Data Management and Sharing Plan
CPS-FR: Physics-Informed Machine Learning for Resilient Microgrid Control

Principal Investigator: Ehsan Reihani

This Data Management and Sharing Plan outlines the comprehensive strategies for managing, preserving, and sharing the research data generated during the proposed CPS-FR project on "Physics-Informed Machine Learning for Resilient Microgrid Control." The plan ensures compliance with NSF policies while promoting open science principles and supporting reproducible research in cyber-physical energy systems.

1 Types of Data and Materials

The project will generate diverse datasets and research materials critical to advancing microgrid control technology. Experimental and simulation data will encompass comprehensive physics-informed neural network training datasets containing microgrid state trajectories, control actions, and system responses under various operating conditions, coupled with real-time control performance data from NVIDIA Jetson AGX Orin hardware-in-the-loop experiments. The research will produce extensive communication network simulation data including delay patterns, packet loss statistics, and IEEE 2030.5 protocol performance metrics, alongside multi-agent consensus convergence data from distributed optimization experiments. Critical safety verification will generate Control Barrier Function datasets with barrier evolution traces and constraint satisfaction metrics, complemented by economic dispatch optimization results and ADMM convergence characteristics.

Microgrid system data will include synthetic microgrid topologies and electrical parameters compliant with IEEE test systems, integrated with grid-forming inverter behavioral models and comprehensive control parameter sets. The project will develop renewable energy generation profiles for solar PV and wind systems with realistic variability patterns, accompanied by load demand profiles specifically designed for critical infrastructure applications including hospitals, emergency services, and research facilities. Communication

network topologies and latency characteristics will be extensively characterized to support distributed control validation across diverse operational scenarios.

Software and code repositories will feature a complete physics-informed neural ODE implementation in PyTorch with embedded power system dynamics, integrated with multi-agent reinforcement learning algorithms providing formal consensus guarantees. The codebase will include graph neural network-enhanced distributed optimization solvers and a comprehensive Control Barrier Function safety layer implementation for real-time constraint enforcement. The vendor-agnostic bump-in-the-wire controller firmware for NVIDIA Jetson AGX Orin platform will be accompanied by hardware-in-the-loop simulation interfaces and real-time control software.

Machine learning models and training materials will encompass trained physics-informed neural networks for microgrid state prediction and control, complemented by graph neural network models for distributed optimization acceleration. The project will produce deep reinforcement learning policies with proven stability guarantees, supported by comprehensive model checkpoints, hyperparameter configurations, and detailed training convergence logs. Validation datasets and performance benchmarking results will provide thorough documentation of model performance across diverse operational conditions.

Documentation and educational materials will include comprehensive technical documentation for all software components and hardware interfaces, accompanied by detailed user guides for BITW controller deployment and configuration. Tutorial materials for physics-informed machine learning in power systems will be developed alongside curriculum modules for undergraduate courses on cyber-physical energy systems. Workshop materials and training resources for industry practitioners will ensure broad dissemination and adoption of the research outcomes.

2 Data and Metadata Standards

The project will adhere to established standards ensuring interoperability and long-term accessibility across multiple technical domains. Power system data standards will include IEEE Common Data Format (CDF) and PSS/E RAW format for power system network data, integrated with IEC 61970 Common Information Model (CIM) for comprehensive system component descriptions. Real-time measurements will utilize IEEE C37.118 synchrophasor data format, while distribution system modeling and analysis will employ OpenDSS format. Detailed electromagnetic transient studies will be documented using PSCAD format to ensure compatibility with industry-standard simulation tools.

Machine learning data standards will incorporate HDF5 format for large-scale numerical

datasets with hierarchical organization, enabling efficient storage and retrieval of complex training data. Trained model serialization will utilize ONNX (Open Neural Network Exchange) format to ensure cross-platform compatibility and deployment flexibility. Experiment tracking and model lifecycle management will employ MLflow format, while structured data analysis will be conducted using NumPy arrays and Pandas DataFrames. Hyperparameter configurations and metadata will be stored in JSON format for human-readable documentation and automated processing.

Communication and control data will adhere to IEC 61850 format for substation automation and smart grid communications, ensuring seamless integration with existing utility infrastructure. IEEE 2030.5 (Smart Energy Profile) will be implemented for demand response and distributed energy resources coordination. Industrial control system integration will utilize established Modbus and DNP3 protocols, while IoT device communication in microgrid applications will employ MQTT and CoAP protocols for efficient data exchange and real-time monitoring.

Metadata and documentation standards will incorporate DataCite Metadata Schema for comprehensive dataset description and citation management, complemented by Dublin Core metadata standard for digital resource description. Web-accessible data documentation will utilize Schema.org vocabulary to enhance discoverability and automated indexing. README files will follow established best practices for computational reproducibility, while Jupyter Notebooks will provide embedded documentation and visualization to facilitate understanding and replication of research methodologies.

3 Access, Sharing, and Privacy Policies

The project implements a comprehensive open access framework where all synthetic datasets, source code, and software packages will be made publicly available through GitHub repositories with comprehensive documentation to ensure broad accessibility and reproducibility. Physics-informed neural network models and training datasets will be shared through public repositories with appropriate licensing to facilitate research advancement and commercial adoption. Educational materials and tutorial content will be freely available under Creative Commons licenses, while research publications will be made available through institutional repositories and preprint servers to maximize scientific impact and knowledge dissemination.

Controlled access components will be carefully managed to balance openness with security and privacy requirements. Hardware-specific performance data from NVIDIA Jetson AGX Orin platforms will be anonymized before public release to protect proprietary information while maintaining scientific value. Industrial collaboration data will be governed by com-

prehensive data sharing agreements with appropriate access controls to respect commercial interests and intellectual property rights. Critical infrastructure load profiles will be synthesized and de-identified to protect sensitive information about power system operations, while communication network vulnerability data will be restricted to authorized researchers through secure access portals to prevent potential security risks.

The licensing and legal framework will promote maximum utility while respecting intellectual property rights. Source code will be licensed under MIT License enabling broad reuse and commercial applications without restrictive constraints. Datasets will be licensed under Creative Commons Attribution 4.0 International (CC BY 4.0) to ensure proper attribution while allowing derivative works and commercial use. Machine learning models will be shared under Apache 2.0 license for maximum flexibility in deployment and modification, while documentation and educational materials will use Creative Commons Attribution-ShareAlike 4.0 (CC BY-SA 4.0) to promote knowledge sharing and collaborative improvement.

Privacy and security measures will be implemented throughout the data lifecycle to protect sensitive information and maintain stakeholder trust. Differential privacy techniques will be applied to protect sensitive operational data while preserving analytical utility for research purposes. Data anonymization protocols will systematically remove personally identifiable information and proprietary system details before public release. Secure data enclaves will be established for collaborative research with industry partners, providing controlled environments for sensitive data analysis. Regular security audits will ensure ongoing compliance with cybersecurity best practices and evolving threat landscapes.

4 Policies for Re-use and Re-distribution

The project implements comprehensive policies to promote responsible re-use and re-distribution of research outputs while maintaining appropriate quality standards and attribution requirements. Attribution and citation requirements establish that users must cite original data sources, publications, and DOIs when using shared resources, with clear attribution guidelines provided with each dataset and software package. Academic citation standards will be enforced through license agreements, while commercial users will be encouraged to acknowledge contributions in product documentation to ensure proper recognition of research contributions.

Derivative works and modifications are explicitly encouraged under CC BY 4.0 and MIT license terms to promote innovation and adaptation of research outputs. Users must document any modifications or extensions to original datasets or code, while version control systems will track contributions and maintain attribution chains to preserve intellectual her-

itage. Improved algorithms and extensions will be welcomed as contributions to main repositories, fostering collaborative advancement of the technology through community-driven development.

Commercial use and technology transfer are actively supported through policies that permit commercial applications under the chosen open source licenses. Technology transfer partnerships will be established to facilitate industry adoption, while startup companies and established utilities will be supported in deploying BITW controllers for real-world applications. Revenue-sharing agreements may be negotiated for significant commercial deployments to ensure sustainable funding for continued research and development activities.

Quality control and standards are maintained through rigorous validation processes that ensure the integrity and reliability of shared resources. Peer review processes will validate contributed datasets and code improvements, while continuous integration testing will ensure code quality and reproducibility across diverse deployment environments. Documentation standards will be enforced for all shared resources, and community governance models will be established for major software projects to maintain long-term sustainability and collaborative development practices.

5 Archiving and Preservation Plans

Comprehensive archiving and preservation strategies ensure long-term accessibility and integrity of research data through multiple redundant storage systems and standardized preservation practices. Institutional repositories provide the foundation for data preservation, with primary storage provided through California State University, Bakersfield institutional repository supported by mirror repositories established at collaborating institutions for redundancy. Long-term preservation commitments will be secured through institutional partnerships, while regular data integrity checks will be performed on archived materials to detect and prevent data corruption or loss over time.

Public archives and DOI assignment facilitate broad accessibility and permanent citation of research outputs through established disciplinary repositories. Final datasets will be deposited in IEEE DataPort for engineering community access, while software packages will be archived in Zenodo for long-term preservation and DOI assignment to ensure persistent identifiers for all major research outputs. Research publications will be deposited in arXiv and institutional repositories to maximize accessibility, and educational materials will be shared through appropriate disciplinary repositories to support ongoing education and knowledge transfer activities.

Version control and change management systems maintain detailed histories of all re-

search outputs and facilitate collaborative development while preserving attribution and modification tracking. Git version control will track all code changes with detailed commit messages, while dataset versions will be managed using data version control (DVC) tools to maintain reproducibility and traceability of analytical results. Semantic versioning will be applied to all software releases to clearly communicate compatibility and feature changes, and migration plans will be developed for evolving data formats and standards to ensure continued accessibility as technologies advance.

Backup and disaster recovery procedures provide robust protection against data loss through multiple layers of redundancy and tested recovery processes. Daily automated backups will be maintained on secure, geographically distributed servers to protect against localized disasters and system failures. Cloud storage services will provide additional redundancy and accessibility, while disaster recovery procedures will be tested annually and updated as needed to maintain effectiveness. Critical data will be stored in multiple formats to ensure long-term accessibility even as software technologies evolve and change over the project lifetime and beyond.

6 Data Management Roles and Responsibilities

Clear definition of roles and responsibilities ensures effective implementation of the Data Management and Sharing Plan through coordinated efforts across all project participants and stakeholders. Principal Investigator responsibilities encompass overall oversight of Data Management and Sharing Plan implementation, ensuring that all activities align with project objectives and maintain compliance standards. The PI ensures compliance with NSF policies and institutional requirements through regular monitoring and documentation of data management activities, while coordinating with collaborating institutions and industry partners to maintain consistent practices across all project sites. Annual review and updating of data management practices ensures that procedures remain current with evolving best practices and technological advances.

The Data Management Coordinator position, filled by a designated graduate student or postdoctoral researcher, provides dedicated expertise and daily oversight of critical data management functions. This coordinator maintains daily oversight of data collection, processing, and documentation activities to ensure consistent application of established protocols and standards. Training team members on data management tools and best practices ensures that all project participants understand and can effectively implement required procedures, while quality assurance and validation of shared datasets and code maintains the integrity and reliability of all public research outputs.

Technical team responsibilities distribute data management tasks among all research participants while maintaining accountability and coordination across the project team. Research team members receive comprehensive training on data management protocols to ensure consistent implementation of established procedures and standards. Individual researchers bear responsibility for documenting their contributions according to established standards, creating comprehensive records of their research activities and outputs. Regular team meetings include data management progress updates to identify and address challenges while maintaining coordination across all project activities, and peer review processes for validating research outputs before sharing ensure quality and accuracy of all public research materials.

Industry collaboration management addresses the complex legal and technical requirements associated with partnerships between academic research and commercial entities. Legal and administrative support for data sharing agreements ensures that all collaborative activities comply with institutional policies and legal requirements while protecting the interests of all parties. Coordination with utility companies and technology partners maintains effective communication and alignment of objectives across diverse organizational contexts, while management of proprietary data and intellectual property considerations balances open science principles with legitimate commercial interests. Facilitation of technology transfer and commercialization activities ensures that research outcomes achieve practical application and societal benefit while maintaining appropriate protections for all stakeholders.

The PI will ensure full compliance with NSF policies and actively promote open science principles throughout the project duration and beyond. This comprehensive plan will be reviewed annually and updated as needed to reflect changes in project scope, technological advances, or evolving best practices in research data management. The plan demonstrates our commitment to advancing the field of cyber-physical energy systems while ensuring maximum societal benefit from publicly funded research.