Simulating ~~multivariate hazards~~ spatial multi-hazards with generative deep learning

Alison Peard[1], Yu Mo[1], and Jim W. Hall[1]

[1]Environmental Change Institute, University of Oxford, Oxford, UK

Correspondence: Alison Peard (alison.peard@ouce.ox.ac.uk)

Abstract. When natural hazards coincide or spread over large areas they can create major disasters. For accurate risk analysis, it is necessary to simulate many spatially resolved hazard events that capture the relationships between extreme variables, but this has proved challenging for conventional statistical methods, particularly in high-dimensional settings. In this article ~~,~~ we show that ~~deep generative models offer a powerful~~ generative deep learning models—when combined with specific transformations to the training data—offer a useful alternative method for ~~creating sets of synthetic hazard eventsdue to their ability to implicitly learn the joint distribution of high-dimensional data~~stochastically sampling realistic multi-hazard events. Our framework combines generative adversarial networks with extreme value theory ~~to construct a hybrid method that captures~~ in a hybrid approach that can capture complex dependence structures in gridded multivariate weather data~~and provides~~ , while providing a theoretical justification for extrapolation to new extremes. We apply our method to model the co-occurrence of strong winds, ~~low pressure, and heavy precipitation~~heavy precipitation, and low atmospheric pressure during storms in the Bay of Bengal, demonstrating that our model learns the spatial and multivariate extremal dependence structures of the underlying data and captures the distribution of storm severities. ~~Validation shows excellent preservation of spatial correlation structures (r = 0.977, MAE = 0.053) and multivariate dependencies (r = 0.817, MAE = 0.096) for wind, precipitation, and pressure fields. In a case studyof storm risk to mangrove forests, we demonstrate that correctly modelling the dependence structures leads to far more realistic estimates of aggregate damages. While our method shows mild underestimation of the damages with a mean absolute error of 93.57 km^2, this remains an order of magnitude lower than errors from independence assumptions (460.54 km^2) and the total dependence assumption (1056.90 km^2) that is implicit when using return period maps. The framework developed in this paper is flexible and applicable across a wide range of data regimes and hazard types.~~ For the Bay of Bengal case study, we validate our approach against a well-known model for multivariate extremes, and demonstrate improved performance in capturing the extremal correlation structure.

1 Introduction

1.1 ~~Background~~

~~Multivariate and spatially compounding hazards can lead to disasters of tremendous severity and complexity (?). Multivariate compound hazards occur when different hazards coincideover the same region. For example, droughts often coincide with~~

heatwaves, while strong winds frequently accompany heavy rainfall (?). During Hurricane Sandy of 2012, disastrous inland flooding was driven by the combined effects of rainfall Hazards that entail several extreme weather variables and extend over large spatial domains are responsible for some of the most damaging natural catastrophes (??). Multivariate or *multi*-hazards (?) occur when multiple hazardous processes coincide, exacerbating impacts via either hazard intensification (*e.g.*, cumulating flood depths from coastal and fluvial flooding) or synergistic damages (*e.g.*, strong winds combining with heavy rainfall increasing damage to structures). Common examples of multi-hazards include droughts and heatwaves, the combination of rainfall and strong winds, tides, and storm surges (?). Furthermore, the vulnerability of natural and man-made assets typically exhibits nonlinear relationships with hazards. Damage to mangroves during storms, for instance, results from the cumulative effects of rainfall, wind speed, and storm speed, alongside other climatological, biological, and geophysical variables (???).

and compound flooding events involving fluvial (river), pluvial (rainfall), and coastal sources (??). Spatially compounding hazards occur when climate hazards spread over large regions and create strain on distributed systems. Examples of this include droughts hitting multiple global breadbaskets and creating food shortages, or widespread a hazard impacts a large region, creating widespread damages and potentially systemic impacts (?). Common examples include droughts causing crop failure in multiple regions simultaneously, leading to food stress; or extensive storm damages stretching emergency response capacity (????). In 2002, droughts hit Europe, Russia, India, and China, leading to significant decreases in the global production of rice and wheat (?). The risk of simultaneous breadbasket failures is estimated to increase with global mean temperature, rising from 6% to 40% for maize between historical (0.85°C) and 1.5°C levels (?). (????).

Hazards and their drivers exhibit highly variable levels of dependence across space and between variables. In spite of this heterogeneous dependence, climate risk to human or natural systems is often assessed by examining hazard variables individually at a specific location or by treating hazard maps like events. This implicitly assumes total dependence across space. A 1-in-50 year flood map, for example, represents a flood with a 50-year return level depth in all locations simultaneously. In reality, however, we expect such a homogeneously distributed flood to be far rarer than a 1-in-50 year event. Such simplifications omit the effects of spatially compounding hazards and generate significant bias (??).; however, assumptions such as these have been employed in high-profile studies. In an article that informed a flagship World Bank report, ? treated flood hazard maps like real eventsand used them in a transport routing analysis for major cities in Mali, Tanzania, Uganda, and Rwanda. Such an assumption may be justified at a city scale, but in this paper we will demonstrate the bias created when employing such assumptions on larger national or continental scales.

A second limitation of hazard maps is that they are univariate and do not account for multi-hazard risk . Thus, any climate risk analysis relying on hazard maps necessarily treats different hazard types independently. Such an approach risks neglecting some of the most potentially disastrous climate risks. How likely is it for high temperatures and drought to hit multiple continents simultaneously? What is the probability of multiple villages along a coastline being hit by storm surge such that emergency services are at capacity? How likely is it that this storm coincides with high tides or pluvial flooding? With univariate approaches, the probability of such disastrous eventsremains unquantified.

As climate change shifts the distribution of climate variables, the probability of co-occurring extremes may also increase (?). The spatially compounding problem has motivated many risk analysts to move towards an event-based approach, estimating the

damages across sets of thousands of plausible hazard events and producing statistics for the resulting damages (???). Synthetic event Risk analysis often uses hazard maps: spatial grids of marginal return levels at specified exceedance probabilities. While hazard maps provide a convenient visualisation of hazard intensity at given return levels, they are uni-hazard and fail to provide information about the spatial extent of potential events, which can bias estimates of tail risk (?). Hazard event sets, which map the hazard intensity of actual or synthetic events, provide the basis for Monte Carlo simulation of damages and losses, which for a large enough sample will yield unbiased loss distributions. Hazard events sets are widely used for catastrophe modelling in the insurance industry by the insurance sector in Cat models and have been used for several in large-scale infrastructure risk assessments in the UK (??). and the USA (?????). Synthetic event sets can be generated from physically based model simulations (?) but are computationally demanding and may entail elaborate model coupling, *e.g.*, to obtain storm surge heights as well as windspeeds and precipitation.

For simulating synthetic hazard event sets , statistical methods have been widely used (see Sect. ??) yet have proven to be limited in their capacity to accurately simulate multivariate extreme events over large domains. In this article, we harness a deep learning framework to produce spatially coherent multi-hazard event ensembles. These event ensembles should preserve the statistical dependence properties of historical climate data in a spatially coherent manner, especially in the extremes. Our criteria for success are that the generated data: (i) replicates the overall distribution of extreme events; (ii) preserves the Alternatively, it is possible to take a purely stochastic approach to hazard modelling by estimating the parameters of models from the statistical theory of multivariate extremes and then simulating large sets spatial events from those models (?????). Classical models for the dependence structure of the training data across space—in the bulk and in the tails; (iii) preserves the dependence structure between different variables—in the bulk and in the tails; (iv) preserves the marginal distributions of the training data; and (v) has a well-justified basis for extrapolation to new, more extreme events.

1.1 Statistical approaches

There is a rich—though at times much-debated (???)—literature concerning the statistics of univariate and multivariate extremes (????), with a dense sub-literature of applications to hydrological and climate data (???). To calculate statistics over possible damages, rigorous climate risk analysis requires samples of thousands of extremes from the joint distribution of potentially compounding climate hazards. Two popular approaches to estimate the joint distribution are copula methods and conditional exceedence models (???). Both approaches decompose the problem into modelling the univariate marginal distributions and learning the dependence structure between the marginals, usually after some probability space transformation. Copulas are based on Sklar's theorem and accurate modelling hinges on choosing the most appropriate copula family for the dependence structure (?). The conditional exceedence model of ? calculates the probability that each variable is extreme, given that some conditioning variable is already extreme. This introduces flexibility into the asymptotic dependence structure between bivariate pairs.

Max-stable processes have also gained popularity in practical application since ?'s development of likelihood-based inference methods,and can be loosely interpreted as an extreme extension of Gaussian random fields for annual maxima (?). They suffer a number of limitations, however, which are reviewed in-depth in ?. In particular, annual maxima almost never represent a

real event over large regions; the max-stability requirement implies that the spatial dependence structure is independent of block size; sub-asymptotic tail dependence is not captured; and most approaches to fitting the likelihood function become computationally expensive as the number of locations modelled increases.

For the specific case of tropical cyclones, various stochastic–empirical methods have been used to simulate realistic cyclone tracks and intensities for risk analysis,mostly based on IBTrACS (**????**). The STORM dataset of **?**, for example, follows a stochastic-empirical approach and models wind speed using an empirical relationship for pressure, whose evolution is predicted using a constrained autoregressive process. The autoregressive process is not an extremal model, multivariate extremes include extremal copulas and the conditional exceedance model (**???**); the latter has become popular in climate applications due to its flexibility and ability to scale to high dimensions. However, the conditional exceedance model is still limited in its ability to scale beyond 2,400 dimensions (*e.g.*, **?**) and cannot be used to generate hazard scenarios at new locations. Spatial models such as max-stable models and r-Pareto processes address this by using geostatistical methods to parametrise dependence structures across the spatial domain. The gradient-based r-Pareto processes of **?** are particularly powerful, capable of modelling up to up to 3,600 dimensions (**?**). Most spatial process models, however, which somewhat limits the justification for extrapolation to unseen extremes. suffer a trade-off between flexibility and scalability: many are either unable to capture a sufficiently wide variety of asymptotic dependence structures, have rigid requirements around event definitions, or struggle to capture spatial nonstationarity (**???**). A further challenge is to simulate high-order dependence structure between multiple variables, which yields the spatial patterns of weather events at large scales. Statistical methods may accurately estimate and reproduce parameterised dependence structures, yet the events do not necessarily 'look like' realistic weather, as might be observed in a rainfall radar.

1.1 Data-driven approaches

Data-driven approaches to modelling multivariate extremes in high dimensional spaces have been evolving over the last two decades (**??????**) but, with the recent strides made in generative artificial intelligence (**?**), data-driven methods are being catapulted to the forefront of research agendas. Since 2021, several innovative empirical methods, which we outline here, have been developed to address the statistical modelling of high dimensional multivariate extremes.

? give an overview of the limitations of classical statistical methods for representing extremes in high-dimensional settings and propose two unsupervised learning approaches, clustering and principal component analysis, to reduce the dimensionality of the problem. **?** use k-means clustering to cluster a global dataset of extreme sea levels and combine this with a conditional exceedence method to generate a global event set. **?**, in a model called DeepGPD, use fully connected and convolutional neural networks to predict the shape and scale parameters of generalized Pareto-distributed asymptotics above a certain threshold, regressing against exceedences and other covariates from the previous time step.

Many authors leverage a deep learning model known as a generative adversarial network (GAN) which has several appealing properties, including the ability to implicitly learn the distribution of the training data (**?**). GANs consist of two deep neural networks Recently, interest has grown in using machine learning methods to generate multivariate climate and weather extremes (**????**). Deep learning approaches such as generative adversarial networks (**??**) have shown particular promise (**????**). GANs

were originally developed for image generation and their adversarial loss formulation naturally lends them to the generation of visually realistic spatial patterns (**??**). GANs place two neural networks in competition: a generator ~~and a discriminator. The networks are trained adversarially: the generator maps a random latent space to an image space, and the discriminator must distinguish between real and generated images. It has been shown, however, that GANs tend to produce samples near the bulk of the distribution (**?**). For this reason, several methods have emerged to adapt GANs to produce more extreme samples, in controlled and theoretically justified ways. We recommend **?** for a good introduction to the fundamentals of GANs.~~

~~**?** propose *ExGAN*, which uses extreme value theory to assign a measure of extremeness to cumulative precipitation over the contiguous United States. They employ a data-shifting routing to filter and shift the data distribution towards the extremes, and use this to generate samples conditional on desired extremeness. **?** prove that, provided it has unbounded activation functions, the tails of data generated by a GAN will have the same shape as those of underlying latent space , which is usually Gaussian. They propose~~ transforms a latent variable (a low-dimensional random variable) into synthetic samples, while a discriminator attempts to distinguish these from real data. The use of GANs for the statistical simulation of extreme events is more recent but shows considerable promise. Most relevant to this work, **?** demonstrated that a GAN with a light-tailed latent space cannot effectively capture the tail behaviour of heavy-tailed datasets. Building on this, **?** developed *ParetoGAN*, which uses a ~~generalised~~ unit Pareto latent space ~~and a root-Euclidean energy distance~~ to better represent extremes—although this required a custom loss function, ~~and demonstrate on synthetic datasets that this better captures power-law behaviour.~~

~~**?** fit generalized extreme value (GEV) distributions to a spatial grid of annual maximum temperature and precipitation over western Europe and perform a probability integral transform (PIT) before training a deep convolutional GAN (DCGAN) on the images with uniform marginals (**?**). They use the fitted GEV parameters to transform data back to the original scale and guide extrapolation beyond the range of the data. Their model *evtGAN* is flexible and can learn complex nonlinear and spatially nonstationary dependence structures. It can also allow for mixtures of heavy and light-tailed marginal distributions. The authors demonstrate that incorporating the PIT step significantly improves the representation of extreme samples versus a DCGAN, whilst maintaining good representation in the bulk of the data. However, evtGAN is trained on annual block maxima, so does not account for compound hazards on smaller time scales, such as storms occurring over a few days. It is also univariate, so it does not cover multivariate compound events.~~ sacrificing some benefits of the adversarial loss. Separately, **?** used methods from extreme value theory to transform weather data to have uniform margins, training a GAN to learn the dependence structure in the transformed space. This approach improved the ability of the GAN to learn extremal dependence structures; however, as we will demonstrate, transforming to uniform margins compresses tail information, and their reliance on annual maxima introduces spatial incoherence, limiting their power to capture shorter-timescale events.

~~Our framework builds on the work of **?** to increase temporal granularity and model multivariate events, thus creating a powerful tool for climate risk analysis of large-scale systems.~~

1.1 ~~Our approach~~

~~This article proposes *hazGAN*, a modelling framework that builds upon the work summarized in Sect. **??** to create datasets which are directly applicable in risk analysis. We begin with the model of **?** and make the following series of modifications:~~

~~We replace the annual block~~ In this manuscript, we develop a model capable of generating spatially coherent multi-hazard event ensembles that preserve both the marginal and joint distributions of the training data. We achieve this by combining insights from **?** and **?** with statistical theories of spatial and multivariate extremes. Building on the conclusions of **?** and **?** that the tail-heaviness of a GAN's latent space and its training data should agree, we demonstrate equivalent improvements in tail representation by transforming the training dataset to have light-tailed margins and training a standard GAN, thereby preserving the benefits of the original adversarial loss. While **?** successfully modelled extremal spatial dependencies by training a GAN on data transformed to have uniform margins, we instead train on data transformed to have light-tailed margins and demonstrate improved representation of both marginal tail behaviour and the dependence structure. We further advance the methodology by replacing the annual maxima approach with a ~~peak-over-threshold (POT) approach , allowing us to use more training data, limit parametric assumption-making to the tails, and draw stronger conclusions about the degree of dependence across space and between variables. The POT approach allows us to work with hazard~~ peaks-over-threshold approach and using domain-wide functions to characterise events, enabling us to capture spatially coherent event footprints. Event *footprints* ~~, which are already~~ are commonly used in ~~risk modelling and have a trivial extension to multivariate events (**?**). We generate hazard footprints from gridded climate data using a runs declustering approach (**?**). We train GANs with multiple channels, such that each channel represents a different climate variable, and thus images represent multivariate hazard footprints. We transform images such that all marginal distributions along the time dimension are Gumbel-distributed, and train the GAN on these, which encourages it to be sensitive to dynamics in the extremes.~~

~~We demonstrate our method 's effectiveness by training the model on historical footprints of wind speed, precipitation, and sea level pressure corresponding to wind storms over~~ catastrophe modelling and post-disaster needs assessments to represent the maximum intensity of a hazard event across a region during its lifetime (*e.g.*, **?**). This provides a two-dimensional representation that can be used to assess the maximum impacts of a hazard event. Finally, we extend the model to multiple channels, allowing us to capture multi-hazard events. Initial benchmarking demonstrates that our method better reproduces the extremal correlation structure of the Bay of Bengal ~~. We then use these to construct risk profiles for the total area of storm damage to mangrove forests across the~~ storm footprints compared to the well-known conditional exceedance model of **?**.

We describe the general theory and methodology in Section **??**. In Section **??**, we demonstrate a practical application, using the model for a case study of storms in the Bay of Bengal. The Bay of Bengal is chosen because it is highly exposed to multi-hazard tropical cyclones (**??**) and existing event sets have been shown to struggle in the region (**?**). In Section **??**, we demonstrate an application in risk analysis: evaluating the risk of wind and rainfall-driven storm damage to mangroves in the region. ~~We choose mangroves as an illustrative example because damage to mangrove forests during storms is driven by a combination of rainfall and wind speed (**???**). This multivariate vulnerability makes it difficult to conduct comprehensive scenario modelling of risk to mangrove forests, especially over large areas. Despite the increasing recognition of the valuable ecosystem services provided by mangroves (**??**), and the documented loss of mangroves over the last decade (**?**), comprehensive large-scale scenario modelling of wind storm risk to mangrove forests has not been attempted to date.~~

2 Theory and methods

We present results from our illustrative example and show that our model learns spatial and multivariate dependence structures over large regions, even in the extremes. We additionally demonstrate that our method produces far more realistic estimates of annual mangrove damages, compared to the total dependence assumption which underlies hazard maps, or assumptions of total independence between variables and space, which are implicit in single-location studies. Thus, we replicate the spatial dependence scenarios assessed for flooding in **?**.

To the best of the authors' knowledge, this framework is the first such framework to facilitate multivariate extreme event set generation over such large scales. It is flexible and can be used for any variables for which there exists suitable data. Though applied to gridded reanalysis data in this paper, the method can be used to augment any gridded dataset. Application to global climate model data instead of reanalysis data, for example, would follow an identical method. By using different deep learning architectures, the method could also be extended to different data topologies, and be applied to networks of river gauges or weather stations.

The remainder of this article is structured as follows: Section **??** reviews concepts in extreme value theory and deep learning fundamental to our approach. Section **??** outlines the method. Section **??** applies the method to a concrete example of wind storms in the Bay of Bengal and evaluates its performance. Section **??** demonstrates a use case for the synthetic event set, analysing storm risk to mangrove forests in the Bay of Bengal. Finally, Sections **??** and **??** summarize the contributions of this paper and outline potential future directions for the research.

3 Theory

2.1 Extreme value statistics

We are particularly concerned with the behaviour of climate variables in the extremes, for example, the strongest winds, heaviest precipitation, or lowest sea-level pressure. For risk analysis, we want to project to values more extreme than previously observed and it is important for a trustworthy model to underpin this extrapolation. As such, the choice of model must be well-justified, either by mathematical result or sufficient empirical evidence. In extreme value statistics, asymptotic models are a popular choice as they have rigorous mathematical foundations (**?**). The Fisher–Tippett–Gndenko theorem proves that, under certain conditions, the limiting distribution for block maxima—as block size approaches infinity—is the generalized extreme value (GEV) distribution. Similarly, the Pickands–Balkema–de Haan theorem proves that the limiting distribution for exceedences over a threshold approaches the Generalized Pareto (GPD) distribution as the threshold approaches infinity. Despite their strong theoretical foundations, however, these models suffer from a bias–variance trade-off, particularly in low-data regimes. The threshold or block size must be high enough that it is appropriate to employ the asymptotic results, but there must still be enough data to limit variance in the fit.

Furthermore, for many climate variables, there is strong empirical evidence for the existence of certain parent distributions, with known asymptotic forms. For wind speed in particular, a wealth of evidence supports the existence of a Weibull parent

~~distribution, which some practitioners refer to as a "stretched exponential" type distribution because it stretches out the tails of the exponential decay. This has been observed for temperate storms, thunderstorms, and tropical cyclones, as well as mixed climates (?).~~

~~According to the Pickands–Balkema–de Haan theorem, as the threshold increases, distributions of~~ This section outlines the general theory and method of our approach. To keep the method general and flexible, we have avoided making specific choices for many of the functions introduced in this section, instead describing the method in more general terms. The method involves a series of steps: (i) extracting a set of multi-hazard footprints from gridded historic weather reanalysis; (ii) fitting extreme value distributions to the ~~stretched exponential family converge to their limiting distributions extremely slowly. Hence a prohibitively large volume of data is required before it is appropriate to apply an asymptotic fit (?). Additionally, the limiting distribution for a Weibull parent distribution is the Gumbel distribution (Type I GEV with zero shape parameter) . ? notes that neither GPD nor GEV methods can produce a zero shape parameter as it corresponds to a singularity in the likelihood function. Fréchet fits are rare, so attempting to fit a GPD or GEV to Weibull parent data usually produces a Type III fit, which corresponds to a negative shape parameter and upper bound and is sometimes referred to as a *reversed Weibull* distribution. This has been considered acceptable due to~~ margins of the ~~belief that wind speeds have a natural upper bound; however, ? writes that this is not sufficient evidence for the adoption of a Type III fit and cautions that any constraint on maximum wind speeds must be rigorously justified, especially in risk analysis. Consequentially, for this work, we will assume either empirically or asymptotically justified distributions to model climate variable extremes, depending on existing evidence in the literature.~~ multi-hazard footprints and standardising them; (iii) training a generative adversarial network (GAN) on the transformed multi-hazard footprints; and (iv) generating synthetic multi-hazard footprints from the trained GAN.

2.0.1 ~~Semiparametric distribution function~~

~~To make the best use of small (fewer than 100 years of) climate datasets, we will choose a peak-over-threshold approach for fitting distributions to climate variables and follow the semiparametric approach used by ? in their conditional exceedence model. However, we will use a generalization that allows for arbitrary parametric distributions to be fitted above the threshold~~ We will use the following terminology throughout: a *variable* refers to a weather hazard variable (*e.g.* wind speed, precipitation, ~~rather than only that of~~ sea level pressure); the *sample dimension* refers to the time dimension filtered to only contain event occurrences; and the ~~GPD distribution. This allows us to use different distributions to model the exceedences in cases where this is deemed more appropriate.~~ *margins* refer to the univariate distributions of each variable at each grid cell along the time (or sample) dimension. We will also use the following notation: latitude is indexed by $i = 1, \ldots, H$; longitude by $j = 1 \ldots W$; time by $t = 1, \ldots T$; sample number by $n = 1, \ldots, N$; and variables by $k = 1, \ldots, 3$. The subscript $|_{ijkt}$ indicates which dimensions of a tensor a function is applied over. Data in physical units are denoted by $\mathbf{x} \in \mathbb{R}^{T \times H \times W \times K}$, variables that have been transformed to have standard uniform margins are denoted by $\mathbf{u} \in \mathbb{R}^{T \times H \times W \times K}$, and variables that have been transformed to have some other distribution are denoted by $\mathbf{y} \in \mathbb{R}^{T \times H \times W \times K}$.

~~For a random variable X and suitably extreme threshold v_X, the three-parameter semiparametric distribution is given by,~~ The method is broadly parametrised by four key choices: (i) the region of interest, defined by a bounding box; (ii) spatiotemporal

data for each variable with the dimensions $H \times W \times T$; (iii) a severity function $r_{|ijk}(\mathbf{x})$, which characterises the severity of the hazard over the spatial domain and is used to select extreme events; and (iv) a temporal aggregation function $h_{k|t}(\mathbf{x})$, which defines how spatiotemporal data for each weather variable is projected into 2-d event footprints. This latter function is employed as the impact from extreme events is typically calculated using a characteristic intensity (*e.g.*, maximum) during the event; accordingly, the GAN is trained on this temporally flattened set of images.

$$\tilde{F}(x) = \begin{cases} 1 - (1 - \hat{F}(v_X))(1 - F_{\xi,\mu,\sigma}(x)) & \text{for } x > v_X \\ \hat{F}(x) & \text{for } x \leq v_X \end{cases}$$

2.1 Event footprint creation

~~where $\hat{F}$ is the empirical distribution function (ECDF) and ξ, μ, and σ are the shape, location, and scale parameters. The meaning of these parameters depends on the distribution in question, but they can be generally understood to describe the tail behaviour, lower bound, and spread of the random variable, respectively.~~ Figure **??** shows the steps to create a set of multi-hazard event footprints. These are: (i) pre-processing the weather data to remove seasonal effects; (ii) identifying hazard events using a severity function and declustering algorithm; and (iii) projecting the spatiotemporal hazard events into 2-d event footprints using a temporal aggregation function.

~~For any random variable X, it is well-known that $F(X)$ is uniformly distributed: $F(X) = U \sim \mathcal{U}(0, 1)$. This fact can be used to transform a random variableto any arbitrary distribution $\mathcal{D}$. Provided the distribution function $F_{\mathcal{D}}$ is known, $F_{\mathcal{D}}^{-1}(U) \sim \mathcal{D}$. This very useful transform is known as the probability integral transform (PIT) and we will make extensive use of it in this work.~~

2.1.1 ~~Extremal dependence metrics~~

Pre-processing the weather data

To handle seasonal effects in weather data, which can bias results and complicate fitting statistical models, deseasonalisation and trend removal is carried out as a pre-processing step. We use the notation $s_{|t} : \mathbb{R}^T \to \mathbb{R}^T$ to denote a generic seasonalisation function. Many methods of varying complexity are available, ranging from simple (subtracting climatology or filtering by season) to complex (fitting generalized linear models with seasonal covariates). The choice of method will depend on the data being modelled and the context (see Sect. **??**).

~~Simply extrapolating every variable to new extremes is insufficient, it is also necessary to capture the dependence between multiple variables in the extremes. If two climate variables are likely to reach extreme values simultaneously, this has important implications in risk analysis. We require extremal dependence metrics to measure and compare the level of extremal dependence between different variables~~

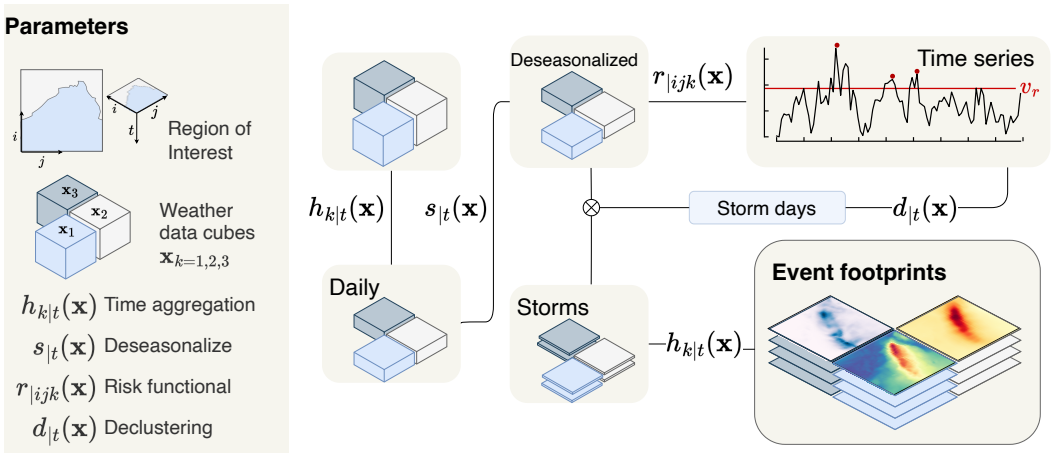


Figure 1. Schematic of the workflow to extract hazard event footprints. Gridded weather data over a region of interest for three variables $\mathbf{x} \in \mathbb{R}^{T\times H\times W\times K}$ (indexed $k = 1, 2, 3$) is deseasonalised according to $s_{|t}(\mathbf{x})$. A risk functional $r_{|ijk} : \mathbb{R}^{T\times H\times W\times K} \to \mathbb{R}^{T}$ constructs a time series from the deseasonalised data and a declustering algorithm $d_{|t}\big(r_{|ijk}(\mathbf{x})\big)$ identifies storm days. Data cubes are extracted for each storm using the storm days and these are aggregated into storm footprints $\mathbf{x} \in \mathbb{R}^{N\times H\times W\times K}$ by applying the temporal aggregation function $h_{k|t}$ along the time dimension.

Identifying hazard events

The deseasonalised weather data is used to identify hazard events. To do this, we use a *severity* function $r_{|ijk} : \mathbb{R}^{H\times W\times T} \to \mathbb{R}^{T}$ to construct a time series of hazard intensities. The definition of this severity function determines the characteristics of the hazards that will be selected, *e.g.* calculating the mean or sum of a variable across a region will prioritize widespread events while selecting the maximum over the region will identify events that reach higher peak intensities and may be more spatially localised. More specialised severity functions could also be used, such as for example hazard indices like the storm severity index for windstorms (**?**) or the fire weather index for fire potential (**??**).

~~Standard dependence measures such as correlation are limited because variables that are correlated near their means won't necessarily exhibit dependence in their extremes . A more specialized metric, the *extremal coefficient* θ, provides a measure of the extremal dependence between any number of variables. For K random variables, θ takes values in $[1, K]$. This has an intuitive interpretation as the effective number of independent variables present, i.e., $\theta = 1$ implies that effectively only one independent variable is present or that the variables are totally dependent—in the extremes. We can construct the extremal correlation, an analogue to the Pearson correlation, as $\chi = (K - \theta)/(K - 1) \in [0, 1]$. In an unpublished but much-cited manuscript, **?** derived an estimator for the extremal coefficient given N samples and $K = 2$ Fréchet-distributed random variables $Y_1, Y_2 \sim \mathcal{F}$,~~

$$\hat{\theta}_{12} = \frac{N}{\sum_{n=1}^{N} \min(y_{n1}^{-1}, y_{n2}^{-1})}$$

which the author calls the "raw estimates" of the extremal coefficient, with a natural extension into K greater than 2. The estimator $\hat{\theta}$ can take values in $[1, \infty)$ To identify independent hazard events, a declustering algorithm $d_{|t} : \mathbb{R}^T \to \mathbb{R}^T$ (**??**) is applied to the time series $r_{|ijk}(\mathbf{x})$. In this framework, hazard days are defined as consecutive days in which $r_{|ijk}(\mathbf{x})$ exceeds some specified threshold v_r, separated by a specified minimum number of non-exceedences days ℓ_r. The choice of v_r and $\hat{\theta} > K$ corresponds to negative extremal dependence. The Smith estimator is quite sensitive; however, and can produce values outside the theoretical bounds $[1, K]$, particularly when applied to finite samples or samples that deviate from the Fréchet assumptionℓ_r will depend on the data being modelled and the context. Since we need to fit parametric models to the margins, the data should be independent along the sample dimension, which favours using high thresholds v_r and extracting fewer storms. However, small sample sizes in the tails will lead to high variance in the parametric fits, so this trade-off must be handled. The simplest solution is to perform a standard grid search over the space of (v_r, ℓ_r) to select the largest number of events while maintaining independence between the extracted $r_{|ijk}(\mathbf{x})$ values. Independence can be verified using a standard Ljung–Box test (**?**).

Finally, a hazard event set is created by extracting the deseasonalised data corresponding to the declustered hazard days. This creates an event set of smaller spatiotemporal data cubes corresponding to each hazard event and variable. Each data cube will have dimensions $H \times W \times T$, re-using the T notation to represent the duration of an arbitrary event. A feature of this approach to event identification is that the extracted variables are sampled conditionally on the occurrence of events, as defined by the severity function. This is intentional: we seek to model the joint behaviour of all variables *during hazard events* rather than the independent natural extremes of each variable. However, the implications of this for fitting statistical models should be considered and will be discussed later.

A second approach uses the *tail dependence coefficient* λ (**?**), which directly measures the conditional probability of joint extreme events. For two variables with marginal distributions F_X and F_Y, the upper tail dependence coefficient is defined as

$$\lambda_u = \lim_{u \to 1^-} P(F_Y(Y) > u \mid F_X(X) > u)$$

where $\lambda_u \in [0, 1]$. This coefficient has a natural interpretation: $\lambda_u = 0.3$ means that where one variable exceeds its 90th percentile, there is a 30% probability the other variable also exceeds its 90th percentile. Unlike extremal coefficients, tail dependence coefficients are naturally bounded, making them more robust to practical applications (**?**)

Creating multi-hazard footprints

Each spatiotemporal hazard event can be projected into a 2-d footprint using a temporal aggregation function $h_{k|t} : \mathbb{R}^{H \times W \times T} \to \mathbb{R}^{H \times W}$. The specific definition of $h_{k|t}$ will depend on the hazard of interest and how its impact materialises over the event. A measure of cumulative precipitation, for example, may be more relevant for assessing flood risk, while the maximum wind speed may

be more relevant for assessing storm damages. Applying a temporal aggregation function to hazard variable creates a set of 2-d event footprints, which can be stacked together to create multi-hazard event footprints with dimensions $H \times W \times K$.

2.2 ~~Generative adversarial networks~~

~~Generative adversarial networks (GANs) were introduced by **?** and a large literature of theoretical advancements, extensions, and modifications has been growing ever since (**?**). The original GAN places two neural networks in competing roles. A data generator—which never sees the training data—produces samples, and the critic—which sees a jumble of real and generated data—guesses which are real and fake. The networks are trained in alternating fashion. In this way, the generator learns to reproduce the distribution of the training data and generate datasets indistinguishable from the training data in appearance and distribution . Classic GANs require tens of thousands of training samples to prevent generator overfitting and stabilize training (**??**). Even then, the training dynamic is extremely sensitive to hyperparameter choice and relies on neither model overpowering the other (**?**). Numerous extensions, such as the introduction of Wasserstein loss functions and gradient penalties have improved this situation (**??**), but ultimately the attention of the scientific community turned towards newer, more stable generative models such as diffusion models (**?**). However, almost a decade of scientific research has resulted in a diverse and well-documented set of GANs. The StyleGAN series of GANs from nVIDIA remains the state of the art. These models have been extensively tuned to require significantly less data (as little as 1, 000 samples) , and further modifications to them have reduced this number to only 100 samples (**??**).~~

2.2 ~~Event footprints~~

~~Event footprints are a useful tool in catastrophe modelling, capturing the maximum intensity of a hazard event over its lifetime for each point in space (**?**). These footprints are created by aggregating hazard-related variables over time at each location. The method for defining peak hazard varies by climate variable—some require accumulation over time, while others use statistical measures such as maximum or minimum values~~

2.2 Transforming the marginal distributions

Figure **??** shows the marginal transformation workflow which takes the set of multi-hazard footprints as input. The marginal distributions of the grid cells are defined along the sample dimension, so there are $H \times W \times K$ marginal distributions. To train a deep learning model we need to standardise these marginals and to properly extrapolate to new extremes, we need to fit an appropriate distribution to their tails.

3 ~~Method~~

~~This section outlines the general methodology of hazGAN, which can be divided into three phases: (i) event footprint extraction, (ii) transformation of the marginal distributions, and (iii) GAN training and inference. The method is flexible and can be applied to different gridded datasets, climate variables, and locations.~~

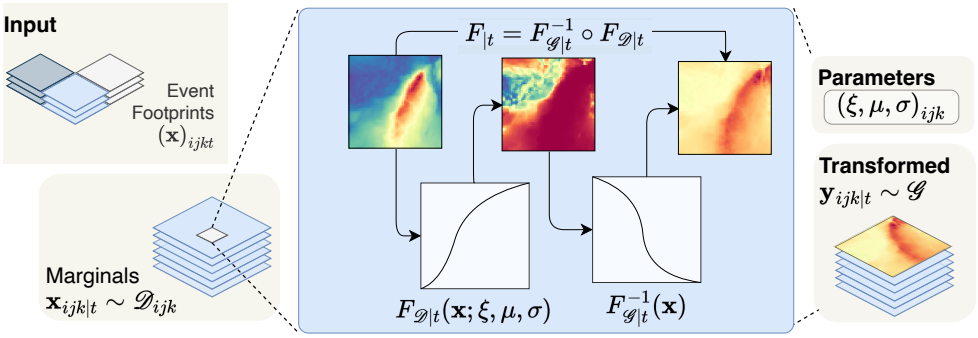


Figure 2. Schematic of the workflow to transform the marginal distributions of the event footprints extracted in Fig. **??** to a standard distribution $\mathcal{G}$. A suitable parametric distribution is fitted to the extremes of each weather variable along the sample dimension for each variable and location (marginal) $\mathbf{x}_{ijk|t} \sim \mathcal{D}_{ijk}$. The semiparametric distribution Eq. (**??**) transforms each marginal to a standard uniform distribution, using the fitted parameters. The quantile function for another distribution $\mathcal{G}$ then transforms each uniformly distributed marginal such that $\mathbf{y} \sim \mathcal{G}$.

~~Only four inputs need to be changed for new studies: (i) a region of interest must be defined by its bounding box; (ii)data cubes with dimensions $H \times W \times T$ must be supplied for each climate variable; (ii) a temporal aggregation function $h_{k|t}(\mathbf{x})$ along the time dimension (t) must be defined for each climate variable; and (iv) $r_{|ijk}(\mathbf{x})$ a hazard definition function or *risk functional* must be defined across all space and variables (ijk).~~

Marginal extreme value fitting

To generate realistic hazard events, we need to learn the multivariate distribution of the multi-hazard footprints. Critically, we need to learn the marginal and joint distributions of the most extreme values. As **?** showed, the ability of a GAN to learn the extremal dependence structure of a dataset can be improved by training it on the empirical distribution functions of its margins. This approach is analogous to classical methods in multivariate extremes, which often disentangle the margins of a multivariate distribution from its dependence structure (**?**). However, **?**'s approach relies on fitting generalised extreme value (GEV) distributions to the margins, requiring data in the form of annual block maxima. The use of block maxima creates spatial incoherence, entails significant loss of information, and is unsuitable for modelling the cumulative impact of hazards, such as storms, which materialise over the duration of the event, not just at the peak.

~~Throughout this paper, latitude will be indexed by $i = 1, \ldots, H$; longitude by $j = 1 \ldots W$; time by $t = 1, \ldots T$; sample number by $n = 1, \ldots, N$; and climate variables by $k = 1, \ldots, 3$. The notation $_{|ijkt}$ will be used to indicate which dimensions of a tensor a function is applied over. Data cubes with their original marginal distributions~~ To overcome the limitations of the block maxima approach, we opt for a peaks-over-threshold (POT) approach, allowing us to make more efficient use of data. **?**, Eq. 1.3 used a semiparametric function, which allowed them to model the entire distribution of the data, using a parametric distribution for the extreme values to guide extrapolation to new extremes, and an empirical distribution for the non-extremes

where there is already sufficient data to provide a good approximation. Considering a random variable X and suitably extreme threshold v_X, the semiparametric distribution function can be written in its most general form as:

$$\tilde{F}(x) = \begin{cases} 1 - (1 - \hat{F}(v_X))(1 - F_{\mathcal{D}}(x)) & \text{for } x > v_X \\ \hat{F}(x) & \text{for } x \leq v_X \end{cases} \tag{1}$$

where $\hat{F}$ is the empirical distribution function (ECDF) and $\mathcal{D}$ is an extreme value distribution used to model the tails. For most applications, $\mathcal{D}$ will be a generalised Pareto distribution (GPD), as this is the only nondegenerate limiting distribution for exceedances over a high threshold (**?**). In this case, the shape, location, and scale parameters of $\mathcal{D}$ are denoted by ~~$\mathbf{x}$, with uniform marginals by $\mathbf{u}$, and with marginals transformed to any other distribution $\mathbf{y}$.~~ ξ, μ, and σ, respectively. However, we have kept the formulation general to allow for alternative parametric distributions where appropriate (see, for example **?**). A semiparametric distribution as in Eq. (**??**) is fitted to all the margins of the multi-hazard event footprints $\mathbf{x}$, transforming them to have standard uniform margins $\mathbf{u} \sim \mathcal{U}(0, 1)$.

2.1 ~~Event footprints~~

~~The first step extracts multivariate hazard footprints from the climate data cubes. Figure **??** gives a schematic of the event footprint extraction stage. For each climate field, a spatiotemporal data cube for the region of interest must be supplied. If observations are sub-daily, $h_{k|t}(\mathbf{x})$ is applied along the time dimension to resample daily observations. Next, a deseasonalization function $s_{|t}(\mathbf{x})$ is applied along the time dimension to remove seasonality from each variable.~~

~~The deseasonalised data cubes are transformed into a time series using a *risk functional* $r_{|ijk}(\mathbf{x})$ (**?**). This function determines what types of hazards are extracted, e.g. calculating the mean or sum of a variable will prioritize widespread events while a maximum or minimum will identify more localized events that reach higher peak intensities. **?** give an in-depth discussion on the effects of different $r_{|ijk}$ choices and the choice of risk functional in our application is discussed further in Sect. **??**.~~

~~A declustering algorithm $d_{|t}(\cdot)$ is applied to the $r_{|ijk}(\mathbf{x})$ time series to identify hazard events. Hazard days are consecutive days in which $r_{|ijk}(\mathbf{x})$ exceeds some threshold v_r, separated by some number of days of non-exceedences ℓ_r (**?**). Finally, the deseasonalised data cubes are intersected with~~ The event identification approach means that, technically, the distribution of all but one of the margins will be a conditional distribution, conditional on $r_{|ijk}(\mathbf{x})$ having exceeding the threshold v_r. While in theory conditioning does not violate the assumptions of a GPD fit, the ~~hazard days to create a hazard event set, which is then collapsed along the time dimension using $h_{|t}(\mathbf{x})$ to generate a set of event footprints.~~ question of whether the conditioned data will be independent and identically distributed is more challenging to address. Although the deseasonalisation and declustering should ensure stationarity and independence of samples, it is still possible that the conditioning will select events arising from different meteorological mechanisms. The POT approach should mitigate this somewhat by isolating the most extreme events, which are more likely to arise from a single dominant mechanism. In Section **??** we will use an automated threshold selection method as a further safeguard, which will fail and revert to empirical distributions for any margins where no suitable GPD

threshold can be identified. The validity of this approach will depend on the weather variables being modelled and need to be assesed on a case-by-case basis.

~~Schematic of the workflow to extract hazard event footprints. Gridded hourly climate data over a region of interest for three variables (indexed $k = 1, 2, 3$) is resampled to daily aggregates according to resampling function $h_{k|t}(\mathbf{x}_k)$. The daily aggregates are deseasonalised according to $s_{|t}(\mathbf{x})$. A risk functional $r_{|ijk} : \mathbb{R}^{T \times H \times W \times K} \to \mathbb{R}^T$ constructs a time series from the deseasonalised data and a declustering algorithm $d_{|t}\big(r_{|ijk}(\mathbf{x})\big)$ identifies storm days. Data cubes of daily aggregates are extracted for each storm using the storm days and these are aggregated into storm footprints by again applying the $h_{k|t}$ functions along the time dimension.~~

Distribution of the training data margins

From a machine learning perspective, transforming the margins to uniform distribution is a natural standardisation step. However, the uniform probability transform means that the extremes occupy a small region at the edge of the domain. Furthermore, work by **?** and **?** demonstrated that a GAN with a light-tailed latent space cannot learn the tail behaviour of a heavy-tailed distribution, leading to underestimation of the tails. **?** developed a GAN with a heavy-tailed latent space to address this issue, but this required replacing the adversarial loss function with a custom loss function.

~~In order to fit parametric models to the marginals,~~ Drawing on these results, we hypothesise that what matters is not the ~~footprint data must be independent along the time dimension, which favours using high thresholds and extracting fewer storms. However, this will not facilitate fitting a distribution to the extremes of each marginal, especially if storms occur in different regions at different times. In other words, one set of days may have lots of extreme values for one location but not capture any of the extremes in another location. Additionally, having more datais important for deep learning. To address this, we can reduce the threshold on $r_{|ijk}\mathbf{x})$ at which storms are defined and extract more storms. In this work, the choice of threshold is optimized using a grid search to identify the largest number of storms possible, while maintaining independence between clusters. To assert storm independence, we only require that the peak $r_{|ijk}(\mathbf{x})$ values are independent between each storm and we use a Ljung–Box test to check this (?)~~specific tail behaviour of the latent space, but rather that it matches that of the data. A sufficient strategy would therefore be to transform heavy-tailed data to a light-tailed distribution before training, and invert the transformation afterwards. This mirrors common practice in the multivariate extremes literature, where margins are transformed to standard distributions such as the Laplace, Gumbel, Fréchet, or unit Pareto to exploit their desirable properties (**???**). We verify this hypothesis by repeating the experiments of **?** in the Supplementary Material. In Sect. **??**, we will explore results of transforming the training data to a standardised distribution $\mathcal{G}$ and compare the model performance for when $\mathcal{G}$ is a uniform distribution or another light-tailed distribution.

2.1 ~~Marginal transforms~~Generative model training and sampling

Figure ~~?? shows the marginal transformation workflow which takes the set of event footprintsfrom Sect. ?? as input. The marginal distributions of the grid cells are defined along the time dimension, so there are $H \times W \times K$ marginal distributions to be fitted. For each marginal, the semiparametric distribution function in Eq. (??) is used to transform it into a uniform~~

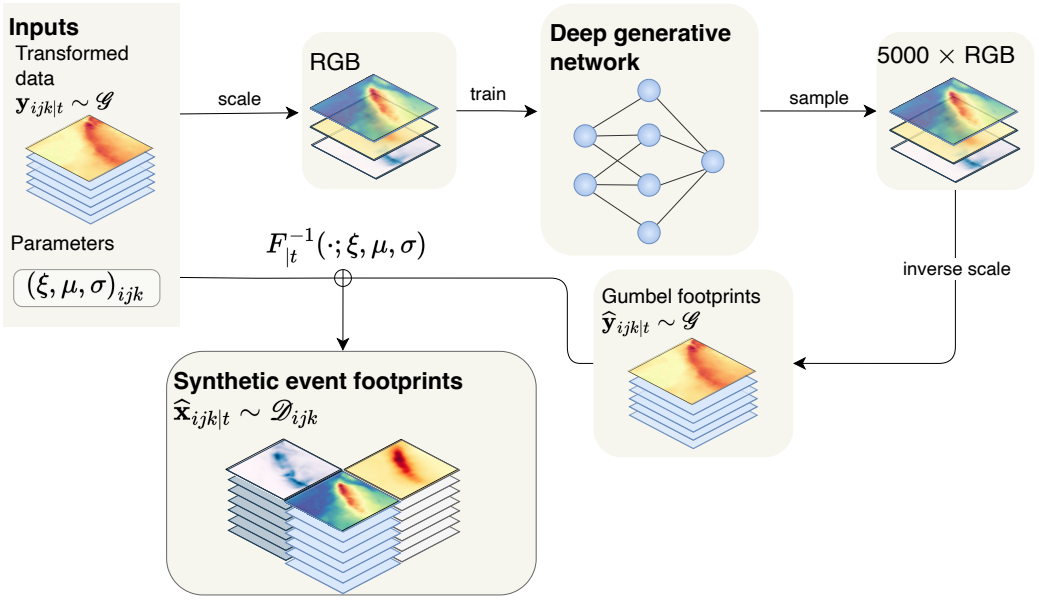


Figure 3. Schematic of the workflow ~~to transform~~ for training the ~~marginal distributions of the storm~~ deep generative model. The transformed hazard footprints from Fig. ~~??~~ **??** are rescaled to ~~a standard Gumbel distribution. A suitable parametric distribution is fitted to~~ take values in the ~~extremes of each climate variable along the sample~~ range $(0, 1)$ using a return period-based scaling method (~~time~~Eq. **??**) ~~dimension for each variable~~ and ~~location (marginal) $x_{ijk|t} \sim \mathcal{D}_{ijk}$~~converted to three-channel RGB images. ~~The semiparametric distribution Eq~~A deep generative network is trained on these images. ~~(??) transforms each marginal to a standard uniform distribution~~To create synthetic storm footprints, new samples are generated from the generative model and these are re-scaled using the ~~fitted parameters~~inverse of Eq.~~The Gumbel quantile function then transforms each uniformly distributed marginal to a standard Gumbel distribution~~ (**??**). The ~~output set~~ inverse of the probability integral transform (Eq. **??**) converts the synthetic footprints ~~$\mathbf{y}$ have a standard Gumbel distribution along~~ back to the ~~sample~~ scale of the original (~~time~~deseasonalised) ~~dimension~~data.

~~distribution. The parameters of each marginal are fitted using maximum likelihood estimation and tested for goodness-of-fit using an Anderson–Darling test. If significant p-values are obtained repeatedly, an empirical distribution functionis used as a fallback. Suitable thresholds are selected using the~~ **??** shows the workflow for training the deep generative model and sampling synthetic multi-hazard footprints. The transformed ~~p–values method of ?.~~ multi-hazard footprints from Fig. **??** are rescaled to take values in the range $(0, 1)$ using a return period-based scaling function and the three hazard variables converted to three-channel RGB images in preparation for GAN training. A deep generative network is trained on these images. To create synthetic storm footprints, new samples are generated from the generative model and these are re-scaled using the inverse of the scaling function. The inverse of the transform described in Fig. **??** is used to convert the synthetic footprints back to the scale of the original (deseasonalised) data. The result is a large multi-hazard event set which can be used in risk analysis applications.

~~At this point, all the marginals have the same standard uniform distribution . This makes them appropriate for traininga neural network, which requires all data to be on the same scale. The uniform distribution, however, places no particular emphasis on extremal values and since we are interested in the tails of the data, we want the network to be very sensitive to variation in high ($\geq$90th) quantiles. Transforming to a Gumbel distribution $\mathcal{G}$, using~~

$$\mathbf{y} = -\log(-\log(\mathbf{x}))$$

~~will stretch out high values of a random variable and encourage the network to pay more attention to data in this range . Thus, we transform all the variables to Gumbel distributions.~~

2.2 ~~Training and sampling~~

~~The footprints are now almost ready to be used for generative model training. However, the Gumbel distribution has an unbounded domain that lies mostly in $(-2, 13)$ and~~

Rescaling footprints be in (0, 1)

Depending on the distribution chosen for training, the transformed multi-hazard footprints ($\mathbf{y} \sim \mathcal{G}$) may require a final, additional rescaling step to ensure they lie in the range $(0, 1)$. The standard rescaling approach is min–max scaling, which maps the largest value of a dataset to 1 and the ~~footprints need to be rescaled to $[0, 1]$ before they are ready for training. This can be done using simple max-min scaling.~~ smallest value to 0. But this would prevent the GAN from generating values outside the range of the training data, which we want to be able to do. An alternative approach that allows us to specify a sensible maximum range for the dataset, without specifying a physical range subjectively, is to rescale using quantiles of the distribution we chose for training. We can specify quantiles that correspond to specific return periods, for example, for a distribution $\mathcal{G}$ and return period R, the return level is $y_R = F_{\mathcal{G}}^{-1}(1 - 1/R)$. Following this approach, the data can be rescaled according to

$$y' = \frac{y - F_{\mathcal{G}}^{-1}(1/R)}{F_{\mathcal{G}}^{-1}(1 - 1/R) - F_{\mathcal{G}}^{-1}(1/R)}, \tag{2}$$

assuming that R is sufficiently large that $F_{\mathcal{G}}^{-1}(1/R)$ is smaller than the smallest value in the training data and $F_{\mathcal{G}}^{-1}(1 - 1/R)$ is larger than the largest value in the training data. This approach allows us to specify a sensible maximum range for the data while still allowing the GAN to extrapolate beyond the maximum values in the training data. After rescaling, the ~~climate~~ weather fields are stacked such that each multivariate footprint is now a three-channel tensor. ~~Each footprint~~ Finally, it is converted to an RGB image, ready to feed into a generative model. ~~A generative model is trained on the images and used to generate thousands of synthetic RGB images with the same distribution. We invert the scaling on these to recover Gumbel-distributed marginal distributions, apply the Gumbel quantile function to obtain uniform samples, and apply the inverse of Eq. (??) to get synthetic dataon the original scale. Depending on the deseasonalization function $s_{|t}(\mathbf{x})$ used, the inverse of this may also be applied. The result at this point is a large multivariate hazard event set ready for risk analysis applications.~~

~~Schematic of the workflow for training the deep generative model. The Gumbel-distributed storm footprints from Fig. ?? are rescaled to take values in the range $(0, 1)$ using min-max scaling and are converted to three-channel RGB images. A deep generative network is trained on these images. To create synthetic storm footprints, thousands of new samples are generated from the generative model and these are re-scaled to have standard Gumbel marginal distributions using inverse min-max scaling. The inverse of the probability integral transform described in Fig. ?? is used to convert the synthetic footprints back to the scale of the original (deseasonalised) data.~~

3 ~~Application: Wind storms in the Bay of Bengal~~

~~This section presents an illustrative example of our hazard simulation method by generating a~~

Generative adversarial network

In theory, this framework is agnostic to the choice of deep generative model, meaning common models like variational auto-encoders (**?**), diffusion models (**?**), or flow-based models (**?**) could be used instead of GANs. In practice, however, historical weather datasets such as ERA5 generally contain fewer than 100 years of data, and for rare events, this does not provide enough extreme events to train standard deep generative models.

Some GANs have been developed to work well with small datasets, such as FastGAN (**?**) and StyleGAN2-ADA (**?**), which use a self-supervised discriminator and differentiable augmentation, respectively, to prevent overfitting. Differentiable augmentation regularises the discriminator by adding semantics-preserving augmentations (*e.g.* additive noise, rotations, isometric scaling, saturation changes) to all samples. When combined with differentiable augmentation, StyleGAN2-ADA successfully learned the distribution of a dataset of only 100 images from scratch (**??**). In Section **??**, we will use StyleGAN2-ADA with differentiable augmentationsto train a generative model on a set of multi-hazard ~~event set of wind speed, precipitation, and sea level pressure in the Bay of Bengal. The Bay of Bengal is the largest bay in the world and its coastline is home to some 500 million people. It has also been the site of 26 of the 35 deadliest tropical cyclones ever recorded. Compounding factors make the bay especially vulnerable: warm seas provide energy for powerful storms while the shallow, concave bay funnels storm surges~~ footprints.

2.1 Evaluation and benchmarking

The quality of generated event footprints will be evaluated against the training data according to several criteria, which can be broadly categorized into three groups: (i) the distribution of event severity; (ii) marginal distributions; and (iii) multivariate dependence structures.

Evaluation metrics

To measure the similarity between any two distributions, we will use the Wasserstein distance. The Wasserstein distance, also known as the Earth Mover's Distance, measures the minimum cost of transporting mass to transform one distribution into

another. For one-dimensional distributions, the Wasserstein distance can be computed from distribution P to distribution Q as

$$W(P,Q) = \int_{-\infty}^{\infty} |P(x) - Q(x)| \, \mathrm{d}x. \tag{3}$$

The Wasserstein distance is non-negative and takes a value of zero if and only if P and Q are the same distribution.

To assess whether the dependence structures are being learned correctly, we will first calculate dependence metrics between pairs of marginal distributions for each dataset. We will then compare the distributions of the generated dependence metrics to those of the training data. We will use Pearson correlation and mean-squared-error to assess the similarity between the calculated dependence metrics.

To measure the dependence between non-extreme values, we will use the Pearson correlation coefficient to measure linear agreement between the variables. To measure the level of extremal dependence between two variables $X_1 \sim F_1$ and $X_2 \sim F_2$, the extremal correlation between them, χ, above a fixed high threshold u can be written as **?**, p. 346,

$$\begin{aligned}\chi(u) &= \frac{\Pr(F_1(X_1) > u, F_2(X_2) > u)}{\Pr(F_1(X_1) > u)} \\ &= \frac{\Pr(F_1(X_1) > u, F_2(X_2) > u)}{1-u}.\end{aligned}$$

We will use the simple empirical estimator for $\chi(u)$, which can be calculated from a sample of size n as

$$\hat{\chi}(u) = \frac{\sum_{i=1}^{n} \mathbb{1}\{\hat{F}_1(X_{1i}) > u, \hat{F}_2(X_{2i}) > u\}}{\sum_{i=1}^{n} \mathbb{1}\{\hat{F}_1(X_{1i}) > u\}},$$

where $\hat{F}_i$ indicates the empirical distribution function of variable X. The true extremal correlation χ is defined as the asymptotic limit of $\chi(u)$ as $u \to 1$. The extremal correlation takes values in $[0,1]$, where $\chi = 0$ indicates asymptotic independence and $\chi > 0$ indicates asymptotic dependence. The higher the value of χ, the stronger the extremal dependence between the two variables. In practice, however, a finite, high threshold u can be used to approximate χ.

The χ metric tells us the strength of extremal dependence, and it is often supplemented by the $\bar{\chi}$ metric, which measures the strength of asymptotic independence—where the variables maintain some dependence at finite, high levels but are ultimately independent. To assess the strength of asymptotic independence, we can use the $\bar{\chi}$ metric, which is defined as **?**, p. 348,

$$\bar{\chi}(u) = \frac{2\log(1-u)}{\log \Pr(F_1(X_1) > u, F_2(X_2) > u)} - 1.$$

Benchmarking

To benchmark the performance of the method, we compare it to a widely used model for multivariate extremes: the **?** model. The model learns the conditional distribution of a set of variables, given that one of them exceeds a high threshold. For two

variables standardised to Gumbel-distributions Y_i and Y_j, the probability that Y_j exceeds a high threshold u given that Y_i exceeds u is given be **?**, Eq. 4.1:

$$Y_j = a(Y_i) + b(Y_i)Z, \quad Z \sim G_{|i}$$

where $G_{|i}$ is the residual distribution after normalising Y_j with the scalars $a(Y_i)$ and $b(Y_i)$. The form of $G_{|i}$ may vary, and depends on whether the margins of the residual distribution are asymptotically dependent. This naturally scales up to the coastline. During Cyclone Bhola in 1970, the deadliest storm in history, the storm surge was estimated at 10.4 m high, and half a million people were killed (**?**). Tropical cyclonesare multi-hazard events; here, we model three of the major drivers of tropical cyclone damages: strong winds, which damage human and natural structures and push waves and storm surge onshore; precipitation, which drives pluvial and fluvial flooding; and low sea-level pressure multivariate case, where we can model the distribution of n other variables $\mathbf{j} = 1, \ldots, n$, modelling $Y_\mathbf{j}$ conditional on Y_i exceeding u.

3 Application: Bay of Bengal storms

We apply the method to a case study of storms in the Bay of Bengal, a region highly exposed to tropical cyclones, ~~which is a major driver~~ which are not well characterised by existing event sets (**?**). We model three variables that determine the impact of tropical cyclones: wind speed, precipitation, and atmospheric pressure at sea-level, which determines the elevation of storm surges.

3.1 ~~Implementation details~~Data

3.1.1 ~~Footprint extraction~~

~~We use historical gridded climate~~ Our dataset consists of hourly gridded weather data from the ERA5 reanalysis product from 1940 to 2022 (**?**). ~~From this, we~~ We extract the northerly and easterly components of 10 m wind speeds (~~ms^{-1}~~ms^{-1}), total precipitation (m), and ~~sea level pressure~~ atmospheric pressure at sea level (Pa) over the region ~~defined by the bounding box in terms of latitude and longitude. We calculate overall~~ and calculate wind speed as the ℓ_2-norm of the northerly and easterly components of the 10 m wind speeds. ~~To aggregate along the temporal dimension, we define a function $h_{k|t}(\mathbf{x})$ for each variable $k = 1, 2, 3$. For wind speeds, $h_{1|t}(\mathbf{x}) = \max_{|t}(\mathbf{x}_1)$, for precipitation $h_{2|t}(\mathbf{x}) = \sum_{|t}(\mathbf{x}_2)$, and for sea level pressure $h_{3|t}(\mathbf{x}) = \min_{|t}(\mathbf{x}_3)$. As a simple deseasonalization~~

3.2 Event identification and footprint creation

Since deseasonalisation and risk estimation are not the central contributions of this work, we opt for simple methods to avoid unnecessary complexity. Seasonal effects are removed from the weather data using the deseasonalisation function $s_{|t}(\mathbf{x})$, ~~we calculate the~~ which computes monthly medians for each ~~field and subtract these from the data, creating~~ margin and subtracts them, yielding a time series of ~~anomalies. This simplistic deseasonalization method risks neglecting multiplicative seasonality or long-term trends but suffices for the current demonstrative purposes.~~

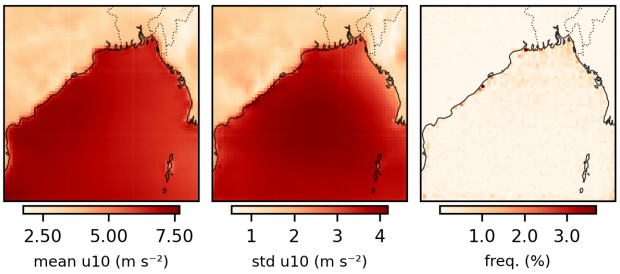


Figure 4. Map of grid cells that triggered storm events, based on being the site of the maximum wind speed over the domain on a given day.

We define a risk functional $r_{|ijk}(\mathbf{x})$ as the peak daily climatological anomalies. Storm events are identified using the daily maximum 10m wind speed. This choice prioritizes strong, localized wind storms , such as tropical cyclones, m wind speeds over the domain $r_{|ijk}(\mathbf{x}) = \max_{k=0|ij}(\mathbf{x})$ (where $k = 0$ indicates the wind speed variable). Taking the domain maximum prioritises strong, localised wind storms (including tropical cyclones) over more widespread, low-intensity storms. Since r is a function of the wind speed only, the extracted precipitation and pressure fields will be conditional on the occurrence of strong winds. If, however, a more nuanced storm definition was required, r could be defined as some function of all three variablesevents.

Figure **??** maps the mean and standard deviation of wind speeds alongside the frequency with which each grid cell contributed the domain maximum on a storm day. While mean winds and wind variance exhibit strong spatial patterns, storm-triggering pixels are distributed relatively uniformly across the domain, predominantly offshore. Over the 1941–2022 period, no single pixel triggered more than 16 separate events (1.2%). While the distribution of storm-triggering pixels is relatively even, some spatial variation in sampling frequency is also expected and acceptable: locations regularly exposed to strong winds should trigger storms more frequently, correctly reflecting the spatial distribution of wind hazard.

3.2.1 Probability integral transformations

To fit parametric distributions to the exceedences, we assume a Weibull distribution for wind speeds (see Sect. **??**) and Generalized Pareto (GPD) distributions for the create create 2-d multi-hazard footprints, we define a function $h_{k|t}(\mathbf{x})$ for each weather variable $k = 1, 2, 3$. For wind speeds, $h_{1|t}(\mathbf{x}) = \max_{|t}(\mathbf{x}_1)$, extracting the peak wind speed per pixel over the duration of each event; for precipitation and mean $h_{2|t}(\mathbf{x}) = \sum_{|t}(\mathbf{x}_2)$, extracting the cumulative precipitation over the events duration; and for sea level pressure . Figures **??**–**??** show the fitted parameters for each $h_{3|t}(\mathbf{x}) = \min_{|t}(\mathbf{x}_3)$, extracting the lowest sea-level pressure over the event's duration.

3.3 Marginal transformations

We fit the semiparametric distribution Eq. (??) to the margins of the wind speed, precipitation, and sea-level pressure ~~event sets over the Bay of Bengal. Suitable thresholds above which to extract the exceedences were selected using the modified p-values~~ footprints and in line with ?, we use a generalised Pareto distribution (GPD) as the parametric tail model. Threshold selection for the GPD for each of the 12,000 margins is done using the *ForwardStop* method of ?. ~~For windand precipitation, a threshold was found for all pixels,~~ The method uses the Anderson–Darling goodness-of-fit test to sequentially test the hypothesis of a GPD fit for a range of candidate thresholds, using a rejection rule that smoothes the p-values of the independent tests, controlling the likelihood of false discoveries.

Figures ????–???? show the fitted parameters for each of the wind, precipitation, and sea-level pressure event sets. For precipitation, the ? method successfully selected a threshold for all margins, while for wind speed and sea level pressure, ~~30 pixels had significant p-values for all thresholds and these were fitted with empirical distributionsinstead. In the figure, we can see strong spatial coherence across the domain~~ it failed to select thresholds for ~~each climate variable. Notably, all variables show higher variability near the coast, and wind and precipitation have higher thresholds offshore. The Weibull shape parameter for wind speeds is in the range of 0.9–1.2 across the domain. This is slightly lower than estimates in other global studies, which have been closer to 2 (??). This would suggest that the distribution of wind speeds follows a heavier-tailed distributionfor storms in the Bay of Bengal. In some grid cells, the parametric fit failed repeatedly, and in these locations, only an empirical distributionfunction was used to transform the data. These are shown as red~~54 and 35 margins, respectively, so these margins were fitted using entirely empirical distributions. These pixels are shown as red/white pixels in the ~~fitted parameter plots. For wind speed, precipitation, and sea-level pressure there were one, zero, and 30 failed fits, respectively.~~ fitted parameter plots.

3.3.1 ~~Generative modelling~~

Figure ?? shows that the GPD parameters vary smoothly across the domain for all three variables, with all variables exhibiting larger scales and variability near the northern coastline. It is possible that the fits could be made more robust by fitting a nonstationary model over the domain, allowing us to effectively 'borrow strength' between pixels (??, pp. 167; pp. 2). Additionally, the shape parameter estimates for all variables are within the range [-0.5, 0.5], which suggests that the marginal distributions are light-tailed, meaning they decay relatively quickly and are less prone to producing extremely large values. This could indicate that the true asymptotic form of the margins is Type I, in which case a direct Gumbel fit could be explored in place of the GPD model. Empirical studies have suggested that the parent distribution for wind speeds is a Weibull distribution, a distribution which indeed converges, albeit slowly, to the Type I (Gumbel) asymptotic form (?). Work by ? has proposed subasymptotic models for wind speeds, which apply a power transformation to raw wind variables, accelerating convergence to the Gumbel distribution. While beyond the scope of the current work, it would be interesting to explore alternative models for the extremes in the future.

~~Using the storm extraction methods~~ Applying the semiparametric distribution functions Eq. (??) with the parameters shown in Fig. ??, the transformed variables $\mathbf{u} = \tilde{F}(\mathbf{x})$ have a standard uniform distribution. But as discussed in Sect. ??, training a

GAN on data with uniform margins can lead to poor representation of tail behaviour. We apply an additional transformation to the data to transform it to a standard light-tailed distribution, using the quantile function of the target distribution. To investigate the effects of this transformation compared to the standard approach of training a GAN on rescaled data or **?**'s approach using uniform margins, we train separate GANs on data using (i) the orignal data, rescaled to $(0,1)$; (ii) uniform margins; (iii) Gaussian margins; and (iv) Gumbel margins. While the Gaussian and Gumbel distributions are both light-tailed distributions that belong to the same (Type I) domain of attraction, their subasymptotic behaviour is different, which we hypothesise may lead to different performance.

To rescale the training data to the interval $(0,1)$ for training the deep learning model, we use the return period-based rescaling method described in Sect. ~~**??**, we obtained 1249 event footprints from the historical data. Of these, 149 have a maximum wind anomaly exceeding 15 ms^{-1} and represent storms with very strong spatial coherence. We choose this wind speed as a threshold above which to define storms as extreme. We initially trained~~ **??**. We choose $R = 10,000$ years as the maximum return level any marginal value can reach. This was converted into a maximum value for rescaling for the uniform distribution $1 - 1/R$ and for the Gaussian and Gumbel distributions using $F^{-1}(1 - 1/R)$, where F_G^{-1} is the quantile function of the distribution. For the data that has not been transformed, the data is rescaled to $(0,1)$ by setting 1 to correspond to the maximum value multiplied by $\log(R)/\log(N)$, where N is the number of independent hazard events.

3.4 Generative modelling

The final training datasets consisted of 1,249 multi-hazard storm footprints. Although we initially attempted to train a Wasserstein GAN with gradient penalty (**??**) on ~~all 1249 storms; however,~~the 1,249-event dataset, we found that the GAN ~~was~~ became biased towards the more common, less ~~coherent, stormsand basically ignored the extreme storms. Various oversampling regimes were attempted to address thisbut training would inevitably collapse once the proportion of resampled extreme storms became too high.~~

~~Instead~~spatially coherent storms. To address this, we used ~~a modification of the StyleGAN2 with differentiable sample augmentations (StyleGAN2-DA) (**??**). Differentiable augmentation adds semantics-preserving augmentations (e.g. additive noise, rotations, isometric scaling, saturation changes) to all samples before the discriminator sees them. These augmentations act as an effective regulariser of the discriminator, forcing it to become invariant to certain distortions and focus on the most essential data features. The augmentations must be differentiable to allow discriminator gradients to back-propagate all the way through the generator to the latent space and to guarantee Jensen–Shannon invariance. The StyleGAN2-DA model has been demonstrated to produce excellent results with~~ the GAN that has been specifically developed for small datasets, the StyleGAN2-ADA model, with additional differentiable augmentation (**??**), which can produce good results on as few as 100 training samples~~, making it a suitable candidate to train on only the 149 extreme storms~~.

We filtered the data to only include storms with a maximum wind speed anomaly $r_{ijk}(\mathbf{x})$ exceeding 15 ms^{-1}, resulting in a dataset of 150 storms. We trained the StyleGAN2-DA on the ~~149 extreme storms until it had seen~~ 150 most extreme storms for 2013 epochs (*i.e.*, so that it had sampled images 300~~,~~,000 ~~footprint images~~times). This took approximately four hours on an nVIDIA GeForce GTX 1080 Ti GPU. ~~We used the trained model to generate 914 multivariate footprints. Extreme storms~~

~~in the training data occurred at a rate of $\lambda = 1.82$ storms per year, so this resulted in~~ The GAN was then used to generate 500 years of synthetic hazard events, which for storms with a yearly rate of $\lambda = 1.82$, corresponds to 914 multi-hazard footprints.

3.5 Results

~~Figure ?? shows~~

Visual appearance

Ranking scores according to the severity function $r_{|ijk}(\mathbf{x})$, Fig. ?? compares the 16 ~~generated wind footprints for~~ most severe storm footprints derived from the ERA5 ~~training GAN-generated samples. These are presented as they were seen by the GAN in Gumbel space (??), in intermediate probability space (??) and in the actual data space (??). The corresponding~~ dataset with the most extreme footprints generated using the GAN for each of the four training configurations: rescaling-only, uniform margins, Gaussian margins, and Gumbel margins. Corresponding figures for sea-level pressure and precipitation fields are also ~~shown~~ provided in the Supplementary Information.

~~In Section ?? we outlined our criteria for success for the hazard generator. These were that the generated data: (i) replicates the overall distribution of extreme events , (ii) preserves the dependence structure of the training data across space—in the bulk and in the tails, (iii) preserves the dependence structure between different variables—in the bulk and in the tails, (iv) preserves the marginal distributions of the training data, and (v) has a well-justified basis for extrapolation to new, more extreme events. With (iv) and (v) satisfied by construction, we turn to statistical methods to assess whether criteria (i)–(iv) are satisfied.~~ The rescaled and Gumbel footprints look most similar to the ERA5 training data. Uniform-trained events are more extreme, overly-widespread, and do not exhibit the gradual decay in intensity with distance from the storm centre that would normally be expected. The footprints generated by standard rescaling also look reasonable, although the track shapes appear more simple and ellipsoidal than those produced by the Gumbel or Gaussian-trained models. The latter exhibit longer tracks and more pronounced changes of direction, better capturing the curved trajectories seen in the ERA5 reference footprints.

~~For validation, we only compare the generated dataset with the training set, and omit a validation set. We do this for two reasons: Firstly, given the image generator in a GAN never actually sees the training data , standard overfitting—where the model has excellent performance on training data but performs poorly on new data—is less of a concern than for other machine learning models. This is because if overfitting occurs in GANs, it typically happens in the discriminator. When the discriminator overfits during training its feedback to the generator becomes meaningless. Without useful feedback, the discriminator begins randomly guessing, leading to diverging gradients and training collapse (?). Because of this dynamic, an overfitted GAN usually performs terribly on both the training and validation data and so a validation set becomes less important. Secondly, given we have only $\mathcal{O}(100)$ samples, any validation set would necessarily contain fewer than 50 samples. Such a small validation set would have high variance, rendering comparison to it relatively meaningless. For these reasons, we consider it most appropriate to only compare descriptive statistics between the training set and the generated samples.~~

Marginal distribution fits

3.5.1 ~~Criterion (i): Event distribution~~

~~Figure ?? compares the distribution of the generated storms with the ERA5~~ To assess the GAN's ability to capture marginal distributions, we calculate the Wasserstein distance between the generated and training data for each margin using Eq. (??), scaling by the standard deviation of the training data distribution. The average rescaled Wasserstein distance across all margins is 0.57 for the rescaled model, 0.21 for the uniform model, 0.13 for the Gaussian model, and 0.14 for the Gumbel model, indicating that the Gaussian and Gumbel models are overall better capturing the marginal distributions of the training data.

~~Wind storms are categorized according to the peak wind speed during each storm. The GAN reproduces the shape of the overall distribution well, but shows some systemic bias towards producing lower intensity storms, predicting more storms in the 15–20 ms^{-1} band (3.28% versus 0.00%) and fewer in the 40–45 ms^{-1} band (0.22% versus 3.36%). On average,~~

To assess how well each model captures the overall distribution of pixel values, Figure ?? shows the ~~GAN underestimates the occurrence of storms in the 30–45 ms^{-1} bands by 4.457% and overestimates the occurrence of storms in the 15–30 ms^{-1} bands by 4.457%. The exact source of this bias remains unclear: it could result from the location or shape parameter of the Weibull parametric fits, however this is unlikely as the Anderson–Darling test returned few significant p-values—indicating good fits—and the low shape parameter estimates would indicate higher (rather than lower) wind speed estimates (see Fig. ??). Equally, the bias may be a result of the StyleGAN training. in future work, it would be interesting to investigate whether this can be mitigated with longer training times.~~ flattened distributions of all pixels, transformed to Gumbel scale to enable cross-comparison. The uniform-trained model shows a pronounced spike at the maximum value, suggesting the GAN attempted to extrapolate beyond its allowed range, saturating at the 10,000-year return period. The Gaussian-trained model exhibits the same behaviour, though far less severely. The Gumbel and rescaled models avoid this saturation: the Gumbel model produces the smoothest tail decay, while the rescaled model shows a more stepped distribution.

~~For additional context, we note that ERA5 has well-documented biases when representing tropical cyclones. In a previous study of TCs impacting Bangladesh, ERA5 peak wind speeds were observed to be approximately 20 ms^{-1} slower than their IBTrACS counterparts. This is largely accepted to be caused by the coarse horizontal grid spacing (approx. 30 km) of ERA5 simulations, which prohibits the resolution of peak winds near the eye of a cyclone—though ? notes that the sea-level pressure–wind speed relationship appears degraded in ERA5 compared to earlier products, suggestive of further issues with the model physics. Additionally, our model's risk functional extracts events based on their peak wind intensity, and the deseasonalization is based on the climatological median for each month. This doesn't account for the North Indian Ocean's bimodal tropical cyclone seasonality. A more specialized deseasonalization method or risk functional to detect tropical cyclones or only using training data from the pre-and post-monsoon cyclones, when TC activity is greatest may yield slightly more higher intensity cyclones. However, the fundamental issue is the training data.~~

Storm event distribution fits

These ERA5 limitations highlight a key constraint of our method: it is only as good as the data it is trained on. Any biases or errors in the input data will necessarily be propagated through the whole model. For this reason, it is important to be familiar with the strengths and limitations of the training data and to consider applying bias correction or downscaling methods to any input data To assess how well the GAN captures the storm intensity distribution, as defined by the domain-maximum wind speed, we calculate $r_{ijk}(\mathbf{x})$ for each storm in the training and generated data and compare their distributions in Figure **??**. The rescaled and uniform-trained models perform poorly here with high Wasserstein distances (3.7 and 16.8, respectively) compared to the Gaussian and Gumbel models (0.71 and 1.0, respectively). There may also be better datasets available for specific regions. In the North Indian Ocean, for example, there is the Indian Monsoon Data Assimilation and Analysis (IMDAA) product, which has a higher horizontal resolution of 0.12° (12 km) . We have also observed it to produce marginally higher winds during Tropical Cyclone Amphan than its ERA5 counterpart. However, since the goal of this work is to develop a flexible methodology applicable to any gridded climate data, we prefer to use a globally available dataset and consider bias correction to be beyond the scope of this paper. Given our model has successfully reproduced the distribution of different wind storm categories, we consider criteria (i)satisfied.

3.5.2 Criterion (ii): Spatial dependence

Spatial dependence structures

To check the model is learning the spatial dependence structures, we calculated Pearson correlation and estimate the tail dependence coefficient for wind speed coefficients $\hat{\chi}(u)$ (choosing $u = 0.8$ based on initial data exploration) between all pairs of pixels across the domainfor the training and generated data , generating for each variable a 4096×4096 matrix for the four generated datasets (Fig. **??**). The plots for both the Pearson correlation and the extremal correlation are almost identical between the training (a) and generated (b) sets. The same result was observed for precipitation and sea-level pressure fields. **??**). We quantify the level of agreement between the ERA5 and GAN-generated correlation structures by calculating the Pearson correlation and mean absolute error (MAE) between the two correlation fields. For wind, a correlation coefficient of 0.971 extremal correlation matrices. The rescaled model performs worst, with a correlation of 0.380 (MAE = 0.062) is achieved between the two Pearson correlation structures, showing excellent overall agreement. Results are similar for precipitation (r = 0.981, MAE = 0.054), 0.309) between the ERA5 and sea-level pressure (r = 0.980, MAE 0.044), with average spatial correlation across three variables of 0.977 GAN-generated correlation fields for wind speed, while the uniform, Gaussian, and Gumbel models perform much better, with correlations of 0.839 (MAE = 0.053). For wind tail dependence coefficient fields, the correlation between the ERA5 and GAN-generated fields is 0.968 0.088), 0.837 (MAE = 0.042). For precipitation tail dependence coefficient fields, the correlation between the ERA5 and GAN-generated fields is 0.906 0.089), and 0.857 (MAE = 0.083). For sea-level pressure tail dependence coefficients, the correlation between the ERA5 and GAN-generated fields is 0.948 (MAE = 0.046), respectively.

For a more detailed view of bivariate spatial relationships, Fig. **??** shows scatter plots of 10 m wind speed anomalies between Chittagong and Dhaka, two cities in Bangladesh. The spatial dependence structure of the training data is reproduced by the

~~GAN. In both cases~~rescaling-only model produces a noticeably simpler, more ellipsoidal dependence structure than is seen in the ERA5 data, ~~dependence between pairs of points is dominated by proximity: points along the diagonal exhibit higher correlations.~~while the uniform-trained model struggles to resolve marginal extremes, producing the artefactual clustering visible at the plot boundaries. Both the Gaussian and Gumbel-based models show considerably better agreement with the observed dependence structure, with the Gaussian model appearing to provide the closest match. Similar results were observed for all other variables and pairs of locations tested.

3.5.3 ~~Criterion (iii): Multivariate dependence~~

~~The final criterion to assess is that of dependence between the different climate fields. To assess this, we again calculate Pearson correlation and the tail dependence coefficient, but between wind and precipitation at each pixel in the domain for the training and generated data(Fig. ??). Again, it is clear that the model captures the dependence structure in the bulk of the data almost perfectly and captures the general pattern of the extremal dependence structure. The correlation between the ERA5 and~~

Multivariate dependence structures

Figure **??** maps the extremal correlation estimates $\hat{\chi}(0.8)$ between 10 m wind speed and total precipitation across the Bay of Bengal for training data and the four GAN-generated ~~correlation fields for wind and precipitation is 0.918 (MAE = 0.083) and the spatial correlation between the tail dependence fields is 0.692~~ datasets. The GANs trained on uniform, Gaussian and Gumbel margins show the best agreement with the ERA5 data, with Pearson correlations of 0.664 (MAE = ~~0.07). For wind speed and sea-level pressure this spatial correlation is 0.647~~ 0.104), 0.666 (MAE = ~~0.083) between the correlation fields and 0.866~~ 0.089) and 0.705 (MAE = ~~0.099)between the tail dependence fields. For precipitation and sea-level pressure the spatial correlation is 0.887~~ 0.08), respectively. The rescaling-only and uniform-trained model performs much worse, with correlation of 0.193 (MAE = ~~0.063)between the correlation fields and 0.875 (MAE = 0.120) between the tail dependence fields. Averaged between the three pairs of variables, the average spatial correlation between correlation fields for variable pairs is 0.817 (MAE = 0.053) and the average spatial correlation between tail dependence fields for variable pairs is 0.817 (MAE = 0.096)~~0.117). These scores are slightly lower than the spatial dependence scores, which may be due to the architecture of StyleGAN prioritising spatial relationships or because the inter-variable dependence structure is more complex and harder to learn than the spatial dependence structure.

Benchmarking against the ? model

~~The extremal dependence structure across the domain is noteworthy because high winds and heavy precipitation appear to show more extremal dependence over land than offshore; however, they show more overall dependence approaching the north-east coastline of the Bay.The increased dependence between wind and precipitation near the coast is reasonable; the strong correlation near the north-east coast corresponds to~~ We choose the Gaussian model as the best-performing model and compare its performance to the **?** conditional exceedance model, which is provided in R's *texmex* package (**?**). In this package,

the *mexDependence* function integrates the GPD fitting of Eq. (3.2) and the ~~prevailing south-westerly winds during the summer monsoon season from June to September. Warm, moisture-laden air blows across the bay from the south-west and~~ fitting of the dependence model **?**, Eq. 4.1 into a single function, but our data has already been transformed to a uniform scale using Eq. (**??**), so we customise the *mexDependence* function to skip the GPD-fitting step and directly accept data on the uniform scale.

Using the dataset of the top-150 storms, we randomly sample 1,~~as it interacts with the friction from the land, creates horizontal convergence, leading to air being pushed upwards. The air cools as it rises, leading to precipitation. The stronger the winds, the more horizontal convergence occurs, leading to increased dependence between wind and precipitation. Orographic lifting and land–sea temperature gradients can also occur at these windward slopes and increase dependence between wind speedand precipitation~~000 points across the domain and fit a conditional exceedance model between (i) different weather variables at that point, and (ii) for the same weather variable between that location and a second, randomly sampled, location. We then use the fitted model to generate 500 years of synthetic data, and calculate $\hat{\chi}(0.8)$ between the generated and training data. Figure **????** plots the distribution of spatial extremal correlations for each weather variable and **????** plots the inter-variable extremal correlations for all combinations of wind speed, precipitation, and sea-level pressure.

~~To examine how the model has learned the pairwise relationships between different points in space, Fig. ?? shows scatter plots of the climate variables at two pairs of points: Chittagong and Dhaka—two cities in Bangladesh, and two RAMA (Research Moored Array for African-Asian-Australian Monsoon Analysis and Prediction Atlas) buoys. When comparing the 149-sample training set with 914 generated samples (500 years of samples), it is clear that the~~ Both models perform less in well in modelling the inter-variable dependence structure than the spatial structure, suggesting that the inter-variable dependence structure is indeed a more challenging joint distribution to learn. Nonetheless, the GAN achieved higher correlations and lower MAE compared to the training data than the **?** model for both spatial and inter-variable dependence. Overall, this result is encouraging, particularly because the hazGAN model learns the full joint distribution of all variables and locations simultaneously, whereas the **?** model is learning ~~the bivariate density distribution between each pair of variables, and (in most cases) extrapolating accordingly~~pairwise relationships separately so does not capture higher-order dependencies. As a result, the GAN may be able to borrow strength across the full dataset, enabling it to learn a more accurate dependence structure, while the **?** model fits many separate models, each with less data. In the future, it would be interesting to compare the GAN to spatial models for extremes, such as r-Pareto processes (**?**), which are designed to model the full spatial dependence structure of extremes. However, we consider this beyond the scope of the current work, which aims only to demonstrate the effectiveness of the hazGAN approach.

4 Use case: Spatial risk assessment for mangrove forests in the Bay of Bengal

~~Here~~ To illustrate an application in risk analysis, we use ~~our~~ the event set generated by the Gaussian-based model to estimate risk to mangrove forests from storms ~~hitting~~ impacting the Bay of Bengal. To model storm risk to mangrove forests, we need to model the joint distribution of both wind and precipitation and to estimate aggregate impacts over a domain as large as the Bay

of Bengal, we need a model that generates spatial events with the right dependence structure. Thus, this problem necessitates a new approach to event set generation that can account for spatial and multivariate dependencies, making it a suitable test case for our model. ~~The rest of this section describes the modelling approach and the results.~~

4.1 Implementation details

To estimate damages to the mangroves from the multivariate storm footprints we use a bivariate logistic regression model trained on global historical mangrove damages and tropical cyclone characteristics (**??**). The model predicts the probability that a mangrove patch is damaged, conditional on local winds and precipitation. A patch is defined as "damaged" if it experiences a drop in enhanced vegetation index (EVI) exceeding 20% in the aftermath of a storm. Further details of the ~~method mangrove vulnerability model~~ mangrove fragility function and relevant calculations are provided in the Supplementary Information. We use the 500 years of wind and precipitation footprints generated in Sect. ~~-~~ **??** as input to the mangrove ~~vulnerability~~ fragility model.

To illustrate the implications of ignoring the spatial dependence structure of climate hazards, two more synthetic datasets are constructed: a dataset that ignores all dependence across the region (independence assumption), and a dataset that assumes total dependence across the region (total dependence assumption). The total dependence assumption is the implicit assumption when return period hazard maps are treated ~~as~~ like true events, while the independence assumption is ~~the assumption~~ implicit when regional risks are modelled separately ~~(?).~~ and the results aggregated (**?**). These assumptions will lead to somewhat exaggerated results: modelling the dependence structure of hazards across space completely independently or dependently is an extreme assumption; however, they illustrate the critical importance of modelling the dependence structure of hazards across space when estimating risk, and the significant potential for bias when this is ignored.

4.2 Results

Figure **??** shows the risk profile for widespread mangrove damages over the Bay of Bengal, plotting ~~expected total damage area against~~ the expected area of mangrove forest damaged against event return period. ~~Return~~ The return period is calculated as a function of the total mangrove ~~damages~~ damage area (see the Supplementary Information for calculation details). The figure also displays risk profiles for hazard events generated under independence and total dependence assumptions, which clearly introduce significant bias even at small return periods.

Applied to the ERA5 data, the logistic model predicts ~~2432.6~~ 2451.21 km^2 (25%) of the 9917 km^2 mangrove forest in the region to be damaged by a five-year storm event and ~~2949.80~~ 2967.11 km^2 (30%) to be damaged by a 100-year storm event. A five-year storm generated by the GAN produces damages of ~~2335.23~~ 2309.12 km^2 (~~24~~23%) and a 100-year storm damages ~~2799.74~~ 2891.09 km^2 (~~28~~29%). A 100-year storm under the total dependence assumption predicts damage to ~~42~~43% of the mangrove forest in the Bay of Bengal, significantly overestimating the risk. For a 500-year event, the GAN-generated data predicts damage to ~~31~~33% of mangrove forest in the region (~~3026.77~~ 3301.93 km^2); the dependence assumption-generated data predicts damage to ~~47~~48% of mangrove forests (~~4643.00~~ 4785.14 km^2); and the independence assumption-generated data predicts damage to only ~~19~~16% of the mangrove forests (~~1903.82~~1660.71km^2).

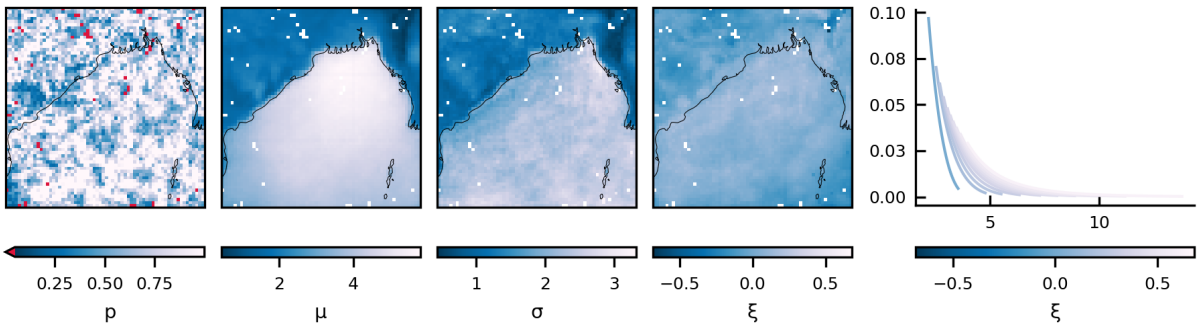
Table 1. Expected damage area and return period deviations for GAN, independent, and dependent generated samples.

Method	Deviation	~~5-year~~ 5 yr	~~10-year~~ 10 yr	~~25-year~~ 25 yr	~~50-year~~ 50 yr	~~100~~
HazGAN	Expected damage area (km^2)	~~-97.37~~ -147.40	~~-160.35~~ -172.92	~~-126.58~~ -154.79	~~-95.87~~ -8.85	~~-~~
	~~(Return period ;~~ Return period (years)	~~(-0.07)~~ -0.11	~~(-0.24)~~ -0.11	~~(-2.31)~~ -2.48	~~(8.97)~~ 8.64	~~(~~
Independent	Expected damage area (km^2)	~~-874.38~~ -1066.51	~~-1042.82~~ -1218.01	~~-1118.72~~ -1277.82	~~-1147.79~~ -1289.83	~~-1219~~
	~~(Return period ;~~ Return period (years)	~~(~~-0.10 ~~)~~	~~(-0.21)~~ -0.19	~~(-2.72)~~ -2.12	~~(8.15)~~ 9.34	~~(~~
Dependent	Expected damage area (km^2)	~~702.60~~ 722.18	~~800.72~~ 832.63	~~958.65~~ 1005.84	~~1153.64~~ 1217.94	~~126~~
	~~(Return period ;~~ Return period (years)	~~(~~-0.45 ~~)~~	~~(~~-0.19 ~~)~~	~~(~~-5.64 ~~)~~	~~(5.64)~~ 5.65	

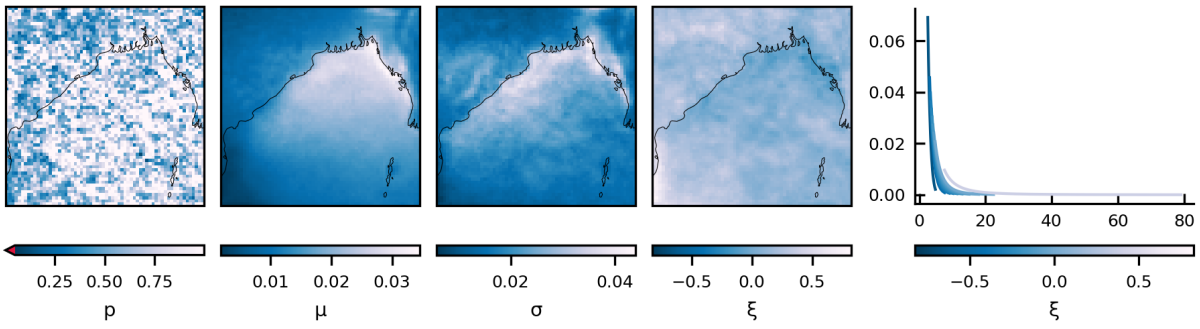
Table **??** reveals that while the GAN-generated dataset ~~systematically underestimates~~ appears to underestimate total expected damages with a mean absolute error of ~~93.57~~ 54 km^2 across all return periods (reaching up to ~~160~~ 173 km^2 for 10-year events), this bias remains an order of magnitude smaller than the independent dataset's systematic underestimation (mean absolute error of ~~460.45~~ 484 km^2) and the dependent assumption dataset's systematic overestimation (mean absolute error of ~~1056.90~~ 1062 km^2). ~~These results reinforce findings from ? and ?, highlighting the critical importance of explicitly modelling spatial dependence of hazards in risk analysis, as treating return period maps as real events leads to massively inflated total damage estimates.~~

To visualise the qualitative difference between modelling the dependence structure and assuming total or no dependence, Fig. **??** shows, for the ERA5 and synthetic datasets, a sample corresponding to a ~~1-in-20~~ 1-in-75 year return period.

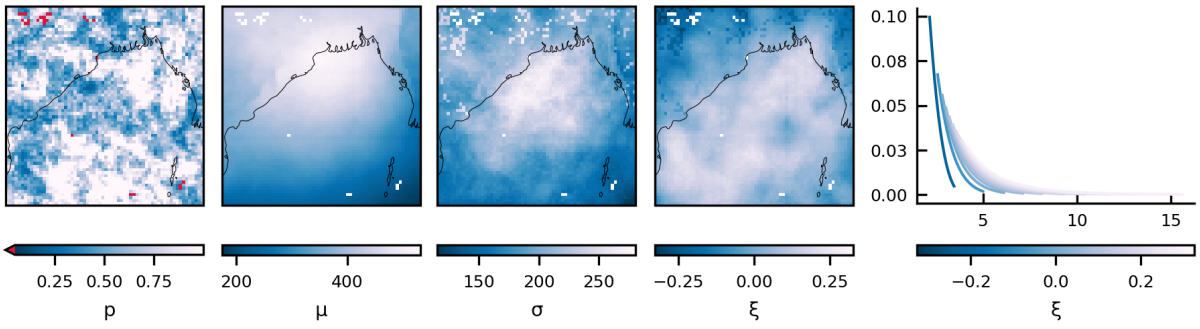
Realistic events have clustering in the extremes while the dependent assumption (hazard maps) distributes extreme winds evenly across the entire region and results in far more widespread and unrealistic disruption estimates. The independence assumption shows no spatial coherence and so underestimates the impacts of spatially coherent events.



(a) Wind speeds at 10 m



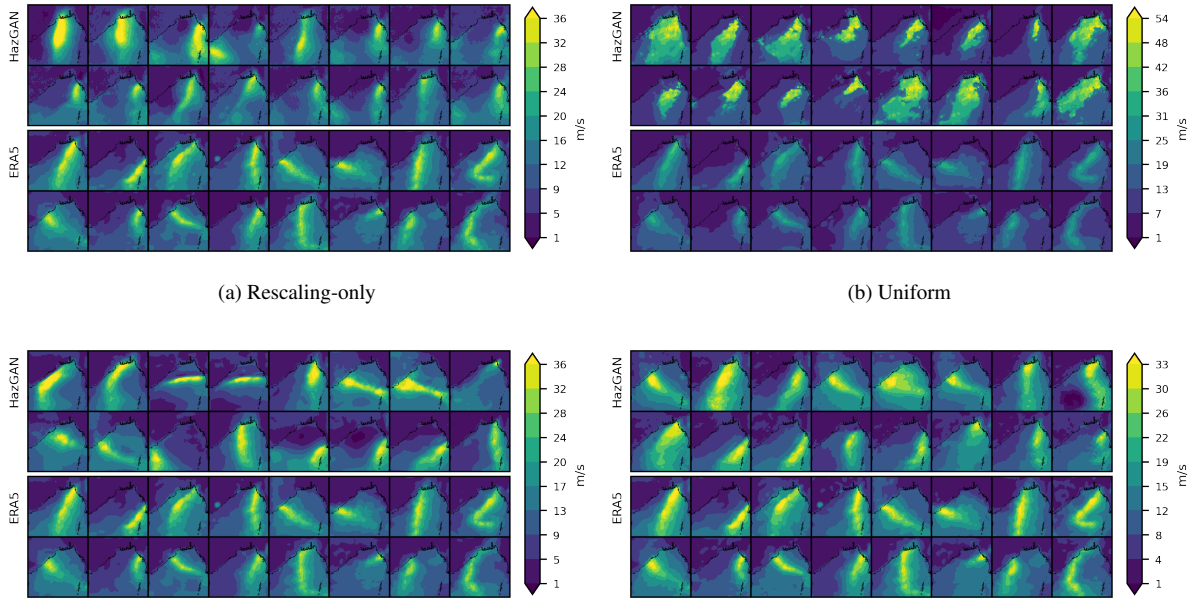
(b) Total precipitation



(c) Sea-level pressure

~~**Left panels** Fitted parameters for the marginal distributions of climate variables over the Bay of Bengal during storms, showing adjusted p-values for Anderson–Darling (AD) goodness-of-fit tests (left column), thresholds μ, scale parameters σ, shape parameters ξ. White pixels indicate locations where the AD test failed and empirical distributions were used instead. **Right column** Density plots of each distribution for a range of tail shapes—with (μ,σ) fixed at (0,1). Shown are: (a) Weibull parameters for peak wind speeds, (b) Generalised Pareto parameters for total precipitation, (c) Generalised Pareto parameters for low pressure.~~

Figure 5. **Left panels** Fitted parameters for the marginal distributions of weather variables over the Bay of Bengal during storms, showing adjusted p-values for Anderson–Darling (AD) goodness-of-fit tests (left column), thresholds μ, scale parameters σ, shape parameters ξ. White pixels indicate locations where the AD test failed and empirical distributions were used instead. **Right column** Density plots of each distribution for a range of tail shapes—with (μ,σ) fixed at (0,1). Shown are: generalised Pareto parameters for (a) peak wind speeds, (b) total precipitation, and (c) low pressure.



(a) Rescaling-only

(b) Uniform

(c) Gaussian

(d) Gumbel

Figure 6. Comparison of wind speed footprints during storms in the Bay of Bengal for ERA5 training data vs. GAN-generated samples. Shown ~~in~~ for a GAN trained on (a) ~~Gumbel space~~margins rescaled to [0,1] in the usual way, (b) margins transformed to uniform~~space~~, ~~and~~ (c) ~~the original data space. The GAN is trained on samples in Gumbel space~~ margins transformed to ~~emphasise variation in the extremes. The probability integral transform is used to transform the marginals of samples into uniform percentiles~~Gaussian, ~~which provide a measure of the extremeness of the wind speed at each point. The fitted parameters from Fig. ?? are used to transform uniform samples back to the scale of the climate variable~~ and (~~anomalies from seasonal median~~d) ~~using the inverse of Eq~~margins transformed to Gumbel. ~~??~~

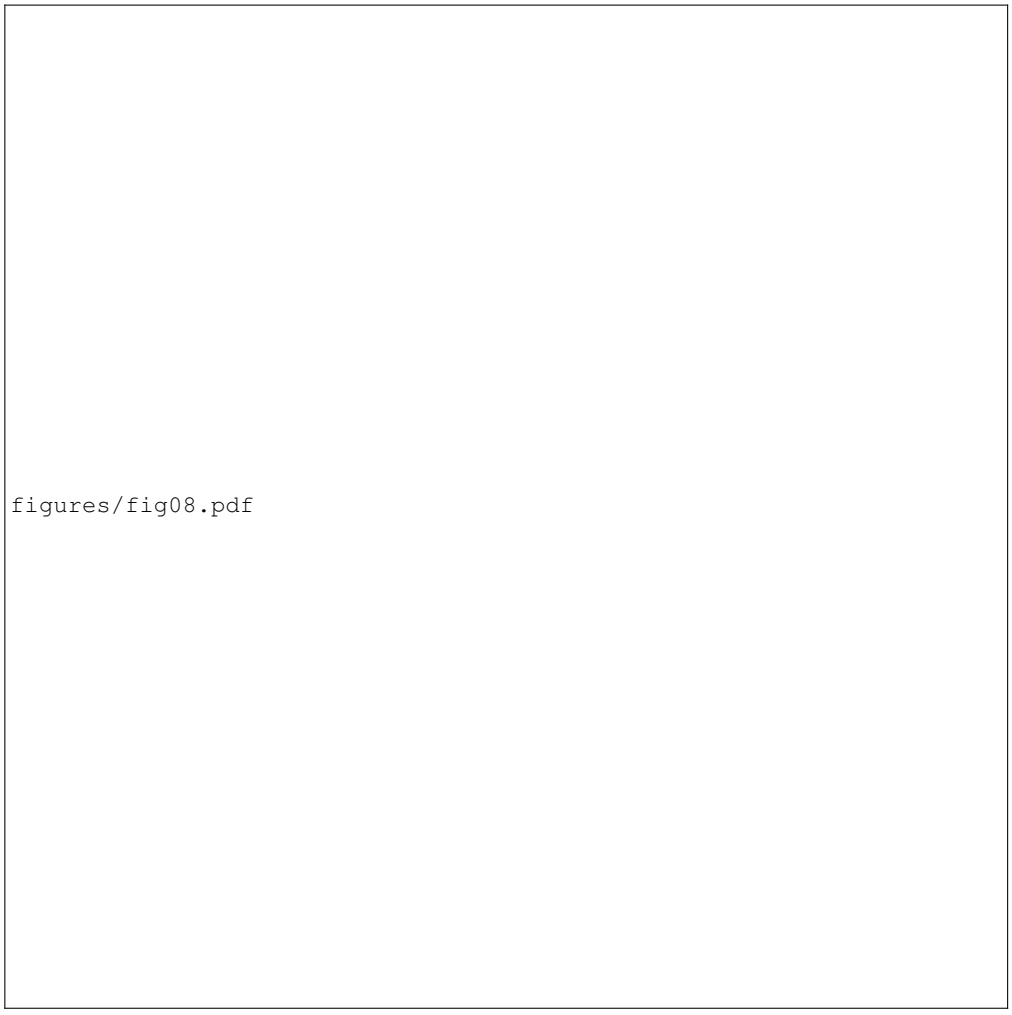
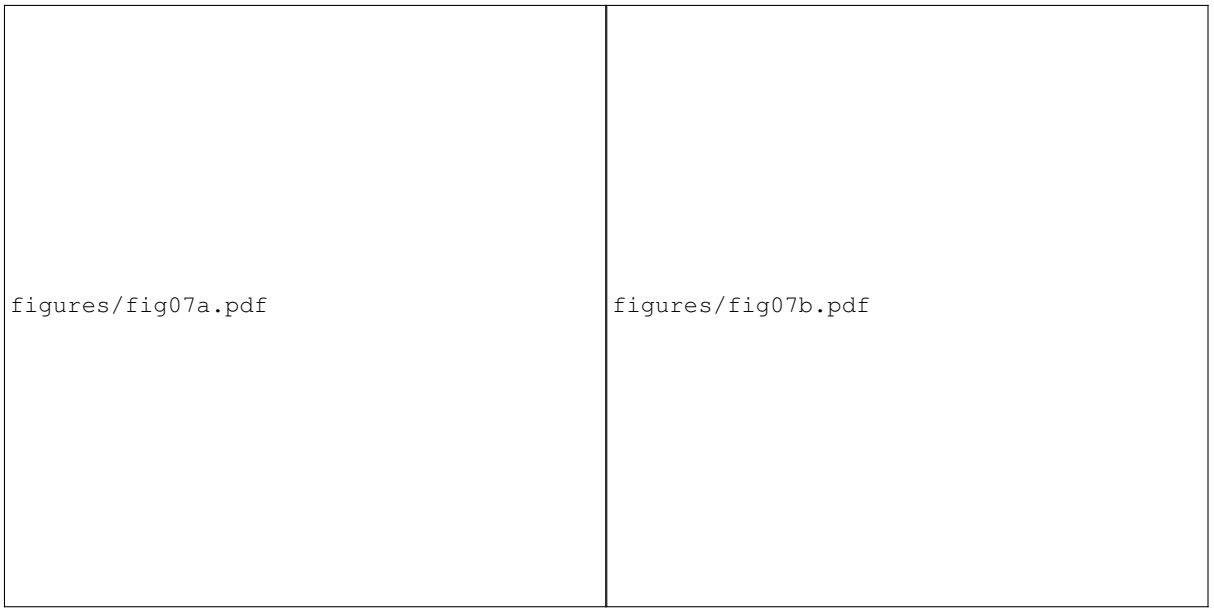
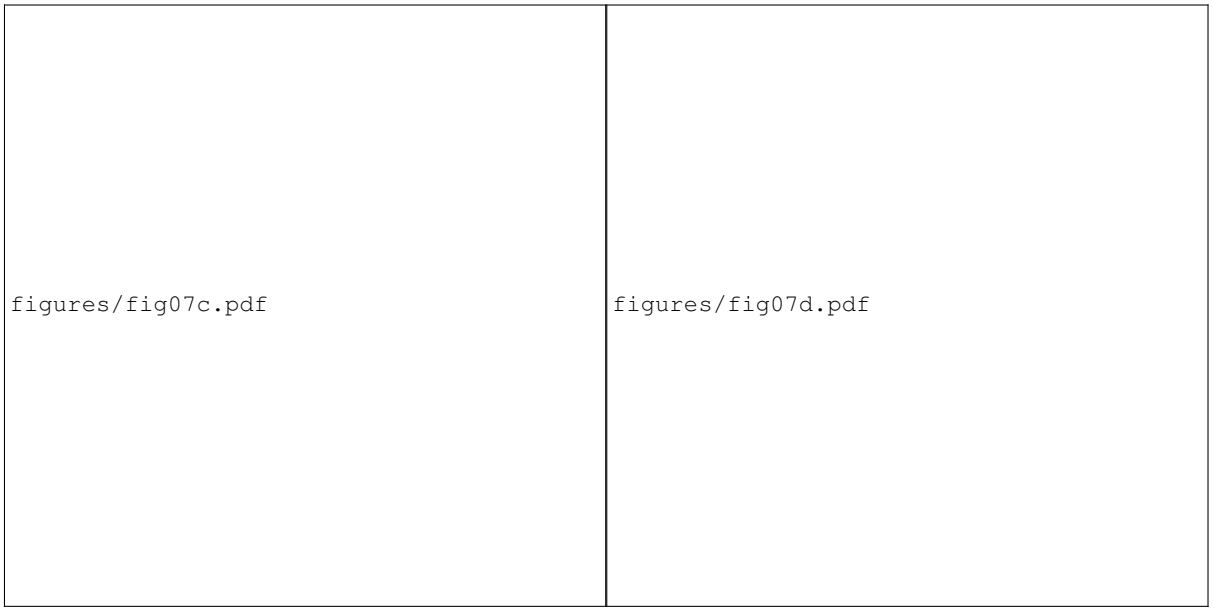


Figure 7. ~~Comparison of the~~ Flattened distribution of ~~storm intensities in the Bay of Bengal between 500 years of GAN-generated and 81 years of ERA5 storm footprints. Storms are categorized according to the peak~~ wind speed ~~during~~ pixels for the data generated using each ~~storm~~model. ~~While the GAN captures the general shape of the distribution, some systemic underestimation is evident~~All variables have been transformed to Gumbel to enable cross-comparison.~~Implications of this are discussed in the text.~~

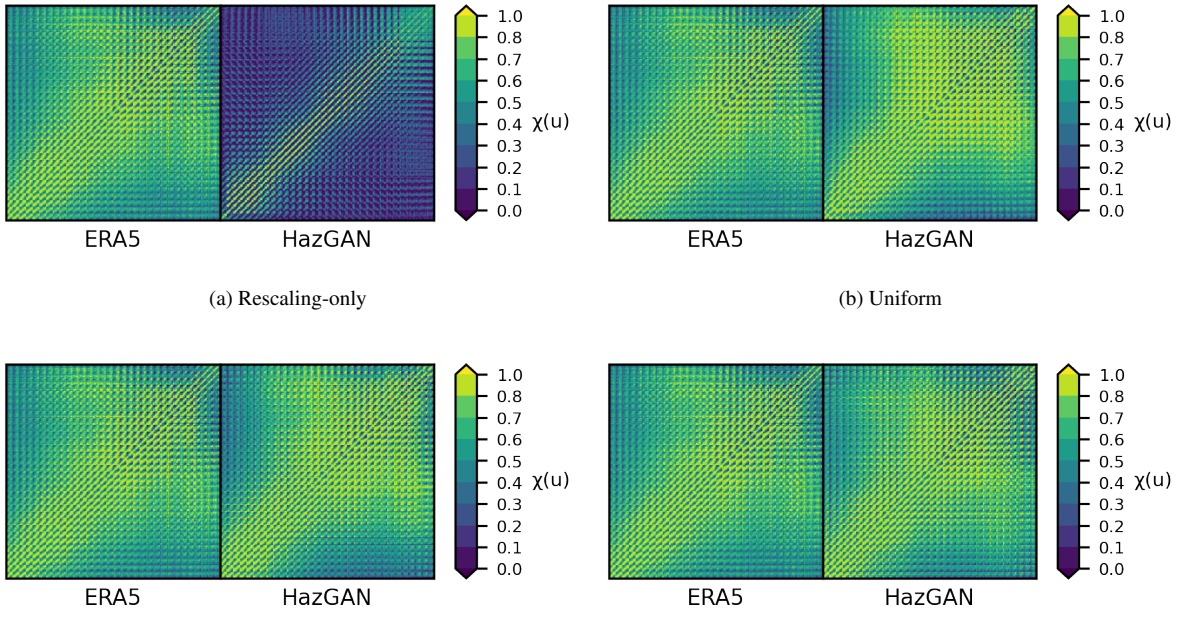


(a) Rescaling-only (b) Uniform



(c) Gaussian (d) Gumbel

Figure 8. Comparison of the distribution of storm intensities in the Bay of Bengal between 500 years of GAN-generated and 81 years of ERA5 event footprints. Storm intensities are assigned according to the domain-wide maximum wind speed during each storm. Shown for GANs trained on (a) data rescaled in the usual way for deep learning; (b) data transformed to a uniform distribution via a probability transform; (c) data transformed to a Gaussian distribution; and (d) data transformed to a Gumbel distribution.



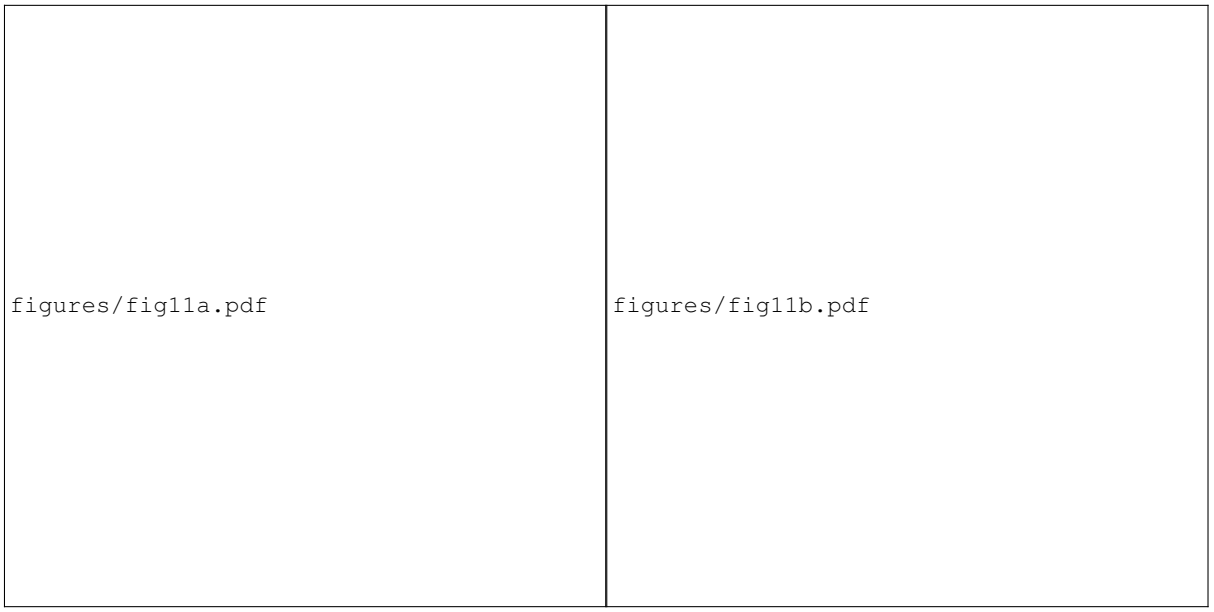
(a) Rescaling-only

(b) Uniform

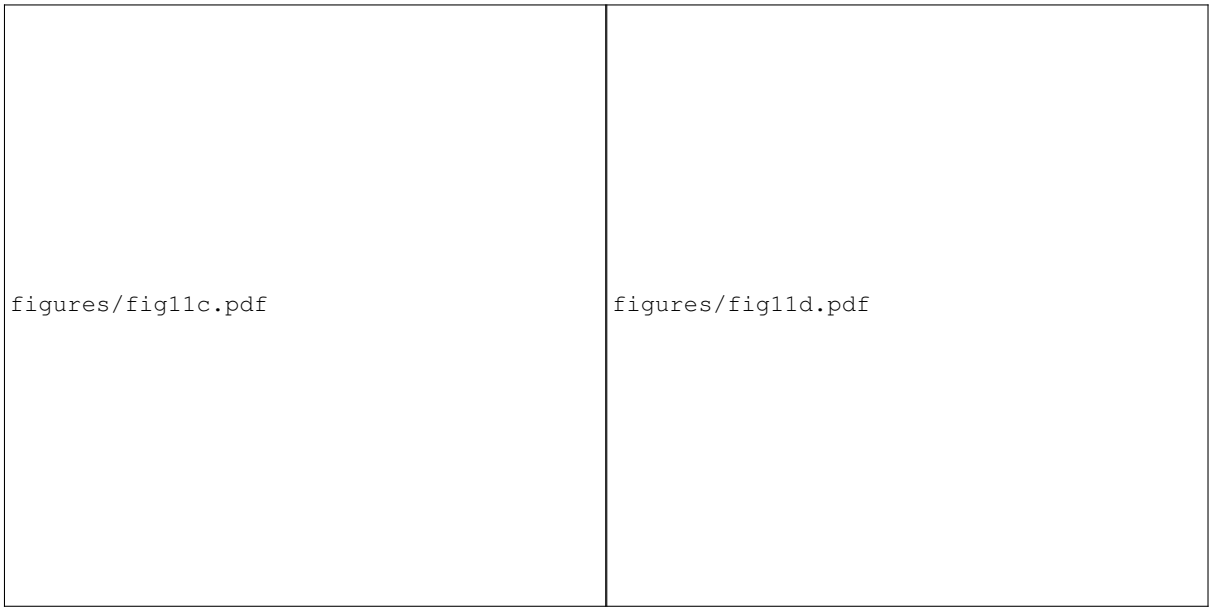
(c) Gaussian

(d) Gumbel

Figure 9. Pairwise spatial ~~(a) Pearson correlations and (b) tail dependence coefficient~~ extremal correlation estimates at $\hat{\chi}(0.8)$ for 10 m wind speed anomalies during storms across the Bay of Bengal.~~For the tail dependence coefficients: values of 1 indicate total dependence and values of 0 total independence in the extremes.~~

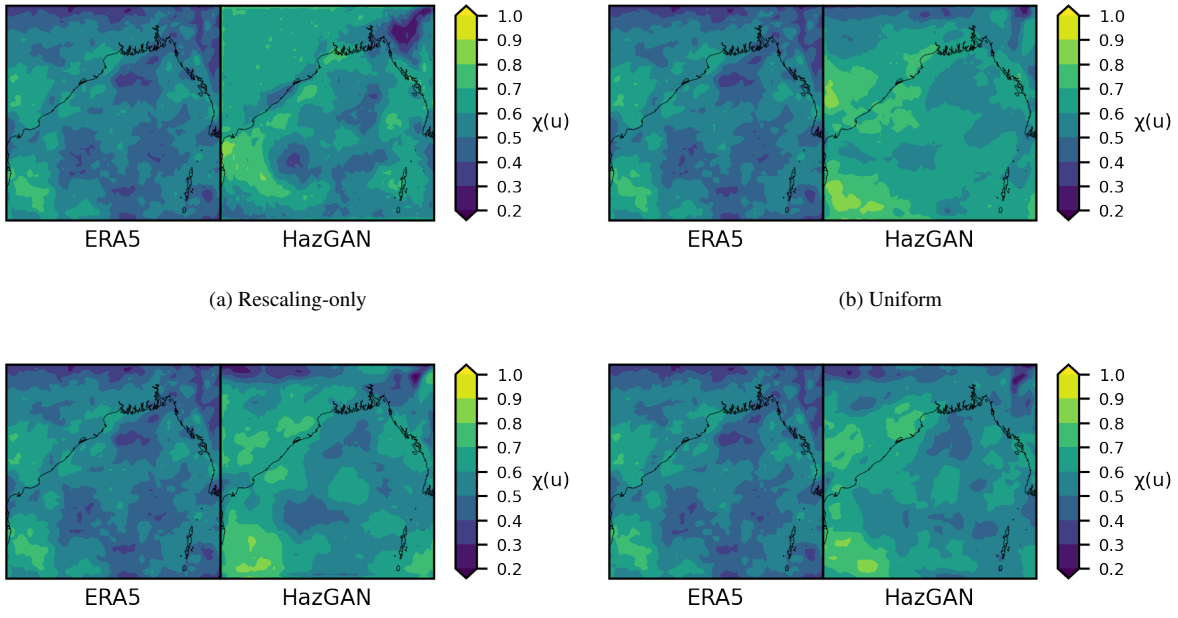


(a) Rescaling-only (b) Uniform



(c) Gaussian (d) Gumbel

Figure 10. Scatter plots comparing the bivariate distribution of 10 m wind speed anomalies between the cities of Dhaka and Chittagong in Bangladesh.



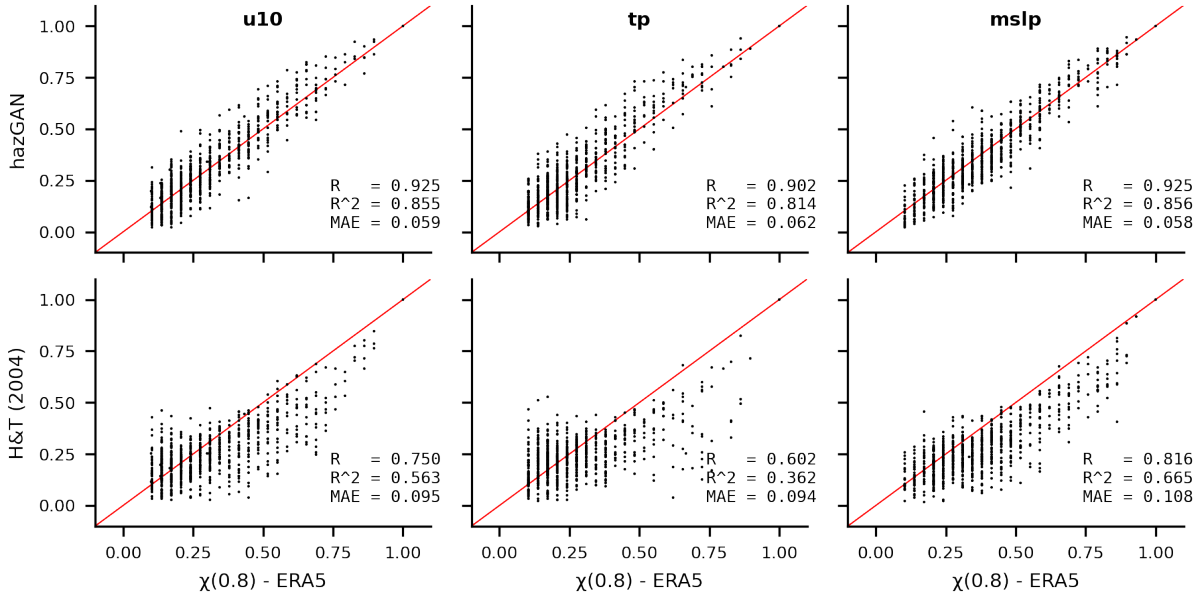
(a) Rescaling-only

(b) Uniform

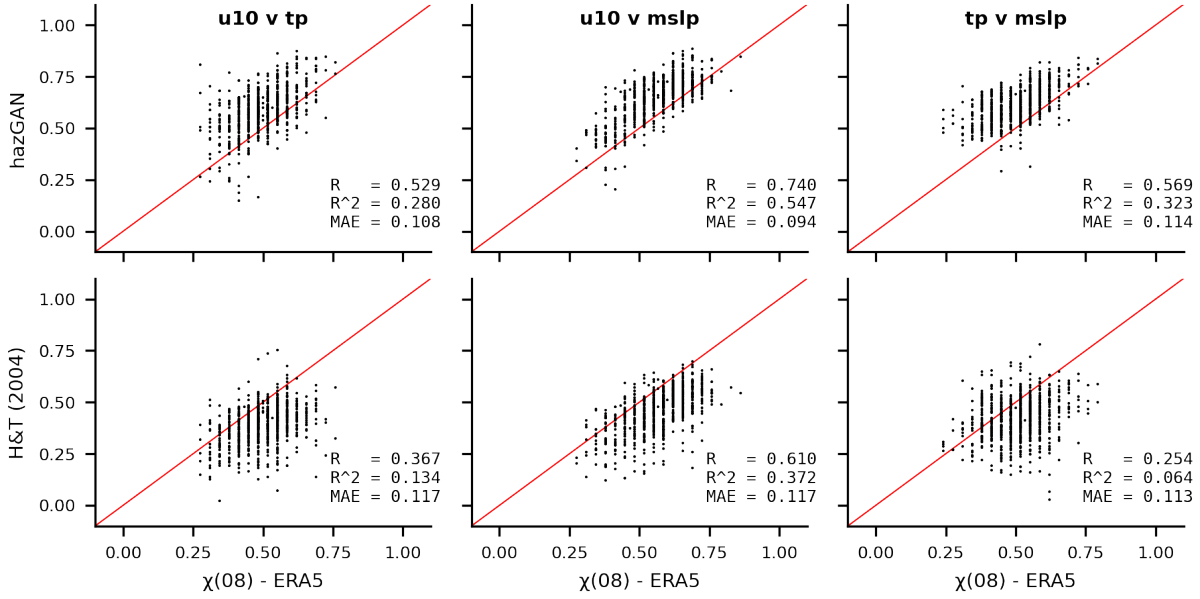
(c) Gaussian

(d) Gumbel

Figure 11. Inter-variable ~~(a) Pearson~~ extremal correlation ~~and (b) tail dependence coefficients~~ estimates $\hat{\chi}(0.8)$ for 10 m wind speed and total precipitation over the Bay of Bengal.~~Similarly to Fig. **??**, the GAN reproduces the dependence structure between variables in both the bulk and the extremes. Winds and precipitation show a high Pearson correlation in the north-east portion of the bay while dependence between the extremes of the two variables increases over land. The GAN shows some overestimation of the correlation between wind and precipitation over the bulk of the bay.~~



(a) Rescaling-only



(b) Uniform

Figure 12. ~~Scatter plots comparing the bivariate distribution of monthly anomalies for al three variable between pairs of points during storms. Results are shown for between Chittagong and Dhaka, two cities in Bangladesh, and two ocean buoys. The GAN learns the overall shape of each bivariate distribution, performing particularly well for mean sea-level pressure for both point-pairs.~~?

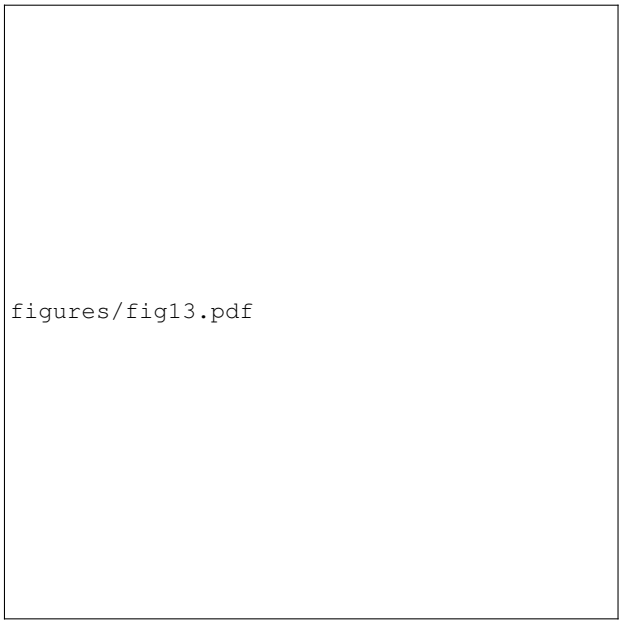


Figure 13. Risk profiles showing the expected total area of mangrove forest damaged by storms of different return periods for ERA5 and GAN-generated storm footprints. Also shown are risk profiles for the two simplifying assumptions of total dependence or independence across the domain, where a total dependence assumption is analogous to using return period maps to calculate risk profiles. With a total area of mangrove forest across the bay of 9,917 km^2 in 2020 (**?**), 2,000 km^2 corresponds to approximately 20% and 4,000 km^2 corresponds to approximately 40% of the total area of mangrove forest present.

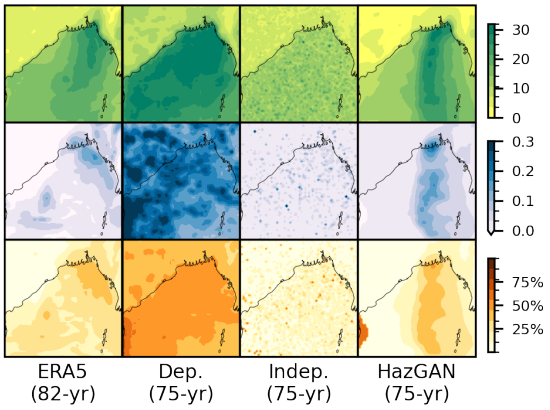


Figure 14. Footprints for events with approximately ~~1-in-20~~ 1-in-75 year return period mangrove damages for the ERA5 and hazGAN-generated datasets, as well as samples from the two simplifying assumptions of total dependence or independence between all variables.

5 Discussion

5.0.1 ~~Method performance and validation~~

In this paper, we have ~~demonstrated an efficient method for generating multivariate and spatially coherent event sets, composed of three-variable hazard footprints. The event sets are suitable for risk analysis over large geographic areas. Our three-phase workflow combines statistical extreme value theory with generative deep learning, creating a comprehensive framework for synthetic hazard generation. The computational workload of this method is front loaded during the parameter-fitting and GAN-training steps, which take less than a day. Once trained and fitted, generating new event sets is fast and efficient, generating 5,000 samples in under 30 minutes.~~ developed a generative deep learning approach for simulating spatially coherent multi-hazard event sets that preserve the tail distributions of the training data. We have done this by building on existing work by **?**, **?** and the statistical theory of multivariate extremes (**?**). We have demonstrated that the ability of a generative adversarial network (GAN) to model both marginal and joint tail behaviour of a dataset can be much improved by transforming the training data to a light-tailed distribution—such as a Gumbel or Gaussian distribution, allowing us to use standard GAN architectures and training methods. We used a peaks-over-threshold modelling approach to model the extreme values in 2-dimensional hazard footprints, allowing us to capture spatially coherent event footprints and cumulative event impacts. In a case study of storms in the Bay of Bengal, we demonstrated that our method better reproduces the extremal correlation structure of the storm footprints compared to the well-known conditional exceedance model of **?**. In an example application to storm risk to mangrove forests, we demonstrated that correctly modelling the dependence structures leads to far more realistic risk profiles compared with approaches that ignore spatial or multivariate dependencies.

~~Our validation against three key criteria demonstrates the method's effectiveness. The model successfully reproduces the training distribution of storm intensities and maintains both spatial and multivariate dependence structure across the bulk of the data and in the extremes. The Anderson–Darling goodness-of-fit test validates the parametric fits used to guide extrapolation in the extremes, ensuring a solid statistical foundation for new, more extreme, synthetic events~~In Sect. **??**, we described the general theory and methodology of our approach, which is broadly parametrised by four key choices: (i) the region of interest; (ii) the weather data from which the hazard footprints were extracted; (iii) a hazard severity function $r_{|ijk}(\mathbf{x})$ which is used to select hazard events; and (iv) a temporal aggregation function $h_{k|t}(\mathbf{x})$, which defined how spatiotemporal hazard data should be collapsed into 2-d event spatial footprints. In the application to storms in the Bay of Bengal in Sect. **??**, we used simple choices for the deseasonalisation, event identification, event severity, and temporal aggregation functions. These were not intended to be prescriptive and more sophisticated choices could be be explored in future work or more applied settings. More specialised event severity functions could be used, such as for example, hazard indices like the storm severity index for windstorms (**?**) or the fire weather index for fire potential (**??**).

5.0.2 ~~Practical applications and flexibility~~

A feature of this approach to event identification was that extracted variables were sampled conditional on the occurrence of hazard events, as defined using the severity function $r_{|ijk}(\mathbf{x})$. Although in the Bay of Bengal case study we validated that

no specific region significantly biased the event selection method (Fig. **??**), in general this conditionality was intentional: we sought to model the joint behaviour of all variables during hazard events rather than the natural marginal extremes of each variable. While this does not violate the assumptions of fitting a generalised Pareto distribution to the marginal exceedances, it remains important to verify that the marginal observations remain independent and identically distributed. In this work, we used deseasonalisation to ensure stationarity and declustering to ensure independence between events. The peaks-over-threshold approach somewhat mitigates the risk of mixed climate effects by isolating the most extreme events, which are more likely to arise from a single dominant mechanism. In future applications, however, the validity of this approach will depend on the specific weather variables and hazard types being modelled, and a more careful treatment of potential mixed climate effects would likely be required (see, for example, **???**).

~~Our method utilises the power of deep learning models in high dimensional settings to push the boundary of the complexity–scope trade-off that exists in risk modelling problems; scientists can use this method to model both multiple variables and large domains, enabling more holistic risk analyses. This is particularly relevant to systems distributedover large areas, where compounding climate hazards represent a significant risk with poorly quantified likelihoods.~~

~~The framework's versatility extends well beyond the case study provided: the modular design allows practitioners to substitute wind speed, precipitation, and sea levelpressure variables with any climate variables for which training datais available. Hence our method can be used to assess risk from various compound hazards including drought, heat stress, fire risk, and compound coastal flooding, each requiring different combinations of meteorological drivers.~~

~~This event-based approach is particularly suited to short-duration, short-memory hazards. Longer duration events, such as droughts, require additional modelling considerations. However, the framework's flexibility accommodates different temporal aggregation periods for variables, so it is possible to include variables accumulated over longer anteceedent periods for more complex modelling~~ Specialised extreme value distributions were fitted to the each margin (location and hazard variable) of the footprints to transform them to standardised distributions. This allowed us to control the tail behaviour of the generated data and to extrapolate to new extremes in a statistically justified manner. In the Bay of Bengal case study, we used a generalised Pareto distribution (GPD) to model the tails of all weather variables. The generalised Pareto distributions were independently fitted to the tails of each margin in the training data and the fitted shape, scale, and location parameters varied smoothly across the domain (Fig. **??**). Fewer than 1% of the 12,288 marginal fits failed the Anderson–Darling goodness-of-fit test at the 5% significance level, indicating that the GPD provided a reasonable fit to the extremes of the training data. This result could potentially be strengthened, however, by fitting a nonstationary model over the domain, allowing us to effectively 'borrow strength' between pixels (**??**, pp. 167; pp. 2). Additionally, the shape parameter estimates for all variables were within the range [-0.5, 0.5], suggesting that the true asymptotic form of the margins was a Type I distribution, in which case a direct Gumbel fit may have been appropriate. Such a fit would align with the existing understanding that wind speeds have a Weibull distribution, a distribution which converges slowly to the Type I (Gumbel) asymptotic form (**?**). In this case a subasymptotic model such as, for example, XIMIS (**?**) may be the most appropriate approach to modelling the margins and we leave this as an interesting avenue for future work.

5.0.3 ~~Current limitations and data quality~~

In Section **??**, we used a StyleGAN2-ADA with differentiable augmentation to train a generative model on a set of 150 multi-hazard footprints (**??**). We chose this model for its demonstrated ability to learn from small datasets, making it appropriate for modelling historical climata data, which generally contains fewer than 100 years of data (*e.g.*, **?**). It would however be interesting to investigate the applicability of alternative data-efficient deep generative models in this framework.

~~The primary limitation of the method remains training dataavailability and quality~~For training data, we used the ERA5 hourly gridded reanalysis from 1940 to 2022 (**?**). ERA5 ~~, despite being~~ data was chosen as it is the most comprehensive ~~global reanalysis dataset available, introduces biases through its~~ and widely used reanalysis product currently available. However, like all reanalysis products, ERA5 has known biases and uncertainties. The 0.25° ~~horizontal resolutionthat smooths extremes. Regional variations in data quality introduce additional uncertainties that propagate through the model. Practitioners should carefully assess their input data's limitations, as any biases will be inherited by the synthetic event set.~~

5.0.4 ~~Extensions~~

~~Several promising extensions could expand the framework's capabilities: different deep learning models, dataset topologies, alternative parametric distributions for the marginals, and various data pre-processing methods could all broaden the method's scope and effectiveness.~~

~~Alternative deep learning models and architectures could provide better or equal performance to the StyleGAN model used in this paper. Normalising flow models allow for the explicit evaluation of a sample's likelihood, allowing for in-built estimates of trainingand generated sample return periods. Additionally, different data topologies could be used. Gridded climate data provides a good illustrative example, as it is high dimensional and can be used with image generator models; however, data with network topologies (river networks) or point data (weather stations) could be used with alternative~~ ° horizontal resolution, for example, cannot resolve many small-scale phenomena such as convective, turbulent, and dissipative processes, leading to parametrisations of varying quality (**??**), a smoothing of extremes, and a loss of fine-scale detail. The deep learning model will learn and propagate these biases into the generated data so the biases of training data should be carefully considered in any applications. It may be desirable to use regionally developed reanalysis or observational products—*e.g.*, IMDAA for the Indian Ocean (**?**) or HadUK for the United Kingdom (**?**)—or to apply bias-correction and downscaling methods before training, although this may introduce additional uncertainties and complexities. Furthermore, this method could in theory be applied to data of any topology, provided a suitable deep generative model ~~types.~~

~~Incorporating alternative parametric distributions could build additional flexibility. Circular distributions are an interesting avenue to explore, and could be used to model periodic variables such as wind direction or seasonality. Discrete distributions, such as the discrete generalised Pareto distribution (**?**) could be used to model extremes in discrete variables such as storm duration~~could be developed. An interesting avenue for future work would be to explore the applicability of this method to non-gridded data, such as point or graph data, which would enable the direct simulation of hazards over stations, river networks, or infrastructure networks.

~~Data pre-processing could also be used to address biases and uncertainties in the training data. Bias correction and downscaling methods could be used to downscale training data to higher resolutions and correct known biases. The StyleGAN architecture has capacity to train up to 1024×1024 resolution images, suggesting potential for global-scale analysis or higher-resolution simulations. These methods could even be used with coarser global climate model simulations to inform scenario modelling for future risk assessments.~~ A key contribution of this work was the insight that training a GAN on data transformed to have light-tailed margins resulted in improved representation of both marginal and joint tail behaviour than standard GAN approaches or training on uniform margins. We demonstrated this by comparing the performance of GANs trained on data transformed to have margins that were: (i) uniform-distributed; (ii) a rescaled version of the original data; (iii) Gumbel-distributed; and (iv) Gaussian-distributed. The quality of generated samples was evaluated against the training data according to the overall distribution of event severity, the marginal distributions, and the multivariate dependence structures. The uniform transformation performed worst, likely because it compressed data in the tails, reducing sensitivity to marginal extremal behaviour. The rescaled transformation also performed poorly, likely because the model had to use more capacity learning the marginal distributions, leaving less capacity to learn the dependence structure. The Gumbel and Gaussian transformations showed comparable performance, with the Gaussian transformation showing slightly better performance on some metrics. All models performed slightly worse at capturing the multivariate dependence structure between hazards compared to capturing the spatial dependence structures. We originally hypothesised that this could be due to the architecture of the StyleGAN2-ADA model used; however, benchmarking against the **?** model showed a similar pattern, indicating that the poorer performance on multivariate dependence structures was indeed due to this being a more challenging task.

6 ~~Conclusion~~

~~This work presents a novel framework that successfully combines extreme value theory with generative deep learning to produce multivariate~~ Overall, the GAN trained on Gaussian margins achieved slightly better scores than the **?** model for capturing the extremal dependence structure. This result is encouraging, particularly because the hazGAN model learned the full joint distribution of all variables and locations simultaneously, whereas the **?** model learned pairwise relationships separately so could not capture higher-order dependencies. The GAN was able to borrow strength across the full dataset, enabling it to learn a more accurate dependence structure, while the **?** model fitted many separate models, each with less data. In the future, it would be interesting to compare the GAN to more recently developed spatial models for extremes, ~~spatially coherent climate hazard event sets. This method addresses a critical gap in climate risk analysis by generating synthetic events that preserve complex dependencies between variables and across space whilst remaining computationally efficient for practical applications.~~

~~The key contribution lies in the three-phase workflow that transforms gridded climate data into multivariate hazard footprints, suitable for input to a deep learning model~~ such as r-Pareto processes (**?**), which are designed to model the full spatial dependence structure of extremes. However, we consider this beyond the scope of the current work, which aims to demonstrate the effectiveness of the hazGAN approach.

In terms of reprodicibility: despite the use of a random seed, StyleGAN2-ADA has irreducible stochasticity, so results will be similar but not exactly repeatable. All other results, such as event identification and marginal fits, should be exactly reproducible. Future work could explore the extent to which the stochasticity of StyleGAN2-ADA affects the results and whether it can be reduced by using a different model or training method. ~~The transformation of the data's marginals to a standard Gumbel distribution places emphasis on the extremes of the climate variables during training, ensuring sensitivity to variation in the extremes. This approach combines the statistical rigour of univariate extreme value theory with the high-dimensional capabilities of deep learning methods. Our validation demonstrates that the method can accurately preserve the storm distributions, spatial correlations, and multivariate dependencies essential for accurate risk modelling.~~

~~The practical value of this methodis evident from its broad applicability across hazard types and its computational efficiency once trained. The Bay of Bengal case study illustrates how practitioners can use this method to generate comprehensive event portfolios for stress-testing vulnerable systems like mangrove forest ecosystems, but the approach extends readily to other hazards, regions, and asset types.~~

6 Conclusion

~~While current limitations centre on training data quality and availability, ongoing advances in the climate data processing and the framework's inherent flexibility position it well for addressing increasingly complex climate risk challenges. The method offers climate risk scientists and decision-makers with a powerful tool for generating the comprehensive event sets necessary for holistic climate risk assessments in an era of increasing climate variability.~~

~~Future applications~~ In this manuscript, we have demonstrated an approach that combines a deep generative model with methods from the statistical theory of multivariate extremes to generate spatially coherent multi-hazard event ensembles. The method is promising: we have demonstrated the ability of the GAN to capture the extremal dependence structures of the training data when certain transformations are made to it, and developed a method that allows us to generate hazard event footprints rather than the more commonly used annual maxima. We demonstrated a simple practical application, modelling wind, precipitation, and atmospheric pressure footprints during storms in the Bay of Bengal, and illustrated a use case in risk analysis: modelling the risk to mangrove forests from storms. While the model has shown promising results, it has only been validated on a single case study region and set of variables, so future work will focus on ~~developing more use cases, developing higher-quality training data , incorporating seasonal or directional variables, and extending the framework to other data topologies and deep learning architectures. As climate risks continue to evolve, this framework offers a scalable foundation for developing more sophisticated and comprehensive hazard modellingcapabilities.~~

further validation by applying the method to a wider range of scenarios and using synthetic training data that will enable more rigorous validation. This method, which is in theory agnostic to choices of region, variable, and deep generative model, has the potential to be used to generate large-scale, spatially coherent, multi-hazard events sets that can be used for a wide range of applications in risk assessments, stress testing, and scenario modelling.

Code and data availability. The code and data to used will be made available at 10.5281/zenodo.15838238.

Author contributions. AP and JH conceptualized the paper and developed the methodology. AP conducted the investigation with supervision from JH and support from YM. YM provided data and code towards the final mangrove damage study. AP prepared the original draft including all code and visualizations. AP, JH, and YM reviewed and edited the manuscript.

Competing interests. The authors declare no competing interests.

Acknowledgements. This work was funded by the UKRI Engineering and Physical Sciences Research Council (grant number: EP/T517811/1).

The authors would like to thank the Geoff Nicholls, Philip Hess, Shruti Nath, Benjamin Walker, and Alberto Fernandez Perez for their advice at various stages along this project.