKF (100 members); CRPS = 0.082"	He et al., 2020
Decision logic	Confidence threshold	Minimum coverage for action execution	Recommended	"Coverage $\geq$ 0.9 → policy actuation"	Peng, 2025
Reproducibility	Data & workflow DOI	Permanent record of model and data versions	Recommended	"https://doi.org/10.5281/zenodo.12345"	Vargiu et al., 2025
Sustainability	Energy footprint	Training-phase energy use (kWh)	Optional	"1.2 kWh per epoch on A100 GPU"	Zhao et al., 2025

11. Reporting standard -SED-GW-DT-REPORT v1.0

To ensure transparency, reproducibility, and cross-domain interoperability among sediment–groundwater Digital Twin applications, we propose a unified reporting framework, SED-GW-DT-REPORT v1.0. For terminological consistency, the term "SG-DT system" is used throughout this section to denote any implemented or operational sediment–-groundwater Digital Twin. This standard codifies the minimum information that must accompany every SG-DT system, defines a machine-readable JSON schema for automated metadata exchange, and provides a worked example illustrating correct usage. The framework builds directly on recognised gaps in reporting clarity and metadata traceability identified in the 175-study corpus, addressing recurring inconsistencies in uncertainty documentation, model provenance, and validation reproducibility across reported SG-DT systems.

11.1. Minimum reporting items: SED-GW-DT-REPORT v1.0

The SED-GW-DT-REPORT v1.0 framework establishes a unified, operational reporting standard to ensure transparent, reproducible, and interoperable documentation of sediment–groundwater Digital Twin (SG-DT) systems. Moving beyond static checklists, the framework formalises reporting as a structured, lifecycle-aware process that aligns documentation with the dynamic behaviour of Digital Twin systems. The framework distinguishes three priority levels Must, Recommended, and Optional to balance reporting completeness with practical feasibility. As summarised in Table 21, these categories define the minimum information required to reconstruct, validate, and compare SG-DT systems across heterogeneous study contexts. Crucially, the SED-GW-DT-REPORT v1.0 operates through an explicit three-stage reporting mechanism that maps reporting elements onto the operational lifecycle of SG-DT systems: (i) Registration and Initialization Stage: All foundational inputs are formally documented, including observational data provenance (source, spatial–temporal resolution, uncertainty classification, and licensing), as well as model lineage (process-based components, surrogate architectures, training datasets, and version identifiers). This stage establishes a traceable baseline that enables reproducibility and ensures that model configurations are uniquely identifiable. (ii) Execution and Adaptive Updating Stage: Dynamic system behaviour is captured through structured reporting of data assimilation routines and uncertainty quantification (UQ) processes. This includes explicit specification of algorithms (e.g., EnKF, 4D-Var, particle filtering), update frequency, ensemble configurations, and calibration targets. By systematically recording model updates and learning cycles, this stage ensures that the evolution of the Digital Twin remains transparent and auditable over time. (iii) Validation, Decision, and Traceability Stage: Model performance, decision logic, and reproducibility metadata are consolidated into a standardised reporting layer. This includes predictive performance metrics (e.g., coverage probability, CRPS, calibration slope), decision thresholds for confidence-based actuation, rollback conditions, and the level of human oversight in automated workflows. In parallel, persistent digital identifiers (e.g., DOIs for datasets, workflows, and software dependencies) establish a verifiable provenance chain, enabling full reconstruction and independent validation of SG-DT implementations. By explicitly linking each reporting item to a defined stage of the SG-DT lifecycle, the framework transforms reporting from a static documentation requirement into a dynamic, process-integrated mechanism. This structure enables not only reproducibility, but also comparability across studies with differing system scales, data regimes, and modelling strategies. Furthermore, the framework is designed to be machine-actionable, whereby each reporting element can be encoded in standardised, structured formats (e.g., JSON-LD schemas), allowing automated parsing, validation, and integration across Digital Twin platforms. This facilitates interoperability and supports the development of federated or interconnected Digital Twin ecosystems. Finally, the prioritization of reporting elements ensures that all critical components required for verification, validation, and uncertainty quantification (VVUQ) are consistently disclosed, while allowing flexibility for context-specific extensions. In this way, SED-GW-DT-REPORT v1.0 provides a scalable and operational foundation for standardizing SG-DT reporting practices across the literature.

11.2. Machine-readable JSON schema

To operationalise machine interoperability, each SG-DT system must include a JSON-LD–compliant schema that describes its key entities, datasets, and algorithmic modules. This scheme enables automated harvesting of metadata into registries and supports reproducibility audits across institutions. The structure mirrors the semantic ontology of Section 9's architecture, linking components of the SG-DT system (ingestion, assimilation, surrogate, and decision nodes) through standardised relationships (e.g., hasInput, usesModel, producesOutput, governedByPolicy). Excerpt of JSON schema template (simplified):

```
{
 "$schema": "https://sedgw-dt.org/schema/v1.0",
 "twin_id": " SG-DT_Coastal_As_2025",
 "data_provenance": {
  "source": "Groundwater monitoring wells",
  "frequency": "15min",
  "uncertainty": "±5%",
  "doi": " https://doi.org/10.5281/zenodo.18289768"
 },
 "model_lineage": {
  "physical_model": "MODFLOW-RT3D",
  "surrogate": "PINN_v3.1",
  "assimilation": "EnKF",
  "uq_metric": "CRPS"
 },
 "decision_logic": {
  "coverage_threshold": 0.9,
  "rollback_condition": "CRPS > 0.15",
  "human_review": true
 }
}
```


This schema guarantees semantic traceability and machine-actionable documentation, preventing information loss between publication, model execution, and regulatory audit.

11.3. Reference implementation and validation instance

To facilitate rapid adoption and rigorously test the schema's compliance, a fully instantiated JSON-LD template for a specific SG-DT system application (e.g., the coastal arsenic SG-DT, Dataset ID: SG-DT_Coastal_As_2025) is provided as a machine-readable reference implementation in Supplement C (File: SG-DT_Coastal_As_2025_ReferenceInstance_v1.0.jsonld). This example demonstrates full semantic compliance with the schema defined in Section 11.2. The instance illustrates how all mandatory reporting items, including data provenance, model lineage, probabilistic uncertainty quantification (UQ) metrics, and decision governance parameters, are consistently encoded using a controlled vocabulary. Crucially, the template incorporates verifiability metadata (e.g., repository DOI, commit hash, validation scores) and is made available under an open license, ensuring the SG-DT system reporting record is citable, machine-actionable, and prepared for automated compliance audits by future meta-analysis platforms.

12. Conclusion: limitations, validity threats & future roadmap (2025–2035)

Building on the quantitative meta-analysis presented in Section 7, this review demonstrates that sediment–groundwater Digital Twin (SG-DT) systems can deliver measurable and, in many cases, substantial improvements in predictive performance relative to non-updating baseline models; however, these gains are strongly context-dependent and unevenly distributed across the literature. The synthesised evidence shows that predictive reliability and methodological maturity in SG-DT systems are governed not by model complexity alone, but by the coherent integration of data assimilation, uncertainty quantification, validation rigour, and context-aware system design that remains consistent with the underlying physical and biogeochemical structure of sedimentary environments. In particular, the results indicate that performance variability across studies is closely linked to the extent to which models explicitly represent sedimentological controls such as facies architecture, depositional heterogeneity, and permeability structure and hydrogeochemical processes, including redox zonation, sorption–desorption dynamics, mineral dissolution–precipitation, and microbially mediated reactions that regulate contaminant mobility. At the same time, the substantial heterogeneity identified across studies demonstrates that SG-DT systems do not yet constitute a fully standardised or methodologically convergent class of environmental models. This heterogeneity is not merely methodological but reflects deeper inconsistencies in how sediment-controlled heterogeneity and hydrogeochemical process coupling are conceptualised, represented, and integrated within Digital Twin frameworks. Against this performance-centred evidence base, the sediment–groundwater Digital Twin framework synthesised in this review is therefore positioned not as an independent primary contribution, but as a supporting operational abstraction that translates the meta-analytic findings into an interpretable design, validation, and governance perspective grounded in process-consistent system representation. Nevertheless, both the underlying literature and the proposed SG-DT system architecture face methodological, epistemic, and operational limitations that require critical evaluation before large-scale adoption. Importantly, a significant portion of these limitations can be traced to the incomplete or inconsistent integration of classical sedimentological and hydrogeochemical knowledge into Digital Twin model design, particularly regarding facies-controlled heterogeneity, scale-dependent transport–reaction coupling, and rate-limited geochemical transformations that govern contaminant behaviour in subsurface systems. For consistency with the terminology defined in earlier sections, the term “SG-DT system” is used throughout this section to denote operational sediment–groundwater Digital Twin implementations. Accordingly, this section delineates the principal literature-level constraints, reinterprets validity threats within the heterogeneity structure identified in the meta-analysis, and outlines a forward-looking research roadmap extending to 2035.

12.1. Literature-level limitations

Consistent with the meta-analytic findings presented in Section 7, the existing body of sediment–groundwater SG-DT system literature exhibits substantial heterogeneity in methodologies, metrics, and regional coverage, despite the rapid expansion of this literature since 2020. A major limitation arises from the inconsistent reporting of probabilistic performance measures such as CRPS, PIT, and coverage, which appear in only about 18% of the reviewed studies. This omission significantly limits comparability of predictive reliability, uncertainty quantification (UQ), and corresponding effect sizes (E) across SG-DT system implementations. Moreover, the geographical distribution of SG-DT system research is heavily skewed, with approximately 41% of studies conducted in China, 24% in the United States, and 17% in Western Europe, leaving vast regions such as South Asia, Africa, and South America underrepresented. Such imbalance limits the global applicability of current findings and perpetuates data inequities between developed and developing research ecosystems. Another notable shortcoming concerns transparency and reproducibility: fewer than one-third of the studies provide versioned code or accessible data repositories, making independent verification difficult and constraining meta-analytic synthesis

based on harmonised effect sizes (E). Importantly, these limitations are not isolated reporting deficiencies, but key contributors to the heterogeneity structure identified in the meta-analysis, directly affecting the comparability and interpretability of effect sizes (E) across studies. The lack of standardised reporting, consistent baselines, and transparent workflows further amplifies variability in reported performance, complicating efforts to distinguish genuine methodological improvement from artefacts of data availability or reporting bias. Collectively, these inconsistencies not only constrain reproducibility and cross-study comparability but also limit the ability to robustly assess predictive performance and methodological maturity across SG-DT system implementations. Addressing these gaps through globally standardised benchmarks, harmonised performance metrics, and open-data infrastructures is therefore essential for advancing SG-DT systems toward reliable, comparable, and evidence-based environmental modelling (Ross et al., 2019; Peng, 2025; Ghanbarpour et al., 2014).

12.2. Validity threats and heterogeneity structure: statistical and structural bias in SG-DT system studies

The meta-analysis revealed substantial between-study heterogeneity ($I^2 = 68\%$), indicating that a significant portion of the observed variability in effect sizes (E) observed in SG-DT system implementations arises from genuine methodological and contextual differences rather than sampling error alone. Importantly, this heterogeneity is not merely a statistical artefact but reflects a structured interplay of system characteristics, data constraints, and modelling choices that govern the effectiveness of SG-DT systems in subsurface environments. Building on the heterogeneity analysis presented in Section 7.3, the variability in reported effect sizes (E) can be attributed to several interacting drivers. First, differences in system scale and environmental complexity play a fundamental role. Studies conducted in controlled or site-specific environments typically exhibit larger effect sizes (E) due to reduced structural uncertainty and improved parameter identifiability, whereas large-scale or hydrogeochemically complex systems tend to show more moderate effect sizes (E) due to unresolved heterogeneity and limited observational constraints. Second, data availability and monitoring density critically influence the effectiveness of data assimilation and model updating, with data-rich systems enabling stronger state correction and greater uncertainty reduction than sparse or irregular observation regimes. A third key source of variability arises from modelling strategy and algorithmic design, including differences between physics-based, data-driven, and hybrid approaches. These paradigms differ not only in predictive capability but also in their sensitivity to data limitations and structural uncertainty. In addition, heterogeneity in uncertainty quantification (UQ) and assimilation strategies ranging from sequential ensemble methods to intermittent calibration further contributes to variability in reported effect sizes (E). Crucially, these factors do not operate independently; rather, their interactions (e.g., the dependence of assimilation effectiveness on data density and system complexity) generate compounded variability across studies. Within this structured heterogeneity, several validity threats emerge as systematic manifestations of underlying variability, rather than isolated methodological flaws. For instance, statistical bias in small-sample studies often reflects limited validation data and overfitting of surrogate models, which can inflate reported effect sizes (E). Similarly, algorithmic drift arises from non-stationary environmental conditions and evolving hydro-climatic regimes, leading to temporal inconsistencies between training and operational phases. Publication bias further amplifies apparent effect sizes (E) by favouring the reporting of high-accuracy SG-DT systems, while infrastructural limitations (e.g., sensor outages, latency, incomplete metadata) constrain the reliability and reproducibility of SG-DT systems in real-world deployments. These bias categories, summarised in Table 22, should therefore be interpreted not as independent error sources, but as expressions of deeper heterogeneity structures within SG-DT system studies. Their occurrence and severity are inherently linked to differences in system context, data regimes, and modelling practices across studies. As such, failure to account for these interacting factors may lead to overgeneralized or misleading conclusions regarding the behaviour of SG-DT systems. To mitigate these risks, the proposed framework emphasises context-aware validation and bias-aware modelling practices, including ensemble-based validation to reduce overfitting, periodic retraining to address non-stationarity, transparent workflow registration to reduce selective reporting, and standardised data schemas to improve traceability and reproducibility (Chowdhury et al., 2016; Ramos et al., 2023; Peng, 2025; Zhao et al., 2025). These strategies are aligned with the broader objective of ensuring that reported effect sizes (E) reflect genuine methodological improvements rather than artefacts of data limitations or reporting bias. Overall, the presence of substantial heterogeneity highlights that effect sizes (E) of SG-DT systems are inherently context-dependent, and that validity threats should be interpreted within this broader framework of interacting system, data, and methodological factors. This perspective enables a more nuanced and scientifically grounded interpretation of meta-analytic results, avoiding oversimplified conclusions and supporting more reliable benchmarking of SG-DT systems in subsurface environments.

12.3. Framework-level constraints

Consistent with the performance variability and heterogeneity identified in the meta-analysis (Section 7), the SG-DT system architecture described in Section 9, while conceptually robust, remains subject to several systemic constraints that limit its operational scalability and predictive reliability across diverse environmental contexts. These constraints should be interpreted within the broader evidence base of SG-DT system performance, where methodological maturity and predictive stability are strongly conditioned by data availability, environmental non-stationarity, and model–data integration strategies. A key challenge arises from the non-stationarity of environmental signals,

Table 22
Mapping of dominant SG-DT validity threats and corresponding mitigation strategies. Risk scoring (Probability × Impact) was derived from a 3 × 3 ordinal matrix calibrated using evidence from 175 reviewed studies. The mitigation framework prioritises continuous validation, ensemble diversity, and transparent workflow registration to minimise systemic bias and improve reproducibility.

Threat type	Description	Probability	Impact	Risk level	Mitigation strategy	Supporting references
Statistical bias	Small-study bias, non-random missing data, over-fitting of surrogates	Medium	High	Severe	Egger regression + Trim-and-Fill correction; multi-seed ensemble validation	He et al., 2020; Ramos et al., 2023
Algorithmic drift	Non-stationarity in surrogate training; evolving hydro-climatic regimes	High	Medium	High	Periodic retraining; online learning with Bayesian regularisation	Du et al., 2023; Panahi et al., 2025
Publication bias	Selective reporting of high-accuracy cases	Medium	Medium	Moderate	Mandated pre-registration of DT workflows (OSF/Zenodo)	Peng, 2025; Ramos et al., 2023
Infrastructure limitations	Sensor outages, latency, incomplete metadata	Low	High	Moderate	Redundant sensing + edge buffering; standard JSON-LD schema	Zhao et al., 2025; L. Lu et al., 2025; B. Lu et al., 2025

where dynamic aquifer responses driven by climate variability, land-use changes, and episodic recharge events can destabilise trained surrogate models and compromise long-term predictive reliability (Bolan et al., 2014; Kohistani et al., 2025). This limitation directly reflects one of the central findings of the meta-analysis: that SG-DT performance degrades when continuous updating and recalibration mechanisms are insufficiently aligned with evolving system dynamics. Equally significant are concerns about the computational energy footprint of continuous model retraining and uncertainty recalibration. GPU-intensive operations can consume over 1.5 kWh per update cycle, underscoring the tension between real-time adaptability and sustainability targets (Zhao et al., 2025). This trade-off highlights a broader methodological constraint, whereby gains in predictive performance through frequent updating must be balanced against computational efficiency and long-term operational feasibility. In addition, interoperability barriers persist due to the continued reliance on legacy hydrological platforms that lack support for modern data-exchange standards such as JSON-LD and NetCDF, thereby constraining seamless DTOps automation, cross-platform integration, and reproducibility (Karki et al., 2024). These limitations are closely linked to the lack of standardised reporting and benchmarking practices identified in the literature synthesis, which further complicates cross-study comparability and system validation. Finally, the architecture's reliance on real-time feedback introduces a trust-calibration latency, in which rapid iterative coupling between model predictions and decision-control systems may generate oscillatory or unstable behaviour unless adaptive damping logic is implemented (Du et al., 2023). This issue reflects the broader challenge of integrating uncertainty-aware decision-making into dynamically evolving environmental systems, where over-responsive control mechanisms can amplify rather than reduce predictive uncertainty. Importantly, these constraints should not be interpreted as isolated architectural limitations but as manifestations of the broader heterogeneity and context dependence identified across SG-DT system studies. As demonstrated in Sections 7.3 and 12.2, system performance is inherently conditioned by the interaction between data regimes, environmental complexity, and methodological design choices. Overcoming these constraints, therefore, requires not only advances in algorithmic design and cyberinfrastructure modernisation but also alignment with the performance-driven and heterogeneity-aware development pathways identified in meta-analysis. This includes the development of adaptive updating strategies for non-stationary systems, energy-efficient surrogate training protocols, interoperable data infrastructures, and stability-aware decision-feedback mechanisms. Collectively, these directions are essential for ensuring that future SG-DT system ecosystems achieve sustainable, reproducible, and self-stabilizing operation across diverse hydro-environmental contexts.

12.4. Decadal research roadmap (2025–2035)

To transform SG-DT systems from experimental research constructs into globally deployable, policy-relevant infrastructures, a coordinated ten-year roadmap is formulated, as illustrated in Fig. 13. Importantly, this roadmap is not proposed as a purely aspirational vision, but as a synthesis-driven response to the quantitative and methodological gaps identified throughout the meta-analysis. In particular, the roadmap is informed by the observed heterogeneity in predictive performance, the uneven maturity of current SG-DT system implementations, and the persistent limitations in uncertainty quantification, validation design, reproducibility, and cross-study comparability documented in Sections 7 and 12.2. The roadmap delineates four interlinked research trajectories that collectively guide the evolution of SG-DT systems toward autonomy, reliability, and long-term operational sustainability. The first trajectory, Hybrid Differentiable SG-DT system frameworks (2025–2028), focuses on the integration of physics-informed neural networks (PINNs) and Fourier Neural Operators (FNOs) with process-based partial differential equation (PDE) solvers to improve physical consistency, interpretability, and stability of model predictions across heterogeneous aquifer systems (Du et al., 2023; Lazar et al., 2024). This trajectory responds directly to a central finding of the meta-analysis: predictive gains are more robust when computational efficiency is achieved without severing the link to governing-process physics. Building on these foundations, the Federated SG-DT system networks trajectory (2027–2030) aims to enable cross-basin collaboration through federated learning architectures that support privacy-preserving data assimilation and distributed knowledge sharing across monitoring networks (Islam, 2025; Singh et al., 2003). This direction addresses current limitations related to data fragmentation, regional imbalance, and uneven observational support, all of which were identified as important sources of heterogeneity and uneven methodological maturity across the reviewed literature. Subsequently, the Self-Calibrating Sensor Ecosystems for SG-DT systems (2028–2032) phase introduces AI-augmented sensing infrastructures capable of real-time bias correction, fault tolerance, and automated drift detection, thereby enabling continuous and adaptive calibration of SG-DT systems under non-stationary environmental conditions (Syed et al., 2024; Zhao et al., 2025). This trajectory is particularly important, given the meta-analytic

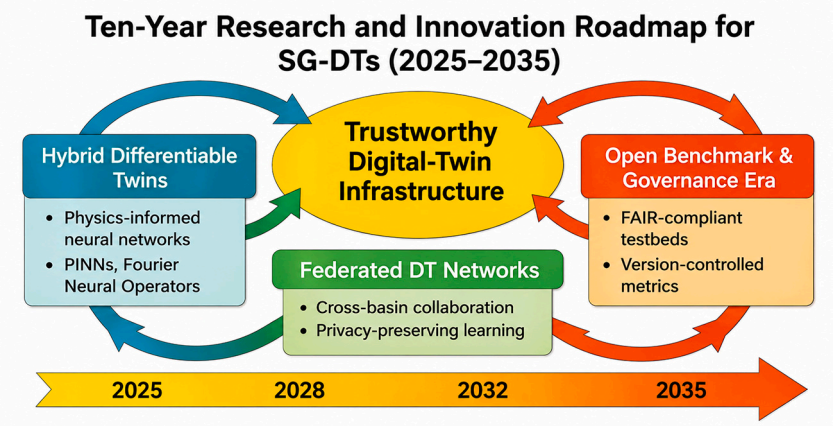


Fig. 13. Ten-year research and innovation roadmap for SG-DTs (2025–2035). Conceptual timeline illustrating the progressive evolution of sediment–groundwater digital twins (SG-DTs) from hybrid differentiable twins integrating physics-informed surrogates (PINNs, FNOs) toward Federated DT Networks enabling cross-basin, privacy-preserving collaboration, culminating in a Trustworthy Digital-Twin Infrastructure and an Open Benchmark & Governance Era anchored in FAIR-compliant testbeds and version-controlled workflows. Prepared by the authors based on a synthesis of the 175-study corpus.

evidence that reliable SG-DT performance depends strongly on data continuity, monitoring density, and the stability of model–data synchronisation over time. Finally, the Open Benchmarks and Governance Standards for SG-DT systems (2032–2035) stage calls for establishing FAIR-compliant global testbeds that integrate version-controlled workflows, standardised evaluation metrics, and reproducible datasets. This phase aims to ensure methodological transparency, comparability, and policy relevance across different application domains (Peng, 2025; Mendil et al., 2010). It also directly addresses one of the clearest limitations identified in this review: the lack of harmonised benchmarking and reporting practices needed to distinguish genuine methodological improvement from artefacts of inconsistent validation, baseline specification, or selective reporting. As illustrated in Fig. 13, these four trajectories converge toward a unified objective: the development of next-generation SG-DT systems that are scientifically robust, operationally reliable, interoperable, and accountable within global environmental governance frameworks. However, the significance of this roadmap lies not only in its technological ambition but in its alignment with the evidence synthesised in this study. The meta-analysis shows that the future advancement of SG-DT systems depends less on increasing architectural complexity in isolation than on improving the coherence among model design, data assimilation, uncertainty quantification, validation practice, and reporting transparency. Accordingly, the proposed ten-year roadmap (2025–2035) should be interpreted as a structured pathway to move SG-DT systems from heterogeneous, unevenly mature implementations toward reproducible, benchmarked, and uncertainty-aware predictive infrastructures. By integrating hybrid physics–AI modelling, federated data assimilation, autonomous sensing, and standardised benchmarking, this roadmap supports the emergence of SG-DT systems that are not only adaptive, transparent, and scalable across diverse environmental contexts, but also increasingly grounded in the performance-based and heterogeneity-aware insights established by the quantitative synthesis. In this sense, the roadmap extends the central conclusion of the meta-analysis: that long-term progress in SG-DT research will depend on systematically advancing predictive reliability, methodological maturity, and cross-context comparability rather than on isolated innovation alone.

Author contributions

Conceptualisation, S.P.; Data curation, S.P., M.M., L.L., F.C. and L.A.D.; Formal analysis, S.P.; Funding acquisition, S.P. and L.A.D.; Investigation, S.P.; Methodology, S.P., M.M., L.L., F.C. and L.A.D.; Supervision, M.M., L.L., F.C. and L.A.D; Visualisation, S.P., M.M., L.L., F.C. and L.A.D; Writing – original draft, S.P.; Writing - review & editing, S.P., M.M., L.L., F.C. and L.A.D. All authors have read and agreed to the published version of the manuscript.

Ethics statement

This research did not involve human participants, animal subjects, or clinical trials. All field studies and data collection followed applicable local regulations and ethical standards. The relevant authorities granted necessary permissions for fieldwork and data acquisition. No sensitive personal data or protected sites were involved in this study.

Funding

This research received support from the Centre of Studies in Geography and Spatial Planning (CEGOT), funded by national funds through the Foundation for Science and Technology (FCT) under the reference UIDB/04084/2025. The first author is supported by a PhD grant from FCT under the reference UI/BD/154881/2023 (https://doi.org/10.5281/zenodo.18289768).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Saeid Pourmorad reports financial support was provided by University of Coimbra. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The editor and anonymous reviewers are gratefully acknowledged for their constructive assessment, which improved the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2026.181953.

Data availability

All data supporting the findings of this review are either included within the main article and its Supplementary materials or are publicly available through an open-access repository. The complete set of supporting materials—including the PRISMA 2020 search strategy, data-extraction templates, the RoB-EnvModel risk-of-bias framework, metadata and reporting schemas (SED-GW-DT-REPORT v1.0), and the full list of reviewed studies (n = 175) has been archived in Zenodo and is accessible via the following DOI: https://doi.org/10.5281/zenodo.18289768. This study did not generate new primary observational or experimental data.

References

Abdullah-Al Mamun, M., Aktar, M.N., Uddin, M.N., Chowdhury, M.H., Islam, M.S., Rahman, M.S., Uddin, M.R., Hossain, M.J., Zahid, A., Senapathi, V., Islam, A.R.M.T., 2026. Hybrid data-driven framework for interpretable prediction of nitrate and sulfate risks in coastal aquifers. Sci. Total Environ. 1011, 181190. https://doi.org/10.1016/j.scitotenv.2025.181190.

Afahnwie, N.A., Yiika, L.P., Marian, C.S., Kachoueiyan, F., Gupta, N., Djibril, K.N.G., Birnila, M.C., Fomekong, B.K., Simou, A., Wochoko, P., 2026. Geospatial distribution of hazardous metals and other elements in stream sediments of Akum, Cameroon: insights into mineral exploration and ecosystem health using indexical and multivariate approaches. Mar. Pollut. Bull. 222 (Part 2), 118761. https://doi.org/10.1016/j.marpolbul.2025.118761.

Akoteyon, I.S., Olatunbosun, J.M., 2026. Sediment quality assessment in coastal wetland of Lagos, Nigeria. Wetl. Ecol. Manag. 34, 2. https://doi.org/10.1007/s11273-025-10105-9.

Alfee, S.L., Bloor, M.C., 2025. A global review of river sediment contamination and remobilisation through climate change-induced flooding. Sustain. Environ. 11 (1), 2440957. https://doi.org/10.1080/27658511.2024.2440957.

Aullón Alcaine, A., Schulz, C., Bundschuh, J., Jacks, G., Thunvik, R., Gustafsson, J.-P., Mörth, C.-M., Sracek, O., Ahmad, A., Bhattacharya, P., 2020. Hydrogeochemical controls on the mobility of arsenic, fluoride and other geogenic co-contaminants in the shallow aquifers of northeastern La Pampa Province in Argentina. Sci. Total Environ. 715, 136671. https://doi.org/10.1016/j.scitotenv.2020.136671.

Bai, T., Tahmasebi, P., 2022. Characterization of groundwater contamination: a transformer-based deep learning model. Adv. Water Resour. 164, 104217. https://doi.org/10.1016/j.advwatres.2022.104217 (ScienceDirect+1).

Bao, J., Li, L., Redoloza, F., 2020. Coupling ensemble smoother and deep learning with generative adversarial networks to deal with non-Gaussianity in flow and transport data assimilation. J. Hydrol. 590, 125443. https://doi.org/10.1016/j.jhydrol.2020.125443.

Bartos, M., Kerkez, B., 2021. Pipedream: an interactive digital twin model for natural and urban drainage systems. Environ. Model Softw. 144, 105120. https://doi.org/10.1016/j.envsoft.2021.105120 (ScienceDirect+1).

Bauer, P., Stevens, B., Hazeleger, W., 2021. A digital twin of Earth for the green transition. Nat. Clim. Chang. 11, 80–83. https://doi.org/10.1038/s41558-021-00986-y.

Bennett, W.W., Canfield, D.E., 2020. Redox-sensitive trace metal dynamics in marine sediments and their links to depositional and diagenetic processes. Earth Sci. Rev. 204, 103175. https://doi.org/10.1016/j.earscirev.2020.103175.

Berg, S.J., 2011. Capturing aquifer heterogeneity: comparison of approaches through controlled sandbox experiments. Water Resour. Res. 47 (9), W09514. https://doi.org/10.1029/2011WR010429.

Berkani, C., Boulabeiz, M., Dali, N., Lakhdari, S., 2026. Evaluation of groundwater quality and its implications for irrigation sustainability in the Bouhmama plain, Northeastern Algeria. Int. J. Energy Water Resour. 10, 45. https://doi.org/10.1007/s42108-025-00470-9.

Boehlert, B., Strzepek, K.M., Chapra, S.C., Fant, C., Gebretsadik, Y., Lickley, M., Swanson, R., McCluskey, A., Neumann, J.E., Martinich, J., 2015. Climate change impacts and greenhouse gas mitigation effects on U.S. water quality. J. Adv. Model. Earth Syst. 7 (3), 1326–1338. https://doi.org/10.1002/2014MS000400.

Bolan, N., Kunhikrishnan, A., Thangarajan, R., Kumpiene, J., Park, J., Makino, T., Kirkham, M.B., Scheckel, K., 2014. Remediation of heavy metal(loid)s contaminated soils - to mobilize or to immobilize? J. Hazard. Mater. 266, 141–166. https://doi.org/10.1016/j.jhazmat.2013.12.018.

Bretzler, A., Lalanne, F., Nikiema, J., Podgorski, J., Pfenninger, N., Berg, M., Schirmer, M., 2017. Groundwater arsenic contamination in Burkina Faso, West Africa: predicting and verifying regions at risk. Sci. Total Environ. 584–585, 958–970. https://doi.org/10.1016/j.scitotenv.2017.01.147.

Chen, J., Dai, Z., Yang, Z., Pan, Y., Zhang, X., Wu, J., Reza Soltanian, M., 2021. An improved tandem neural network architecture for inverse modeling of multicomponent reactive transport in porous media. Water Resour. Res. 57 (12), e2021WR030595. https://doi.org/10.1029/2021WR030595 (AGU Publications+1).

Chen, J., Dai, Z., Dong, S., Zhang, X., Sun, G., Wu, J., 2022. Integration of deep learning and information theory for designing monitoring networks in heterogeneous aquifer systems. Water Resour. Res. 58, e2022WR032429. https://doi.org/10.1029/2022WR032429.

Cherlinka, J., Gishti, F., Vráblik, P., Hlavcová, R., Šúri, M., Pauková, Ž., Brychta, M., Sasík, Ľ., Rybanský, J., 2025. Predictive soil map of Slovakia: a digital twin for decision makers. Int. J. Environ. Stud. https://doi.org/10.1080/00207233.2025.2539654.

Choi, S., Yoon, S., 2025. AI agent-based intelligent urban digital twin (I-UDT): concept, methodology, and case studies. Smart Cities 8 (1), 28. https://doi.org/10.3390/smartcities8010028.

Chowdhury, S., Mazumder, M.J., Al-Attas, O., Husain, T., 2016. Heavy metals in drinking water: occurrences, implications, and future needs in developing countries. Sci. Total Environ. 569–570, 476–488. https://doi.org/10.1016/j.scitotenv.2016.06.166.

Ciliberti, F.G., Berardi, L., Laucelli, D.B., Giustolisi, O., 2021. Digital transformation paradigm for asset management in water distribution networks. In: Proceedings of the 10th International Conference on Energy and Environment (CIEM 2021). IEEE, pp. 1–6. https://doi.org/10.1109/CIEM52821.2021.9614866.

Ciliberti, F.G., Berardi, L., Laucelli, D.B., Giustolisi, O., 2023. Digital water services using the digital twin paradigm. IOP Conf. Ser. Earth Environ. Sci. 1136, 012002. https://doi.org/10.1088/1755-1315/1136/1/012002.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46. https://doi.org/10.1177/001316446002000010.

Conejos Fuertes, P., Martínez-Alzamora, F., Hervás-Carot, M., Alonso-Campos, J.C., 2020. Building and exploiting a Digital Twin for the management of drinking water distribution networks. Urban Water J. 17 (8), 704–713. https://doi.org/10.1080/1573062X.2020.1771382.

Dai, L., Wang, L., Li, L., Liang, T., Zhang, Y., Ma, C., Xing, B., 2018. Multivariate geostatistical analysis and source identification of heavy metals in the sediment of Poyang Lake in China. Sci. Total Environ. 621, 1433–1444. https://doi.org/10.1016/j.scitotenv.2017.10.085.

Dai, Z., Zhan, C., Yin, H., Chen, J., Xu, L., Xia, Y., Songlin, Y., Wei, C., Mingxu, C., Zhengyang, D., Zhang, X., Bicheng, Y., Ma, Y., Wang, H., Moeini, F., Soltanian, M.R., Thanh, H.V., Carroll, K.C., 2025. Incorporating deep learning into hydrogeological modeling: advancements, challenges, and future directions. J. Geophys. Res. Mach. Learn. Comput. 2, e2025JH000703. https://doi.org/10.1029/2025JH000703.

Davis, J.A., Kent, D.B., 1990. Surface complexation modeling in aqueous geochemistry. Rev. Mineral. 23, 177–260.

Dawid, A.P., 1984. Present position and potential developments: some personal views—statistical theory: the prequential approach. J. R. Stat. Soc.: Ser. A (Gen.) 147 (2), 278–292. https://doi.org/10.2307/2981683.

Du, H., Zhao, Z., Cheng, H., Yan, J., He, Q.Z., 2023. Modeling density-driven flow in porous media by physics-informed neural networks for CO_2 sequestration. Comput. Geotech. 159, 105433. https://doi.org/10.1016/j.compgeo.2023.105433.

Essamlali, I., Nhaila, H., El Khaili, M., 2024. Advances in machine learning and IoT for water quality monitoring: a comprehensive review. Heliyon 10, e27920. https://doi.org/10.1016/j.heliyon.2024.e27920 (PMC+1).

Faal, F., Ghafouri, H.R., Ashrafi, S.M., 2022. Monitoring and predicting saltwater intrusion via temporal aquifer vulnerability maps and surrogate models. Water Resour. Manag. 36, 785–801. https://doi.org/10.1007/s11269-021-02970-9.

Falasca, E., Zuliani, T., Hoti, S., Militaru, M., 2025. Artificial intelligence and digital twin for bioclimatic building design: a review. Energies 18 (19), 5230. https://doi.org/10.3390/en18195230.

Fendorf, S., Michael, H.A., van Geen, A., 2010. Spatial and temporal variations of groundwater arsenic in South and Southeast Asia. Science 328 (5982), 1123–1127. https://doi.org/10.1126/science.1172974.

Fleckenstein, J.H., Krause, S., Hannah, D.M., Boano, F., 2010. Groundwater–surface water interactions: new methods and models to improve understanding of processes and dynamics. Adv. Water Resour. 33 (11), 1291–1295. https://doi.org/10.1016/j.advwatres.2010.09.011.

Fogg, G.E., Zhang, Y., 2016. Debates—stochastic subsurface hydrology from theory to practice: a geologic perspective. Water Resour. Res. 52 (12), 9235–9245. https://doi.org/10.1002/2016WR019699.

Frogner-Kockum, P., Göransson, G., Haeger-Eugensson, M., 2020. Impact of climate change on metal and suspended sediment concentrations in urban waters. Front. Environ. Sci. 8, 588335. https://doi.org/10.3389/fenvs.2020.588335.

Gao, W., Qu, B., Yuan, H., Song, J., Li, W., 2023. Heavy metal mobility in contaminated sediments under seawater acidification. Mar. Pollut. Bull. 192, 115062. https://doi.org/10.1016/j.marpolbul.2023.115062.

Ghafari, R., Samaei, S.R., 2025. Integrated AI and digital twin technologies for green project management in resilient coastal and port infrastructure systems. In: Proc. 1st International Conference on Civil Engineering, Architecture, Urban Planning & Environment (capelconf), Wolfsburg, Germany.

Ghanbarpour, M.R., Goorzadi, M., Vahabzade, G., 2014. Spatial variability of heavy metals in surficial sediments: Tajan River Watershed, Iran. Sustain. Water Qual. Ecol. 1–2, 48–58. https://doi.org/10.1016/j.swaqe.2014.04.002.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102 (477), 359–378. https://doi.org/10.1198/016214506000001437.

González-Macías, C., Schifter, I., Lluch-Cota, D.B., Méndez-Rodríguez, L., Hernández-Vázquez, S., 2006. Distribution, enrichment and accumulation of heavy metals in coastal sediments of Salina Cruz Bay, México. Environ. Monit. Assess. 118 (1–3), 211–230. https://doi.org/10.1007/s10661-006-1492-8.

Halpern, O., 2025. The geo-politics of resilience: on the historical convergence between ecology, artificial intelligence, and corporate strategy. New Media Soc. 27 (8), 4581–4605. https://doi.org/10.1177/14614448251336420.

Han, Z., Lu, W., Fan, Y., Lin, J., Yuan, Q., 2020a. A surrogate-based simulation–optimization approach for coastal aquifer management. Water Supply 20 (8), 3404–3418. https://doi.org/10.2166/ws.2020.259.

Han, Z., Lu, W., Lin, J., 2020b. Uncertainty analysis for precipitation and sea-level rise of a variable-density groundwater simulation model based on surrogate models. Environ. Sci. Pollut. Res. 27, 28077–28090. https://doi.org/10.1007/s11356-020-09177-2.

Hansel, C.M., 2015. Manganese in marine microbiology. In: Ehrlich, J.J., Ehrlich, H.C., Newman, D.K. (Eds.), Geomicrobiology, 6th ed. CRC Press, pp. 623–648.

Hauser-Davis, R.A., Wosnick, N., 2022. Climate change implications for metal and metalloid dynamics in aquatic ecosystems and its context within the Decade of Ocean Sciences. Water 14 (15), 2415. https://doi.org/10.3390/w14152415.

He, Q.Z., Barajas-Solano, D., Tartakovsky, G., Tartakovsky, A.M., 2020. Physics-informed neural networks for multiphysics data assimilation with application to subsurface transport. Adv. Water Resour. 141, 103610. https://doi.org/10.1016/j.advwatres.2020.103610.

He, L., Cheng, H., Nan, Z., Gong, Y., Guo, H., Mao, J., Zhang, J., 2025. Improving joint identification of groundwater contaminant source and non-Gaussian distributed conductivity field using a deep learning-based ensemble smoother. J. Hydrol. 658, 133202. https://doi.org/10.1016/j.jhydrol.2025.133202.

Hou, Y., Elwakil, E., Jiang, Y., 2026. Advancing smart construction management: a comprehensive analysis of digital twin applications and challenges. J. Struct. Des. Constr. Pract. 31 (2), SCENG-1953. https://doi.org/10.1061/JSDCCC.SCENG-1953.

Igbokwe, E.C., Oliha, A.O., 2024. Potential for the integration of digital twin and AI technology in gully mapping and prediction. Int. J. Eng. Res. Dev. 20 (5), 199–207 (HAL id: hal-04601963).

Iordache, A.M., Nechita, C., Voica, C., Pluháček, T., Schug, K.A., 2022. Climate change extreme and seasonal toxic metal occurrence in Romanian freshwaters in the last two decades-case study and critical review. npj Clean Water 5 (2). https://doi.org/10.1038/s41545-021-00147-w.

Islam, F.A.S., 2025. A multi-dimensional AI framework for sustainable drinking water management: integrating federated learning, digital twins, and blockchain. J. Eng. Res. Rep. 27 (6), 466–492. https://doi.org/10.9734/jerr/2025/v27i61558.

Islam, M.S., Ahmed, M.K., Habibullah-Al-Mamun, M., 2015a. Geochemical speciation and risk assessment of heavy metals in sediments of a river in Bangladesh. Soil Sediment Contam. Int. J. 24 (6), 639–655. https://doi.org/10.1080/15320383.2015.997869.

Islam, M.S., Ahmed, M.K., Raknuzzaman, M., Mamun, M.H.A., Islam, M.K., 2015b. Heavy metal pollution in surface water and sediment: a preliminary assessment of an urban river in a developing country. Ecol. Indic. 48, 282–291. https://doi.org/10.1016/j.ecolind.2014.08.016.

Islam, M.S., Ahmed, M.K., Habibullah-Al-Mamun, M., Hoque, M.F., 2015c. Preliminary assessment of heavy metal contamination in surface sediments from a river in Bangladesh. Environ. Earth Sci. 73, 1837–1848. https://doi.org/10.1007/s12665-014-3538-5.

Jiang, W., Sheng, Y., Mao, H., 2026. Challenges and advances in soil and groundwater quality and resources assessment. Water 18(2), Article 160. https://doi.org/10.3390/w18020160.

Jones, D., Snider, C., Nassehi, A., Yon, J., Hicks, B., 2020. Characterising the Digital Twin: a systematic literature review. CIRP J. Manuf. Sci. Technol. 29, 36–52. https://doi.org/10.1016/j.cirpj.2020.02.002.

Jørgensen, C.J., Jacobsen, O.S., Elberling, B., Aamand, J., 2012. Microbial oxidation of pyrite coupled to nitrate reduction in anoxic groundwater sediment. Environ. Sci. Technol. 43 (13), 4851–4857. https://doi.org/10.1021/es803417s.

Karki, B.K., Lamichhane, K., Joshi, L., KC, R., Sah, M.K., Pathak, M., Karki, K.R., 2024. Risk assessment of heavy metals in the major surface water system of Nepal with potential remediation technologies. Environ. Challenges 14, 100865. https://doi.org/10.1016/j.envc.2024.100865.

Kim, M.-G., Bartos, M., 2024. A digital twin model for contaminant fate and transport in urban and natural drainage networks with online state estimation. Environ. Model Softw. 171, 105868. https://doi.org/10.1016/j.envsoft.2023.105868.

Kohistani, A.J., Faqiry, F.M., Hashimzai, I.A., 2025. Role of artificial intelligence in groundwater management: challenges and opportunities in Afghanistan. Cognizance J. Multidiscip. Stud. 5 (3), 354–361. https://doi.org/10.47760/cognizance.2025.v05i03.027.

Kolditz, O., Jacques, D., Claret, F., Holt, E., Szöke, R., Garcia, D., Montoya, V., Churakov, S.V., Szöke, I., Laikari, A., Chen, M., Zheng, T., 2026. Digitalisation in geosciences for environmental protection. Environ. Earth Sci. 85, 51. https://doi.org/10.1007/s12665-025-12750-y.

Koltermann, C.E., Gorelick, S.M., 1996. Heterogeneity in sedimentary deposits: a review of structure-imitating, process-imitating, and descriptive approaches. Water Resour. Res. 32 (9), 2617–2658. https://doi.org/10.1029/96WR00025.

Kuleshov, V., Fenner, N., Ermon, S., 2018. Accurate uncertainties for deep learning using calibrated regression. In: Proceedings of the 35th International Conference on Machine Learning (ICML 2018), 80, pp. 2796–2804. https://doi.org/10.48550/arXiv.1807.00263.

Kwon, S., Kang, Y., Nam, S.H., Kim, Y.D., 2025. Water quality monitoring using hybrid physical–soft sensors for river digital twins: a comprehensive review. Water Sci. Technol. 92 (9), 1286–1307. https://doi.org/10.2166/wst.2025.145.

Lazar, N.-N., Simionov, I.-A., Petrea, S.-M., Iticescu, C., Georgescu, P.-L., Dima, F., Antache, A., 2024. The influence of climate changes on heavy metals accumulation in Alosa immaculata from the Danube River Basin. Mar. Pollut. Bull. 200, 116145. https://doi.org/10.1016/j.marpolbul.2024.116145.

Leong, W.Y., 2025. Artificial intelligence-driven analytics for monitoring and mitigating climate change impacts. Eng. Proc. 108, 7. https://doi.org/10.3390/engproc2025108007.

Li, X., Luo, J., Li, Y., Wang, W., Hong, W., Liu, M., Li, X., Lv, Z., 2022. Application of effective water-energy management based on digital twins technology in sustainable cities construction. Sustain. Cities Soc. 87, 104241. https://doi.org/10.1016/j.scs.2022.104241. ScienceDirect.

Liu, W., Du, Y., Ning, J., Zheng, S., Qiu, W., Wang, Y., 2026. Geogenic ammonium enrichment in alluvial-lacustrine aquifer systems: coupled controls of microbial pathways and organic matter composition. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.5c11156 (Advance online publication).

Long, X., Liu, F., Zhou, X., Pi, J., Yin, W., Li, F., Huang, S., Ma, F., 2021. Estimation of spatial distribution and health risk by arsenic and heavy metals in shallow groundwater around Dongting Lake plain using GIS mapping. Chemosphere 269, 128698. https://doi.org/10.1016/j.chemosphere.2020.128698.

Long, K., Ma, C., Li, H., Li, Z., Huang, H., Shi, H., Huang, Z., Sheng, Z., Shi, L., Li, P., Chen, S., Li, X., 2025. AI-enabled digital twin framework for safe and sustainable intelligent transportation. Sustainability 17, 4391. https://doi.org/10.3390/su17104391.

Lu, L., Yin, M., Xie, Y., Pan, Y., Wang, M., Brilakis, I., 2025a. Development of a trustworthy AI-supported digital twin framework for road operation and maintenance. Smart Constr. 2025 (1), 0001. https://doi.org/10.55092/sc20250001.

Lu, B., Francescutto, L., Howie, S., Lin, H., Wu, Q., Hedley, N., Jamali, A., McDonald, I., 2025b. Exploring the concept of digital twins of wetlands for supporting ecosystem monitoring and management. Big Earth Data. https://doi.org/10.1080/20964471.2025.2480446.

Lumley, D., Wanninger, D.J., Magnusson, Å., I'Ons, D., Gustafsson, L.-G., 2024. Implementing a digital twin for optimized real-time control of Gothenburg's regional sewage system. Water Pract. Technol. 19 (3), 657–670. https://doi.org/10.2166/wpt.2024.060.

Majee, U., Ghosh, P., Filippelli, G.M., 2026. Seasonal dynamics of groundwater pollution and health risks in municipal solid waste–affected urban settlements of Bengaluru, Kolkata and Durgapur, India. Environ. Geochem. Health 48, 47. https://doi.org/10.1007/s10653-025-02928-5.

Malvandi, H., 2017. Preliminary evaluation of heavy metal contamination in the Zarrin-Gol River sediments, Iran. Mar. Pollut. Bull. 117, 547–553. https://doi.org/10.1016/j.marpolbul.2017.02.035.

Manocha, A., Bhatia, M., Sood, S.K., 2025. Analyzing groundwater quality in healthcare: a digital twin framework approach. IEEE Internet Things J. https://doi.org/10.1109/JIOT.2025.3620845.

Marcussen, H., Dalsgaard, A., Holm, P.E., 2008. Content, distribution and fate of 33 elements in sediments of rivers receiving wastewater in Hanoi, Vietnam. Environ. Pollut. 155, 41–51. https://doi.org/10.1016/j.envpol.2007.11.001.

Mendil, D., Ünal, Ö.F., Tüzen, M., Soylak, M., 2010. Determination of trace metals in different fish species and sediments from the River Yeşilırmak in Tokat, Turkey. Food Chem. Toxicol. 48, 1383–1392. https://doi.org/10.1016/j.fct.2010.03.006.

Miao, T., Lu, W., Lin, J., Guo, J., Liu, T., 2019. Modeling and uncertainty analysis of seawater intrusion in coastal aquifers using a surrogate model: a case study in Longkou, China. Arab. J. Geosci. 12 (1). https://doi.org/10.1007/s12517-018-4128-8.

Middelburg, J.J., 2019. Marine Carbon Biogeochemistry: A Primer for Earth System Scientists. Springer. https://doi.org/10.1007/978-3-030-10822-9.

Minasny, B., McBratney, A.B., 2025. Machine learning and artificial intelligence applications in soil science: toward a soil digital twin. Geoderma 453, 116679. https://doi.org/10.1016/j.geoderma.2025.116679.

Mo, K., Zabaras, N., Shi, X., Wu, J., 2019. Deep convolutional encoder-decoder networks for uncertainty quantification of dynamic multiphase flow in heterogeneous media. Water Resour. Res. 55 (1), 703–728. https://doi.org/10.1029/2018WR023528.

Mo, S., Zabaras, N., Shi, X., Wu, J., 2020. Integration of adversarial autoencoders with residual dense convolutional networks for estimation of non-Gaussian hydraulic conductivities. Water Resour. Res. 56 (2), e2019WR026082. https://doi.org/10.1029/2019WR026082.

Mok, W.J., Ghaffar, M.A., Noor, M.I.M., Lananan, F., Azra, M.N., 2023. Understanding climate change and heavy metals in coastal areas: a macroanalysis assessment. Water 15 (5), 891. https://doi.org/10.3390/w15050891.

Morillo, J., Usero, J., Gracia, I., 2002. Partitioning of metals in sediments from the Odiel River (Spain). Environ. Int. 28 (4), 263–271.

Morlot, M., Rigon, R., Formetta, G., 2024. Hydrological digital twin model of a large anthropized Italian alpine catchment: the Adige river basin. J. Hydrol. 629, 130587. https://doi.org/10.1016/j.jhydrol.2023.130587.

Mucke, N.T., Pandey, P., Jain, S., Bohté, S.M., Oosterlee, C.W., 2023. A probabilistic digital twin for leak localization in water distribution networks using generative deep learning. Sensors 23 (13), 6179. https://doi.org/10.3390/s23136179.

Nasrabadi, T., Nabi Bidhendi, G., Karbassi, A., Mehrdadi, N., 2010. Partitioning of metals in sediments of the Haraz River (Southern Caspian Sea basin). Environ. Earth Sci. 59 (6), 1111–1117. https://doi.org/10.1007/s12665-009-0101-x.

Nawab, J., Din, Z.U., Ahmad, R., Khan, S., Zafar, M.I., Faisal, S., Raziq, W., Khan, H., Rahman, Z.U., Ali, A., Qayash Khan, M., Ullah, S., Rahman, A., 2021. Occurrence, distribution, and pollution indices of potentially toxic elements within the bed sediments of the riverine system in Pakistan. Environ. Sci. Pollut. Res. 28, 54986–55002. https://doi.org/10.1007/s11356-021-14783-9.

Nearing, G.S., Yatheendradas, S., Crow, W.T., Zhan, X., Liu, J., Chen, F., 2018. The efficiency of data assimilation. Water Resour. Res. 54 (9), 6374–6392. https://doi.org/10.1029/2017WR020991.

Nguyen, B.T., Do, D.D., Nguyen, T.X., Nguyen, V.N., Nguyen, D.T.P., Nguyen, M.H., Truong, H.T.T., Dong, H.P., Le, A.H., Bach, Q.-V., 2020. Seasonal, spatial variation, and pollution sources of heavy metals in the sediment of the Saigon River, Vietnam. Environ. Pollut. 256, 113412. https://doi.org/10.1016/j.envpol.2019.113412.

O'Donnell, J.A., Carey, M.P., Koch, J.C., Baughman, C., Hill, K., Zimmerman, C.E., Sullivan, P.F., Dial, R., Lyons, T., Cooper, D.J., Poulin, B.A., 2024. Metal mobilization from thawing permafrost to aquatic ecosystems is driving rusting of Arctic streams. Commun. Earth Environ. 5 (268). https://doi.org/10.1038/s43247-024-01446-z.

Ohta, A., Imai, N., Terashima, S., Tachibana, Y., 2005. Influence of surface geology and mineral deposits on the spatial distributions of elemental concentrations in the stream sediments of Hokkaido, Japan. J. Geochem. Explor. 86, 86–103. https://doi.org/10.1016/j.gexplo.2005.04.002.

Ojuederie, O.B., Babalola, O.O., 2017. Microbial and plant-assisted bioremediation of heavy metal polluted environments: a review. Int. J. Environ. Res. Public Health 14 (12), 1504. https://doi.org/10.3390/ijerph14121504.

Osae, R., Sablah, T.Y., Abasiya, C., Peasah, M.Y., Agyare, E.A., Loh, S.K., Owusu, M., Owusu, F., 2025. Toxicity assessment of heavy metals in leachate from dredged sediments deposited near residential areas: implications for environmental and human health risk. Environ. Qual. Manag. https://doi.org/10.1002/tqem.70264 (Advance online publication).

Pan, Z., Lu, W., Fan, Y., Li, J., 2021. Identification of groundwater contamination sources and hydraulic parameters based on Bayesian regularization deep neural network. Environ. Sci. Pollut. Res. 28, 16867–16879. https://doi.org/10.1007/s11356-020-11614-1 (SpringerLink+1).

Pan, Z., Lu, W., Bai, Y., 2022. Groundwater contamination source estimation based on a refined particle filter associated with a deep residual neural network surrogate. Hydrogeol. J. 30, 881–897. https://doi.org/10.1007/s10040-022-02454-z.

Panahi, M., Porta, G.M., Riva, M., Guadagnini, A., 2025. Modeling parametric uncertainty in PDEs models via physics-informed neural networks. Adv. Water Resour. 195, 104870. https://doi.org/10.1016/j.advwatres.2024.104870.

Pandey, S., Chu, X., Laurien, E., 2017. Investigation of in-tube cooling of carbon dioxide at supercritical pressure by means of direct numerical simulation. Int. J. Heat Mass Transf. 114, 944–957. https://doi.org/10.1016/j.ijheatmasstransfer.2017.06.089.

Park, D., You, H., 2023. A digital twin dam and watershed management platform. Water 15 (11), 2106. https://doi.org/10.3390/w15112106.

Peng, J., 2025. Digital-twin technology and its application in water governance: China's practices and achievements. Water Resour. Dev. Rep. 38 (4), 205–218.

Postma, D., Larsen, F., Minh Hue, N.T., Duc, M.T., Viet, P.H., Nhan, P.Q., Jessen, S., 2007. Arsenic in groundwater of the Red River floodplain, Vietnam: controlling geochemical processes and reactive transport modeling. Geochim. Cosmochim. Acta 71 (21), 5054–5071. https://doi.org/10.1016/j.gca.2007.08.020.

Rahman, M.S., Saha, N., Molla, A.H., 2014. Potential ecological risk assessment of heavy metal contamination in sediment and water body around Dhaka export processing zone, Bangladesh. Environ. Earth Sci. 71, 2293–2308. https://doi.org/10.1007/s12665-013-2631-5.

Raissi, M., Perdikaris, P., Karniadakis, G.E., 2019. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. Comput. Phys. 378, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045.

Ramos, H.M., Morani, M.C., Carravetta, A., Fecarotta, O., Adeyeye, K., López-Jiménez, P. A., Pérez-Sánchez, M., 2022. New challenges towards smart systems' efficiency by digital twin in water distribution networks. Water 14 (8), 1304. https://doi.org/10.3390/w14081304.

Ramos, H.M., Kuriqi, A., Besharat, M., Creaco, E., Tasca, E., Coronado-Hernández, O.E., Pienika, R., Iglesias-Rey, P., 2023. Smart water grids and digital twin for the management of system efficiency in water distribution networks. Water 15 (6), 1129. https://doi.org/10.3390/w15061129.

Rasheed, A., San, O., Kvamsdal, T., 2020. Digital twin: values, challenges and enablers from a modelling perspective. IEEE Access 8, 21980–22012. https://doi.org/10.1109/ACCESS.2020.2970143 (Ramos, H. M., Morani, M. C., Carravetta, A., Fecarotta, O., Adeyeye, K., López-Jiménez, P. A., & Pérez-Sánchez, M. (2022). New challenges towards smart systems' efficiency by digital twin in water distribution networks. Water, 14(8), 1304. doi:10.3390/w14081304).

Rath, P., Panda, U.C., Bhatta, D., Sahu, K.C., 2009. Use of sequential leaching, mineralogy, morphology and multivariate statistical technique for quantifying metal pollution in highly polluted aquatic sediments—a case study: Brahmani and Nandira Rivers, India. J. Hazard. Mater. 163, 632–644. https://doi.org/10.1016/j.jhazmat.2008.07.048.

Ross, M.R.V., Topp, S.N., Appling, A.P., Yang, X., Kuhn, C., Butman, D., Simard, M., Pavelsky, T.M., 2019. AquaSat: a data set to enable remote sensing of water quality for inland waters. Water Resour. Res. 55, 10012–10025. https://doi.org/10.1029/2019WR024883.

Roy, T., Schütze, N., Grundmann, J., Brettschneider, M., Jain, A., 2016. Optimal groundwater management using state-space surrogate models: a case study for an arid coastal region. J. Hydroinf. 18 (4), 666–686. https://doi.org/10.2166/hydro.2016.086.

Samaei, S.R., Ghafari, R., 2025. An AI-Driven and Digital Twin-Based Framework for Holistic Risk and Green Project Management in Sustainable Urban Megaprojects: The Case of Tehran under Climate and Population Pressures. , April. (Conference paper).

Sharafat, A., Tanoli, W.A., Zubair, M.U., Mazher, K.M., 2025. Digital twin-driven stability optimization framework for large underground caverns. Appl. Sci. 15, 4481. https://doi.org/10.3390/app15084481.

Sharifi, A., Tarlani Beris, A., Sharifzadeh Javidi, A., Nouri, M., Gholizadeh Lonbar, A., Ahmadi, M., 2024. Application of artificial intelligence in digital twin models for stormwater infrastructure systems (review). Adv. Eng. Inform. 61, 102485. https://doi.org/10.1016/j.aei.2024.102485.

Sheng, D., Lou, Y., Sun, F., Xie, J., Yu, Y., 2023. Reengineering and its reliability: an analysis of water projects and watershed management under a digital twin scheme in China. Water 15 (18), 3203. https://doi.org/10.3390/w15183203.

Sheykhi, V., Moore, F., 2012. Geochemical characterization of Kor River water quality, Fars Province, southwest Iran: application of chemometric models. Int. J. Environ. Res. 6 (4), 913–926.

Singh, D., Sharma, V., 2025. A framework for smart-village for sustainable groundwater management using digital twin. In: Proceedings of the 2025 International Conference on Emerging Trends in Industry 4.0 Technologies (ICETI4T). IEEE, pp. 1–6. https://doi.org/10.1109/ICETI4T63625.2025.11132268.

Singh, M., Müller, G., Singh, I.B., 2003. Geogenic distribution and baseline concentration of heavy metals in sediments of the Ganges River, India. J. Geochem. Explor. 80 (1), 1–17. https://doi.org/10.1016/S0375-6742(03)00016-5.

Singh, H., Pandey, R., Singh, S.K., Shukla, D.N., 2017. Assessment of heavy metal contamination in the sediment of the River Ghaghara, a major tributary of the River Ganga in Northern India. Appl Water Sci 7, 4133–4149. https://doi.org/10.1007/s13201-017-0572-y.

Skierszkan, E.K., Dockrey, J.W., Lindsay, M.B.J., 2025. Metal mobilization from thawing permafrost is an emergent risk to water resources. ACS EST Water 5, 20–32. https://doi.org/10.1021/acsestwater.4c00789.

Song, J., Yang, Y., Chen, G., Sun, X., Lin, J., Wu, J., 2019. Surrogate-assisted multi-objective robust optimization for groundwater monitoring network design. J. Hydrol. 577, 123994. https://doi.org/10.1016/j.jhydrol.2019.123994.

Sprocati, R., Rolle, M., 2021. Integrating process-based reactive transport modelling and machine learning for electrokinetic remediation of contaminated groundwater. Water Resour. Res. 57, e2021WR029959. https://doi.org/10.1029/2021WR029959.

Steefel, C.I., Appelo, C.A.J., Arora, B., Jacques, D., Kalbacher, T., Kolditz, O., Lagneau, V., Lichtner, P.C., Mayer, K.U., Meeussen, J.C.L., Molins, S., Moulton, D., Shao, H., Šimůnek, J., Spycher, N., Yabusaki, S.B., Yeh, G.T., 2015. Reactive transport codes for subsurface environmental simulation. Comput. Geosci. 19 (3), 445–478. https://doi.org/10.1007/s10596-014-9443-x.

Syed, T.A., Khan, M.Y., Jan, S., Albouq, S., Alqahtany, S.S., Naqash, M.T., 2024. Integrating digital twins and artificial intelligence multi-modal transformers into water resource management: overview and advanced predictive framework. AI 5 (4), 1977–2017. https://doi.org/10.3390/ai5040098.

Syvitski, P.M., Saito, Y., 2007. Morphodynamics of deltas under the influence of humans. Glob. Planet. Chang. 57 (3–4), 261–282. https://doi.org/10.1016/j.gloplacha.2006.12.001.

Tanazefti, L., Boussada, R., Azouz Saidane, L., 2025. Enhancing digital twins for smart agriculture. In: 2025 International Wireless Communications and Mobile Computing Conference (IWCMC). IEEE. https://doi.org/10.1109/IWCMC65282.2025.11059528.

Tao, F., Zhang, M., Liu, Y., Nee, A.Y.C., 2019. Digital twins and cyber–physical systems toward smart manufacturing and Industry 4.0: correlation and comparison. Engineering 5, 653–661. https://doi.org/10.1016/j.eng.2019.01.014.

Thelen, A., Zhang, X., Fink, O., Lu, Y., Ghosh, S., Youn, B.D., Todd, M.D., Hu, Z., 2022. A Comprehensive Review of Digital Twin – Part 2: Roles of Uncertainty Quantification and Optimization, a Battery Digital Twin, and Perspectives. arXiv Preprint arXiv:2208.12904.

Uma Maheswara Rao, R., Waila, P., Mammen, P.C., Muthu, R., 2026. Enabling artificial intelligence (AI) and machine learning (ML) techniques for managing forest fires. In: Vyas, S., Jawak, S.D., Deb Burman, P.K., Patel, H., Kandekar, A., Sawant, S. (Eds.), Application of Machine Learning in Earth Sciences. Springer, pp. 429–459. https://doi.org/10.1007/978-3-032-11426-6_20.

Vargiu, E., Munaretto, S., Antzoulatos, G., Eliades, D.G., 2025. Towards interconnected digital twins for inland water. In: Proceedings of CCWI 2025—21st Computing & Control for the Water Industry Conference (Sheffield, UK) (IDEATION project paper).

Varol, M., Şen, B., 2012. Assessment of nutrient and heavy metal contamination in surface water and sediments of the upper Tigris River, Turkey. Catena 92, 1–10. https://doi.org/10.1016/j.catena.2011.11.011.

Visser, A., Kroes, J., van Vliet, M.T.H., Blenkinsop, S., Fowler, H.J., Broers, H.P., 2012. Climate change impacts on the leaching of a heavy metal contamination in a small lowland catchment. J. Contam. Hydrol. 127, 47–64. https://doi.org/10.1016/j.jconhyd.2011.04.007.

Vosoogh, A., Saeedi, M., Lak, R., 2016. Heavy metals relationship with water and size-fractionated sediments in rivers using canonical correlation analysis (CCA): case study, rivers of south western Caspian Sea. Environ. Monit. Assess. 188, 603. https://doi.org/10.1007/s10661-016-5611-x.

Wang, S.-L., Xu, X.-R., Sun, Y.-X., Liu, J.-L., Li, H.-B., 2013. Heavy metal pollution in coastal areas of South China: a review. Mar. Pollut. Bull. 76, 7–15. https://doi.org/10.1016/j.marpolbul.2013.08.025.

Wang, Q., Hong, H., Yang, D., Li, J., Chen, S., Pan, C., Lu, H., Liu, J., Yan, C., 2020. Health risk assessment of heavy metal and its mitigation by glomalin-related soil protein in sediments along the South China coast. Environ. Pollut. 263, 114565. https://doi.org/10.1016/j.envpol.2020.114565.

Wang, L., Kurihana, T., Meray, A., Mastilovic, I., Praveen, S., Xu, Z., Memarzadeh, M., Lavin, A., Wainwright, H., 2022. Multi-scale digital twin: developing a fast and physics-informed surrogate model for groundwater contamination with uncertain climate models. In: NeurIPS 2022 Workshop on ML and Physical Sciences. arXiv: 2211.10884.

Wang, L., Shen, Z., Bank, M.S., Kawasaki, S., Wu, W.-M., Alessi, D.S., Yu, K., 2026. Sedimentary rocks as analogues of coastal pollution and sustainable solutions building future resilience. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.5c16109 (Advance online publication).

Wu, Z.Y., Chew, A., Meng, X., Cai, J., Pok, J., Kalfarisi, R., Lai, K.C., Hew, S.F., Wong, J. W., 2023. High fidelity digital twin-based anomaly detection and localization for smart water grid operation management. Sustain. Cities Soc. 91, 104446. https://doi.org/10.1016/j.scs.2023.104446.

Xu, J., Zhang, Y., 2025. AI-powered digital twin technology for highway system slope stability risk monitoring. Geotechnics 5 (1), 19. https://doi.org/10.3390/geotechnics5010019.

Yang, D., Liu, J., Wang, Q., Hong, H., Zhao, W., Chen, S., Yan, C., Lu, H., 2019. Geochemical and probabilistic human health risk of chromium in mangrove sediments: a case study in Fujian, China. Chemosphere 233, 503–511. https://doi.org/10.1016/j.chemosphere.2019.05.245.

Yang, Z., Li, C., Chen, H., Shan, X., Chen, J., Zhang, J., Liu, S., Liu, Q., Wang, X., 2023a. Source-oriented ecological and resistome risks associated with geochemical enrichment of heavy metals in river sediments. Chemosphere 336, 139119. https://doi.org/10.1016/j.chemosphere.2023.139119.

Yang, C., Zeng, Z., Wang, Y., He, G., Hu, Y., Gao, D., Dai, Y., Li, Q., Zhang, H., 2023b. Ecological risk assessment and identification of distinct microbial groups in heavy metal-polluted river sediments. Environ. Geochem. Health 45, 1311–1329. https://doi.org/10.1007/s10653-022-01343-4.

Yang, X., Li, Z., Zhang, Q., Liu, Y., Wang, H., 2025. Policy evolution and framework design for China's irrigation-management informatization: from fundamentals to AI-driven Digital-Twin irrigation districts. China Irrig. Drain. J. 44 (2), 1–15.

Yin, W., Hu, Q., Liu, W., Liu, J., He, P., Zhu, D., Kornejady, A., 2024. Harnessing game engines and digital twins: advancing flood education, data visualization, and interactive monitoring for enhanced hydrological understanding. Water 16, 2528. https://doi.org/10.3390/w16172528.

Yu, X., Sreekanth, J., Cui, T., Pickett, T., Xin, P., 2021. Adaptative DNN emulator-enabled multi-objective optimization to manage aquifer–sea flux interactions in a regional coastal aquifer. Agric. Water Manag. 245, 106571. https://doi.org/10.1016/j.agwat.2020.106571.

Zhang, Z., Li, J., Mamat, Z., Ye, Q., 2016. Sources identification and pollution evaluation of heavy metals in sediments of Bortala River, Northwest China. Ecotoxicol. Environ. Saf. 126, 94–101. https://doi.org/10.1016/j.ecoenv.2015.12.025.

Zhao, T., Song, C., Yu, J., Xing, L., Xu, F., Li, W., Wang, Z., 2025. Leveraging immersive digital twins and AI-driven decision support systems for sustainable water reserves management: a conceptual framework. Sustainability 17, 3754. https://doi.org/10.3390/su17083754.

Zhu, Y., Zabaras, N., 2018. Bayesian deep convolutional encoder–decoder networks for surrogate modelling and uncertainty quantification. J. Comput. Phys. 366, 415–447. https://doi.org/10.1016/j.jcp.2018.04.018.

Zitoun, R., Marcinek, S., Hatje, V., Sander, S.G., Völker, C., Sarin, M., Omanović, D., 2024. Climate change-driven effects on transport, fate and biogeochemistry of trace element contaminants in coastal marine ecosystems. Commun. Earth Environ. 5 (560). https://doi.org/10.1038/s43247-024-01679-y.

Zou, S., Ju, H., Zhang, J., 2025. Water quality management in the age of AI: applications, challenges, and prospects. Water 17, 1641. https://doi.org/10.3390/w17111641.