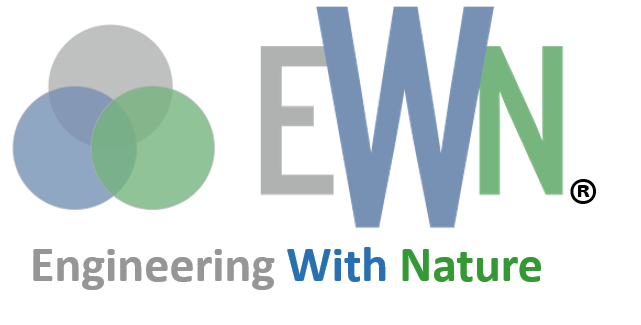
Text

Description automatically generatedA picture containing water, grass, outdoor, herd

Description automatically generated

|  |  |  |
| --- | --- | --- |
| *Engineering With Nature® is the intentional alignment of natural and engineering processes to efficiently and sustainably deliver economic, environmental and social benefits through collaboration.* |  | *The Network for Engineering With Nature is a community of researchers, practitioners, and educators working together to advance Engineering With Nature®* |

Data Science to Advance Nature-based Solutions for Water Resource Management

October 2024

|  |  |
| --- | --- |
| PRINCIPAL INVESTIGATORS | |
| UNIVERSITY PI POC  Name: K. Selcuk Candan  Email: candan@asu.edu  Phone Number: 480-965-2770  Department: SCAI | ERDC PI POC  Name: Todd Steissberg  Email: Todd.E.Steissberg@usace.army.mil  Phone Number: 530-574-5572  Org.Code: U433D90 |
| OTHER TEAM MEMBERS | |
| Name 1: Rebecca Muenich  Name 3: Ting Liu  Name 5: Daniel Siegel | Name 2: Huan Liu  Name 4: Giuseppe Mascaro  Name 6: Amber Wutich |
| FUNDING REQUESTED ($K) | |
| UNIVERSITY:  ERDC:  PROJECT TOTAL: | |

Project Background

## Project Abstract

The Cybersecurity and Infrastructure Security Agency within the U.S. Department of Homeland Security defines and manages national security and resilience of 16 sectors of critical infrastructure. Natural infrastructure tends to be built in a haphazard way, often without reference to basin-scale objectives for flood-control and water supply. Natural Water Infrastructure (NWI) under this definition can include wetlands, healthy soils, forest ecosystems, and snowpack and the water they provide through runoff. NWI provides services that include flood protection, erosion control, water storage, and purification. Built water infrastructure (BWI) comprises two entire sectors – the dams sector and water and wastewater systems sector – and the resilience of at least nine other sectors has critical dependency on water infrastructure. Our nation’s NWI and BWI are diverse, cross-sectoral, and are characterized by highly complex and dynamic systems, requiring data and model-driven situational awareness and decision making at different scales. Enabling the resilience of such systems brings forth various critical challenges that are difficult, or even out-right impossible, with the technologies available today. Real-time and online machine learning needed to support local and global inferences and predictions are lacking. Models need to be constructed in the presence of sensed data, along with applicable physical models, from multiple sources, characterized by varying levels of accuracy, which together overwhelm today’s data science and Artificial Intelligence (AI) techniques. Moreover, these systems are often lacking the centralized coordination and data and model integration capabilities to support today’s forecasting and decision-making techniques. Given these and other critical shortcomings, we argue that to have transformative impact, complementary innovations in foundational and use-inspired data science and engineering for resilience of NWI and BWI are needed. We further argue that successful development and deployment of robust, reliable, and effective data science solutions must be complemented with an inclusive community of academic, government, and NGO partnerships, and innovations in education and workforce development.

## Project Goal

The science science foundation for Engineering With Nature (EWN) needs to be developed as a strategic, outcomes-based approach to the design and operations of integrated Natural and Built Water Infrastructure (NABWI) systems. The goal of this project is to answer the following three research questions:

* Q1: The state-of-the-art integrated physically based models we propose to develop will provide predictions of watershed resilience to various EWN scenarios under climate change and associated shifting patterns of extreme events such as floods and droughts (Klos et al., 2014; Godsey et al., 2014). While they provide valuable information, applications of these physically based models to operational decision making have been limited by substantial efforts needed to set up and calibrate the models for a given study area and high computational expense. *Given this, how can we achieve transparent and efficient decision making via causal learning and transfer learning?*
* Q2: The management of water infrastructure is a pressing contemporary challenge given a changing climate (Arias et al. 2012; Fletcher et al. 2019) and increasing water demand for agriculture and human consumption (Guy, 1996; Lienert et al. 2015). This mismatch is exacerbated by extreme events like droughts or floods (Salas et al. 2018). Computational models can support the design of operational plans to mitigate the negative impacts of such adverse events (Kurtz et al. 2017; Tian et al. 2017; Al-Jawad, 2019). *Under these conditions, how can computational models inform the operation of built and natural infrastructure to increase the resiliency of a water infrastructure network? Can these models provide accurate operational plans, ultimately providing holistic mitigation actions against droughts and floods?*
* Q3: As detailed in our recent report titled “*Data Integration in the Service of Synthetic Research”* [Kintigh et al. 2018], research of the scale that grand challenges require, demands for the ability to explore multiple, large sets of primary data and models to provide the opportunity to discover important cross-dataset and cross-community patterns that could never be seen when comparing higher-level interpretations. Yet, building a single monolithic model to understand the dynamics of a complex natural/engineered process, such as those commonly seen in water resource management, and its effects over the affected systems is generally not possible. Moreover, the component data, models, and simulators may be at different spatio-temporal scales, rely on different mathematical models, and be implemented using different computational frameworks. Therefore, the proposed work is aimed to answer the question whether it is possible to effectively stitch together multiple, potentially independently developed and implemented, off-the-shelf, component data sources, models, and simulators, each capturing a different natural, human, or built component of the complex system.

## Statement of Need

## The ability to leverage data science, engineering, AI, and machine learning to understand and provide resilience to natural and built water systems (NWI and BWI) at different scales is critical in many urgent contexts, exemplified by the first three PMPs. Data technologies provide unmatched potential to explain human behavior and to revolutionize our lives by enabling smarter and better informed decisions. While the promise is apparent, the core technologies needed to achieve this in the context of NWI and BWI goals of the project are in the early stages and lack a framework to help realize their potential.

## Water systems and simulations generate massive time series data representing complex and dynamic physical processes operating at varying spatial and temporal scales. Natural and built water systems involve heterogeneous multi-modal (temporal, spatial, networked) data and models, 100s of inter-dependent parameters, spanning multiple layers and geo-spatial frames, affected by complex dynamic processes operating at different scales and resolutions. Many of these dynamically evolve over time, due to how the ecosystems develop and due to the preventive and reactive actions taken by individuals and public interventions, requiring continuous adaptation and re-modeling. Scientists, planners ,and decision makers need extensive time and labor to manually sort through these data to understand how buildings are functioning. Both human decision making and model discovery can be significantly improved if this massive amount of data can be analyzed for key causal features to discover the underlying latent structure and dependencies in the data critical for modeling, optimization, and control. If effectively leveraged, data and physically based models can be collectively used for advanced control, commissioning, or retrofitting existing systems. But, these models suffer from two key challenges that reduce their wide-spread usage and prevents their potential impact: Cost of Modeling: The simulation model for a typical NWI and BWI can be complex, and creating a complex model from scratch requires significant manpower. Forecasting Accuracy: due to the complexity, tight coupling, and various temporal scales of the natural and built water systems, calibrating a simulation model with the level of granularity for control and fault diagnosis is very challenging. Using such a model for simulation, whose outcomes often highly rely on a specific region’s characteristics, is even more complicated. Cost of Simulations: Given the unpredictability of the natural and human factors and unpredictability of the actions of various independent agencies, decision makers need to generate many thousands of simulations, each with different parameters corresponding to different, but plausible scenarios. Running and interpreting simulations to generate timely actionable results is computationally costly and, consequently, data and simulations have been sparse.

Therefore, tackling the key domain challenges necessitates a novel framework built on computational advances in big data and model integration, causal learning and discovery, large scale data- and model-driven simulations, emulations, and forecasting, data-driven and model centric operational recommendations, and effective visualization and explanation.

## Problem or Opportunity

## Natural water infrastructure tends to be built in a haphazard way, often without reference to basin-scale objectives for flood-control and water supply. This necessitate a strategic, outcomes-based approach to the design and operations of integrated NABWI systems. This EWN approach must supports design of both constructed and restored natural assets that are strategically placed and operated to improve the function of existing (and failing) built water infrastructure. To achieve this, we propose to develop a data science toolkit that will allow for rapid, scenario-based assessment of outcomes of combinations of natural and built infrastructure. This data science (data integration, artificial intelligence, operations research and visualization) toolkit will be build on domain science (hydrology and geohydrology) PMP and will answer the question: “What combinations of data science are needed to advance EWN solutions within a particular hydrologic and water resources context?” In other words, what is the missing data science ingredient necessary to bring EWN solutions to decision makers such that EWN solutions are implemented, replicated and scaled strategically?

## Project Value Statement

This cross-cutting project will deliver new data tools and platform and for natural and built water infrastructure management, with the goal of reducing overall inefficiencies with respect to environmental outcomes and infrastructure investment. We will achieve our objectives through the combination of data science, engineering, optimization, and water resources modeling and management, in partnership with the other domain projects that constitute the overall effort.

## 

## A picture containing diagram Description automatically generated

## How Does The Project Align with the EWN Program

At a high level, this project aims to develop data science, engineering, AI, and optimization technologies to identify and inform scientists and decision makers, including USACE Districts and other water managers, to better maintain NNBF and traditional water resources infrastructure such that the overall system can better respond to large scale natural processes.

## How Does The Project Satisfy EWN Elements

|  |  |
| --- | --- |
|  | Leverage Natural Processes |
| In partnership with the three domain PMPs, this project will leverage natural processes: we will work with nature-based solutions on upland (agricultural) lands and nature-based solutions for existing infrastructure (e.g., planted levees, reservoir release schedules that prevent harmful algal blooms). We will also target understanding where, spatially, the physical conditions are suitable for MAR (i.e., where is there excess runoff water available for capture, and does it overlap a groundwater recharge zone where it can be stored). | | |
|  | Produce Efficiencies |
| Our project will work to develop data- and model-driven, generalizable tools to identify and quantify inefficiencies within BWI: where they exist, as well as where and how they can be managed to produce more synergistic outcomes. Leveraging data science methods, the knowledge and tools developed for this project will be transferable across larger scales to expand EWN. | | |
|  | Broaden Benefits |
| This project directly links natural and built water infrastructure systems with contributing area management, helping to identify solutions that improve water resource management for water quantity and quality. There will also be benefits beyond the environmental outcomes, in terms of identifying opportunities for more synergies among federal spending across multiple (often siloed) domains. | | |
|  | Collaboration |
| This project works across a wide variety of projects happening within the USACE and ERDC system, from modeling studies to applied implementations of nature-based solutions. We are connecting to ERDC’s Water Quality and Contaminant Modeling Branch (Todd Steissberg, Lauren Melendez), studying how reservoir management influences harmful algal blooms and reservoir water quality; and to ERDC’s Ecohydrology Team (Dr. Steissberg, team Lead), developing emulated water quality models to better simulate these connected systems. | | |

## Objectives

Our primary objective is to answer this question: “How can we effectively empower decision makers, including USACE and other local water infrastructure managers, to find nature-based solutions that make the most sense for their local context?” This project will achieve its goals through 5 primary objectives:

1. Developing spatio-temporal causal discovery algorithms and causally-informed, generalizable data and physically-based models of impact for characterizing complex natural and built systems
2. Developing a scalable and modular cloud-based platform for data and model integration and complex system simulation and emulation
3. Developing spatio-temporal multi-objective and high-dimensional optimization frameworks
4. Developing methods to apply transfer learning and emulation modeling techniques to simulate and understand processes driving and controlling integrated NWI and BWI systems.
5. Implementing, testing, and monitoring the performance of the proposed data frameworks within the context of the three domain PMPs

# Project Objectives and Outcomes

## Technical Approach

*Machine learning emulation of physically based model predictions*

Organizing Question: How can we achieve transparent and efficient decision making via causal learning and transfer learning?

Background: The state-of-the-art integrated physically based models we propose to develop can provide predictions of watershed resilience to various EWN scenarios under climate change and associated shifting patterns of extreme events such as floods and droughts (Klos et al., 2014; Godsey et al., 2014). While they provide valuable information, applications of these physically based models to operational decision making have been limited by the substantial efforts needed to set up and calibrate the models for a given study area along with high computational expense. To overcome these barriers, there is a need to develop surrogate models that are easier and faster to run and also transferable across applications (Cai et al., 2015; Akhtar and Shoemaker, 2016; Xu and Liang 2021). Machine learning including deep learning methods are well suited to develop fast surrogates to enable decision-speed analytics (Xu and Liang 2021). These methods have proven equal or more effective than physically based models in various hydrologic applications (Kratzert et al., 2018; Damavandi et al., 2019; Damavandi et al., 2020; Stampoulis et al., 2020). In addition, when trained using data representing a wide range of variability, machine learning models can be generalizable and transferable to new locations and scenarios (Nearing et al., 2021). The Data Mining and Machine Learning Lab (DMML) in the School of Computing and Augmented Intelligence (SCAI) at Arizona State University, led by Professor Huan Liu, develops computational methods for data mining, machine learning, and social computing and designs efficient machine learning algorithms to enable effective problem-solving with a focus on real-world applications. DMML’s research projects (i) analyzing high-dimensional data via feature selection and feature discretization; (ii) causal learning with observational data; (iii) integrating multi-modal data to overcome ambiguity and uncertainty, (iv) low-resource learning and transfer learning, and (v) explainable deep learning. The DMML team includes dozens of PhD students, who regularly receive scholarly awards and continue their careers at research universities and prominent IT companies.

Methods: Conventional physically based models have long yielded promising results, as they have been the main tool to depict the underpinnings of the physics governing the hydrological events. These models, however, suffer from certain issues such as the intense calibration time or the uncertainty in the estimation of hydrological variables. The development of sophisticated data-driven techniques, and machine learning models, combined with rapid increases in computational abilities (graphics processing units, computer clusters. etc.), has enabled hydrologists to utilize data driven models in tandem with the well-established hydrological models to simulate miscellaneous environmental processes nimbly, and therefore circumvent the conundrums associated with the physically based models. Aiming for transparent and efficient decision making, we propose to deepen our research in the following two directions: (1) causal learning to leverage physically based models’ advantages and (2) transfer learning to make data-driven learning more generalizable across different applications with similarity. We aim to move beyond both physically based models and data-driven machine learning models by calling for a transparent and nimble solution to unravel the underlying intricate inter-relationships of hydrological variables.

We have successfully demonstrated the power of transfer learning in [Damavandi et al 2019, Damavandi et al 2020]. In this WP, we propose to extend transfer learning with low-resource learning methods (e.g., few-shot learning [Ding et al 2021]) to overcome data shortage problems in new application domains and improve learning efficiency. This approach will enable transferring learned knowledge from our four study areas to less-studied areas. For transparent decision making, we propose to leverage physical based models to generate interventional data. More specifically, we will use intermediate variables of physically based models (e.g., hydrological variables) to train machine learning models that are interpretable and robust against spurious correlations [Lee et al 2021, Kaushik et al 2019, Ravfogel et al 2021] via causal learning research [Guo et al 2020a, Moraffah et al 2020a, Moraffah et al 2020c]. Domain experts can guide and help design interventions. For example, we can answer the following two questions when designing an interventional dataset: (1) what types of intermediate variables should be intervened such that we can avoid spurious correlations, and (2) what proper values or distributions would be used to create the interventions? With interventional data from the physically based models, we can augment the original data to train interpretable and robust machine learning models.

*Optimization and operations in space and time—Siting of natural infrastructure to optimize operations and coordination with built infrastructure*

Organizing Question: How can computational models inform the operation of built and natural infrastructure to increase the resiliency of a water infrastructure network? Can these models provide accurate operational plans, ultimately providing holistic mitigation actions against droughts and floods?

Background: The management of water infrastructure is a pressing contemporary challenge given a changing climate (Arias et al. 2012; Fletcher et al. 2019) and increasing water demand for agriculture and human consumption (Guy, 1996; Lienert et al. 2015). This mismatch is exacerbated by extreme events like droughts or floods (Salas et al. 2018). Computational models can support the design of operational plans to mitigate the negative impacts of such adverse events (Kurtz et al. 2017; Tian et al. 2017; Al-Jawad, 2019). These plans can be in the form of release schedules from dams and reservoirs (Pan et al. 2015; Li et al. 2015; Uen et al. 2018; Feng et al. 2019), groundwater storage decisions (Yeh et al. 2015; Singh et al. 2016; Wu et al. 2016), and location of wetlands (Dai et al. 2016; Scroggie et al. 2019), among others. Physically based models capture critical hydrology dynamics and are useful to model specific aspects of a river basin. However, they typically study the operation of a basin’s built infrastructure in isolation and do not exploit the properties of the networked infrastructure (Rani and Moreira, 2010; Shourian et al., 2008; Hatamkhani and Moridi, 2019). The traditional focus is on the built infrastructure, with less emphasis on the coordination with existing natural infrastructure or its development (e.g., wetland restoration). Modeling the operation of multiple components, including dams, reservoirs, wetlands, among other infrastructure, allows for a holistic analysis of the system, expanding the set of possible mitigation actions against floods and droughts. For instance, the synchronized operation of wetlands and dams can provide an additional source of water storage that can buffer against drought and can smooth flood peaks, delaying not only their maxima but reducing their instantaneous effects. To address these needs, this work package will develop scalable models to: (1) identify candidate locations for wetlands as a strategy for water management, and (2) prescribe coordinated operational plans between built and natural infrastructure to mitigate the adverse impact of floods and droughts. We will use the methods developed here in other WPs and NABWI contexts (e.g., with Flood MAR and reservoir re-operations).

Methods: Wetlands are critical natural assets to mitigate the effects of climate change (Erwin, 2019). Optimal location and restoration of wetland green infrastructure is a high-dimensional, complex multi-objective problem. Wetland restoration is costly and predicting the dynamical effects of a restored wetland requires the use of sophisticated, spatially explicit models of surface and ground water as well as driving functions of flood events. Existing restoration studies focus on wetlands in isolation, ignoring their interrelationship with other natural and built water infrastructure (Maleki et al., 2019; Sieben et al., 2011; Reis et al., 2019; Grand, 2020). When considering basin-scale coordinated operations between built infrastructure and wetlands, the space of possible wetland configurations (i.e., locations to focus cost-limited restoration efforts) is very large. The per-scenario time required to simulate responses to possible flood events is computationally impractical, making it impossible to pre-calculate all possible outcomes. As there may be multiple metrics of interest (e.g., storage needs vs. flood mitigation), an automated search through the space of possible operational plans is non-trivial as one metric may be improved at the cost of other metrics.

To address these challenges, we will develop a simulation-optimization method for the identification of a set of optimal (or near optimal) candidate wetlands and their coordinated operations with the existing built infrastructure. The prescribed operational plans will satisfy spatiotemporal water needs across a basin; for instance, water demands for agriculture and human consumption. The proposed simulation–optimization component will use a combination of VIC (Liang et al., 1994), SWAT (Neitsch et al. 2011) , CaMa-Flood (Yamazaki et al., 2011; Yamazaki et al., 2013) and WAM (Wurbs, 2005) to perform high-resolution basin-level analysis. The combination of these numerical simulation models will predict groundwater and surface water flows that result from forcing functions based on historical flooding data, which can also be used to extrapolate to plan for future floods that might vary in frequency and magnitude due to climate change. The simulation models will also incorporate the infrastructure dynamics from WPs 1-4, related to wetland and reservoir operations, agriculture, forest management, and groundwater. The simulation will provide basin performance metrics reflecting the quality of the given operational plans under external factors, including storage goals and response to extreme events (e.g., flood-peak reduction). We will utilize NSGA-II (Deb et al., 2002), a leading tool for multi-objective simulation optimization, to determine candidate operational plans and invoke the simulation. Because the set of possible wetland configurations is prohibitively large, we will complement the search procedure of NSGA-II with surrogate models from mixed-integer linear programming and deterministic optimization not only to increase the computational tractability of this optimization problem but also to increase the quality of the solution. We will use the results from WP6 as an alternative to the more complicated physically based models. Our modeling approach will also allow what-if analyses to evaluate the impact of external shocks affecting the basin, including droughts, floods, and increasing water demands, among others.

*Data and model integration via development of seamless application program interfaces (APIs)*

Organizing Question: As detailed in our recent report titled “Data Integration in the Service of Synthetic Research” [Kintigh et al. 2018], research of the scale that major challenges require demands the ability to explore multiple, large sets of primary data and models to provide the opportunity to discover important cross-dataset and cross-community patterns that would not be seen when comparing higher-level interpretations. Yet, building a single monolithic model to understand the dynamics of complex natural/engineered processes, such as those commonly seen in water resource management, and their effects over the affected systems, is generally not possible. Moreover, the component data, models, and simulators may be at different spatio-temporal scales, rely on different mathematical models, and be implemented using different computational frameworks. Therefore, the proposed work is aimed to answer the question whether it is possible to effectively connect and integrate multiple, potentially independently developed and implemented, off-the-shelf, component data sources, models, and simulators, each capturing a different natural, human, or built component of the complex system.

Background: Analysis and decision support tools for water resources management require the integration of large volumes of geological, hydrological, climate, and other data and models. The proposed work must integrate a wealth of spatio-temporal data of varying granularity. For example, social data from our participatory work, may be collected or available at the household, census tract, community, or county level, derived either from statistical samples or complete data. Environmental data may derive from irregularly positioned stream-flow gauges or weather stations. Topographic information may have a much higher spatial resolution than soil or geological substrate data. Satellite-based data may offer regional-scale information on vegetation communities but species-level data may only be available from dispersed small quadrats.

The key characteristics of the many of these data sets and models include the following: (a) voluminous, (b) multi-variate, (c) multi-resolution, (d) multi-modal, (e) spatio-temporal, and (f) often incomplete/imprecise. Once integrated, these data and models can enable researchers, analysts, and planners to both monitor and explore potential environmental and social consequences of engineering operations. Therefore, the ability to acquire, integrate, and share data and models made available in a variety of formats is key, and sharing data across models covering diverse aspects of the natural, economic, and environmental systems requires a semantic platform supporting adaptive contextualization of data collections and knowledge-bases (Cavalo et al. 2018, 2019; di Mauro et al. 2016; Cataldi et al. 2010). In an integrated complex coupled model, the component models/simulators should be able to feed data from each other to produce alternative (integrated) timelines, combining compatible states from each of these component models/simulators. Yet, data and model integration in its many forms (including mediator systems (Papakonstantinou et al. 1996; Adali et al. 1996), data lakes (Nargesian et al. 1989; Hai et al. 2016). peer-to-peer pay-as-you-go systems (Paton et al. 2012; Sarma et al. 2008; Salles et al. 2007; Jefferey 2008), query and user-driven solutions (Milo and Somech 2019; Zhang et al. 2018; Dragisic et al 2016; Miller et al. 2000; Cao et al. 2019; Candan et al. 2008) is becoming a significant bottleneck in the discovery of important cross-dataset patterns for grand challenge problems and in harnessing the available data to address the needs of data consumers (scientists, enterprises, governments). According to recent reports (Abadi et al. 2020; Stonebreaker and Ilyas 2018; Thirumuruganathan et al. 2018), data scientists spend 80-90% of their efforts on data integration and data wrangling.

Methods: We will, therefore, design and develop algorithms, tools, APIs, and services for integrated, unified access to the relevant data and models from existing databases and those created from other project tasks: the system will integrate structured and unstructured data. The developed tools and services will allow shared access to information, with appropriate safeguards, will provide a unified view of the available databases, and will share the global information structure for the private databases of project partners. The proposed work will require research and development into source description (Doan 2000), data and query translation across diverse and incompatible schemas (Shraga et al. 2020; de Una et al 2018; Bellahsene et al 2011; Halevy et al. 2003;Kementsietsidis et al. 2003; Madhavan et al. 2001; Pottinger and Bernstein 2003), entity resolution (Simonini et al. 2019; Efthymiou et al 2015; Vieira et al. 2019; ), and data and query processing across uncertain, heterogeneous sources and models (Behrens et al 2018; Poccia et al 2017; Gal et al. 2019; Gal 2011; Dong et al. 2009; Marie and Gal 2007; Doan et al. 2004; Bertossi 2006; Taylor and Ives 2006; Cao et al. 2010).

When needed to support the requirements of different tasks, data and observations will be rescaled or aggregated across spatio-temporal, semantic or quantitative hierarchies with minimal information loss. Results of semantic alignment or data re-scaling will be presented to the user, and feedback from individual or groups of users will enable semantic realignment and re-scaling of the data and scientific workflows. Building on outcomes of our prior NSF supported work, “III One Size Does Not Fit All: Empowering the User with User-Driven Integration” (#1016921), which led to “the Digital archaeological Record”(https://core.tdar.org) since supported by a number of organizations, including the Andrew W. Mellon Foundation and the National Endowment for the Humanities) and “DataStorm: A Data Enabled System for End-to-End Disaster Planning and Response” (#1610282), the proposed framework will provide an innovative architecture that facilitates the transformation of data into knowledge and enables the assessment of the robustness of analytical sequences.

Our proposed work will manage the acquisition of scientific data from diverse static and dynamic sources and process these data through complex scientific workflows. In a transdisciplinary research environment where scientists with different goals consume raw data and synthetic results, data streamed through the underlying workflows must be processed, filtered, and fused at different semantic or quantitative scales simultaneously. Accounting for differences in data scale or precision is relatively straightforward. However, scientific workflows that begin with the same input data but produce outputs for consumers from different disciplines may differ in crucial structural aspects and may rely on oversight by human technicians to different degrees. Consequently, a novel, multiscale scientific, workflow-management system is needed that can not only support needs of different disciplines, but support interdisciplinary interactions; they must bridge and integrate workflows of different semantic and quantitative scales and of varying structures. This involves developing a scientific data and workflow processing middleware that can process large amounts of data through workflows consisting of possibly distributed components and can support a range of visualization and presentation applications.

Contextualization, through appropriate metadata at multiple levels, will permit the semantic alignment of data, recognizing that the alignment operations will differ depending on the queries, models, and the implicit and explicit assumptions of the collaborative group, as well as the data sources at issue. Feedback from scientists will enable semantic re-aligning and re-scaling of the data as well as refining the scientific workflows for continuous improvement of end-to-end scientific reasoning.

The proposed group of services, therefore, consists of (a) Data/Model Description: Heterogeneous data and models for various types of disaster specific data feeds are described; (b) Data/Model Assessment: Data and models are validated and assessed. Note that model assessment metrics are critical in helping researchers and decision makers to select, among the alternatives, the most appropriate model for a task. The commonly used metrics to determine the quality of a given model and its applicability in a usage context include “correctness”, measuring the degree to which the model performs the required functions accurately, and “consistency,” quantifying the level of contradictions involved in the model (Mohagheghi et al. 2009). While these metrics are useful in general applications, they are not sufficient to quantify the quality and applicability of a model in all use scenarios. We will, therefore, develop formal and informal metrics (including, but not limited to, conciseness, complexity, interpretability, modularity, resilience, maintainability, reusability, freshness, precision, specificity, changeability, and confinement – Table 1) to compare models within a decision context. We will define an ontology to describe models, evaluation criteria to compare them, and mappings of the factors that are captured within the model; (c) Data/Model Ingress: Data and models are ingested, validated, stored, and indexed.

## Collaboration

The University and ERDC teams are naturally complementary and synergistic. The ASU team brings expertise in data science, engineering, AI, and optimization along with new approaches for the use of machine learning, while the ERDC team brings extensive knowledge and experience on modeling and management of water resources infrastructure. Both partners have experience in understanding how management influences water quality outcomes. The ERDC team also brings expertise to enable modelling and assessing drivers of water quality impacts in these complex and dynamic systems.

ERDC will participate in this project by providing technical review, data, and analysis advice such that the data scoping, modeling, and data science activities are aligned with USACE strategic goals. ERDC will also work closely with the ASU team on site selection based on data scoping and stakeholder engagement. Narrowing down potentially hundreds of candidate watersheds to one that could have good engagement from the USACE District, as well as enough local data to advance the modeling and data science initiatives will be a key role for ERDC researchers.

ERDC will further participate in this project by providing technical review, data, and analysis recommendations so that the MAR workflow is congruent with USACE Engineering References regarding Flood Frequency, Project Development and Risk analysis. This ensures that conclusions from the MAR screening tool conforms to accepted analysis approaches and has the best opportunity to be accepted at District levels.

ASU will also bring strong expertise in participatory modeling (detailed in other PMPs) to help strengthen the methods in these domains.

## Activities, Deliverables, POCs, and Schedule

## Communication and Technology Transfer Plan

Early on, our outcomes will be focused on communicating new research results with the academic and research community through conference publications, journal articles, and conference presentations. Major results and findings will be packaged into continuing credit courses for NEWN, EWN fact sheets, and podcasts. Progress will also be updated monthly on ASU’s EWN website blog, highlighting collaborations and ongoing activities. As relevant, policy white papers will be developed and distributed.

The stakeholder workshop planned for the end of year one will help to determine the major barriers as well as opportunities related to the integration of agricultural land management and water resources infrastructure. The outcomes of this workshop will be written into a policy brief to advance the longer-term goal of more integrated activities at the federal, state, and local levels. Similar outcomes will be developed from the local stakeholder participatory modeling that occurs in year 2 at the modeled and monitored site.

Where possible, data and modeling tools will be made publicly available for further use. This practice is common among both the ASU and ERDC collaborators (see: <https://github.com/EnvironmentalSystems/CE-QUAL-W2>)

## Expected Benefits and Impact

This project will not only advance the understanding of what aspects of the NWI and BWI are most likely affecting outcomes, but it will also work to identify and develop methods and tools to better understand and address to core challenges of optimality in design and maintenance. The modeling, simulation/emulation, and causal discovery outcomes can suggest ways to secure better water quality outcomes while still effectively managing water resources. In addition, the project has potential to increase aquifer resiliency, reduce flood risks to developed areas, reduce strains on existing flood control infrastructure, and improve ecosystem processes.

## Cited References

# Schedule and Funding

## Overall Requested Funding

## Requested Funding by Cost Type

## Requested Funding for Project Activities

The budget requested in this PMP does not include budgets from the domain PMPs and other cross-cutting. So here we describe only the budgets for activities associated with this PMP.

Costs for developing scientific articles, white papers, blog posts, and other communication products are included in the related activity budgets. Travel costs were split evenly across each activity.

## Cost-Sharing or In-Kind Contributions

The project team will leverage this collaboration to pursue additional funding opportunities together through agencies such as the USDA and NSF.

## Risk Management

Workflow: We have purposely included modeling, data synthesis, and new data collection throughout the project to provide opportunities to work both independently and synergistically as needed. While the activities are all connected to each other, there is little direct dependency which could potentially delay the timeline.

USACE Receptiveness: Another potential issue that could arise is that sites we think may have the best chance for integrated system management (through scoping) may not have willing district leads to help champion this integration. However, we are already working with an existing USACE-led project so we don’t anticipate this to be a major issue. Also, given the prevalence of levees and reservoirs in agricultural basins, we expect there to be many sites to choose from for our implementation purposes.

Application to NNBF: We are proposing to work with existing USACE projects that will install planted levees, such that we can monitor and model the system in our watershed-site based work. If we are unable to coordinate timing or otherwise with this existing project, we have two alternatives: (1) find another USACE or EWN team focused on implementation of a NNBF in one of our potential working sites to work with. Funds will be reallocated for monitoring and modeling of that NNBF; (2) work with teams studying the impact of reservoir management as a nature-based solution and use that for our NNBF. Funds will be reallocated for monitoring the effectiveness of changes in strategies.

# Project Delivery Team Members

## Principle Investigator Point of Contact

|  |  |
| --- | --- |
| University POC | ERDC POC |
| Name: K. Selcuk Candan | Name: Todd E. Steissberg |
| Title: Professor | Title: Research Environmental Engineer |
| Email: candan@asu.edu | Email: Todd.E.Steissberg@usace.army.mil |
| Phone: 480-965-2770 | Phone: 530-574-5572 |
| Department: SCAI | Org.Code: U433D90 |

## Project Delivery Team

|  |  |  |
| --- | --- | --- |
| PDT Member | Institution | Project Role |
| K. Selcuk Candan, [candan@asu.edu](mailto:candan@asu.edu), 480-965-2770 | ASU | University Co-PI and PMP lead. Data and model integration and causality aware ML. Will mentor graduate students and postdocs to perform data science and engineering tasks. Will plan and organize project meetings |
| Todd Steissberg, Todd.E.Steissberg@uasce.army.mil, 530-574-5572 | ERDC | ERDC Lead and project Co-lead. Will supervise ERDC staff to collaborate on data scoping task and lead model integration application tasks. |
| Huan Liu, huan.liu@asu.edu | ASU | Causal discovery and causality-driven generalizable machine learning and AI and data science. |
| Ted Pavlic, tpavlic@mainex1.asu.edu | ASU | Multi-objective decision making given different wetland portfolios; enhanced data on flood damage (water and $) through robust estimates of the value of enhanced storage for water supply (in water and $) |
| John Sabo, jsabo1@tulane.edu | Tulane | Project and Brazos PMP Lead. Will serve as the lead contact with the Brozos PMP. |
| Reepal Shah, rshah3@tulane.edu | Tulane | Integration of non-USACE hydrology with USACE hydraulics and reservoir operations through the lens of VIC-4L CaMaFlood |
| Qi Deng, qdeng6@asu.edu | ASU | Data/model integration (creating more efficient mashups of the many domain models involved). |
| Rebecca Muenich, rmuenich@asu.edu, 480-965-1077 | ASU | University Lead and project Co-lead. Will serve as the lead contact with the Agriculture PMP. |
| Amber Wutich, Amber.Wutich@asu.edu | ASU | Participatory modeling PMP lead; will mentor postdocs on completing outlined stakeholder engagement activities |
| Billy Johnson, Billy.E.Johnson@erdc.dren.mil | ERDC | ERDC collaborator working with Todd on advancing the integrated flow and water quality models |
| Ting Liu, [tliu154@asu.edu](mailto:tliu154@asu.edu) | ASU | ASU postdoctoral scholar who will work with Todd on advancing the SWAT-CE-QUAL-W2 model and help coordinate the data science activities with Selcuck |
| Monica Williams, Monica.S.Williams@asu.edu, 602-543-5075 | ASU | Communications lead |
| Tara Seaton, Tara.Seaton@asu.edu, 480-965-6804 | ASU | Financial Lead |

# Appendix A: Summary Table of Activities, Deliverables, Communication Plan, and Budget

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Activity | Deliverable | Communication/Technical Transfer Plan | Lead POC | Anticipated Start Date (Month and Year) | Anticipated End Date (Month and Year) | Estimated Budget for Completion | Percent Completed |
| TOTAL BUDGET: | | | | | | \*\*\* | |

# Appendix B: PDT Biographies



K. Selcuk Candan (Ph.D.) is a professor of computer science and engineering at Arizona State University (ASU), co-Director of ASU’s School of Computing and Augmented Intelligence (SCAI), and the director of ASU’s Center for Assured and Scalable Data Engineering (CASCADE). His primary research interest is in the area of management and analysis of non-traditional, heterogeneous, and imprecise (such as multimedia, web, and scientific) data. He has published over 200 journal and peer-reviewed conference articles, one book, and 16 book chapters. He has 9 patents. Prof. Candan served as an associate editor of for the Very Large Databases (VLDB) journal and IEEE Transactions on Multimedia. He is currently in the editorial boards of the ACM Transactions on Database Systems, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Cloud Computing, and the Journal of Multimedia. He has served in the organization and program committees of various scientific conferences. In 2006, he served as an organization committee member for SIGMOD’06, the flagship database conference of the ACM. In 2008, he served as a PC Chair for another leading, flagship conference of the ACM, this time focusing on multimedia research (MM’08). More recently, he served as a program committee group leader for ACM SIGMOD’10. He also serves in the review board of the Proceedings of the VLDB Endowment (PVLDB). In 2011, he served as a general co-chair for the ACM MM’11 conference. In 2012 he served as a general co-chair for ACM SIGMOD’12. In 2015, he served as a general co-chair for the IEEE International Conference on Cloud Engineering (I2CE). In 2017, he served as a program chair for the International Conference on Database Systems for Advanced Applications (DASFAA) and in 2019, he served as a program chair for the ACM Int. Conference on Multimedia Retrieval (ICMR). In 2022 Candan served as a General co-Chair for the ACM Web Search and Data Mining (WSDM) conference and is serving as a Program co-Chair for ACM SIGMOD’23. He has successfully served as the PI or co-PI of numerous grants, including from the National Science Foundation, Air Force Office of Research, Army Research Office, Mellon Foundation, and various industrial partners. He also served as a Visiting Research Scientist at NEC Laboratories America for over 10 years. He was a member of the Executive Committee of ACM SIGMOD and is currently an ACM Distinguished Scientist.

A person wearing glasses

Description automatically generated with medium confidenceHuan Liu (Ph.D.) is a professor of computer science and engineering at Arizona State University (ASU) and is affiliated with ASU’s Center for Assured and Scalable Data Engineering (CASCADE). Prof. Liu is a pioneering leader in Artificial Intelligence (AI), data mining, and data science. He has made seminal contributions to AI, social media mining and social computing, recorded in his over 500 peer-reviewed research publications. Prof. Liu has over 77,700 citations and an h-index of 122. He has been recognized as an IEEE Fellow (2012), ACM Fellow (2018), and AAAI Fellow (2019) for his contributions to AI, as a AAAS Fellow (2018) for broad contributions to data science, and, most recently, he was ranked #56 in the world and #37 in the US as one of the world’s top 1000 computer scientists; to place these numbers into further perspective, there are ~28,000 university professors in Computer Science (CS) in the US alone. He has served in numerous organizing roles as a Conference Chair, Program Chair, Senior Program Committee, etc. for many of the premiere data mining, AI, web mining, and big data conferences including AAAI, SIGKDD, IJCAI, ICDM, WSDM, SDM, etc. He is also the Editor in Chief of ACM Transactions on Intelligent Systems and Technology, the founding Chief Editor of Frontiers in Big Data and its Specialty Chief Editor on Data Mining and Management.



Rebecca Muenich (Ph.D.) is an Assistant Professor in the School of Sustainable Engineering and the Built Environment at Arizona State University. Dr. Muenich holds a B.S. in Biological Engineering from the University of Arkansas in 2009, an M.S. and Ph.D. in Agricultural & Biological Engineering from Purdue University in 2011 and 2015, respectively, and completed a postdoctoral position at the Graham Sustainability Institute at the University of Michigan from 2015- 2017. She joined Arizona State University as an Assistant Professor in the School of Sustainable Engineering and the Built Environment in the Fall of 2017. She has published 41 peer-reviewed articles and her work has been presented over 70 times at national and international conferences, including 5 invited presentations. She was a recipient of the ASABE Robert E. Stewart Engineering and Humanities Award from ASABE in 2011, named a New Face of ASABE in 2020, inducted into the Arkansas Academy of Agricultural and Biological Engineering in 2020, received an Outstanding Faculty Mentor award from the ASU Faculty Women’s Association in 2021, and was awarded an Early Career Award from the University of Arkansas College of Engineering in 2022.

A person smiling for the camera

Description automatically generated with medium confidence

Todd Steissberg (Ph.D., PE) is a Research Environmental Engineer at the U.S. Army Engineer Research and Development Center's Environmental Laboratory (ERDC-EL). Dr. Steissberg leads a team at the U.S. Army Engineer Research and Development Center (ERDC) that develops and applies water quality and environmental systems models for rivers, reservoirs, and watersheds. The objective of his research is to provide interdisciplinary teams with the tools and methods needed to perform integrated watershed-scale environmental impact assessments, improve water quality management and real-time operations, and design and implement ecosystem restoration projects that incorporate natural and nature-based features (NNBF) to improve the health and resiliency of ecosystems and communities. Dr. Steissberg obtained his B.S. in Civil Engineering from Washington State University, where he researched air pollution chemistry and transport processes and aquatic ecosystem restoration. Dr. Steissberg obtained his M.S. and Ph.D. in Civil and Environmental Engineering from University of California, Davis, while serving as a NASA Earth System Science fellow at NASA/JPL, researching satellite-based remote sensing, physical limnology, and water quality. As a Postdoctoral Researcher at the Tahoe Environmental Research Center, John Muir Institute of the Environment, University of California, Davis between 2008 and 2010, he developed methods to characterize nearshore and offshore water quality and its spatial-temporal variability using satellite and field measurements. Dr. Steissberg lead development and application of water quality models and geospatial tools as a Senior Hydraulic Engineer at the U.S. Army Corps of Engineers' Hydrologic Engineering Center (USACE-HEC) between 2008 and 2019. In 2019, he transitioned to ERDC to continue his research and build a team of researchers to address complex issues in water quality and water resource modeling, ecosystem restoration, and environmental resiliency and adaptation of freshwater and coastal ecosystems, civil works infrastructure, and military installations under the threat of climate change. Dr. Steissberg is the lead developer of ERDC’s CE-QUAL-W2 model and the Corps Library for the Environmental Analysis and Restoration of Watersheds (ClearWater) and continues to lead water quality capability development for the HEC models (HEC-RAS, HEC-ResSim, and HEC-HMS).



Ting Liu (Ph.D.) is a Postdoctoral Research Scholar who joined ASU Environmental Engineering Dr. Rebecca Muenich’s research team at the beginning of 2021. His current research focuses on locating and tracking non-point source pollutants and their impact on river water qualities. Before joining ASU, he was on a one-year Post-doc appointment with Purdue University researching industrial water use in the US. He finished a master’s degree in Civil Engineering and a Ph.D. degree in Environmental Engineering, both at Stevens Institute of Technology, Hoboken, NJ. His research interests mainly focus on water resources topics: water use, surface hydrology, machine learning in the application of water-environmental problems, water collection systems, the interaction between surface water and groundwater, and the transport of contaminants in both surface water and groundwater.



Giuseppe Mascaro is an Associate Professor in the School of Sustainable Engineering and the Built Environment at Arizona State University (ASU). Dr. Mascaro holds a “Laurea” (B.S. and M.S.) in Civil Engineering and a Ph.D. in Hydrology from the University of Cagliari, Italy, and a Master in Environmental Management at Scuola Superiore Sant’Anna in Pisa. Before joining ASU, he worked in the industry for two years and as a postdoc and research scientist at New Mexico Tech and the University of Cagliari. Dr. Mascaro has 53 published papers since 2007; he is first author on 19 papers and single author on three. Dr. Mascaro’s research experience that is relevant to this project includes: application of physics-based hydrologic models at hyper-resolutions; verification of climate model outputs; and statistical analyses and modeling of hydrologic variables, with focus on precipitation.

Daniel Siegel is a geospatial scientist with Earth Genome, a San Francisco based non-profit. He holds a bachelor degree in Physics from New York University and a Masters of Science in Environmental and Water Resource Engineering from the University of Texas at Austin. His research in graduate school focused on using remote sensing and GIS technology to quantify and visualize the water balance of river basins. Previously, he worked at ESRI, where he built GIS applications for the Living Atlas of the World and supported development of the National Water Model, at Jupiter Intelligence, where he engineered data pipelines for climate hazard models, and at the Pacific Disaster Center, where he led the development of hazard and vulnerability maps. At Earth Genome, he works on managed aquifer recharge (MAR) projects, and is currently supporting a study on MAR feasibility in California’s San Joaquin Valley.

A person with blonde hair

Description automatically generated with low confidence

Amber Wutich (Ph.D.) is a President’s Professor of Anthropology and Director of the Center for Global Health at Arizona State University. Her two decades of community-based fieldwork are concerned with how inequitable and unjust resource institutions impact people’s well-being, especially under conditions of poverty. An expert on water insecurity and mental health, she directs the Global Ethnohydrology Study, a cross-cultural study of water knowledge and management. Wutich maintains longstanding ties in her field sites in Paraguay and Bolivia, and manages a strategic alliance between la Universidad Nacional de Itapúa (Paraguay) and ASU. An ethnographer and methodologist, Wutich has authored 150+ peer-reviewed publications, co-authored 4 books, edits the journal Field Methods, and directs the NSF Cultural Anthropology Methods Program. Her teaching has been recognized with awards such as Carnegie CASE Arizona Professor of the Year. Wutich has raised over $44 million in research funds, as part of collaborative research teams, from the National Science Foundation, USDA, and other funders.

# Appendix C: Model Brief

This project uses, among others, two process-based models to help simulate the integrated management of upland agricultural management (SWAT) and downstream reservoir management (CE-QUAL-W2). We provide very brief overviews of the two models for further context for this proposal.

SWAT/SWAT+:

SWAT and SWAT+ are semi-distributed hydrologic models driven by daily or sub-daily climate inputs that then simulate surface hydrologic and biogeochemical processes. The model was developed by the United States Department of Agriculture and has been widely applied across many geographies, climates, and agricultural systems globally. The model uses spatially-explicit data on soils, land use, topography, climate, and land management to simulate the movement of water and pollutants throughout a watershed system. Interactions with groundwater, ponds, wetlands, streams, reservoirs are simulated, but with simple, reduced-order models. The model is free and open source, allowing easier modifications such as with CE-QUAL-W2. SWAT versions were built with fortran, while SWAT+ versions are C based.

More information about SWAT: <https://swat.tamu.edu/>

CE-QUAL-W2:

CE‐QUAL‐W2 is a two‐dimensional (2D), longitudinal/vertical, hydrodynamics and water quality model that characterizes vertical and longitudinal dynamics and changes in reservoir and riverine systems. USACE and other organizations around the world have developed more than 300 CE-QUAL-W2 water quality model applications on reservoirs, lakes, rivers, and estuaries, generating fundamental insight into current and future reservoir and river system flow and water quality conditions and aiding environmental impact assessments, reservoir operations decision-making, and the planning, design, and evaluation of water resources systems and infrastructure. Recent enhancements in CE-QUAL-W2 Version 4.5 enable more granular modeling of Total Dissolved Gas (TDG), sediment diagenesis, and the buoyancy, rise, and fall of algae. Prediction of variable velocity of cyanobacteria, one especially useful model enhancement, advances the simulation of Harmful Algal Bloom (HAB) conditions. Together these new features make CE-QUAL-W2 an even more powerful tool for addressing national and global environmental challenges.

More information can be found on our GitHub site: [https://github.com/EnvironmentalSystems/CE-QUAL-W2](https://urldefense.com/v3/__https:/github.com/EnvironmentalSystems/CE-QUAL-W2__;!!IKRxdwAv5BmarQ!YMn8ITqGwEOSAOr_l6j1D_ktGFTc7O0kdtqHbFBJoFgWXBVg09xxOdt5bSQ-HmtYutk8PBZNlM67_fZq3zalpYek6UsaJWYkFTIm$)

ClearWater

The Corps Library for Environmental Analysis and Restoration of Watersheds (ClearWater) is a modular library developed by the Environmental Laboratory, U.S. Army Engineer Research and Development Center (ERDC). ClearWater provides environmental modeling capabilities that leverage existing hydrologic and hydraulic (H&H) models. ClearWater's water quality kinetics and vegetation simulation modules include the following:

* NSM: Nutrient Simulation Module (NSM-I and NSM-II)
* TSM: Temperature Simulation Module
* GCSM: General Constituent Simulation Module
* CSM: Contaminant Simulation Module
* MSM: Mercury Simulation Module
* SSM: Solids Simulation Module
* RVSM: Riparian Vegetation Simulation Module

The ClearWater water quality engines compute transport (advection and diffusion) of heat and mass in water bodies. The reservoir water quality engine is integrated with HEC-ResSim and is the foundation of reservoir water quality capabilities in HEC-ResSim. The river water quality engine computes transport using hydrodynamic data provided by the two-dimensional (2D) HEC-RAS river model. Integration of environmental capabilities with H&H models enables efficient development and application of new capabilities and linking water quality, vegetation, and other environmental data and processes with existing visualization and reporting capabilities.

More information can be found on our GitHub site:

https://github.com/EcohydrologyTeam/ClearWater

ClearWater-Riverine

The ERDC Environmental Laboratory (ERDC-EL) is developing a state-of-the-art water quality modeling system, Clearwater-Riverine, that simulates temperature and advanced nutrient cycling in branching river systems and floodplains, incorporating hydrodynamic, water quality, and meteorologic inputs from multiple data sources and models. ClearWater-Riverine enables the evaluation of system vulnerabilities and identification of adaptation pathways to improve the resilience of floodplain ecosystems to environmental stresses, which include increasing frequency and intensity of extreme precipitation events and decreasing freshwater flows. Water quality kinetics, heat budget, and transport simulation capabilities in ClearWater-Riverine are furnished by ERDC's ClearWater modules. The ClearWater-Riverine framework links the capabilities provided by the ClearWater modules and advection-diffusion engine with the two-dimensional (2D) HEC-RAS model. HEC-RAS models have been developed for most of the watersheds around the world. By leveraging these existing models, ClearWater-Riverine provides a cost-effective, data-driven tool for impact assessment, planning studies, and the restoration and management of aquatic ecosystems.

More information can be found on our GitHub site:

https://github.com/EcohydrologyTeam/ClearWater-Riverine