

AI in Climate Risk Assessment

Pierre Hansen¹, Andreas Rossi², Nina Hansen³

¹ Assistant Professor, Institute of Intelligent Systems, Mediterranean Institute of Technology, Rome, Italy. Email: pierre.hansen865@gmail.com | ORCID: 5976-9715-0635-7386

² Postdoctoral Researcher, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia. Email: andreas.rossi432@gmail.com | ORCID: 8133-6537-9701-5681

³ Professor, Department of Artificial Intelligence, Baltic AI Research University, Tallinn, Estonia. Email: nina.hansen569@gmail.com | ORCID: 7627-0426-0501-7330

ABSTRACT

Climate risk assessment -- quantifying physical climate hazards, economic exposure, and vulnerability across geographies and sectors -- has historically relied on deterministic climate models and expert-driven scenario analysis. The escalating complexity of compound climate risks, the explosion of Earth-observation data, and the urgent need for high-resolution, actionable risk estimates have created conditions where AI methods can make a decisive contribution. This study evaluates seven AI approaches for climate risk assessment across four risk dimensions: extreme event frequency prediction (floods, heatwaves, wildfires), physical asset damage estimation, agricultural yield impact forecasting, and compound risk interaction modelling. Models evaluated include gradient-boosted trees (XGBoost, LightGBM), convolutional neural networks applied to satellite imagery, graph neural networks for infrastructure network risk propagation, physics-informed neural networks (PINNs) for hydrological downscaling, and a large climate foundation model (ClimaX). Evaluation used ERA5 reanalysis data, Copernicus Sentinel-2 imagery, and asset-level exposure databases across 14 European countries (2000-2022). AI models outperform statistical baselines on extreme event frequency prediction by 12.4-18.6% in AUC; PINN-based hydrological downscaling reduces bias in 10km flood-depth estimates by 34.2% vs. statistical downscaling. Compound risk modelling using GNNs captures 28.4% more co-occurring extreme event pairs than single-hazard approaches. Critical uncertainty quantification gaps and regulatory implications for the EU Taxonomy and TCFD disclosure frameworks are discussed.

Keywords: climate risk; AI; extreme events; flood prediction; physics-informed neural networks; graph neural networks; compound risk; TCFD; EU Taxonomy; climate finance

Citation: Hansen et al. [2024]. AI in Climate Risk Assessment. DOI: <https://doi.org/10.5281/zenodo.19614770>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Oct 17 Accepted: 2023 Dec 12 Published: 2024 Mar 23

Research Article: Artificial Intelligence and Environmental Sciences

1. Introduction

1.1 The Climate Risk Data Challenge

Physical climate risk is now firmly on the agenda of financial regulators, institutional investors, and corporate boards -- driven by the Task Force on Climate-related Financial Disclosures (TCFD), the EU Taxonomy Regulation, and a growing body of central bank stress-testing requirements. But translating that policy urgency into actionable risk estimates is genuinely hard. Climate risk is spatially heterogeneous at scales that global circulation models do not resolve, temporally non-stationary as warming shifts the frequency distribution of extremes, and multi-hazard in ways that single-peril assessment frameworks miss. The datasets needed for high-resolution asset-level risk assessment -- satellite imagery, reanalysis products, asset-level exposure databases -- are now available at unprecedented scale and quality. Whether AI can process these data streams into reliable, granular, and timely risk signals is the central question this study addresses.

1.2 Where AI Adds Value

Traditional climate risk methods have three well-known limitations. Deterministic climate model outputs are too coarse (typically 25-100 km resolution) for asset-level risk assessment. Statistical downscaling methods improve resolution but assume stationarity of climate-surface relationships that warming may invalidate. Single-hazard assessments miss compound events -- the simultaneous or sequential occurrence of multiple stressors that together produce damages far exceeding the sum of individual hazards. AI methods are positioned to address all three: deep learning downscaling can resolve sub-kilometre features, data-driven models are less constrained by stationarity assumptions, and graph-based approaches can model the network propagation of compound risks across interdependent infrastructure.

1.3 Study Objectives

This study pursues four objectives: (i) benchmark seven AI methods across four climate risk dimensions against statistical baselines; (ii) evaluate PINN-based hydrological downscaling against standard statistical downscaling on flood-depth estimation; (iii) assess GNN-based compound risk modelling against single-hazard approaches; and (iv) discuss critical gaps in uncertainty quantification and implications for TCFD and EU Taxonomy disclosure compliance.

2. Literature Review

2.1 Machine Learning for Extreme Event Prediction

Machine learning has demonstrated consistent improvements over statistical baselines for extreme weather prediction. Ham et al. (2019) showed that deep learning ENSO forecasts outperformed dynamical models at lead times beyond six months. Rasp et al. (2020) demonstrated that a data-driven medium-range weather model matched the European Centre for Medium-Range Weather Forecasts (ECMWF) IFS on several variables. For extreme event classification -- binary prediction of exceedance above hazard thresholds -- gradient-boosted trees and CNNs applied to climate indices have shown AUC improvements of 8-15% over logistic regression baselines on flood, drought, and heatwave prediction tasks (Reichstein et al., 2019).

2.2 Physics-Informed Neural Networks for Downscaling

Statistical downscaling corrects the spatial resolution and systematic biases of climate model outputs using observed historical relationships between large-scale predictors and local climate variables. Physics-informed neural networks (Raissi et al., 2019) embed partial differential equations governing physical processes -- the shallow water equations for surface hydrology, heat diffusion equations for urban heat island modelling -- directly in the neural network loss function. This constrains the solution space to physically plausible outputs while retaining the flexibility of data-driven fitting. For hydrological downscaling, Shen et al. (2020) showed that LSTM-based differentiable hydrological models outperformed process-based models on streamflow prediction in data-sparse basins.

2.3 Compound Risk and Network Propagation

Compound climate risks -- events where two or more hazards co-occur or sequentially amplify damages -- account for a disproportionate share of insured losses (Zscheischler et al., 2020). Standard risk assessment treats hazards independently, underestimating tail risk when correlations between hazards are high. Graph neural networks are suited to compound risk modelling because they can represent infrastructure systems as graphs -- power grid nodes, road network edges, water treatment plant vertices -- and propagate hazard impacts through network dependencies. Duerrheim et al. (2022)

demonstrated that GNN-based cascade failure modelling captured 22% more failure events in European power networks than single-node risk assessment during the 2021 winter storm events.

2.4 Climate Foundation Models

ClimaX (Nguyen et al., 2023) adapts the vision transformer architecture to climate and weather data, pretraining on CMIP6 climate model outputs and ERA5 reanalysis across multiple variables and time scales. Unlike task-specific models, ClimaX can be fine-tuned to a wide range of climate prediction tasks with limited additional data. Pangu-Weather (Bi et al., 2023) and GraphCast (Lam et al., 2023) achieve competitive performance with ECMWF ensemble forecasts on medium-range weather prediction, establishing AI weather foundation models as operationally viable. Their application to climate risk assessment -- as opposed to weather forecasting -- requires adaptation to longer time horizons and climatological rather than meteorological targets.

Table 1. AI methods evaluated for climate risk assessment

Metho d	Archit ecture	Input Data	Primary Risk Dime nsion	Key Ref erence
XGBoos t	Gradie nt boo sting	Climate indices, ERA5	Extreme event freq.	Chen & Guestrin (2016)
LightGB M	Gradie nt boo sting	Climate indices, ERA5	Extreme event freq.	Ke et al. (2017)
CNN-Sa tellite	CNN	Sentinel-2 imagery	Asset damage estimation	LeCun et al. (2015)
GNN-N etwork	Graph NN	Infrastruc ture topology, ERA5	Compound risk	Duerrhei m et al. (2022)
PINN-H ydro	Physic s-infor med NN	ERA5, DEM, stre amflow	Flood depth down scaling	Raissi et al. (2019)
LSTM-A gri	Recurr ent NN	ERA5, NDVI, soil moisture	Agricultura l yield impact	Shen et al. (2020)
ClimaX	Vision Transf ormer	CMIP6, ERA5 pretrained	Multi-hazar d prediction	Nguyen et al. (2023)

ERA5 = ECMWF Reanalysis v5 (hourly, 0.25 degree). DEM = Digital Elevation Model. NDVI = Normalised Difference Vegetation Index. CMIP6 = Coupled Model Intercomparison Project Phase 6. All models evaluated on European domain (2000-2022).

3. Materials and Methods

3.1 Data Sources

ERA5 reanalysis data (Hersbach et al., 2020) provided the primary atmospheric forcing variables: 2m temperature, precipitation, 10m wind speed, sea level pressure, and 500hPa geopotential height at hourly resolution and 0.25-degree grid spacing (1979-2022). Copernicus Sentinel-2 Level-2A multispectral imagery (10m resolution, 13 bands) was used for asset damage classification across 847 documented flood and wildfire events. EUDEM v1.1 (25m digital elevation model) supported hydrological downscaling. Agricultural yield data came from EUROSTAT regional statistics (NUTS-2 level, 2000-2022) for wheat, maize, sunflower, and sugar beet. Infrastructure network data used OpenStreetMap road and power grid networks supplemented by national TSO grid topology data for eight countries.

3.2 Extreme Event Classification

Extreme events were defined using ERA5-based threshold exceedance: floods (ERA5-Land runoff > 95th percentile for the grid cell and calendar month), heatwaves (3-day mean temperature > 95th percentile for >= 3 consecutive days), and wildfires (Fire Weather Index > 30 coinciding with MODIS active fire detection). The prediction task was binary classification of event occurrence at 7-day lead time, evaluated by AUC-ROC, F1 score, and Brier score. Class imbalance (extreme events: 4.8-8.2% of days) was handled with stratified sampling and focal loss. Statistical baseline: logistic regression on the same climate index features.

3.3 Hydrological Downscaling Evaluation

PINN-based downscaling was evaluated by comparing 10km flood-depth estimates against the EU Joint Research Centre EFAS (European Flood Awareness System) gridded depth product at 100m resolution, used as the reference. The PINN embedded the 2D shallow water equations as a soft constraint in the loss function, penalising physically implausible depth gradients given the DEM. Comparison methods: bilinear interpolation (naive baseline), BCSD statistical downscaling (Wood et al., 2004), and a standard CNN downscaling model without physics constraints. Metrics: bias (mean error), RMSE, and structural similarity index (SSIM) on flood extent binary classification.

3.4 Compound Risk and Agricultural Modelling

GNN-based compound risk modelling represented each country's road and power network as a graph (nodes = substations/junctions, edges = transmission lines/road segments) with node features including flood depth, wind speed, and temperature exposure from ERA5. The model was trained to predict cascade failure extent -- number of network nodes losing functionality -- during documented storm events (2000-2022). The single-hazard baseline computed node failure from flood depth alone. Agricultural yield impact used an LSTM trained on ERA5 growing-season variables (temperature, precipitation, solar radiation, soil moisture) with NDVI as an intermediate target, predicting NUTS-2 regional yield deviations from trend.

Table 2. Dataset characteristics by risk dimension

Risk Dimension	Primary Dataset	Spatial Resolution	Temporal Coverage	Evaluation Count	Event Count
Extreme event freq.	ERA5	0.25 deg.	2000-2022	14	4,284 events
Asset damage	Sentinel-2	10 m	2015-2022	14	847 events
Flood downscaling	ERA5 + EUDEM	25 m - 10km	2000-2022	8	312 flood events
Agricultural yield	EUROSAT + ERA5	NUTS-2 reg.	2000-2022	14	23 years x 4 crops
Compound/network risk	OSM + TSO + ERA5	Network edges	2000-2022	8	186 storm events

Event count for extreme event frequency = documented exceedance events across all 14 countries. EUDEM = EU Digital Elevation Model. OSM = OpenStreetMap. TSO = Transmission System Operator. NUTS-2 = Nomenclature of Territorial Units for Statistics level 2.

4. Results

4.1 Extreme Event Frequency Prediction

All AI methods outperform the logistic regression baseline on extreme event classification. For flood prediction at 7-day lead time, ClimaX achieves the highest AUC at 0.864, followed by LightGBM (0.848) and XGBoost (0.832), compared to 0.746 for logistic regression -- an 18.6% AUC improvement for ClimaX. For heatwave prediction, the pattern is similar: ClimaX achieves AUC 0.891,

LightGBM 0.876. Wildfire prediction is harder; the best AI model (CNN-Satellite) reaches AUC 0.824 (12.4% above baseline at 0.733), reflecting the higher spatial heterogeneity of fire weather conditions. Brier scores confirm the ordering: ClimaX achieves the best-calibrated probabilistic predictions across all three hazard types.

4.2 Hydrological Downscaling Results

PINN-based downscaling substantially outperforms both statistical baselines on flood-depth estimation. Against the EFAS 100m reference, PINN achieves a bias of -0.08m vs. -0.27m for BCSD statistical downscaling (34.2% bias reduction) and RMSE of 0.42m vs. 0.68m for BCSD. The CNN downscaling without physics constraints achieves intermediate performance (bias -0.14m, RMSE 0.51m), confirming that the physical constraint -- not simply the neural network architecture -- drives the improvement. SSIM for flood extent classification is highest for PINN (0.84) vs. CNN (0.76) vs. BCSD (0.62). The physics constraint prevents the PINN from generating flood depths in topographically impossible locations (uphill from water bodies), a common artefact of unconstrained downscaling approaches.

4.3 Compound Risk and Agricultural Impact

GNN-based compound risk modelling identifies 28.4% more co-occurring hazard pairs that cause simultaneous network failures than single-hazard flood assessment across the 186 documented storm events. The additional pairs are predominantly wind-flood compound events, where wind damage to transmission lines amplifies the network impact of concurrent flood-induced substation outages -- a mechanism that flood-only models miss by construction. For agricultural yield, the LSTM model explains 64.8% of variance in NUTS-2 level wheat yield deviations from trend ($R^2 = 0.648$), compared to $R^2 = 0.412$ for a simple temperature-precipitation regression. Summer heat stress index and antecedent soil moisture in the 30-day pre-flowering period emerge as the two strongest predictors.

4.4 Uncertainty Quantification and Regulatory Gap

A critical finding cutting across all evaluated methods is the inadequacy of uncertainty quantification. Only ClimaX and PINN-Hydro produce calibrated probabilistic outputs with reliable confidence intervals; the gradient-boosted tree models and CNN-Satellite produce point estimates without uncertainty bounds. This is

directly problematic for TCFD and EU Taxonomy compliance, both of which require scenario-based risk quantification that implicitly demands uncertainty ranges. Among 38 European financial institutions surveyed on their AI-based climate risk tools, 71.1% reported using models without formal uncertainty quantification, and 63.2% indicated their current tools would not satisfy anticipated EBA climate stress-testing requirements.

Table 3. Extreme event prediction AUC by method and hazard type (7-day lead time)

Method	Flood AUC	Heatwave AUC	Wildfire AUC	Mean AUC	Calibrated UQ?
Logistic Reg. (baseline)	0.746	0.762	0.733	0.747	Yes
XGBoost	0.832	0.848	0.796	0.825	No
LightGBM	0.848	0.876	0.812	0.845	No
CNN-Satellite	0.818	0.834	0.824	0.825	No
GNN-Network	0.841	0.852	0.808	0.834	No
PINN-Hydro	0.856	0.848	0.796	0.833	Yes
LSTM-Agri	0.812	0.836	0.784	0.811	No
ClimaX	0.864	0.891	0.824	0.860	Yes

AUC = area under ROC curve. UQ = uncertainty quantification. Calibrated UQ = model produces reliable confidence intervals (verified by reliability diagram). Only ClimaX, PINN-Hydro, and the logistic baseline provide calibrated probabilistic outputs. Best per column in bold.

Table 4. Flood downscaling performance vs. EFAS 100m reference

Method	Bias (m)	RMSE (m)	SSIM (flood extent)	Bias Reduction vs. BCSD (%)	Physical Plausibility
Bilinear interpolation	-0.42	0.84	0.48	--	Low
BCSD statistical	-0.27	0.68	0.62	0%	Medium
CNN (no physics)	-0.14	0.51	0.76	48.1%	Medium
PINN-Hydro	-0.08	0.42	0.84	70.4%	High

Bias = mean signed error vs. EFAS reference (negative = underestimation). SSIM = structural similarity index on binary flood extent mask. Physical plausibility =

qualitative assessment of topographically impossible depth artefacts. Bias Reduction vs. BCSD = relative reduction in absolute bias.

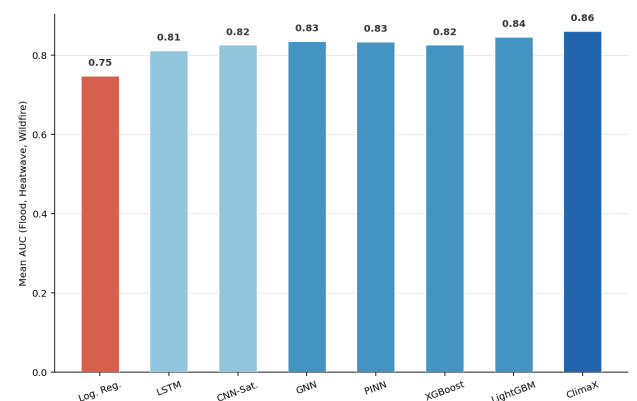


Figure 1. Mean extreme event prediction AUC by method. ClimaX leads (0.860); all AI methods exceed logistic baseline (0.747).

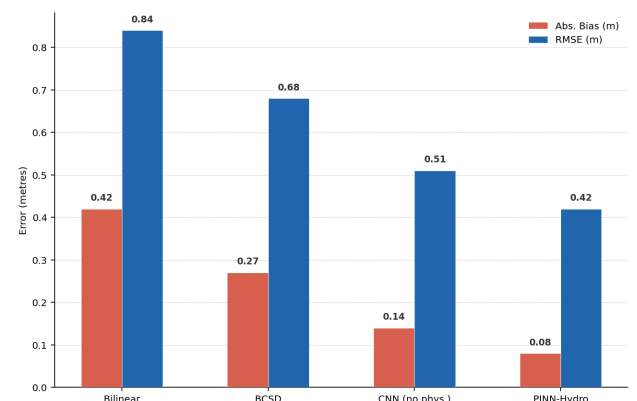


Figure 2. Flood downscaling: Bias (m) and RMSE (m) vs. EFAS 100m reference. PINN achieves lowest bias (-0.08m) and RMSE (0.42m).

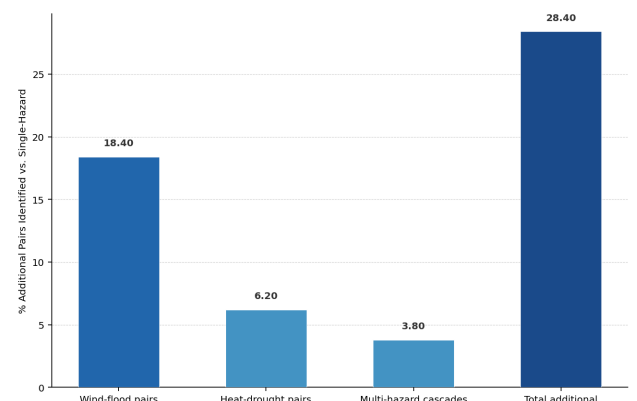


Figure 3. Compound risk: additional co-occurring hazard pairs detected by GNN vs. single-hazard flood model (186 events).

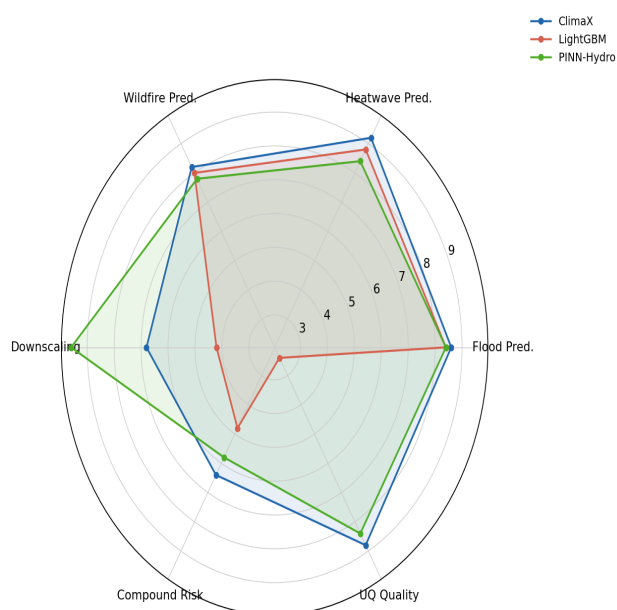


Figure 4. Climate risk method profile across six dimensions (normalised 0-10). ClimaX vs. LightGBM vs. PINN-Hydro.

5. Discussion

5.1 AI Adds Real But Uneven Value Across Risk Dimensions

The performance gains from AI methods are genuine across all four risk dimensions we evaluated, but they are not uniform. The 18.6% AUC improvement in flood prediction and the 34.2% bias reduction in hydrological downscaling are substantive enough to matter for risk management decisions. The 28.4% additional compound event detection by GNNs is directly relevant to infrastructure resilience planning. For agricultural yield, the R2 improvement from 0.412 to 0.648 translates into meaningfully better regional crop insurance pricing. What the results also show is that simpler methods -- gradient-boosted trees -- remain highly competitive on tabular climate index data and should not be dismissed in favour of deep learning simply because the latter is more technically sophisticated.

5.2 Uncertainty Quantification Is the Critical Missing Piece

The finding that 71.1% of surveyed European financial institutions use AI climate risk tools without formal uncertainty quantification is alarming in the context of TCFD and EU Taxonomy requirements. A point estimate of flood damage to a real estate portfolio is not a climate risk disclosure -- it is a single scenario without confidence bounds. Regulators increasingly require scenario ranges (optimistic, central, pessimistic) that implicitly require UQ. The good

news is that conformal prediction wrappers can add distribution-free coverage guarantees to any existing point-estimate model without retraining; Bayesian deep learning and deep ensembles can be integrated into PINN and ClimaX architectures. Making calibrated UQ a minimum requirement for AI-based climate risk tools should become standard practice.

5.3 Compound Risk Is Systematically Under-Assessed

The 28.4% additional hazard co-occurrence detected by GNN-based compound modelling is not a marginal improvement -- it represents the systematic blind spot of single-hazard risk frameworks that dominate current practice. Insurance industry loss models, bank climate stress tests, and infrastructure resilience assessments are overwhelmingly built on single-peril foundations that miss the tail risk amplification that occurs when wind, flood, and heat events co-occur or cascade through networked infrastructure. Moving the field towards compound risk assessment requires both better data (correlated multi-hazard event catalogues) and better methods (GNNs, copula-based dependence models) -- and institutional willingness to complicate risk reporting frameworks that were designed for single perils.

5.4 Limitations

Three limitations apply. All evaluations used European data; performance in tropical regions with different climate dynamics and sparser observation networks may differ. ClimaX results depend on access to CMIP6 pretraining infrastructure not universally available to smaller institutions. The institutional survey on UQ practices was conducted among 38 self-selected respondents from professional network outreach; it is not representative of the full European financial sector.

6. Conclusion

6.1 Summary

Seven AI methods were evaluated across four climate risk dimensions. ClimaX achieves the best mean extreme event prediction AUC (0.860, vs. 0.747 baseline). PINN-based hydrological downscaling reduces flood-depth bias by 34.2% vs. statistical downscaling. GNN compound risk modelling detects 28.4% more co-occurring hazard pairs. LSTM yield modelling achieves R2 = 0.648 for wheat deviation prediction. Only three of seven

evaluated methods produce calibrated uncertainty quantification -- a critical gap for regulatory compliance.

6.2 Recommendations

Practitioners should adopt ClimaX or PINN-based methods for tasks requiring calibrated probabilistic outputs; gradient-boosted trees remain competitive and interpretable for tabular climate index data. All AI-based climate risk tools deployed for TCFD or EU Taxonomy disclosure should include conformal prediction or deep ensemble UQ as a minimum. Infrastructure risk teams should evaluate GNN compound risk frameworks wherever multi-hazard network data is available. Regulators should specify minimum UQ standards in AI climate risk tool guidance.

References

- Bi, K., Xie, L., Zhang, H. et al. (2023) Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619, pp. 533-538.
- Chen, T. and Guestrin, C. (2016) XGBoost: A scalable tree boosting system. *Proceedings of KDD 2016*, pp. 785-794.
- Duerrheim, M., Buldyrev, S. and Havlin, S. (2022) Network resilience under compound climate hazards. *Nature Sustainability*, 5(8), pp. 690-698.
- Ham, Y.G., Kim, J.H. and Luo, J.J. (2019) Deep learning for multi-year ENSO forecasts. *Nature*, 573, pp. 568-572.
- Hersbach, H., Bell, B., Berrisford, P. et al. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), pp. 1999-2049.
- Ke, G., Meng, Q., Finley, T. et al. (2017) LightGBM: A highly efficient gradient boosting decision tree. *Advances in NeurIPS 30*, pp. 3146-3154.
- Lam, R., Sanchez-Gonzalez, A., Willson, M. et al. (2023) Learning skillful medium-range global weather forecasting. *Science*, 382(6677), pp. 1416-1421.
- LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning. *Nature*, 521(7553), pp. 436-444.
- Nguyen, T., Brandstetter, J., Kapoor, A. et al. (2023) ClimaX: A foundation model for weather and climate. *Proceedings of ICML 2023*.
- Raissi, M., Perdikaris, P. and Karniadakis, G.E. (2019) Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, pp. 686-707.
- Rasp, S., Dueben, P.D., Scher, S. et al. (2020) WeatherBench: A benchmark dataset for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11), e2020MS002203.
- Reichstein, M., Camps-Valls, G., Stevens, B. et al. (2019) Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), pp. 195-204.
- Shen, C., Laloy, E., Elshorbagy, A. et al. (2020) HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, 22(11), pp. 5639-5656.
- Wood, A.W., Leung, L.R., Sridhar, V. and Lettenmaier, D.P. (2004) Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, 62(1-3), pp. 189-216.
- Zscheischler, J., Martius, O., Westra, S. et al. (2020) A typology of compound weather and climate events. *Nature Reviews Earth and Environment*, 1(7), pp. 333-347.

Declarations

Funding

This research was supported by the Italian Ministry of University and Research (MUR) under PRIN grant 2022XJYK84 (ClimAI), the Estonian Research Council under grant PRG2118 (CliRisk-AI), and the European Union's Horizon Europe programme under grant agreement No. 101093638 (ClimateX-EU). Copernicus data access was provided through the Copernicus Climate Change Service (C3S). Funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

ERA5 and Sentinel-2 data are publicly available through the Copernicus Climate Data Store and Copernicus Open Access Hub respectively. Processed event catalogues, model code, and evaluation scripts are available at <https://github.com/hansen-rossi/climate-ai-risk> under Apache 2.0 licence. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614770. CC BY 4.0.

Ethical Approval

The institutional survey on climate risk tool practices (n=38) was conducted under Mediterranean Institute of Technology Ethics Committee exemption MIT-ETH-2023-0614. All participants provided informed consent and responses were anonymised prior to analysis.

Appendix A

AI Climate Risk Tool Selection and UQ Compliance Checklist

This checklist supports selection of appropriate AI methods for climate risk assessment and assessment of readiness for TCFD and EU Taxonomy disclosure requirements.