HINT: An Educational Dataset of U.S. East Coast Hurricanes and Tide Gauge Data for Climate Resilience

Dr. Thabet Kacem, University of the District of Columbia

Thabet Kacem is a tenured associate professor of computer science at the University of the District of Columbia. He obtained his PhD in information technology at George Mason University in May 2016. He obtained his master's in computer science at the University of the District of Columbia in January 2010. His research interests include cyber security, intelligent transportation systems, smart cities, cyber-physical systems, blockchains, and climate change. He is the principal inventor of the US Patent number 11022696 entitled "ADS-BSec: A holistic framework to secure ADS-B". He has served as principal and co-principal investigator in several sponsored projects from NSF and NSA. His research sits at the intersection of Artificial Intelligence, Cybersecurity, and Cyber-Physical Systems, with a focus on developing trustworthy, intelligent, and resilient technologies for safety-critical infrastructures. His work integrates machine learning, deep learning, secure system design, and data-driven modeling to address challenges in intelligent transportation, mobile systems, and emerging digital infrastructures. He is particularly interested in building robust and explainable AI systems, improving digital trust and security, and advancing AI driven analytics for complex, real-world environments. Through interdisciplinary collaborations, his research aims to create practical, high-impact solutions that enhance safety, reliability, and decision-making in modern cyber-physical and socio-technical systems.

Dr. Bryan James Higgs, University of the District of Columbia

Dr. Bryan Higgs is a native of southwest Virginia. Currently, he is an Assistant Professor in the Department of Civil Engineering at the University of the District of Columbia. Prior to this position, he was an Intelligence Community Postdoctoral Researcher.

HINT: An Educational Dataset of U.S. East Coast Hurricanes and Tide Gauge Data for Climate Resilience

Abstract

Climate change and global warming have intensified the frequency and impact of hurricanes and coastal flooding, posing major challenges to communities along the U.S. East Coast. Understanding these risks requires integrated datasets that capture both meteorological and hydrological factors. However, existing data sources are fragmented, and many coastal tide gauge stations contain gaps or incomplete records. In this work, we present a novel dataset entitled HINT, which merges hurricane historical records from the International Best Track Archive for Climate Stewardship (IBTrACS) with water level measurements from NOAA tide gauge stations. Each storm is augmented with measurements from all available tide gauges within its path or near landfall, producing multiple storm-station pairs when several gauges are matched. We decided to retain the storms without any valid station data due to distance from land, inactive or nonexistent stations, but we explicitly added a corresponding flag to indicate the absence of tide gauge augmentation. This approach preserves the full historical coverage of IBTrACS while enabling enriched analysis wherever observational data exist. Beyond research applications, the dataset is designed as an educational resource for engineering and data science programs. It provides a platform for students to explore data cleaning, integration, and analysis tasks while studying the intersection of climate resilience, coastal engineering, and applied machine learning. Faculty can incorporate the dataset into project-based learning activities, allowing students to practice technical skills in a real-world, societally relevant context. This contribution bridges the gap between data science and climate change education, providing the ASEE community with an open, reproducible dataset that supports both research and teaching in climate resilience and engineering education.

Introduction

Global warming[1] is a worldwide phenomenon manifested by raising earth surface temperatures. It is mainly caused by deforestation and a combination of concentrated gases from fossil fuels such as oil, gas and coal. This in turn contributes in making sea temperatures higher, which melts ice faster threatening coastal communities. For instance, the frozen fresh water in the glaciers around the world have lost an average of 273 billion tons per year in the period between 2000 and 2023[2], accounting for an average of 3.4 mm per year in sea level rise. An example of the face pace of melting ice in recent years, Antarctica lost an average of 107 Giga tons of ice per year from 1979 to 2023[3], while losing 54 Giga tons in 2023 alone.

It also has an effect in the more frequent and more intense storms that generally have devastating economic causalities. For example, Song et al.[4] reported that the global planet temperatures raised by about 1.2 degrees Celsius compared to pre-industrial era. This in turn contributed to doubling the number of weather disasters during 2005 to 2014 period compared to 1985 to 1994 period. In this context, global warming is just one facet of climate change[5] that characterize long-term shift of climate patterns that also includes other phenomena such as droughts, rainstorms, and wind. Droughts may cause soil shrinkage affecting oil infrastructure while heavy rain can cause floods damaging roads and bridges in coastal communities. Extreme sea level rise can have a devastating effect on off-shore platforms and coastal infrastructure such as the damage that Fukushima nuclear power plant endured as a result of the Tsunami waves following the strong earthquake off Japan's coast in 2011.

As far as the US is concerned, powerful hurricanes such as Katrina in 2005 and Sandy in 2012 left havoc throughout Louisiana and New Jersey/New York, respectively. Balaguru et al.[6] argue that future projections of hurricane activity in the US east coast in the period from 1980 to 2100 show increased frequency. This projection obtained by synthesizing several hurricane models explain the main cause to be adiabatic heating. There exist a wide of climate change related datasets covering global temperature, greenhouse gases, sea level rise, precipitations, and weather. However, these various data sources provide a fragmented view and often fail to capture the relation of hurricane data with storm surge. Moreover, NOAA tide gauge stations in US east coast contain several coverage gaps making them incomplete.

In this paper we bridge this gap by proposing a novel dataset entitled Hurricane IBTRACS with NOAA Tide gauge measurements (HINT) dataset. The dataset is publicly available for public download from Zenodo[7]. HINT merges hurricane data from IBTRACS filtered to the US east coast with the corresponding tide gauge stations within a specific range of the landfall of a given hurricane. In the process of the data creation, we employed data augmentation techniques by associating each storm to more than one tide gauge station. In addition, we follow a transparent strategy for handling missing data. We also conducted several experiments to assess the quality of the produced dataset.

The main contributions of this work are:

- We created a novel dataset, HINT, that pairs the storm surge data with sea level rise in the US east coast.
- We proposed a data augmentation techniques pairing each storm with several tide gauge station within a threshold distance of its landfall.
- We propose a transparent method for handling missing data.
- We provide a platform to bridge the gap between data science and climate change in academic environments.

The rest of the paper is organized as follows. We start by discussing the background on climate and coastal impacts and available datasets followed by an analysis of the related work and the identification of key knowledge gaps. Then, we describe the methodology we followed when creating this dataset before presenting some of the results evaluating our dataset. We finish by summarizing the key findings of our work, discussing the strengths, limitations and future work.

Background

The global climate is undergoing a profound transformation, characterized by rising surface temperatures and accelerated melting of polar ice[3]. Between 2000 and 2023, glaciers worldwide lost an average of 273 billion tons of ice per year[8], contributing significantly to sea-level rise and threatening coastal communities. In 2023 alone, Antarctica lost 54 gigatons of ice, emphasizing the rapid pace of change. These shifts have led to an increase in the frequency and intensity of weather disasters; the number of such events doubled in the period from 2005 to 2014 compared to the previous decade[9].

Impacts on Coastal Communities

Hurricanes and their associated storm surges represent one of the greatest natural threats to coastal infrastructure, economies, and public safety. While high winds cause significant structural damage, storm surge—the abnormal rise of water generated by a storm over and above the predicted astronomical tide—is historically the leading cause of hurricane-related fatalities and the primary driver of catastrophic flooding.

The impacts on coastal communities are multi-faceted:

- **Infrastructure Failure:** Storm surge destroys critical lifelines, including coastal roads, bridges, and power grids, often isolating communities during rescue operations[10].
- **Economic Disruption:** Beyond immediate property damage, coastal flooding causes long-term economic shifts through the loss of business continuity, soaring insurance premiums, and the potential for "managed retreat" from vulnerable zones[9].
- **Human Development and Health:** Platforms like **Human Climate Horizons (HCH)**[11]—a collaboration between the UNDP and the Climate Impact Lab—highlight that climate shifts do not just damage physical structures but profoundly impact human mortality, labor productivity, and energy consumption. By visualizing hyper-local impacts on the Human Development Index (HDI), such data underscores that engineering resilience must be inherently linked to social and economic stability.

Historical Damage from Major U.S. Hurricanes

The economic toll of these events has increased significantly over the last two decades, driven by both the rising intensity of storms and increased coastal development[9]. Table 1 summarizes the estimated cost of some of the most destructive hurricanes to strike the United States.

The staggering magnitude of these losses emphasizes a critical evolution required in engineering education. Future engineers can no longer rely solely on deterministic design codes based on historical stationarity; they must be equipped with a broader set of skills to address this growing societal threat[3]. This necessitates a transition toward **systems thinking**, where students learn to evaluate the intersection of physical infrastructure with human vulnerability, economic policy,

Table 1: Historical Damage from Major U.S. Hurricanes

Hurricane	**Year**	**Primary Impact Areas**	**Damage (USD)**
Katrina	2005	LA, MS, FL	˜$190 Billion
Harvey	2017	TX, LA	˜$155 Billion
Ian	2022	FL, SC	˜$115 Billion
Maria	2017	PR, USVI	˜$110 Billion
Sandy	2012	NY, NJ, CT	˜$85 Billion
Ida	2021	LA, MS, NJ, NY	˜$78 Billion
Irma	2017	FL, GA, SC	˜$62 Billion

and environmental health[10]. To design for a climate-resilient future, the next generation of engineers must be provided with tools that perform the complex task of harmonizing disparate climate and hydrological datasets. By utilizing an already integrated framework like the HINT dataset, students can bypass the prohibitive technical barriers of raw data fusion and instead focus on quantifying risk under high levels of uncertainty, ensuring that technical solutions are both robust and socially equitable.

Available Climate Datasets by Collection Methodology

To address these challenges, the scientific community produces data using diverse methodologies. Understanding these sources is critical for evaluating the uncertainty and resolution of the resulting datasets, as categorized in Table 2.

In-Situ Observations (Point-Specific Ground Truth)

- **Global Climate Observing System (GCOS):** Provides a cross-cutting framework for the monitoring of Essential Climate Variables (ECVs)[12].
- **Global Temperature (GISTEMP, MLOST, BEST):** Analyzes surface temperature anomalies using land-based weather stations[13,14,15].
- **Atmospheric CO2 (Mauna Loa/Scripps):** Foundation for the “Keeling Curve”[3].
- **Coastal Water Levels (NOAA Tide Gauge Network):** Ground-truth for sea-level rise and storm surge events[16].
- **GESLA-2:** Unified high-frequency sea-level database[17].
- **Oceanic Monitoring (Argo Floats):** Measures temperature, salinity, and velocity of the upper 2,000 meters of the ocean[18].

Remote Sensing (Satellite and Aerial Imagery)

- **Sea Level & Altimetry (NASA/NOAA):** Measures ocean surface height relative to the Earth’s center[19].

- **Oceanic Temperatures (ERSST):** Uses satellite observations and buoy data for historical records[20].
- **Precipitation (GPM IMERG):** Provides global precipitation estimates[21].
- **Geostationary Data (GOES-R):** Real-time monitoring of hurricane structure and lightning[22].
- **Digital Elevation Models (DEMs):** Terrain maps needed for flood risk assessment[23].

Model-Based and Integrated Products

- **SURGEDAT:** Manual curation of peak surge height for over 700 global events since 1880[24].
- **Global Downscaled Projections (GDPCIR):** Bridges global climate models and local-scale impact assessments[25].
- **Human Climate Horizons (HCH):** Maps hyper-local climate impacts on human development[11].
- **Emissions Tracking (UNFCCC, EDGAR, Climate TRACE):** Tracks emissions drivers; Climate TRACE uses AI for real-time tracking[26].
- **Atmospheric Reanalysis (ERA5 and MERRA-2):** Blended observations into consistent global grids[27,28].
- **Hurricane Track Records (IBTrACS):** Aggregates global tropical cyclone trajectories[29].
- **Numerical Models (SLOSH, P-Surge, ESTOFS):** Outputs used to forecast and reconstruct storm surge[30,31,32].
- **GSSR:** Model-based reconstruction of global storm surges since 1950[33].

Related Work

The development of the **HINT (Hurricane IBTrACS with NOAA Tide gauge)** dataset sits at the intersection of three critical domains in engineering education: the integration of multi-source climate data, the shift toward resilience-oriented civil engineering curricula[10], and the application of machine learning (ML) within project-based learning (PBL) environments[36].

Integration of Multi-Source Climate Data

The HINT dataset follows a philosophy of "merging and augmentation" similar to established global products, which can be categorized into three levels of complexity and required technical skill:

Reanalysis and High-Resolution Numerical Modeling (Expert Level)

Products like **ERA5**[27], **MERRA-2**[28], **GSSR**[33], **GDPCIR**[25], and **Satellite Altimetry**[19] reconstructions use numerical models or complex orbital mechanics. Working with these

Table 2: Summary of Environmental Data Sources

Dataset / Model	Primary Use Case	Collection Methodology	Complexity	Ref
Forecasting				
GOES-R Series	Real-time Monitoring	Satellite (Remote Sensing)	Intermediate	[22]
P-Surge	Probabilistic Risk	Numerical (Ensemble)	Expert	[31]
ESTOFS	Tide & Surge Forecast	Numerical (Hydrodynamic)	Expert	[32]
Historical				
SURGEDAT	Peak Surge History	Integrated (Manual Curation)	Accessible	[24]
GCOS Framework	ECV Monitoring	Integrated Framework	Intermediate	[12]
GISTEMP / BEST	Temp. Trends	In-Situ (Blended)	Intermediate	[13,15]
ERA5 / MERRA-2	Reanalysis	Model (Assimilation)	Expert	[27,28]
Mauna Loa (CO2)	GHG Trends	In-Situ (Atmospheric)	Intermediate	[3]
UNFCCC / EDGAR	Emissions	Integrated (Reporting)	Accessible	[34]
Argo Floats	Ocean Heat	In-Situ (Robotic)	Expert	[18]
IBTrACS	Hurricane Tracks	Integrated (Meta-data)	Intermediate	[29]
GSSR	Surge History	Model Reconstruction	Expert	[33]
GDPCIR	Impact Research	Model (Downscaling)	Expert	[25]
Both				
Human Horizons	Social Impact	Integrated (Impact Lab)	Accessible	[11]
Climate TRACE	Real-time Emissions	Satellite + AI	Intermediate	[26]
Satellite Altimetry	Sea Level Rise	Satellite (Altimetry)	Expert	[19]
ERSST	Sea Surface Temp	Satellite + Buoys	Intermediate	[20]
NOAA Tide Gauges	Water Levels	In-Situ (Sensors)	Accessible	[16]
GESLA-2	Extreme Sea-Level	In-Situ (Aggregated)	Intermediate	[17]
SLOSH	Surge Prediction	Numerical (Deterministic)	Expert	[30]
CERA	Surge Viz	Integrated (Decision Support)	Accessible	[35]

typically requires proficiency in high-performance computing (HPC) and multi-dimensional formats like NetCDF.

Aggregated Observational Datasets (Intermediate Level)

These merge different physical measurements using statistical methods. Examples include the temperature records in **GISTEMP/BEST**[13,15], atmospheric measurements from **Mauna Loa**, monitoring via **GCOS**[12], and the sensor-agnostic harmonization in **GESLA-2**[17]. Strong GIS and statistical skills are required.

Accessible Impact and Integrated Products (Pedagogical Level)

These layer climate data on top of "non-climate" datasets, such as **Human Climate Horizons**[11], **UNFCCC inventories**, and **SURGEDAT**[24]. These are often distributed in tabular formats, significantly lowering the barrier to entry.

Resilience-Oriented PBL through Machine Learning

Traditional engineering curricula have historically relied on stationary records, yet the **American Society of Civil Engineers (ASCE)** emphasizes a shift toward **Resilience Thinking** to address modern environmental uncertainties[10]. HINT supports this shift by providing an "ML-ready" platform that bridges the gap between purely statistical learning and practical climate adaptation. Unlike standard "toy" datasets, HINT provides a **societally relevant context** where students practice feature engineering using storm-station pairs to predict surge impacts. This approach addresses the documented "competency gap" in students' ability to design systems for uncertainty, aligning with the entrepreneurial **KEEN framework**[36] by encouraging the use of data-driven insights to protect critical infrastructure. Furthermore, the dataset provides the raw observational data needed to validate specialized tools like **Inundation Maps** or visualize broader socio-economic risks highlighted by platforms such as **Human Climate Horizons**[11].

Methodology

The methodology we followed to create the HINT dataset is depicted in Figure 1. The data generation process is fully automated, reproducible, and designed to support both research analysis and instructional use in data science and engineering curricula. Our data augmentation strategy relies on picking the top 5 nearest tide gauge stations within 300 km for each storm row. Then, we retrieve the related water level datum, Mean Sea Level (MSL), and its standard deviation which are temproaly aligned with the storm record in question.

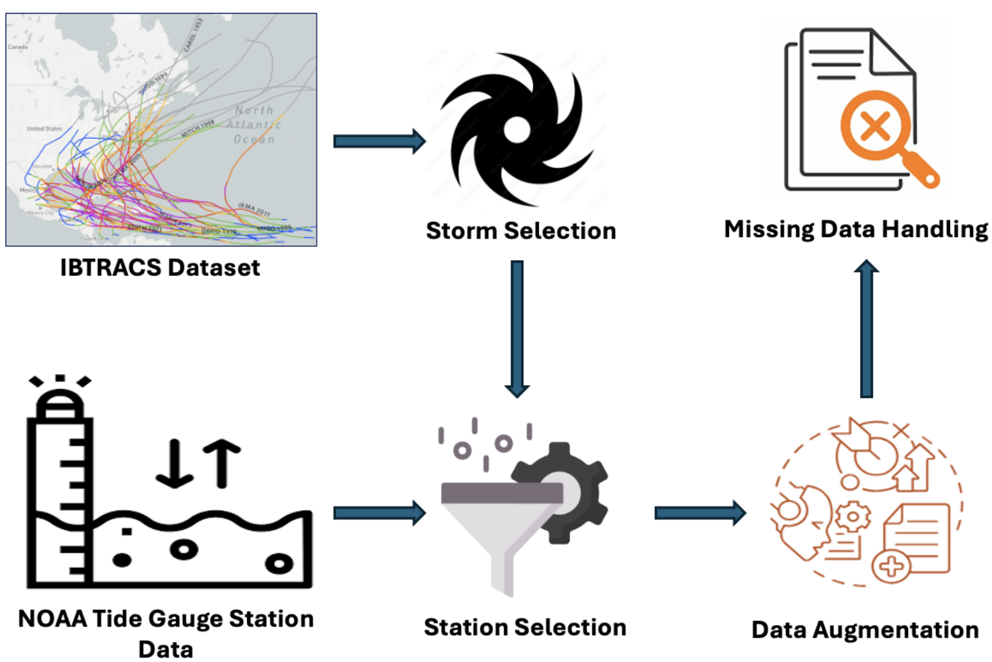

Figure 1: Methodology Overview

There are two inputs, notably, the IBTRACS dataset that contains historical storm data and NOAA tide gauge station data measuring the water level. Since our HINT dataset focuses on the

US east coast, we filtered the storm data to include those affecting the US east coast, and similarly we considered the NOAA stations in the US east coast as well. IBTRACS provide storm metadata correlated with geographic coordinates at regular time intervals. Storm metadata include the nature of the storm, its longitude and latitude, timestamp of the collected data, wind speed, and pressure. Water level observations are obtained from NOAA's Center for Operational Oceanographic Products and Services (CO-OPS), which maintains long-term tide gauge measurements along the U.S. coastline. There are several vertical datums such as MSL and Mean Low Lower Water (MLLW). However, the main problem is that not all stations supports all datums. For instance, some may support MSL while other support MLLW and in some occasions some stations only collect meteorological data without any reported water level. Another issue with NOAA stations is that some of them are active while other stopped working, not all stations started at the same time. These outlined differences in temporal resolution, station availability, and datum support pose integration challenges which we explicitly address during the generation process of the HINT dataset.

The station selection process is outline in Algorithm 1. The inputs are (1) the storm record representing a line in the IBTRACS, (2) list of NOAA Tide Gauge Stations in the US east coast, and (3) NOAA Tide Gauge Water Level Repository. In line 3, we initialize the dictionary O that holds the selected stations, where the key is the storm record's timestamp and the mapped content is a list of top 5 nearby stations. In lines 4-9 we start the for loop according to the existence of station in the list of NOAA tide gauge stations in the east coast, before computing the distance in kilometers between the storm record location and the station location using the Haversine formula. Then, we search if the associated station file containing the water level data in our directory exist. If it does, the station is added to the station list corresponding to the mapped content of the element whose key is the storm timestamp. In line 10, we return the output dictionary.

Algorithm 1: Station Selection

1 **Inputs:** I_1: Storm Record, I_2: NOAA Tide Gauge Station List, I_3: NOAA Tide Gauge Water Level Repository;
2 **Outputs:** O: Dictionary of nearby stations for each given storm record;
3 initialize(O);
4 **for** *station in* I_2 **do**
5   distance = haversine_km($I_1.lat$, $I_1.lon$, station.lat, station.lon);
6   exist = searchStationFile(station);
7   **if** *exist* **then**
8   O[$I_1.timestamp$].add(station);
9   **end**
10 **end**
11 return O;

Fetching water level from NOAA tide gauge stations was one of the most challenging tasks since it involved several considerations. To start with, not all stations return water level since some of them are for various purposes such as meteorological. Also, older stations, before the digital era, reported very basic info and did not support MSL or MLLW datums. In addition, even for modern stations, not all of them report the same datum, with MSL and MLLW being the most popular choices. Conversely, using NOAA's CO-OPS Data API has quota restrictions rendering the fetching of water level for stations within a large geographical region, such as US east coast, an error-prone task. To overcome these difficulties, we decided to download the water level data for each station in the US east coast per year separately into a separate file since 1970 until 2024. This choice was motivated by the fact that before 1976 that was the pre-satellite or early satellite, and with lack of digitized NOAA hourly water level being reported. That way, we did not exceed the quota restriction of CO-OPS API and prevented situation when the server would freeze due to the big amount of fetched data resulting in the whole download failing. Algorithm 2 describes this process. The inputs are the storm list in the us east coast and the list of NOAA tide gauge stations identifiers, while the output is a collection of files saved into a separate directory. In lines 3-5, there is a nested for loop, by station id then by storm, where the URL to be passed to CO-OPS get is first prepared. This is done by attaching all the necessary parameters including the storm time, station id, water level rate, datum type, units, time zone, and response format. in lines 6-7, the GET request with the given URL and the timeout set to 10s is sent before the repsonse is saved into json format. In line 8, we prepare the file name that consists of a prefix being the station id, and a suffix being the year, while in line 9 we save the data in a CSV file with the provided filename under the provided folder.

Algorithm 2: Water Level Retrieval from NOAA Tide Gauge Stations

```
1  Inputs: I_1: Storm List, I_2: NOAA Tide Gauge Station List ;
2  Outputs: O: Collection of Files of Water Level by station Id by year;
3  for station in I_2 do
4      for storm in I_1 do
5          url = prepare_URL(storm.date, station.id, 'hourly_height','MSL','metric','gmt','json');
6          response = request(url, 'GET', timeout);
7          data = response.json();
8          filename = prepareFileName(station, storm.year);
9          saveData(data,filename);
10     end
11 end
```

The data augmentation is the last step of our methodology, which is depicted in algorithm 3. The inputs are (1) IBTRACS storm list in the east coast, (2) NOAA Tide Gauge Station List in the east coast, and (3) NOAA Tide Gauage Water Level Repository. The only output is the completed HINT dataset. In line 3, we iterate over each storm in the filtered IBTRACS dataset, before invoking the station selection method, described in algorithm 1, for the given storm record at that given timestamp in line 4. In lines 5-7, we iterate over each item in the list of nearby stations and we search for the corresponding file in the repository of downloaded water level data. Then we compare the timestamps: the one from the storm record versus the one from the station water level. The last step is invoking the appendAndUpdate function that duplicates each considered strom record as many times as there are nearby stations and in each duplicated record it records the corresponding water level data including the station id, the distance to the storm location, the observed water level, and the water standard deviation.

Algorithm 3: Data Augmentation

Inputs: I_1: Storm List, I_2: NOAA Tide Gauge Station List, I_3: NOAA Tide Gauge Water Level Repository ;
Outputs: O: HINT Dataset;
for *storm in* I_1 **do**
    nearbyStation = stationSelection(I_1,I_2,I_3);
    for *station in nearbyStation* **do**
        waterData = searchWaterLevel(station, storm.timestamp);
        appendAndUpdate(storm, station, waterData);
    end
end

The HINT dataset comprises 31292 rows and includes all the features of the IBTRACS dataset in addition to the following ones that we added as part of our data augmentation strategy:

- **Station Id:** identifier of the station in question.
- **Distance to Storm:** distance of the station to the storm location at the given timestamp.
- **Water Level:** observed water level in metric units at the given timestamp.
- **Water Standard Deviation:**
- **Datum:** MSL is used for all stations.
- **Water Flag:** binary flag indicating if water level is found or not.

Results

To evaluate our dataset we conducted the following experiments. First, we measured how many selected stations associated with storm records observed an empty water level data. This is depicted in Figure 2 that shows 79% availability. This shows that our decision to discard storms before 1976 where water level was very scare due to lack of digitized reporting is justified and beneficial. The 1976 cutoff applies to the entire dataset.

Our second experiment focused on measuring the temporal storm coverage by year, as shown in Figure 3. It is observed that the number of storm records fluctuates per year with higher tendencies in the 2020s and 2020s decades.

Our third experiment focused on the spatial coverage in our HINT dataset. It shows the storm track points versus the tide gauge stations.

The last experiment shows the evolution of the number of missing water level data per year in HINT dataset as shown in Figure 4. It is observed that the number kept decreasing considerably, which is understood since more and more tide gauge station have been deployed and started collecting data.

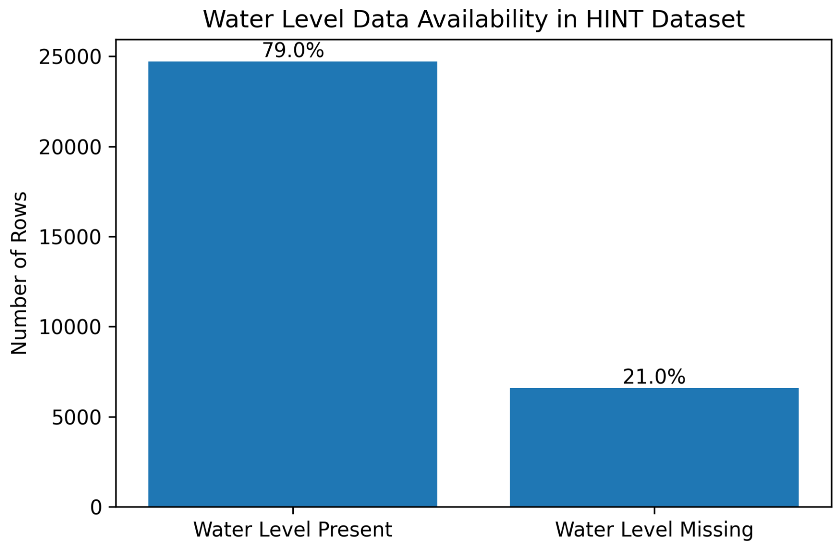

Figure 2: Overall Water Level Availability

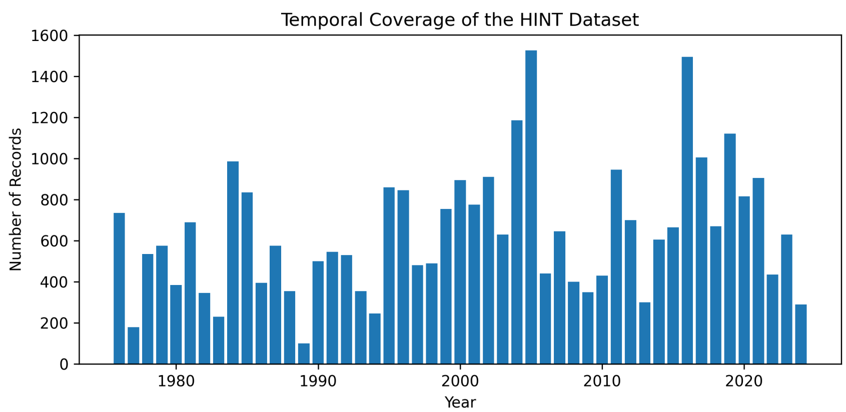

Figure 3: Temporal Storm Coverage per Year

Discussion

The creation of the HINT dataset addresses a fundamental challenge in climate data science: the reconciliation of dynamic, moving meteorological phenomena with stationary hydrological observation networks. By systematically pairing the Lagrangian trajectories of hurricanes from the IBTrACS database with the Eulerian fixed-point sensors of the NOAA tide gauge network, HINT provides a unique spatiotemporal bridge.

Technical Challenges in Data Integration

Achieving this integration required overcoming several technical and logistical hurdles. One of the most significant was datum heterogeneity. NOAA stations utilize various vertical datums, including Mean Sea Level (MSL), Mean Lower Low Water (MLLW), and NAVD88. Because

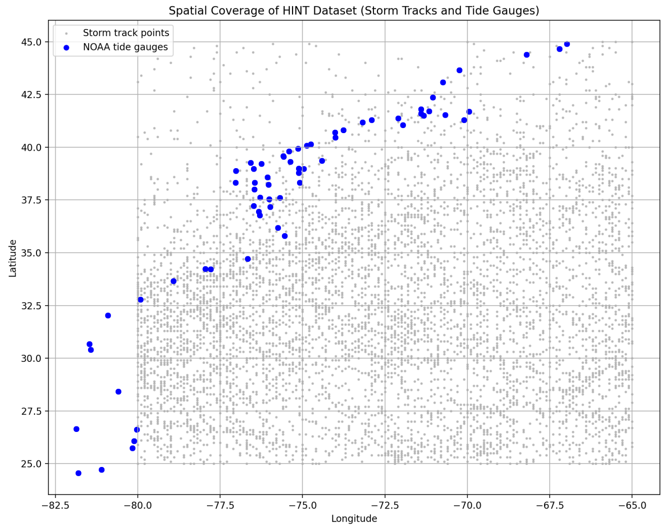

Figure 4: Dataset Spatial Coverage

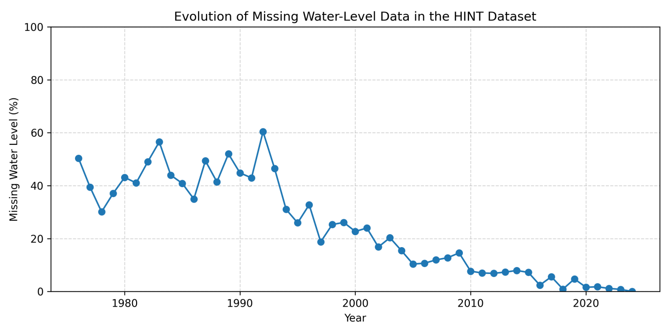

Figure 5: Missing Water Level Data per Year in HINT dataset

many older records lack the necessary metadata for conversion, we standardized HINT on MSL. This required a rigorous filtering process that discarded stations with incompatible datums, preventing ”silent” errors that often occur in simpler student-led climate projects.

Furthermore, API and Quota Management posed a substantial obstacle. NOAA's CO-OPS Data API imposes strict rate limits. Our response—the creation of a localized repository partitioned by station and year—represents a strategic move that not only circumvented API throttling but also ensured that the resulting dataset is reproducible and stable for classroom environments.

Spatiotemporal Alignment and Historic Constraints

The alignment of the two sources also required complex synchronization. IBTrACS typically reports storm positions at 3-hour or 6-hour intervals, whereas tide gauges provide hourly data. Syncing these required precise timestamp matching and handling time zone discrepancies (GMT vs. local standard time).

Finally, our decision to begin the dataset in 1976 is validated by the historical evolution of the

NOAA National Water Level Observation Network (NWLON). Prior to 1976, the network was relatively sparse, with fewer than 100 primary stations, many of which relied on manual paper-strip recorders. The current network has expanded to over 210 primary NWLON stations along the U.S. coast. This increase in spatial density explains the observed decrease in missing water level data per year shown in our results. Pedagogically, HINT transforms the student experience by handling these prohibitive ”data wrangling” phases, allowing students to focus on high-level resilience analysis.

Implications of Development Decisions

Each architectural choice in the development of HINT carries significant implications for its final utility and pedagogical impact:

- **Standardization on MSL:** Substantially reduces the technical barrier to entry. Engineering students can focus on surge magnitude and trend analysis without the confusion of reference frame offsets.
- **Localized Repository Architecture:** Prioritizes classroom scalability. An entire cohort of students can concurrently query and train models on the dataset without the risk of API-related ”quota fails” or high-latency network bottlenecks.
- **Multi-Station Pairing Strategy (Top 5 / 300 km):** Provides data redundancy and spatial context. In the event of a sensor failure—frequent during extreme weather—the dataset provides nearby alternatives and the high-dimensional richness required for ML models.
- **Lagrangian-to-Eulerian Pairing:** Centering the dataset on the moving storm track represents a fundamental shift in perspective. Students learn to view infrastructure risk as a dynamic interaction between a meteorological ”driver” and a physical ”receiver,” reinforcing systems-thinking.
- **Temporal Threshold (Post-1976):** Limiting scope to the digital era ensures students are training models on high-fidelity, high-frequency data, which is essential for accurate machine learning outcomes.

Potential Use Cases of the HINT Dataset

Beyond its primary function as a training set for storm surge prediction, the HINT dataset enables a wide array of research and instructional use cases:

- **Surrogate Modeling for Coastal Engineering:** Students can use HINT to develop ML-based ”surrogate models” that replicate the behavior of complex numerical solvers like SLOSH or ESTOFS at a fraction of the computational cost, facilitating rapid ”what-if” scenario analysis in design projects.
- **Spatiotemporal Gradient Analysis:** By leveraging the five-station proximity pairing, researchers can analyze the spatiotemporal gradients of surge attenuation. This allows for the study of how coastal topography influences the rate at which surge heights diminish as a storm moves inland or along the littoral zone.

- **Longitudinal Sea-Level Rise Sensitivity:** Because the dataset spans nearly 50 years of observations, it can be used to analyze how historical surge events would manifest under future sea-level rise scenarios. Students can ”shift” the MSL datum to evaluate the increasing vulnerability of existing infrastructure over time.
- **Feature Importance for Climate Adaptation:** HINT provides a platform for students to practice feature engineering—determining which meteorological variables (e.g., minimum central pressure vs. maximum wind speed) are the most significant predictors of surge for different segments of the U.S. East Coast. This directly informs climate adaptation strategies by identifying localized risk drivers.

Potential Future Expansion

Future development of the HINT framework aims to broaden its analytical scope and impact through two primary focus areas:

- **National Geographic Scaling:** Future iterations could incorporate NOAA stations from the gulf coast and the Pacific Coast. This continental scaling would enable students to perform comparative studies across diverse coastal geomorphologies, analyzing how regional topography influences surge dynamics.
- **Multidisciplinary Dataset Integration:** The framework could be expanded to ingest high-resolution bathymetric data, satellite-derived Sea Surface Temperatures (SST), and socio-economic indicators from the U.S. Census Bureau. These additions would allow for a more holistic exploration of the relationship between thermodynamic ocean drivers, coastal geometry, and community-scale social vulnerability.

Conclusion

This paper presented the Hurricane IBTrACS with NOAA Tide gauge (HINT) dataset, an automated and reproducible framework designed to facilitate climate impact research and engineering education. The methodology employed a rigorous algorithmic sequence: first, filtering global meteorological and sensor networks for relevance post-1976; second, applying the Haversine formula to dynamically pair moving storm coordinates with the top five nearest Eulerian sensors; and third, standardizing heterogeneous vertical datums onto a unified Mean Sea Level (MSL) reference frame. This automated pipeline ensures that the complex task of multi-source data fusion is handled algorithmically, preserving raw observational integrity while resolving the spatiotemporal discrepancies inherent in historical climate records.

The impact of this work is twofold. Scholarly, it provides a validated, ”ML-ready” dataset that avoids the smoothing effects of model-based reanalysis, allowing for high-fidelity surge impact modeling. Pedagogically, it removes the ”wrangling” barriers that often derail undergraduate projects, empowering students to shift their focus from data cleaning to risk interpretation and systems thinking. By transforming complex, multi-source raw observations into a structured spatiotemporal bridge, this work generates significant value for coastal resilience research and provides a versatile foundation for use cases ranging from predictive surrogate modeling to hyper-local climate adaptation planning.

References

[1] James Hansen, Makiko Sato, Paul Hearty, Reto Ruedy, Maxwell Kelley, Valerie Masson-Delmotte, Gary Russell, George Tselioudis, Junji Cao, Eric Rignot, et al. Ice melt, sea level rise and superstorms: evidence from paleoclimate data, climate modeling, and modern observations that 2 c global warming could be dangerous. *Atmospheric Chemistry and Physics*, 16(6):3761–3812, 2016.

[2] University of Arizona News. As glaciers melt faster, freshwater sources dwindle and sea-level rise accelerates — news.arizona.edu. `https://news.arizona.edu/news/glaciers-melt-faster-freshwater-sources-dwindle-and-sea-level-rise-accelerates`. [Accessed 08-01-2026].

[3] Rebecca Lindsey. Climate change: mountain glaciers — climate.gov. `https://www.climate.gov/news-features/understanding-climate/climate-change-mountain-glaciers`, 2025. [Accessed 08-01-2026].

[4] Fengfei Song, Guang J Zhang, V Ramanathan, and L Ruby Leung. Trends in surface equivalent potential temperature: A more comprehensive metric for global warming and weather extremes. *Proceedings of the National Academy of Sciences*, 119(6):e2117832119, 2022.

[5] Climate Change. Mitigating climate change. *Working Group III contribution to the sixth assessment report of the intergovernmental panel on climate change*, 2022.

[6] Karthik Balaguru, Wenwei Xu, Chuan-Chieh Chang, L Ruby Leung, David R Judi, Samson M Hagos, Michael F Wehner, James P Kossin, and Mingfang Ting. Increased us coastal hurricane risk under climate change. *Science advances*, 9(14):eadf0259, 2023.

[7] Thabet Kacem and Bryan Higgs. Hint dataset for climate resilience, 2026. URL `https://doi.org/10.5281/zenodo.18600756`.

[8] J. Hugonnet et al. Accelerated global glacier mass loss in the early twenty-first century. *Nature*, 592:726–731, Apr. 2021.

[9] NOAA NCEI. U.S. Billion-Dollar Weather and Climate Disasters, 2024. URL `https://www.ncei.noaa.gov/access/billions/`.

[10] ASCE. *Civil Engineering Body of Knowledge for the 21st Century: Preparatory Steps for the 21st Century*. ASCE Press, Reston, VA, USA, 3rd edition, 2019.

[11] UNDP. Human climate horizons, 2023. URL `https://horizons.hdr.undp.org/`.

[12] GCOS. The 2022 GCOS implementation plan. Technical Report GCOS-244, World Meteorological Organization, Geneva, Switzerland, 2022.

[13] NASA GISS. GISS surface temperature analysis (GISTEMP), version 4, 2023. URL `https://data.giss.nasa.gov/gistemp/`.

[14] NOAA. NOAA global surface temperature (NOAAGlobalTemp), version 5, 2023. URL `https://www.ncei.noaa.gov/products/land-based-station/noaa-global-temp`.

[15] Berkeley Earth. Land and ocean summary data, 2023. URL `BerkeleyEarth.org/data/`.

[16] NOAA CO-OPS. Station selection - tides & currents, 2024. URL `https://tidesandcurrents.noaa.gov/stations.html`.

[17] P. L. Woodworth et al. GESLA-2: the second version of the global extreme sea level analysis database. *Geosciences Data Journal*, 4(1):26–37, Jun. 2017.

[18] Argo. Argo float data and metadata from Global Data Assembly Centre (Argo GDAC), 2023.

[19] NASA JPL. Physical oceanography distributed active archive center (PO.DAAC), 2024. URL `https://www.google.com/search?q=https://podaac.jpl.nasa.gov/SeaLevel`.

[20] B. Huang et al. Extended reconstructed sea surface temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *Journal of Climate*, 30(20):8179–8205, 2017.

[21] G. J. Huffman et al. Integrated multi-satellite retrievals for GPM (IMERG) technical documentation. Technical report, NASA/GSFC, 2019.

[22] S. J. Goodman et al. *The GOES-R Series: A New Generation of Geostationary Environmental Satellites*. Elsevier, 2019.

[23] USGS. 3d elevation program (3DEP), 2024. URL `https://www.usgs.gov/3d-elevation-program`.

[24] H. F. Needham and B. D. Keim. A storm surge database for the US Gulf Coast. *International Journal of Climatology*, 32(14):2108–2123, Nov. 2012.

[25] Climate Impact Lab. Global downscaled projections for climate impacts research (GDPCIR), 2023. URL `https://impactlab.org/`.

[26] Climate TRACE. Climate TRACE emissions database, 2023. URL `https://climatetrace.org`.

[27] H. Hersbach et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146 (730):1999–2049, Jul. 2020.

[28] R. Gelaro et al. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate*, 30(14):5419–5454, 2017.

[29] K. R. Knapp et al. The International Best Track Archive for Climate Stewardship (IBTrACS). *Bulletin of the American Meteorological Society*, 91(3):363–376, Mar. 2010.

[30] B. Jurewicz and A. Mehra. SLOSH: Sea, lake, and overland surges from hurricanes. Technical report, NWS Technical Procedures Bulletin, 2020.

[31] NOAA. P-surge 2.5: Probabilistic storm surge. Technical report, National Weather Service, 2021.

[32] NOAA. Extratropical surge and tide operational forecast system (ESTOFS). Technical report, National Ocean Service, 2023.

[33] M. G. Tadesse et al. Data-driven modeling of global storm surges. *Frontiers in Marine Science*, 7:260, Apr. 2020.

[34] UNFCCC. National inventory submissions 2023. Technical report, 2023.

[35] CERA. Coastal emergency risks assessment, 2024. URL `https://cera.coastalrisk.live/`.

[36] J. S. Kern et al. The KEEN framework: Entrepreneurial mindset in engineering education. In *Proc. ASEE Annu. Conf. Expo.*, 2015.