Pacific Northwest
NATIONAL LABORATORY

Generative AI for Environmental Data Synthesis, Generation, and Augmentation

Z. Jason Hou
PNNL

U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

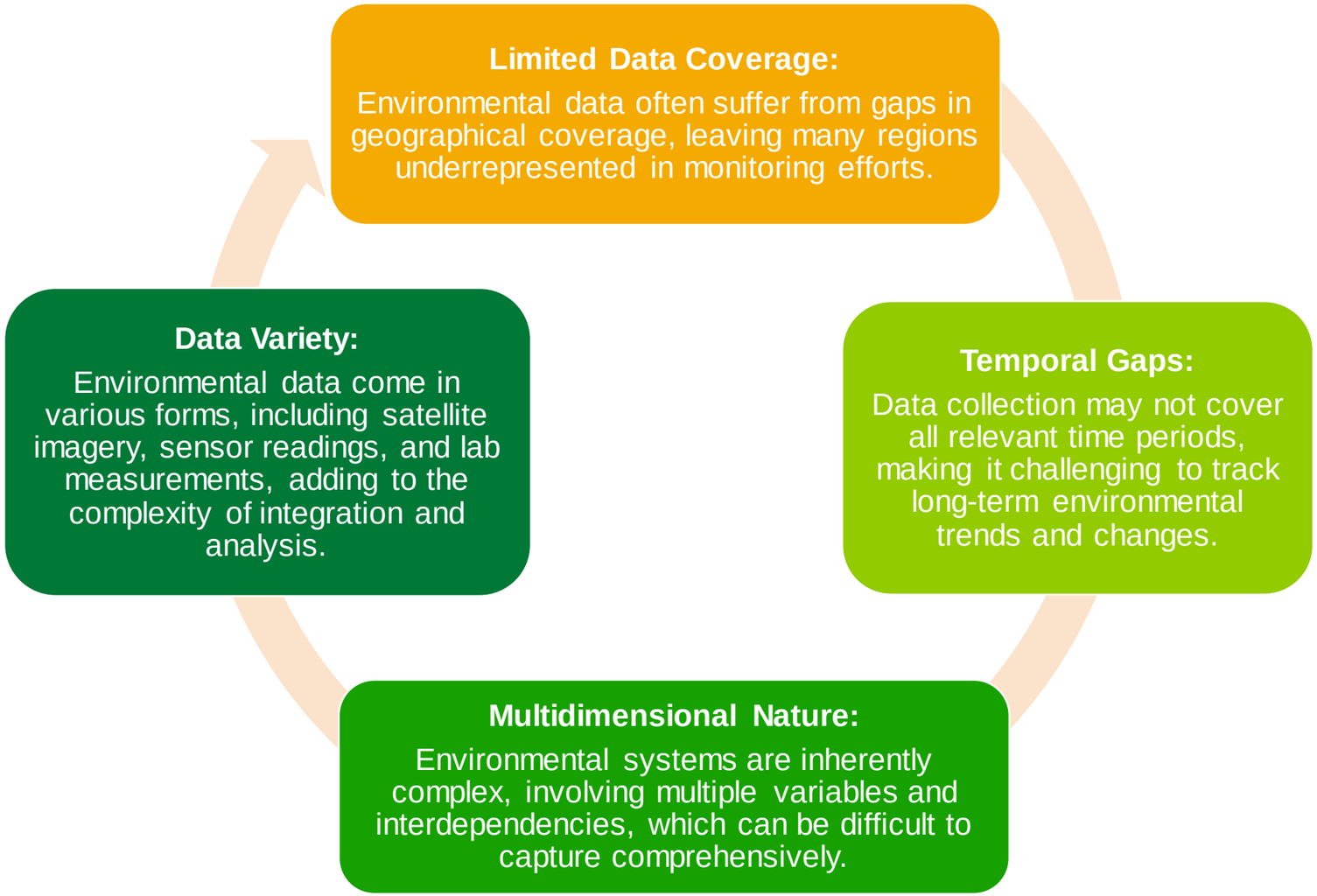
Introduction

Generative AI – a subset of artificial intelligence techniques that focus on creating new data, images, or content rather than just analyzing or classifying existing data.

These algorithms (e.g., Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs)) are driven by the idea of learning patterns and structures within existing data and using that knowledge to generate novel, often highly realistic, data points.

It can help bridge gaps in environmental data, contributing to environmental remediation, monitoring, understanding and predicting biogeochemical processes, among other applications.

Challenges in Environmental Data



Motivation for Generative AI

- Produce data in regions where traditional monitoring is challenging or impossible, providing a more comprehensive view of environmental systems.
- Generate high-quality, realistic data points that closely resemble real-world observations.
- Bridge temporal gaps, enabling the analysis of environmental changes at finer time scales.
- Offer a cost-effective alternative, reducing the financial burden associated with data acquisition.
- Model intricate relationships can lead to more accurate data generation.
- Adapt to different environmental contexts and data types, making them versatile tools for various applications.

Generative AI Algorithms

➢ Key Algorithms

- **Generative Adversarial Networks (GANs)**
  - GANs consist of two neural networks, a generator, and a discriminator, engaged in a competitive game.
  - The generator creates data samples, while the discriminator tries to distinguish between real and generated data.
  - Training continues until the generator produces data indistinguishable from real data.
- **Variational Autoencoders (VAEs)**
  - VAEs are probabilistic models that learn to represent data in a lower-dimensional latent space.
  - They generate new data by sampling from this latent space, allowing for the creation of novel data samples.

➢ Common Characteristics

- Both GANs and VAEs learn patterns and structures from existing data to create new, often highly realistic data points.
- These algorithms have been widely adopted in various domains, including computer vision, natural language processing, and environmental sciences.

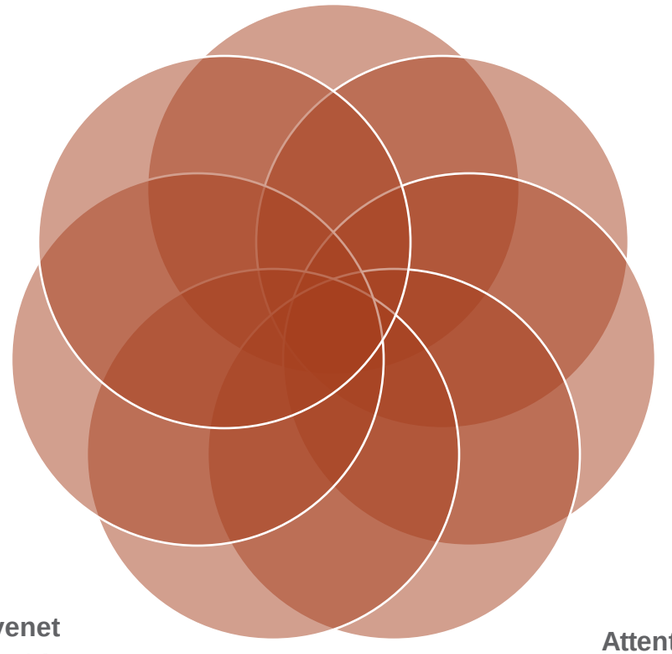
Other Generative AI Algorithms

PixelRNN and PixelCNN

These models are used for generating images pixel by pixel. They are designed to capture the dependencies between pixels in an image to produce highly realistic images.

Recurrent Neural Networks (RNNs)

Though not exclusively generative models, RNNs can be used to generate sequences of data in various domains, such as natural language processing and music generation.



Autoregressive Models

These include models like LSTM-based and Transformer-based autoregressive models. They generate sequences of data one element at a time, considering the previous elements in the sequence.

Deep Convolutional Inverse Graphics Networks (DC-IGN): These models are used for generating 3D scenes from 2D images, making them valuable for applications in computer graphics and computer vision.

Flow-Based Models

Models like RealNVP (Real Non-Volume Preserving) and Glow use invertible transformations to generate data. They are particularly useful for generating high-resolution images.

WaveGAN and Wavenet

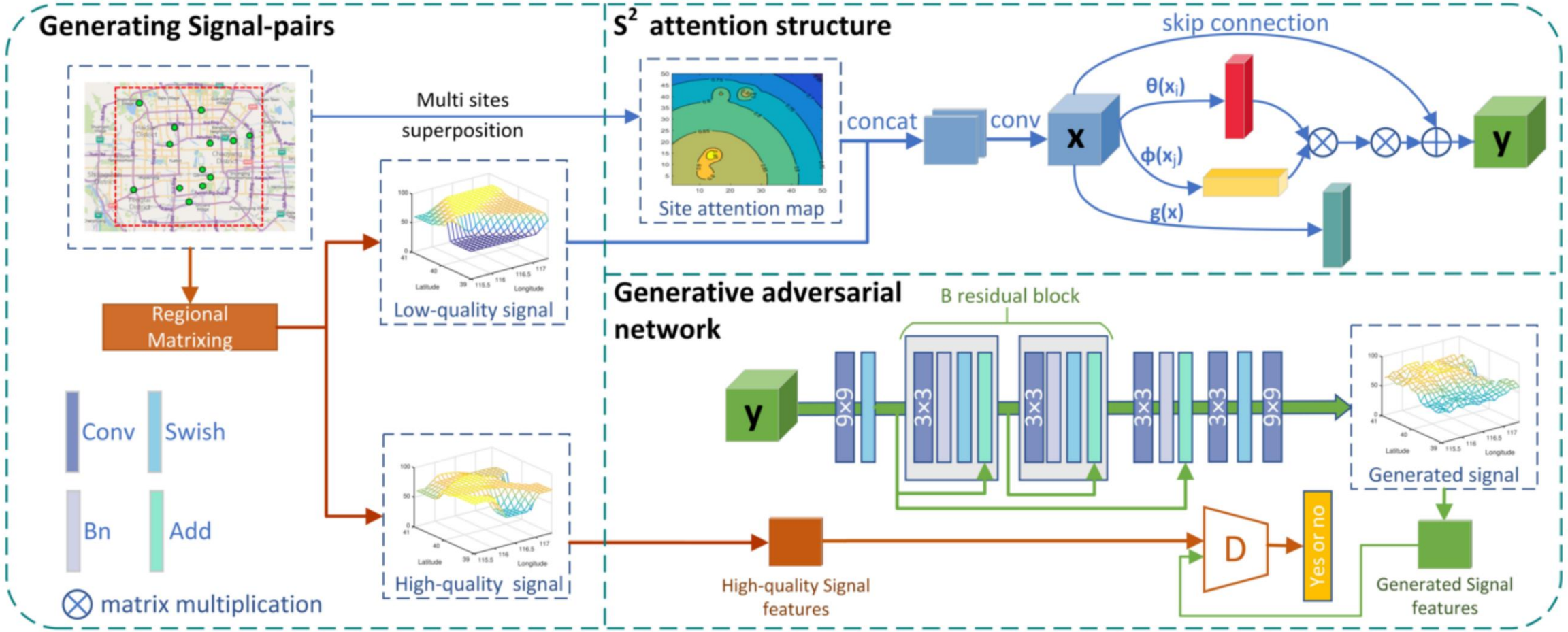
These models are used for generating audio waveforms. WaveGAN focuses on generating audio samples directly, while Wavenet models the conditional probability distribution over audio waveforms.

Attention-Based Models

Models like the BigGAN and DALL-E use attention mechanisms to generate images and text, respectively. They are known for their ability to generate highly detailed and coherent content.

Interpolation

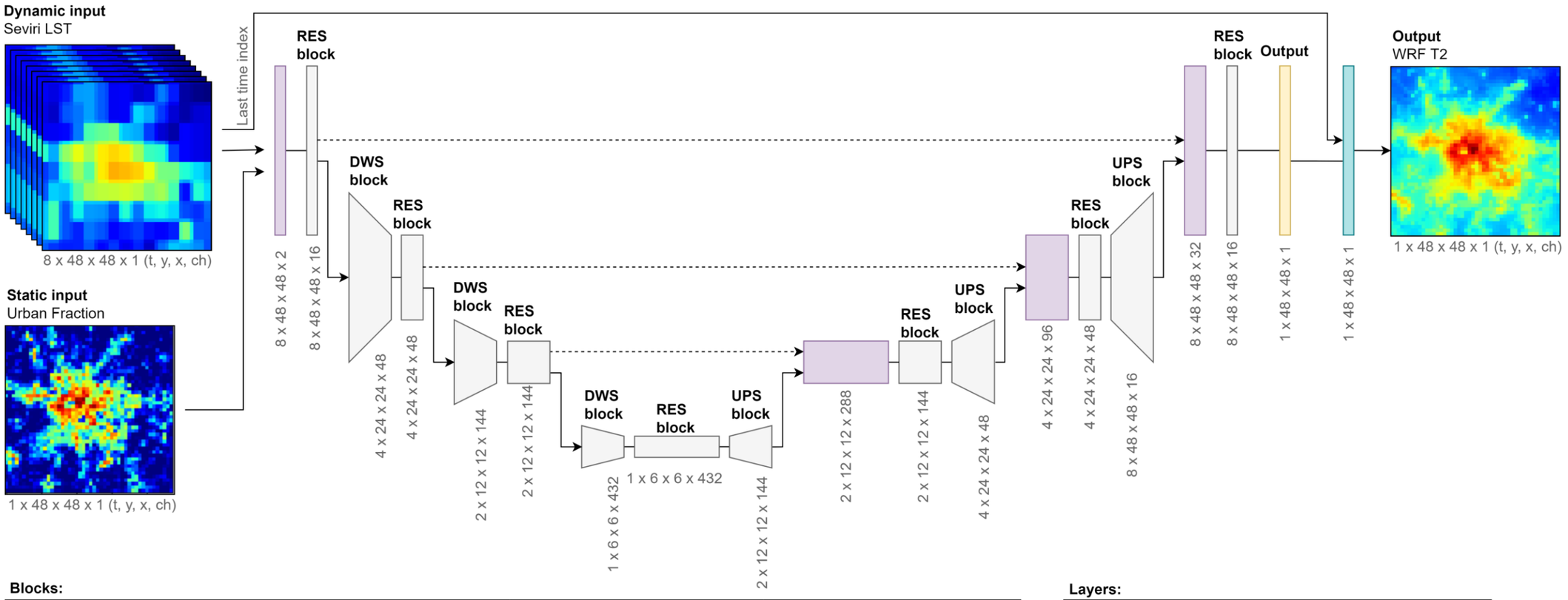
- Generative AI can analyze patterns and relationships within the available data (e.g., coarse land surface temperature) to create new data points that fit seamlessly within the observed range.

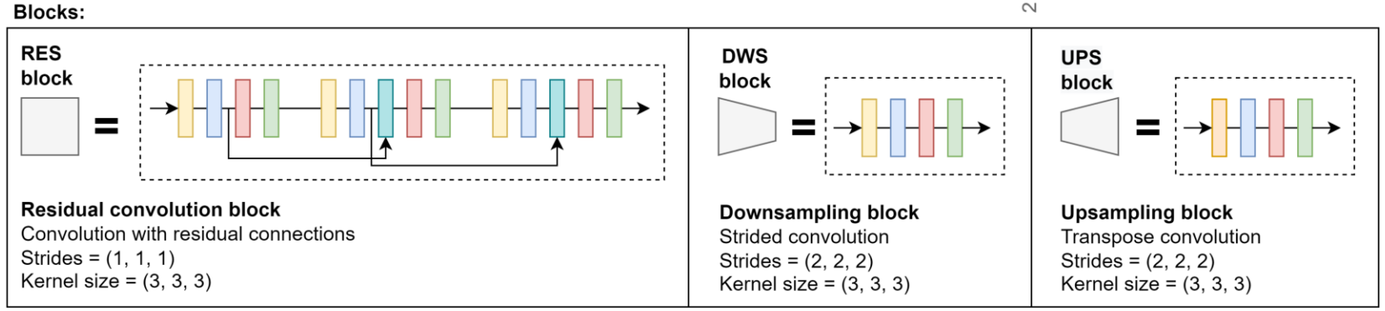


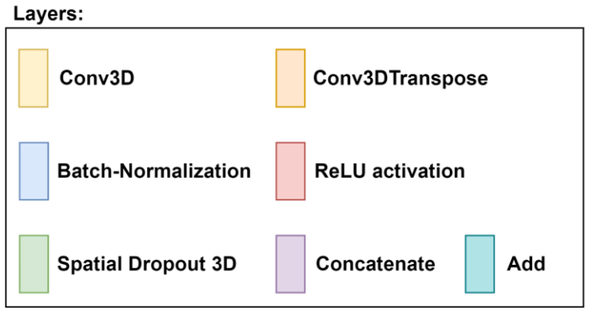
(Gao et al., 2023, SI-AGAN)

Interpolation

- Generative AI can analyze patterns and relationships within the available data (e.g., coarse land surface temperature) to create new data points that fit seamlessly within the observed range.



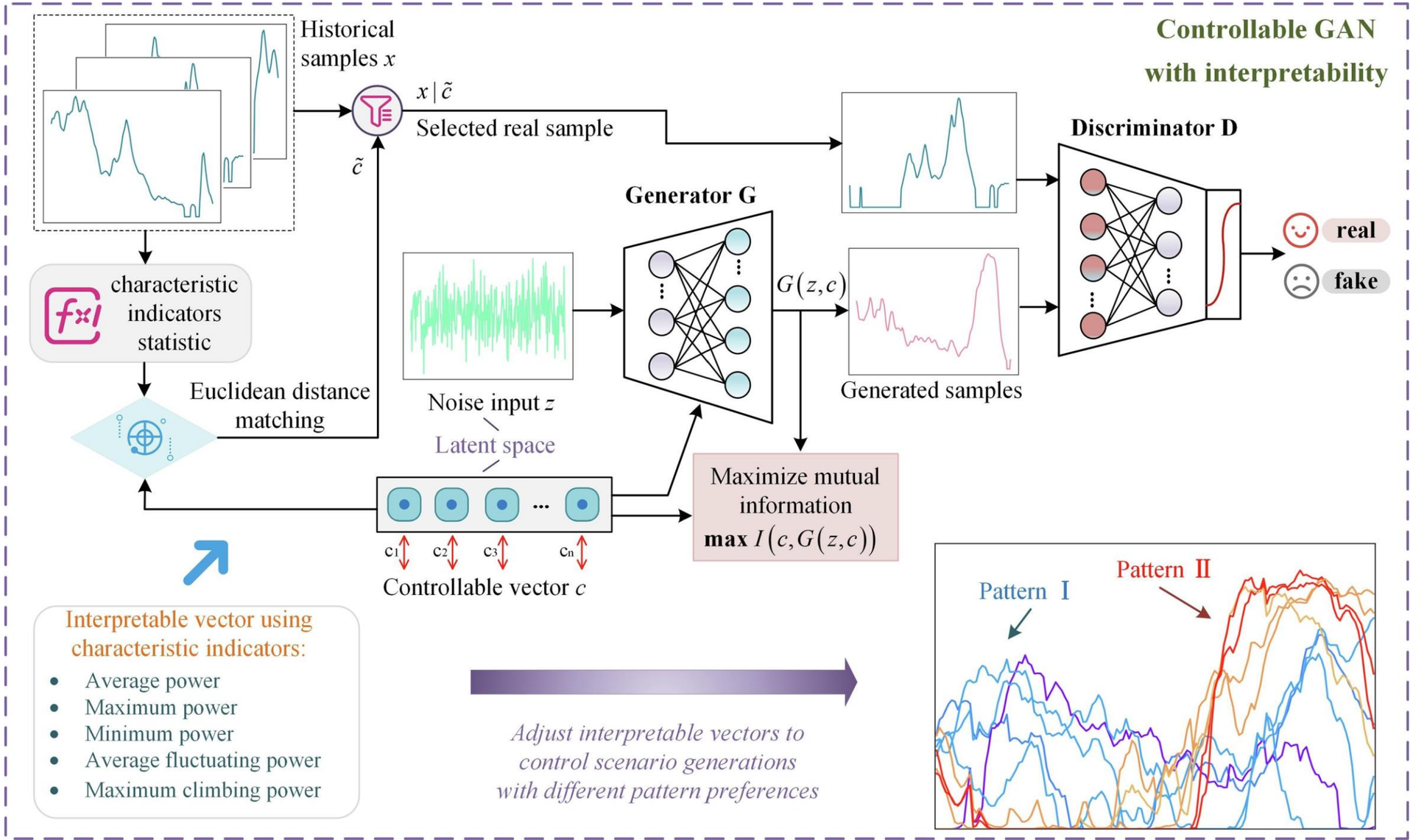




(Afshari et al., 2023, Remote Sensing)

Extrapolation

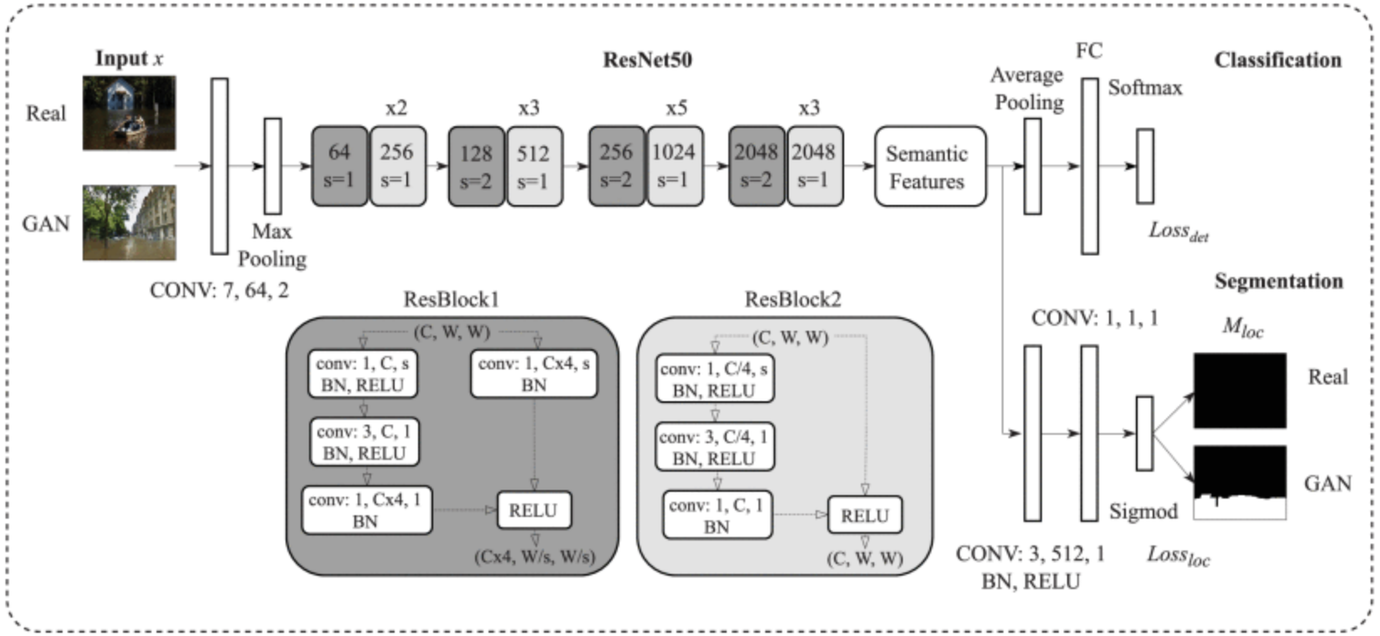
- Generative AI leverages learned patterns to make educated predictions about future data points. An example of generating future trends and outcomes in wind and solar energy is shown below.

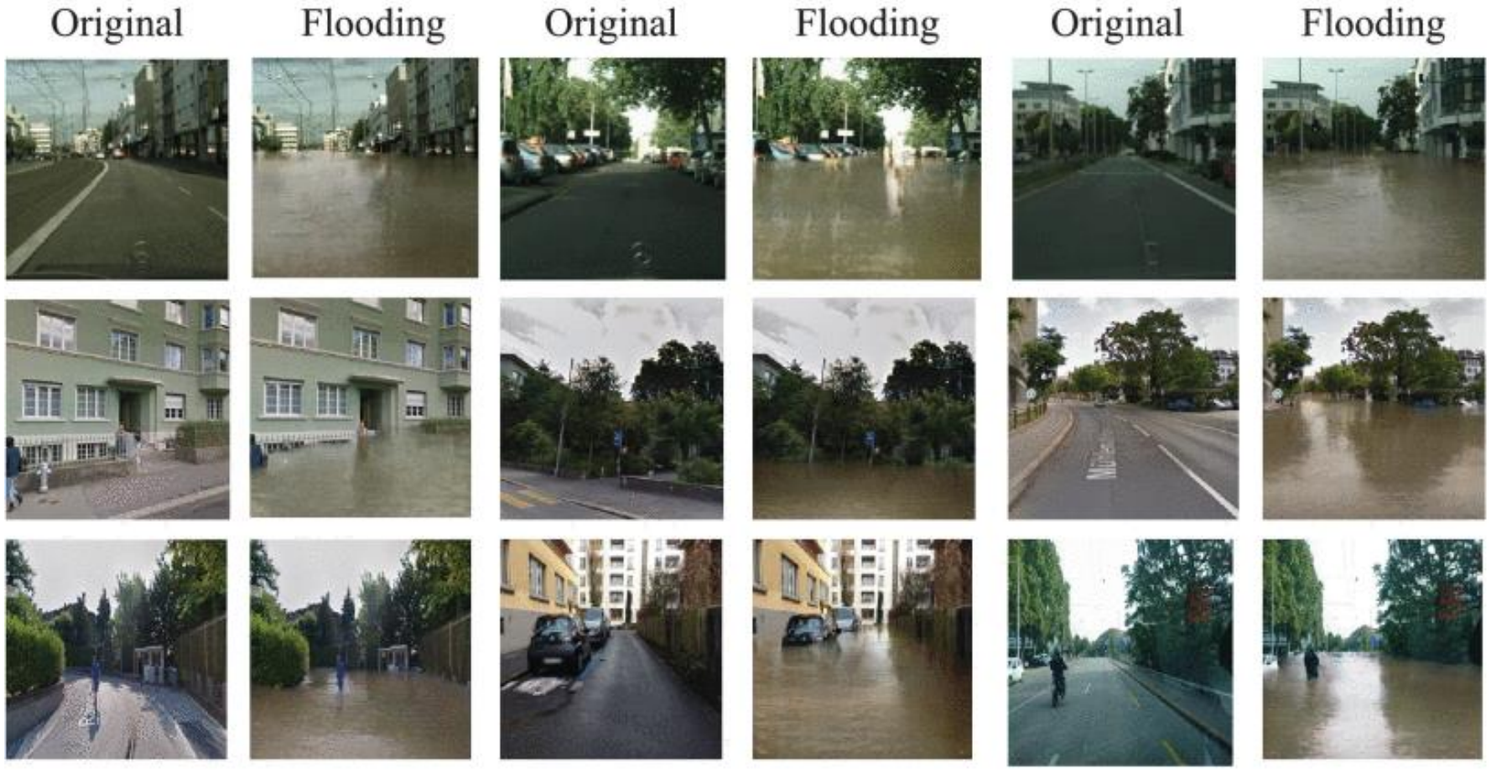


(Dong et al., 2022, Applied Energy)

Scenario Generation

- Generate AI can generate synthetic environmental scenarios (e.g., flood), including various factors and variables. These scenarios help test and refine mitigation strategies under different conditions.

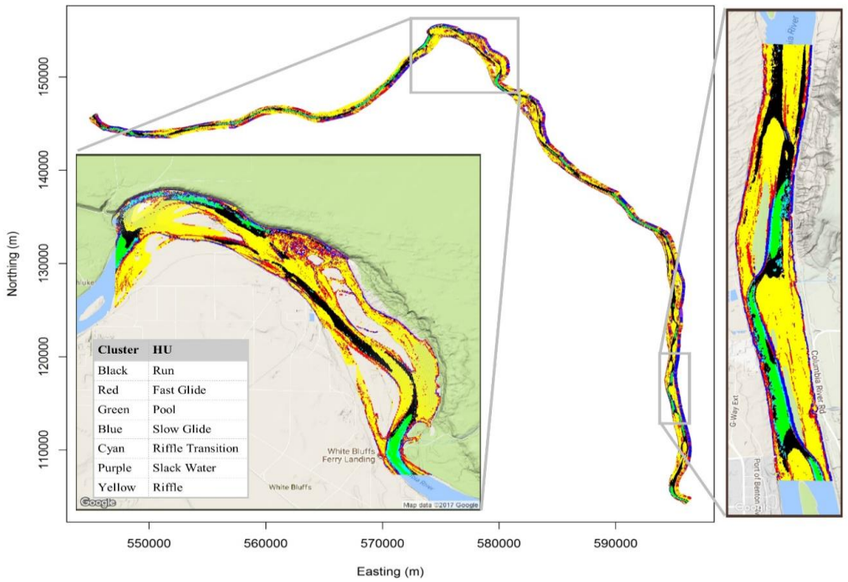




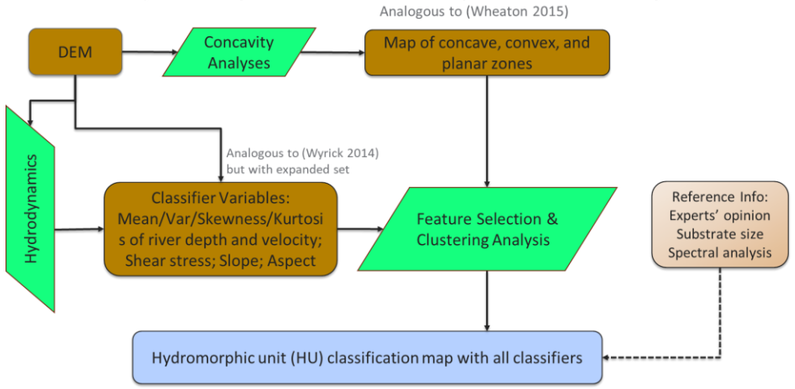
(Wang et al., 2022, IEEE IVMSP)

Enhancing Predictive Models

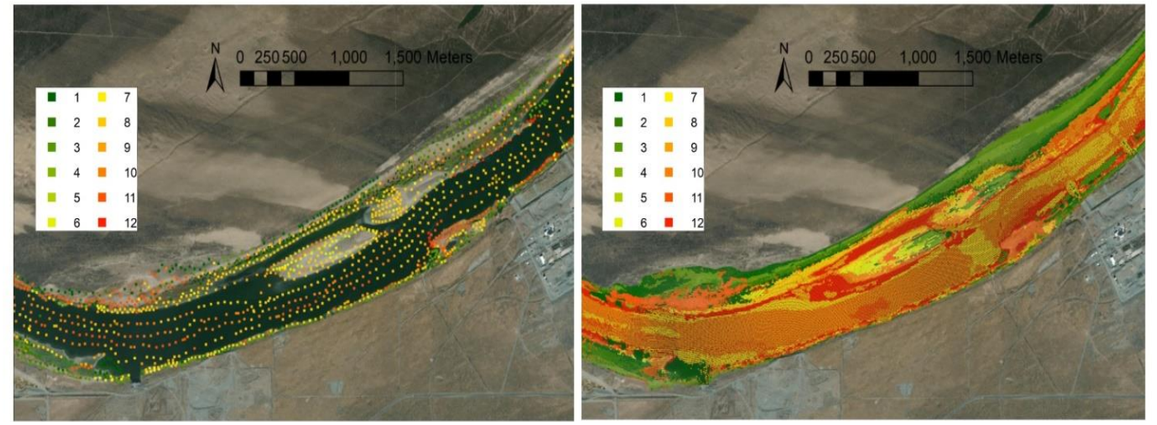
- Generative AI can augment datasets used for more reliable environmental model training or simulation. It generates additional data points, filling gaps and providing a more diverse and representative dataset.

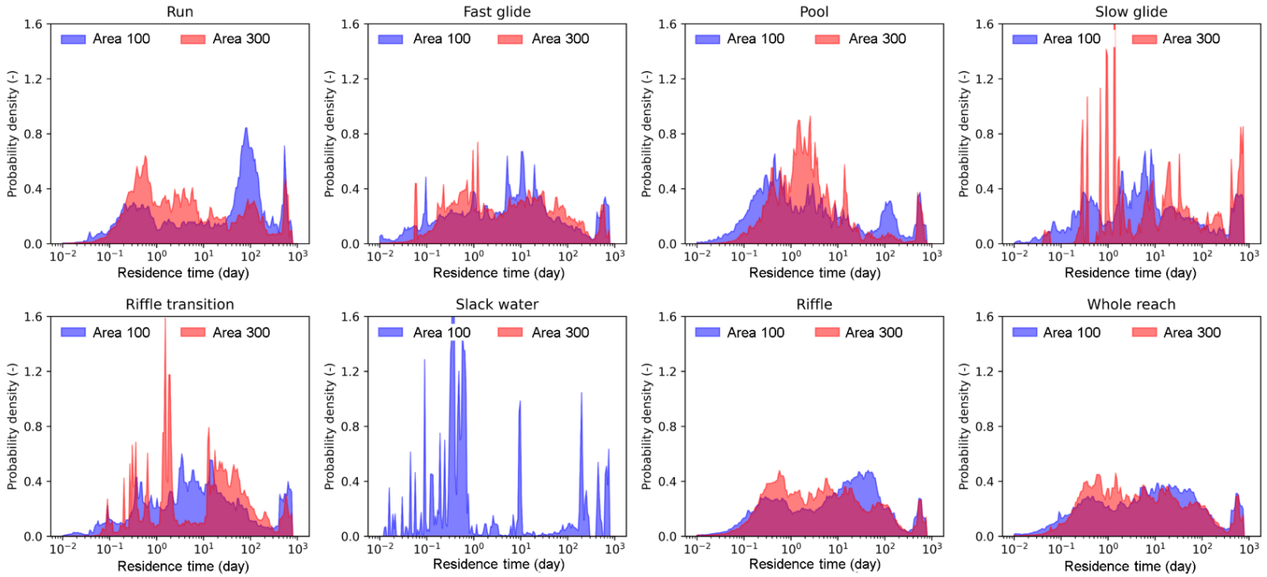


Hydromorphic classification (Hou et al., 2021)



Hydraulic property augmentation and mapping (Ren et al., 2020)





Simulated RTD distributions per HUs (Bao et al., 2023)

Challenges and Concerns

Accuracy Challenges

- Generative AI-generated data may not always perfectly represent real-world observations.
- Factors such as model limitations and incomplete training data can affect accuracy.

Bias in Data Generation

- Generative AI models can inadvertently introduce biases present in the training data.
- Addressing and mitigating these biases is crucial to avoid skewed or unfair representations.

Data Privacy Concerns

- Generating synthetic data often involves using real data as the basis, raising privacy issues.
- Striking a balance between data utility and privacy protection is essential.

Ethical Considerations

- The ethical use of generative AI requires careful attention.
- Ensure transparency, accountability, and fairness in data generation and application.

Validation and Verification

- Implement robust validation techniques to assess the accuracy and reliability of generated data.
- Verification processes help ensure that data aligns with real-world observations.

Ethical Guidelines

Responsible AI Principles	• Adhere to responsible AI principles, including transparency, fairness, accountability, and privacy.
Data Ethics	• Prioritize ethical data collection, use, and sharing throughout the generative AI process. • Ensure data sources are reputable, and respect privacy and consent.
Bias Mitigation	• Implement bias mitigation strategies to prevent discrimination and skewed representations in generated data.
Transparency and Explainability	• Make generative AI processes transparent and explainable to users and stakeholders. • Provide information about the data sources, model architecture, and decision-making criteria.
Stakeholder Engagement	• Involve relevant stakeholders, including environmental experts, communities, and regulators, in decision-making and validation processes.
Continuous Monitoring	• Continuously monitor and evaluate the impact of generative AI on environmental systems. • Be prepared to adapt and correct course as needed.
Compliance with Regulations	• Ensure full compliance with data protection and environmental regulations relevant to the application of generative AI.
Ethical Review	• Consider conducting ethical reviews of generative AI projects to identify and address potential ethical concerns.
Public Awareness	• Educate the public about the use of generative AI in environmental systems and its ethical considerations.

Data Validation Techniques

Cross-Validation	• Employ cross-validation, where the generative AI model is tested on data it hasn't seen during training. • This helps evaluate the model's ability to generalize to unseen data.
Comparison with Real Data	• Compare generated data with real-world observations to identify discrepancies. • This direct comparison provides insights into the model's performance.
Statistical Metrics	• Use statistical metrics like mean absolute error, root mean square error, or correlation coefficients to quantify the difference between generated and real data.
Spatial and Temporal Patterns	• Analyze spatial and temporal patterns in generated data to ensure they align with known environmental processes. • Identify anomalies or inconsistencies.
Expert Evaluation	• Involve domain experts and environmental scientists in the validation process. • Their expertise can provide valuable insights and judgment.
Sensitivity Analysis	• Perform sensitivity analysis to assess how variations in model parameters impact data quality. • This helps identify areas for model improvement.
Iterative Model Refinement	• Continuously refine the generative AI model based on validation feedback. • Iteration enhances data reliability over time.

Prospects

Advanced Models

- Expect more sophisticated generative AI models to capture finer environmental data details.

Multi-Modal Synthesis

- Generative AI will evolve to combine various data sources, offering holistic environmental insights.

Real-Time Monitoring

- Real-time generative AI data generation will revolutionize environmental monitoring.

Bias Mitigation

- Research will prioritize robust bias detection and mitigation techniques for fair and equitable data generation.

Ethical Frameworks

- Development of standardized ethical frameworks for generative AI in environmental systems.

Cross-Domain Applications

- Generative AI's adaptability will lead to cross-domain applications beyond environmental systems.

Conclusion

Summary

- Generative AI transforms environmental data synthesis and modeling, addressing data gaps and enhancing understanding.
- Ethical considerations are vital, including bias mitigation and privacy protection.
- Validation techniques ensure the reliability of generative AI-generated data.
- Generative AI has already made significant impacts in environmental science.

Needs

- Embrace generative AI's potential for environmental applications.
- Collaborate across disciplines to ensure responsible and ethical AI use.
- Stay updated with the latest research and advancements.
- Join interdisciplinary partnerships to tackle environmental challenges with generative AI.

Pacific Northwest
NATIONAL LABORATORY

Thank You!

U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy