MODULAR TOOLING AND ADAPTIVE SPATIAL MLOPS: AN MCP-ENABLED AGENTIC DELIVERY OF NATURAL HAZARDS INSIGHTS AND PREDICTIONS TO MULTIDISCIPLINARY STAKEHOLDERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Extreme weather events expose human and economic frailty, of which is reliant upon a stable physical environment. To minimize the societal risk incurred from these natural bouts of instability, multi-disciplinary resilience efforts, spanning science, engineering, policymaking, and more, are essential. The issue, however, arises when the needed collaborative efforts are impeded by the inability to derive findings outside of one's own expertise or to discern the outputs and outcomes of another. Hence, this work novelly proposes a shared space solution, an adaptable, flexible and automated agentic-enabled application, to facilitate collaboration among experts and stakeholders by the provision of practical climate resilience content. The application presently affords physical and social characteristics, extreme-event risk information, interactive and downloadable maps, predictive models, and resilience-planning tools, along a modular architecture designed for the scalable incorporation of additional capabilities. Operationally, via the interface, the end-user may either directly access an instrument or select an agent to orchestrate instruments via a Model Context Protocol (MCP) tool layer, where the Large Language Model (LLM) interprets requests, selects and invokes tools, and synthesizes outputs for the end-user. Further, two implemented capabilities are highlighted in this paper. First, a flood susceptibility prediction model, running via a Machine Learning Operations (MLOps) lifecycle, is demonstrated. The automation includes the weekly consumption of new radar and sensor observations, triggering the retraining and evaluations of graph convolutional neural network and logistic models. Models and sensors are then dynamically selected based on predefined performance criteria. At the daily scale, same- and next-day precipitation forecasts are ingested, producing localized 1km x 1km predictions. A second highlight is the LLM-enabled multi-risk tool, which integrates physical, extreme-event risk, and sociodemographic information, allowing users to identify areas exhibiting combinations of vulnerabilities relevant to their particular resilience inquiries. Ultimately, by providing the stakeholder facing application, the immediate aim is to assist New Yorkers in preparing for and mitigating the consequences of extreme weather events, while providing a foundational framework that may be extended to other urban environments.

1 INTRODUCTION

Severe weather events may exert considerable havoc and destruction upon surrounding communities. Particularly, in urban environments, marked by high population and building densities, property and infrastructural loss may be substantial, and human calamity measurable and tragic. Thus, as urbanization rises, resilience efforts are all the more essential (Agonafir et al., 2023). Resilience, the ability to mitigate harm and to overcome disaster, of which vulnerability and hazard intersect, preliminarily, necessitates the identification of susceptibilities, which encompass social, physical, psychological and financial aspects (Pelling, 2012). For instance, in the dire event of Hurricane Katrina, there were 1,833 fatalities and $108 billion in damages (National Weather Service, 2025); of those displaced, 88,000 were elderly, 183,000 were school-aged children, and 30% were 1.5 times

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

below the poverty limit (Falk et al., 2005); more, it has been shown that many residents suffered immediate and prolonged post-traumatic stress disorder and other psychological concerns (Weems et al., 2007). Correspondingly, natural disasters enact multifaceted effects, and the expertise and energies of various domains, including engineering, the sciences, economics, and policy, are needed to recognize the scale of vulnerabilities, assess risk, and create preventative and restorative measures.

While action and resilience implementations induced by the collective insights of multiple domains are desired, challenges in computational alignment often thwart these fruitful unions. Firstly, computational models may employ specific technical stacks and coding practices, creating difficulties when attempting to build upon another's work; moreover, semantic differences may pervade even when using the same programming language (Fan & Sun, 2012). While real-time collaborative strategies have been developed to resolve such discrepancies during model construction (Williams et al., 2000; Fan & Sun, 2012; Mankawade et al., 2026), these approaches often presume that collaborators are known and present during development, limiting accessibility by future or unidentified users. Furthermore, the integration of varying models, particularly across disciplines, require specialized processing skills. For example, General Circulation Models (GCMs), which use numerical methods and parameterizations of unresolved physical processes to simulate weather conditions, oftentimes, produce multidimensional grid-based outputs (Kochkov et al., 2024; USDA Northwest Climate Hub, 2023). These outputs may serve as useful inputs to impact models, which estimate system conditions following a physical phenomenon (Gilardi & Sharf, 2002); however, seamless incorporation is difficult due to formatting inconsistencies. Specifically, a climate-based impact model, such as one for flood prediction, may use tabular, categorical, or topographical inputs. Hence, to reconcile the differing formats, processing skills woud be required. On a last note, barriers to data discovery also affect data reusability (Panwar & Sen, 2020; Fakhruddin et al., 2022). For instance, the data associated with domain-specific models are often maintained within their respective hosts' repositories, and outside researchers may be unaware of their existence. Ultimately, there remains a need for a city-focused environment, which would provide obtainable and usable outputs and analyses, preprocessed data, and insights, within an accessible language.

To address these challenges, this work proposes the **NYC Resilience Application**. The application was initially conceived and developed for use during the city-wide, three-day hackathon, *Urban Futures: Co-Creating Climate Resilience in NYC*, hosted by the American Museum of Natural History and the Learning the Earth with Artificial Intelligence and Physics (LEAP) NSF Science and Technology Center at Columbia University, in collaboration with NVIDIA. The event aimed to convene community and civil groups to innovate climate resilience solutions. To support participants in developing these solutions, an initial Large Language Model (LLM)-enabled application was created to provide neighborhood-level access to climate insights. The approach facilitated access to multiple datasets, with nearly half of participants discovering datasets previously unknown to them. The application's value extended beyond the hackathon itself, as one participant later used the tool to advocate for fire-hydrant sprinkler caps in a high heat risk area. Presently, the expanded application provides a stakeholder-facing space for exploring urban and social characteristics, extreme-event risks, geospatial information, localized flood predictions, multi-vulnerability locations, and resilience-planning tools. Capabilities are accessible through direct interfaces, with selected scientific functions additionally exposed as Model Context Protocol (MCP) tools for LLM-based agent orchestration. Automated data and modeling workflows incorporate new observations into the historical dataset and retrain models, supporting the integration, continued operation, and evolution of multidisciplinary resilience information.

2 Related Works

While current literature is sparse, there are agentic AI frameworks and MLOps pipelines that have been applied in environmental and climate research. Involving the MCP standard, Jalalzadeh Fard et al. (2026) demonstrate an agentic AI architecture, allowing for the retrieval of National Weather Service weather data. Essentially, by exposing weather consumption through a collection of independently callable MCP tools, agents are able to obtain location-specific observations (Jalalzadeh Fard et al., 2026). Additionally, in regards to MLOps and automation, Bhaskar et al. (2026) employ an automated modeling architecture towards the detection of urban water scarcity. With the inclusion of a dashboard, their pipeline automates daily data ingestion, model retraining and performance monitoring for operational use (Bhaskar et al., 2026). Lastly, and perhaps the most

related work to this paper, Malamas et al. (2026) propose an Agentic Climate Intelligence architecture for climate-resilient smart cities. Their proposed framework employs specialized agents to support climate forecasting, risk assessment, and coordinated decision-making, while incorporating human oversight (Malamas et al., 2026). Overall, the current literature remains limited; yet, agentic AI, MCP, and automated MLOps in earth system sciences are rising implementations, expected to continue expanding.

This paper extends beyond the previous works of the existing environmental and climate related automation and agentic AI. Indeed, the application here uniquely allows for the incorporation of varied scientific capabilities, including predictive models, geoprocessed data, risk maps, urban features and automated MLOps pipelines, accessible directly or via agentic orchestration, thereby exceeding a single, modular data retrieval tooling capability or an architecture centered on one predictive workflow, as seen in Jalalzadeh Fard et al. (2026) and Bhaskar et al. (2026), respectively. In comparison to the broader agent autonomy proposed by Malamas et al. (2026), where decision-making is intentionally distributed among specialized agents, this work's system, firstly, contrasts by the establishment of a more explicit agentic boundary. Instead of replacing expertise with autonomous agent-generated scientific analyses, this paper's agentic-enabled design allows for the selection of appropriate MCP tools, determination for when additional tool calls are needed, and the synthesis of outputs. Decidedly, the underlying scientific computations of this work remain expert-developed implementations. Secondly, Malamas et al. (2026) presents an architectural formalization and evaluation design; whereas, this work demonstrates an executed operational system within a metropolis, utilizing localized datasets and features. In summary, while the related works provide important elements of agentic AI, MLOps, and modeling, this work is distinguished by a multi-faceted integration within an applied, extensible-tooling, stakeholder-facing application designed with bounded agent autonomy to provide immediate facilitation of multidisciplinary, climate resilience collaboration. An overview of the NYC Resilience Application, from raw-data processing to stakeholder-facing capabilities, is shown in Fig. 1.

3 NYC Resilience Application: Resilience Capabilities

3.1 Geoprocessed Characteristics, Risk and Planning Tools

There are three dashboard functions providing urban, sociodemographic, and disaster-risk information. Physical characteristics include elevation, slope, building footprint area, building count, and building footprint density, selected based on their relevance to flood and heat event occurrence or impact (Li et al., 2020; Macintyre et al., 2018; Agonafir et al., 2022; 2023; An et al., 2025). The application additionally provides 141 Census sociodemographic characteristics and FEMA National Risk Index (NRI) coastal flooding, inland flooding, and heat risk scores. These data originated from multiple publicly accessible sources and required specialized geospatial processing to harmonize their differing representations at the NYC census-tract level, as detailed in Appendix Table A1. Fig. 2a–c demonstrate samples of the processed features on the production-level dashboard.

The two interactive cost estimator tools provided in the application are the rain garden and green roof planners. Both estimators utilize general cost estimates from the Environmental Protection Agency (EPA): `https://www.epa.gov/`. However, the user may optionally enter in a cost basis from their own source. Then, the cost estimate will be calculated within the application. For the rain garden estimating tool, the urban characteristics (processing described above) of slope and elevation are also incorporated.

4 NYC Resilience Application: Intelligent and Automated Capabilities

4.1 LLM-Enabled Multi-Tract Query

The Multi-Tract Query uses Anthropic Claude Sonnet 4.6, accessed through Amazon Bedrock services, to understand a natural-language request and translate into structured criteria for deterministic census-tract filtering. The LLM is called at a temperature of 0.0 with a system prompt defining the available variables and a four-class ordinal schema. Specifically, the data available to the tool are

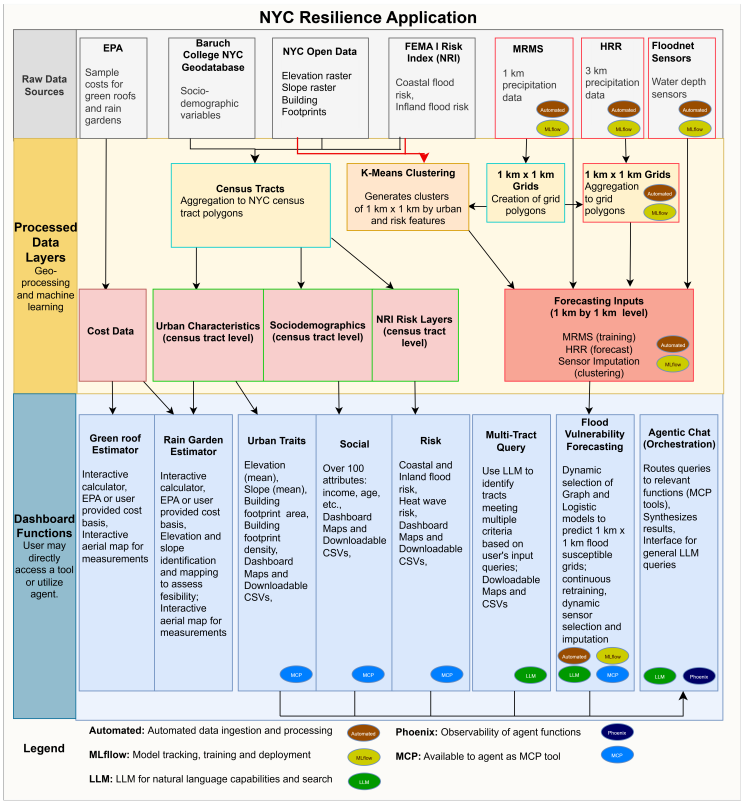

Figure 1: The high-level overview of the NYC Resilience Application, depicting the flow from raw data to processing to user accessible functions.



Figure 2: Urban, Social, and Hazard-Risk Dashboard Traits. Mean slope (percent rise), education (no diploma), NRI heat risk depicted in Fig. 2a, Fig. 2b, and Fig. 2c, respectively. The maps are in a graduated scheme, where darker red colors indicate higher values.

the previously processed urban and sociodemographic traits and natural-hazard risk attributes at the NYC census-tract level. Claude does not perform the geospatial filtering or tract selection; solely, it converts the natural-language request into structured JSON specifying each selected field and its target class, where 0–3 represent the lowest through highest relative classes. For multi-variable queries, the resulting classes are combined using an AND operation, retaining only tracts satisfying every LLM-generated criterion. Finally, matching tract identifiers and geometries are assembled into a GeoJSON FeatureCollection and rendered through a PyDeck GeoJsonLayer, separating matching and nonmatching tracts in the resulting spatial visualization. The Multi-Tract Query flow is depicted in Fig. 3, with snapshots of the live dashboard shown in Fig. 4.

4.2 Adaptive Flood Forecasting

4.2.1 Predictive Modeling and Spatial Representation

Flooding is one of the most frequently occurring hazards in NYC, with impacts ranging from deaths to economic losses (Depietri & McPhearson, 2018). Accordingly, flood forecasting was selected as the initial case study for implementing automated data processing, full MCP enablement, and

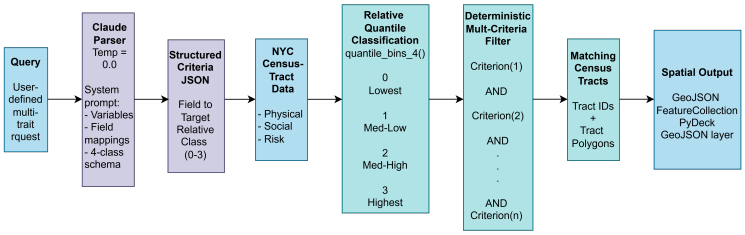

Figure 3: Multi-Tract Query Flow.



Figure 4: Snapshot of the Multi-Tract Query Tool. Fig. 4a depicts the query, where the user enters a prompt, "Show me tracts with high heat risk, low income and elderly population." Fig. 4b shows the findings matching the qualities, which are 22 tracts, highlighted in blue.

the MLflow model lifecycle. Ideally, for forecasts to translate into specific actions, such as avoiding travel, predictions should provide localized information, within a reasonable time window, depicting where and when flooding may occur. With this reasoning, the aim for the flood forecasting model was to provide hourly flood susceptibility forecasts at the 1 km grid level. The proceeding describes the variable representations.

For the response variable, flood-depth observations were obtained from the FloodNet sensor network in New York City (FloodNet NYC). FloodNet sensors use ultrasonic range finders to measure the distance to the surface below the sensor at approximately one-minute intervals, from which flood-water depth is derived. Here, the observations were temporally aggregated to an hourly scale for model development. For the predictors, precipitation, particularly rainfall, has been shown to be the most influential time-series predictor of flood occurrence in NYC (Agonafir et al., 2022). For this purpose, precipitation observations were obtained from the NOAA Multi-Radar/Multi-Sensor (MRMS) system using its Surface Precipitation Rate product, at approximately 1 km spatial and two-minute temporal resolution (Zhang et al., 2016; National Oceanic and Atmospheric Administration). The MRMS spatial coverage over NYC was used to establish the approximately 1 km $\times$ 1 km grid structure for the flood analysis. Ultimately, three precipitation predictors were constructed at the hourly scale: current-hour precipitation, cumulative precipitation over the previous six hours, and daily total precipitation.

Graph Convolutional Neural Network Model

Agonafir & Zheng (2026) found that graph-based model architectures, especially, the graph convolutional neural network (GCN), performed well in localized flood prediction in NYC in comparison to other deep learning and statistical techniques. The GCN, implemented here for flood vulnerability prediction, represented each sensor as a graph node and connected to four (k=4) nearest neighbor by haversine distance. For each hourly timestep, t, the node feature matrix contained the three standardized precipitation predictors.

The GCN was formulated as a regression model. The response variable, y, for each sensor, i, may be described as flood duration, which is defined as the number of minutes, m, during which the flood depth, d, exceeds 25.4 mm:

$$y_{i,t} = |\{m : d_{i,t,m} > 25.4\,mm\}|$$

For event-based classification, a predicted sensor hour was classified as a flood event when the GCN predicted flood duration exceeded one minute. As an additional note, a main motivation for

formulating the GCN firstly a regression model wass to collect and track R^2, where, as the sparse sensor record builds over time, the GCN may prove suitable towards predicting flood water depth in the future, as opposed to solely classifying.

Logistic Regression Model

Logistic Regression, a traditional classification algorithm, has also been employed in urban flood modeling (Li & Willems, 2019; Viteri López & Morales Rodriguez, 2020; Tella et al., 2023). The response variable, an observed flood event, y, was defined as occurring per sensor-hour if the depth exceeded 25.4 mm for any duration:

$$y_{i,t} = \begin{cases} 1, & |\{m : d_{i,t,m} > 25.4\,mm\}| > 0, \\ 0, & |\{m : d_{i,t,m} > 25.4\,mm\}| = 0, \end{cases}$$

Here, $y_{i,t} = 1$ represents an observed flood event, while $y_{i,t} = 0$ represents the converse.

A predicted probability greater than or equal to 0.50 was classified as a flood event.

Unsupervised Clustering

Unsupervised clustering was used to group the approximately 1 km flood-analysis grids according to similar urban and inland flood-risk characteristics. This extends the approach of Agonafir & Zheng (2026) , in which areas with similar physical characteristics are assumed to have similar susceptibility to ponding under comparable rainfall conditions. The urban and risk attributes described in Section 3.1 were aggregated to the 1 km grid structure using ArcGIS Pro. Five features were used for clustering: mean slope, mean elevation, mean inland flood risk, total building footprint area, and building count. They were standardized prior to clustering and then partitioned using K-Means (K =5; 10 centroid initializations). For grid, i, assignment to cluster, C, may be represented as follows:

$$C_i = \underset{k \in \{1,\ldots,5\}}{\arg\min} \; \|z_i - \mu_k\|_2^2$$

Here, z_i represents the five grid characteristics, and μ_k is the centroid of cluster k.

4.2.2 Adaptive MLOps and Operational Forecasting

Historical and Forward Training-Data Construction

Flood-depth observations were obtained from the FloodNet sensor network, with ongoing acquisition automated through the FloodNet REST API. Precipitation observations were obtained from NOAA MRMS and associated with FloodNet deployments through a persistent sensor-to-MRMS grid mapping. The historical modeling record begins in October 2020 and is preserved separately from newly acquired observations.

Automated Model Retraining, Evaluation, and Selection

Following each weekly data refresh, the GCN and Logistic Regression models are retrained using the combined historical and append-only forward canonical observations and then evaluated.

Model and Sensor Eligibility For each sensor, the GCN and Logistic Regression models were independently required to achieve a minimum precision of 0.25 and recall of 0.30. A sensor was considered eligible when at least one model satisfied both thresholds. If only one model qualified, that model was assigned to the sensor; if both qualified, the model with the higher F1 was selected, with recall, precision, and false alarms used sequentially as tie-breakers. Sensors for which neither model qualified were excluded from the eligible sensor/model registry.

Spatial Model Assignment and Imputation Support

Following per-sensor model selection, each eligible FloodNet sensor is spatially assigned to the approximately 1-km analysis grid and inherits the corresponding K-Means cluster. Eligible sensors are ranked by selected-model F1, followed by recall and precision as tie-breakers. For grids containing eligible sensors, the highest-ranked sensor within the grid is designated as the primary support sensor. For grids without an eligible sensor, support is assigned from the highest-ranked eligible

sensor within the same cluster, thereby extending predictions to grids with similar urban and inland flood-risk characteristics. If no eligible sensor remains within a cluster following a future model refresh, the geographically nearest eligible sensor is used as a cross-cluster fallback.

The resulting spatial reference retains the selected GCN or Logistic Regression model associated with each support sensor and provides one support assignment for every 1-km grid. The reference is refreshed following model retraining and selection, allowing changes in sensor eligibility and performance to propagate automatically to subsequent forecasting. Spatially imputed grid predictions remain downstream products and are not incorporated into model retraining. Fig. 5 and Fig. 6 depict the above processes for the weekly training and grid imputation workflow and the daily forecasting workflow, respectively.

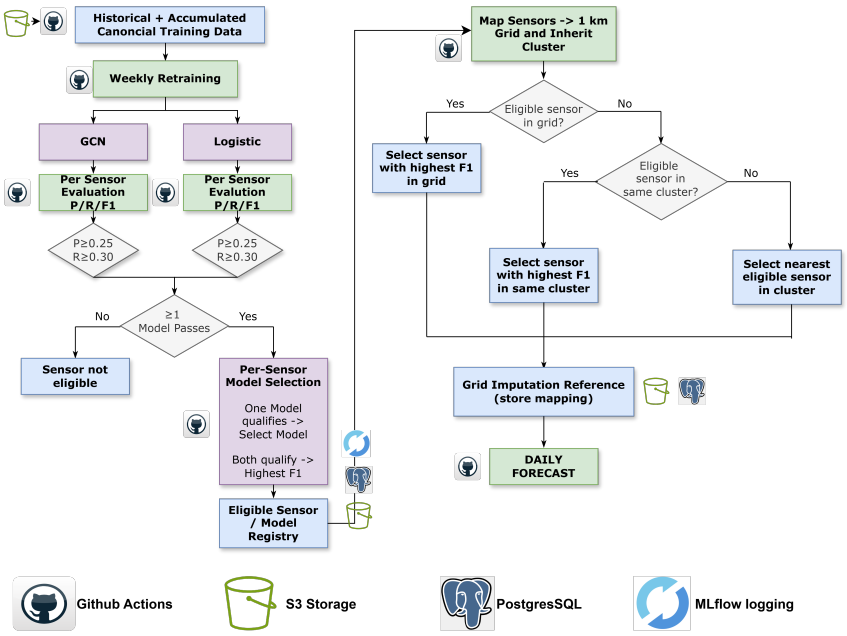

Figure 5: Weekly Forecasting and Imputation Flow.

Operational Forecast Ingestion and Daily Prediction.

Operational flood forecasts were generated using precipitation forecasts from NOAA's High-Resolution Rapid Refresh (HRRR) system, while MRMS precipitation remained restricted to model evaluation and retraining. HRRR surface forecast files were retrieved programmatically from NOAA's public noaa-hrrr-bdp-pds Amazon S3 archive using anonymous unsigned S3 access. HRRR precipitation was attained at 3-km resolution and associated to each approximately 1-km NYC analysis grid using the nearest HRRR forecast point. Importantly, the model identity was inherited from the assigned support sensor, whereas precipitation remained specific to the target grid; rainfall at the support sensor location was not transferred to grids requiring spatial support. Finally, the automated daily workflow generates separate same-day and day-ahead products, validates complete grid-hour coverage and predictor/model contracts prior to publication, and archives successful forecasts.

The resulting dashboard allows users to be able to select the hour of the same day or next forecast, of which they would like to view precipitation and flood susceptibility predictions at a granular level. Fig. 7a and Fig. 7b show images of the precipitation and flood forecasting for 10 AM Eastern Standard Time (EST) on September 13, 2026. Fig. 7c depicts an image of NOAA National Water Prediction Service observed hydrograph for the Bronx River at Bronx Botanical Gardens (BBGN6), captured September 13, 2026 (National Oceanic and Atmospheric Administration, 2026). The flood model shows high flood vulnerability prediction (with an event probability of 0.97) for the area, of which the NOAA gauge is located. This corresponds with the hydrograph that depicts an observed ponding of 2.05 feet. Future work will entail in-depth validation, as the historical record grows.

MLflow Tracking

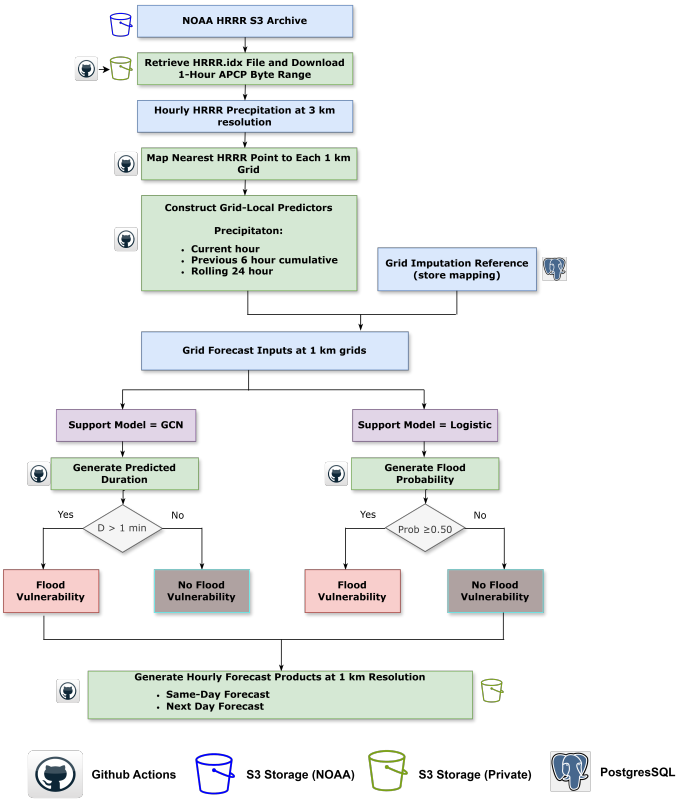

Figure 6: Daily Forecasting Flow.

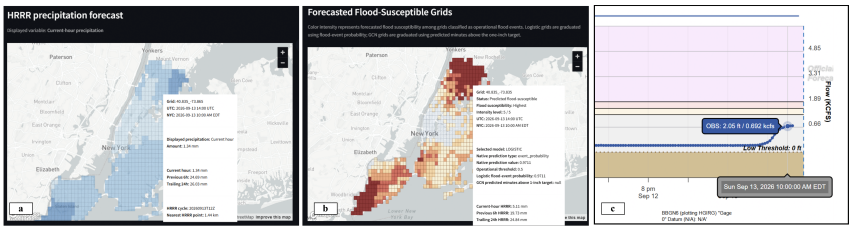

Figure 7: Flood Forecasting Tool. Precipitation and flood susceptible grid forecasting are shown in Fig. 7a and Fig. 7b, respectively, with a hover over a grid. Fig. 7c shows the NOAA observed flooding for same grid area.

MLflow monitoring is executed as part of the weekly retraining cycle to track predictive performance, model eligibility, and downstream spatial deployment as new observations are incorporated. Detailed tracked metrics and a representative weekly-cycle snapshot are provided in Appendix Table A2.

4.3 MCP-Enabled Agentic Orchestration

The stakeholder-facing agentic layer was implemented as a FastAPI orchestration service using Anthropic Claude Haiku 4.5 through the Amazon Bedrock Converse API, and capabilities were exposed as MCP tools using FastMCP. At runtime, the orchestrator discovers the available tools and their input schemas from five MCP services: geospatial, calculators, data access, model operations, and forecasting: This tool catalog is then given to Claude. Claude uses the user's request and the tool descriptions to select only the required capabilities; the orchestrator subsequently dispatches each requested tool call to the MCP server that owns it and returns the structured result to Claude for synthesis.

The MCP layer currently provides access to geospatial flood-grid summaries, reference methodologies for urban features and NRI flood risk and production model metadata and execution. The forecasting service additionally retrieves operational forecast summaries, available forecast hours, daypart summaries, grid- and hour-specific forecasts, highest-risk grids, cluster-specific forecasts, and predicted flood-event grid-hours. Operational values originate from the underlying tools rather than being generated or inferred by the LLM. Therefore, capabilities remain independently implemented deterministic functions and workflows, and the MCP provides a standardized interface through which selected capabilities may be invoked by the agentic orchestrator. Finally, Phoenix, integrated through OpenTelemetry, provides observability of the agentic orchestration layer, allowing execution traces associated with stakeholder requests and the resulting orchestration workflow to be inspected through the local Phoenix environment.

5 Conclusion

From time immemorial, the ability to foresee, withstand, or adapt to extreme events has driven the movements and actions of humans. In the present era, there are experts and stakeholders continuing this self-preservation inherent via technology and scientific progress. To that end, this work aims to contribute to the climate resilience effort with the presentation of an operational and extensible shared space for multidisciplinary collaboration and advancement. Initially developed to support participants during a NYC climate-resilience hackathon, the application has expanded to incorporate additional scientific and planning tools, an automated MLOps lifecycle, localized same- and next-day flood forecasting, and MCP-enabled agentic access to selected capabilities. Highlighting the flood-forecasting implementation, it provides the first fully operational predictive model of this architecture. Expressly, it couples continuously acquired observations and meteorological forecasts with model retraining, sensor- and model-specific performance selection, spatial support, and daily 1-km predictions. In a highly populated and topographically diverse city, such as NYC, localized outputs and insights are essential towards preventative actions. Ultimately, with MCP incorporation and the flexible architecture, facilitating additional incorporations, this framework will allow for multiple climate resilience connections.

The current application has opportunities for defining the next stages of evaluation and development. Regarding the flood model, there are multiple intended extensions. Firstly, the flood-event depth threshold is provisional, and its effect on event classification, model performance, and model selection is planned to be experimented across alternative thresholds. Secondly, the current GCN and logistic-regression implementations are not only intended for current operational outputs; indeed, they will also provide baselines against of which other predictive algorithms may be explored by training and testing. Thirdly, validation of flood conditions with observed satellite imagery is a plan for future research. Concerning the agentic layer, the evaluation of tool selection, parameter extraction, multi-tool reasoning, grounding, and response quality remains an open area of investigation. Also, as MCP coverage is also currently limited to a subset of application capabilities, existing functions that are not yet fully exposed through the agentic interface, such as the Multi-Tract Query, are planned to be integrated. In distinction to further climate foci, the next purposed application capability is extreme heat, which has been identified as the leading weather-related cause of fatalities in New York City (Depietri & McPhearson, 2018). Fundamentally, the progression of this work from a hackathon application to an operational system integrating automated MLOps, predictive forecasting, modular MCP services, agent observability, and multidisciplinary resilience capabilities demonstrates a capacity to evolve alongside emerging research and stakeholder needs.

6 Ethics Statement

This research adheres to the ethical guidelines outlined by the ICLR conference.

7 Reproducibility Statement

To support reproducibility, the implementation and supporting code for the resilience application described in this work are provided as anonymized supplementary material. The manuscript and appendix document the associated data sources and processing procedures.

8 AI USE STATEMENT

We have not used Generative AI for research ideation, design, or modeling suggestions or selections. We have not used Generative AI to write sections of this manuscript. We have used Generative AI to assist in software integration of methods pre-determined by the our researchers. AI was used for code assistance. AI was used to assist in polishing minimal textual elements of this manuscript. We have reviewed all AI assistance and take responsibility for the final content of this work.

9 ACKNOWLEDGEMENTS

This work was supported by NSF grant 2019625.

REFERENCES

Candace Agonafir and Tian Zheng. Structured exploration of machine learning model complexity for spatio-temporal forecasting of urban flooding. *Urban Water Journal*, 23(3):380–407, 2026. doi: 10.1080/1573062X.2025.2589078. URL https://doi.org/10.1080/1573062X.2025.2589078.

Candace Agonafir, Tarendra Lakhankar, Reza Khanbilvardi, Nir Krakauer, Dave Radell, and Naresh Devineni. A machine learning approach to evaluate the spatial variability of new york city's 311 street flooding complaints. *Computers, Environment and Urban Systems*, 97:101854, 2022. doi: 10.1016/j.compenvurbsys.2022.101854.

Candace Agonafir, Tarendra Lakhankar, Reza Khanbilvardi, Nir Krakauer, Dave Radell, and Naresh Devineni. A review of recent advances in urban flood research. *Water Security*, 19:100141, 2023. ISSN 2468-3124. doi: https://doi.org/10.1016/j.wasec.2023.100141. URL https://www.sciencedirect.com/science/article/pii/S2468312423000093.

Zihao An, Yujia Ming, Yong Liu, and Guangyu Zhang. Investigating 2d/3d factors influencing surface urban heat islands in mountainous cities using explainable machine learning. *Urban Climate*, 59:102325, 2025. doi: 10.1016/j.uclim.2025.102325.

Baruch College. New york city data. Newman Library, 2022. URL https://guides.newman.baruch.cuny.edu/nyc_data/nbhoods. Baruch College.

Navaneeth Bhaskar, Priyanka Tupe-Waghmare, Trisha S Shetty, Amrith R Naik, Pranav R Bhat, and B J Hrutik. Probabilistic multi-horizon urban water crisis prediction using ensemble machine learning and hydrological feature engineering. In *2026 7th International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 1904–1909, 2026. doi: 10.1109/ICIRCA69024.2026.11570539.

Yaella Depietri and Timon McPhearson. Changing urban risk: 140 years of climatic hazards in new york city. *Climatic Change*, 148(1):95–108, 2018. doi: 10.1007/s10584-018-2194-2.

Bapon Fakhruddin, Jenty Kirsch-Wood, Dev Niyogi, Li Guoqing, Virginia Murray, and Nina Frolova. Harnessing risk-informed data for disaster and climate resilience. *Progress in Disaster Science*, 16:100254, 2022. doi: 10.1016/j.pdisas.2022.100254.

Gene Falk, Thomas Gabe, and Maggie McCarty. Hurricane katrina: Social-demographic characteristics of impacted areas. Technical Report RL33141, Congressional Research Service, November 2005. URL https://www.congress.gov/crs-product/RL33141.

Hongfei Fan and Chengzheng Sun. Supporting semantic conflict prevention in real-time collaborative programming environments. *SIGAPP Applied Computing Review*, 12(2):39–52, June 2012. doi: 10.1145/2340416.2340420. URL https://doi.org/10.1145/2340416.2340420.

FloodNet NYC. About floodnet. https://www.floodnet.nyc/about. Accessed September 16, 2026.

G. Gilardi and I. Sharf. Literature survey of contact dynamics modelling. *Mechanism and Machine Theory*, 37(10):1213–1239, 2002. doi: 10.1016/S0094-114X(02)00045-9.

Babak Jalalzadeh Fard, Sadid A. Hasan, and Jesse E. Bell. Facilitating ai-driven sustainability: A service-oriented architecture for interoperable environmental data access. *Sustainability*, 18(5), 2026. ISSN 2071-1050. URL https://www.mdpi.com/2071-1050/18/5/2445.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024. doi: 10.1038/s41586-024-07744-y.

Xiaohan Li and Patrick Willems. Probabilistic flood prediction for urban sub-catchments using sewer models combined with logistic regression models. *Urban Water Journal*, 16(10):687–697, 2019. doi: 10.1080/1573062X.2020.1726409. URL https://doi.org/10.1080/1573062X.2020.1726409.

Yunfei Li, Sebastian Schubert, Jürgen P. Kropp, and Diego Rybski. On the influence of density and morphology on the urban heat island intensity. *Nature Communications*, 11(1):2647, 2020. doi: 10.1038/s41467-020-16461-9.

H. L. Macintyre, C. Heaviside, J. Taylor, R. Picetti, P. Symonds, X.-M. Cai, and S. Vardoulakis. Assessing urban population vulnerability and environmental risks across an urban area during heatwaves – implications for health protection. *Science of The Total Environment*, 610-611:678–690, 2018. doi: 10.1016/j.scitotenv.2017.08.062.

Vangelis Malamas, Panagiotis Giannopoulos, and Dimitris Koutras. Agentic climate intelligence: A multi-agent architecture for human-centered climate-resilient smart cities. In *2026 22nd International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, pp. 1027–1034, 2026. doi: 10.1109/DCOSS-IoT69657.2026.00159.

Amruta Mankawade, Pratik Ravasaheb Thete, Aryan Firoj Shaikh, Shreya Kishor Patil, Gauri Anil Kumar Mehami, and Om Ganeshkumar Sonawane. Enhancing collaborative code editors with semantic awareness. In *2026 International Conference on Computing Theory and Wireless Communications (ICCTWC)*, pp. 1–5, 2026. doi: 10.1109/ICCTWC68241.2026.11583625.

National Oceanic and Atmospheric Administration. Surface precipitation rate (spr). https://vlab.noaa.gov/web/wdtd/-/surface-precipitation-rate-spr-1. Accessed September 16, 2026.

National Oceanic and Atmospheric Administration. Bronx river at bronx botanical gardens (bbgn6). National Water Prediction Service, 2026. URL https://water.noaa.gov/gauges/bbgn6#hg=official. Observed hydrograph screenshot captured September 13, 2026.

National Weather Service. Hurricane katrina - august 2005. National Weather Service, Mobile/Pensacola, 2025. URL https://www.weather.gov/mob/katrina. Last updated July 2025; accessed September 9, 2026.

Vikrant Panwar and Subir Sen. Disaster damage records of EM-DAT and DesInventar: A systematic comparison. *Economics of Disasters and Climate Change*, 4(2):295–317, 2020. doi: 10.1007/s41885-019-00052-0.

M. Pelling. *The Vulnerability of Cities: Natural Disasters and Social Resilience*. Taylor & Francis, 2012. ISBN 9781136551475. URL https://books.google.com/books?id=Y1euBAAAQBAJ.

A. Tella, M. R. U. Mustafa, A. O. Balogun, C. J. Okolie, I. Bello Yamusa, and M. B. Ibrahim. Spatial prediction of flood in kuala lumpur city of malaysia using logistic regression. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-4/W6-2022:363–369, 2023. doi: 10.5194/isprs-archives-XLVIII-4-W6-2022-363-2023. URL https://isprs-archives.copernicus.org/articles/XLVIII-4-W6-2022/363/2023/.

USDA Northwest Climate Hub. Basics of global climate models. USDA Climate Hubs, 2023. URL https://www.climatehubs.usda.gov/hubs/northwest/topic/basics-global-climate-models. Accessed: 2026-09-12.

Andrea Salomé Viteri López and Carlos Augusto Morales Rodriguez. Flash flood forecasting in são paulo using a binary logistic regression model. *Atmosphere*, 11(5):473, 2020. doi: 10.3390/atmos11050473. URL https://www.mdpi.com/2073-4433/11/5/473.

Carl F. Weems, Sarah E. Watts, Monica A. Marsee, Leslie K. Taylor, Natalie M. Costa, Melinda F. Cannon, Victor G. Carrion, and Armando A. Pina. The psychosocial impact of hurricane katrina: Contextual differences in psychological symptoms, social support, and discrimination. *Behaviour Research and Therapy*, 45(10):2295–2306, 2007. doi: 10.1016/j.brat.2007.04.013.

Laurie Williams, Robert R. Kessler, Ward Cunningham, and Ron Jeffries. Strengthening the case for pair programming. *IEEE Software*, 17(4):19–25, July 2000. doi: 10.1109/52.854064.

Jian Zhang, Kenneth Howard, Carrie Langston, Brian Kaney, Youcun Qi, Lin Tang, Heather Grams, Yadong Wang, Stephen Cocks, Steven Martinaitis, Ami Arthur, Karen Cooper, Jeff Brogden, and David Kitzmiller. Multi-radar multi-sensor (mrms) quantitative precipitation estimation: Initial operating capabilities. *Bulletin of the American Meteorological Society*, 97(4):621–638, 2016. doi: 10.1175/BAMS-D-14-00174.1.

A Appendix

A.1 Geospatial Data Processing

Table A1: Sources and geospatial processing of urban, sociodemographic, and hazard-risk characteristics provided through the NYC Resilience Application.

Data Type	Source	Processing
Physical/urban characteristics	NYC Open Data	Elevation, slope, building footprint area, and building count were processed into NYC-bounded census-tract polygons using ArcGIS Pro tooling. Mean elevation and slope, total building footprint area, and total building count were calculated by tract. Building footprint density was derived as the sum of building footprints divided by tract area.
Census tract boundaries	U.S. Census Bureau	Census tract shapefiles provided the common spatial units for integrating and representing the processed characteristics at the NYC census-tract level.
Sociodemographic characteristics	Baruch College database (Baruch College, 2022)	The 141 Census-derived sociodemographic characteristics were aggregated to census-tract polygons by taking the mean.
Hazard-risk characteristics	FEMA National Risk Index for Natural Hazards (NRI)	Composite coastal flooding, inland flooding, and heat risk scores were integrated with census-tract polygons by matching census tract identification numbers.

A.2 MLflow Model-Selection and Spatial-Support Governance

MLflow monitoring is executed as part of the weekly retraining cycle. For the GCN, monitoring includes training and validation loss; pooled and sensor-level regression performance, including R^2, RMSE, and MAE; and flood-event precision, recall, and F1 under the ± 24-hour event-matching evaluation. For Logistic Regression, held-out classification performance and sensor-level accuracy, precision, recall, F1, ROC-AUC, and average precision are tracked, together with the number of sensors satisfying the operational precision and recall thresholds. Following evaluation, MLflow records the number of sensors assigned to each model, the eligible sensor population, and mean selected-model precision, recall, and F1. Spatial monitoring subsequently records direct versus imputed grids, GCN versus Logistic support assignments, support-sensor counts and distances, support precision/recall/F1, same- versus cross-cluster support, and missing, duplicate, or inconsistent grid assignments.

Table A2: Representative MLflow-tracked model-selection and spatial-support governance metrics from the September 14, 2026 weekly MLOps cycle. Timestamps are reported in UTC.

Record Type	**Metric**	**Value**
Per-sensor Model Selection	Sensors evaluated	445
	Eligible sensors	109
	Selected GCN	47
	Selected Logistic	62
	Mean selected-sensor precision	0.440
	Mean selected-sensor recall	0.627
	Mean selected-sensor F1	0.492
Spatial/Imputation Governance	Operational grids	837
	Direct grids	87
	Imputed grids	750
	GCN-supported grids	185
	Logistic-supported grids	652
	Mean grid-weighted support F1	0.748
	Missing support sensors	0
	Model-assignment mismatches	0