Simulating 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~~ across 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 generative deep learning models—when combined with specific transformations to the training data—offer a useful alternative method for stochastically sampling realistic multi-hazard events. Our framework combines generative adversarial networks with extreme value theory in a hybrid approach that can capture complex dependence structures in gridded multivariate weather data, while providing a theoretical ~~justification~~ basis for extrapolation to new extremes. We apply our method to ~~model the co-occurrence~~ jointly model fields of strong winds, heavy precipitation, and low atmospheric pressure (~12,000 variables) 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. For the Bay of Bengal case study, we validate our approach against a ~~well-known~~ popular model for multivariate climate extremes, and demonstrate improved performance in capturing the extremal correlation structure.

1 Introduction

Hazards that entail several extreme weather variables and extend over large spatial domains are responsible for some of the most damaging natural catastrophes ~~(Zscheischler et al., 2018, 2020).~~ (Zscheischler et al., 2018). Compound hazards can be broadly grouped into four typologies: preconditioned, temporally compounding, multivariate, and spatially compounding (Zscheischler et al., 2020). This article focuses on the multivariate and spatially compounding types, specifically on the co-occurrence of hazard extremes across space or across variables. Multivariate or *multi*-hazards (UNDRR, 2017) 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, and compound flooding events involving fluvial (river), pluvial (rainfall), and coastal sources (Yin et al., 2023; Zscheischler et al., 2018). Spatially compounding hazards occur when a hazard impacts a large region, creating widespread damages and potentially systemic impacts (Zscheischler et al., 2020). Common examples include droughts causing crop failure in multiple regions simultaneously, leading to food stress; or extensive storm damages stretching emergency response capacity (Zscheischler et al., 2018; Boulaguiem et al., 2022; Gaupp et al., 2019; Bailey et al., 2015).

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~~ model single hazard types and fail to provide information about the spatial extent of potential events, which can bias estimates of tail risk (Bates et al., 2023). 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 by the insurance sector in Cat models and have been used in large-scale risk assessments in the UK and the USA ~~(Lamb et al., 2010; Bates et al., 2023; Cross et al., 2020; **?**; Quinn et al., 2019)~~(Lamb et al., 2010; Bates et al., 2023; Cross et al., 2020; Qu . Synthetic event sets can be generated from physically based model simulations (Guillod et al., 2018) but are computationally demanding and may entail elaborate model coupling, *e.g.*, to obtain storm surge heights as well as windspeeds and precipitation.

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~~ can be taken by fitting statistical models to weather data and simulating large events sets from those models ~~(Lamb et al., 2010, 2019; Speight et al., 2017; Becher et al., 202 .~~ (Wilks and Wilby, 1999). Recent advances in stochastic weather generation have enabled the generation of multivariate and spatiotemporal fields (Obakrim et al., 2025); however, adequately representing the extremal behaviour of multiple weather variables generally requires the use of specialised methods from multivariate extreme value theory (Davison et al., 2019; Lamb et al., 2010, . Classical models for the dependence structure of multivariate extremes include extremal copulas and the conditional exceedance model (Nelson, 2006; Heffernan and Tawn, 2004; Davison et al., 2012); 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~~ remains limited in its ability to scale ~~beyond~~ to high dimensions (see, for example, Quinn et al. 2019, who required >800 CPU cores to model the dependence between 2,400 ~~dimensions (*e.g.*, Quinn et al., 2019)~~ river gauges over the U.S.) 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 de Fondeville and Davison (2018) are particularly powerful, ~~capable of modelling up to up to~~ and have successfully modelled up to 3,600 dimensions, although this does not necessarily represent an upper bound on their potential (Huser and Wadsworth, 2022). Most spatial process models, however, 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 (Huser and Wadsworth, 2022; Huser et al., 2025; Engelke and Ivanovs, 2021). A further challenge is ~~to simulate~~ that the majority of spatial statistical methods model only pairwise dependence structures (Davison et al., 2012), entailing a loss of information (Serinaldi et al., 2015) and making it challenging to capture complex 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~~meteorological features such as rainbands, spiralling vortices, and fronts. Papalexiou et al. (2021) for example, couple random fields with velocity fields in order to model such features in a spatiotemporal setting.

Recently, interest has grown in ~~using~~ the use of machine learning methods to generate multivariate climate and weather extremes ~~(Bhatia et al., 2018; Engelke and Ivanovs, 2021; Boulaguiem et al., 2022; Huser et al., 2025)~~(Bhatia et al., 2021; Engelke and Ivan
. Deep learning approaches such as generative adversarial networks (GANs) (Goodfellow et al., 2014; Radford et al., 2015) have shown particular promise ~~(Bhatia et al., 2018; Wilson et al., 2022; Girard et al., 2025; Lhaut et al., 2025)~~(Bhatia et al., 2021; Wilson et al.,
. GANs were originally developed for image generation and their adversarial loss formulation naturally lends them to the generation of visually realistic spatial patterns (Ledig et al., 2017; Stengel et al., 2020). They have recently been used, for example, to successfully to downscale CMIP6 projections of hourly precipitation data (Abdelmoaty et al., 2025). GANs place two neural networks in competition: a generator 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, Wiese et al. (2019) demonstrated that a GAN with a light-tailed latent space cannot effectively capture the tail behaviour of heavy-tailed datasets. Building on this, Huster et al. (2021) developed *ParetoGAN*, which uses a unit Pareto latent space to better represent extremes—although this required a custom loss function, sacrificing some benefits of the adversarial loss. Separately, Boulaguiem et al. (2022) 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.

In this ~~manuscript~~article, 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 Huster et al. (2021) and Boulaguiem et al. (2022) with statistical theories of spatial and multivariate extremes. Building on the conclusions of Huster et al. (2021) and Wiese et al. (2019) 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 Boulaguiem et al. (2022) 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 peaks-over-threshold approach and using domain-wide functions to characterise events, enabling us to capture spatially coherent event footprints. Event *footprints* are commonly used in catastrophe modelling and post-disaster needs assessments to represent the maximum intensity of a hazard event across a region during its lifetime (*e.g.*, Lloyd's, 2025). 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 storm footprints compared to the ~~well-known~~ popular conditional exceedance model of Heffernan and Tawn (2004).

~~We~~ In n Section 2, we describe the general theory and methodology ~~in Section 2~~of our method. In Section 3, 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 (Islam and Peterson, 2009; Hunt and Bloomfield, 2025) and existing event sets have been shown to struggle in the region (Meiler et al., 2022). In Section 4, we demonstrate an application in risk analysis: evaluating the risk of wind and rainfall-driven storm damage to mangroves in the region.

2 Theory and methods

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 margins of the 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.

We will use the following terminology throughout: a *variable* refers to a weather hazard variable (*e.g.* wind speed, precipitation, sea level pressure); the *sample dimension* refers to the time dimension filtered to only contain event occurrences; and the *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 $i = 1, \dots, H$; longitude by $j = 1 \dots W$; time by $t = 1, \dots T$; sample number by $n = 1, \dots, N$; and variables by $k = 1, \dots, 3$. 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 $\mathbf{y} \in \mathbb{R}^{T \times H \times W \times K}$.

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.

2.1 Event footprint creation

Figure 1 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.

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~~

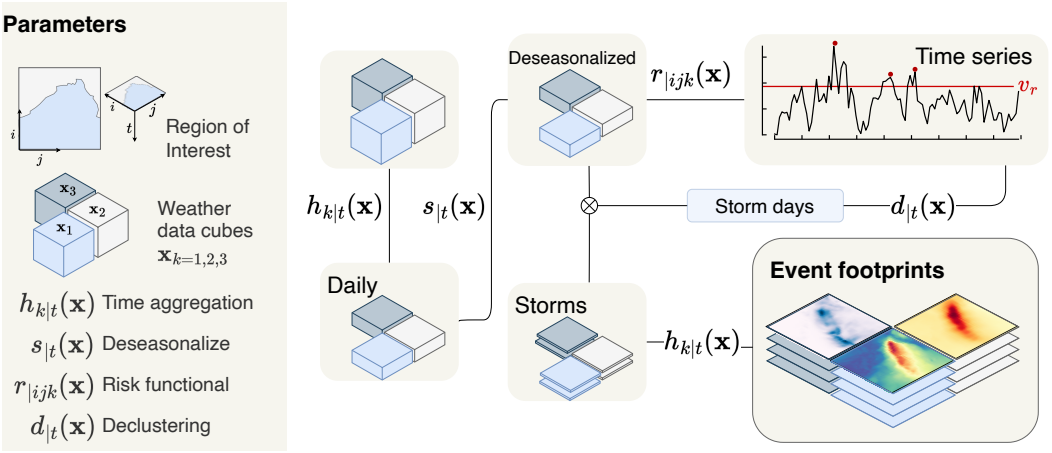


Figure 1. Schematic of the workflow to extract hazard event footprints. Gridded weather data over a region of interest for three variables $\mathbf{x} \in \mathbb{R}^{T\times H\times W\times K}$ (indexed $k = 1, 2, 3$) is deseasonalised according to $s_{|t}(\mathbf{x})$. A risk functional $r_{|ijk} : \mathbb{R}^{T\times H\times W\times K} \to \mathbb{R}^{T}$ 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.

deseasonalisation 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. 3).

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 (Dunlop, 2008) or the fire weather index for fire potential (Thompson et al., 2025; van Wagner et al., 1974).

To identify independent hazard events, a declustering algorithm $d_{|t} : \mathbb{R}^{T} \to \mathbb{R}^{T}$ (Gilleland and Katz, 2016; Coles et al., 2001) 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 ℓ_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 (Ljung and Box, 1978).

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~~ re-purposing 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. Broadly, this approach to storm detection bears similarities to the Method of Independent Storms (Cook, 1982) for univariate wind speeds and to the methods of de Fondeville and Davison (2018, 2022) in the spatial setting.

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 Transforming the marginal distributions

Figure 2 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.

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 Boulaguiem et al. (2022) 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 (Davison et al., 2012). However, Boulaguiem et al. (2022)'s approach relies on fitting generalised extreme value (GEV) distributions to the margins, requiring data in the form

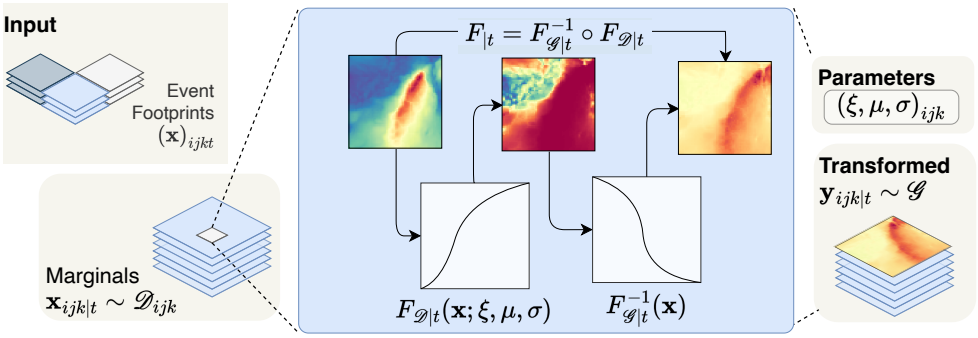


Figure 2. Schematic of the workflow to transform the marginal distributions of the event footprints extracted in Fig. 1 to a standard distribution $\mathcal{G}$. A suitable parametric distribution is fitted to the extremes of each weather variable along the sample dimension for each variable and location (marginal) $\mathbf{x}_{ijk|t} \sim \mathcal{D}_{ijk}$. The semiparametric distribution Eq. (1) 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 $\mathbf{y} \sim \mathcal{G}$.

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.

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. Heffernan and Tawn (2004, 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 (Coles et al., 2001). In this case, the shape, location, and scale parameters of $\mathcal{D}$ are denoted by ξ, μ, and σ, respectively. However, we have kept the formulation general to allow for alternative parametric distributions where appropriate (see, for example Harris, 2009). A semiparametric distribution as in Eq. (1) is fitted to all the margins of the multi-hazard event footprints $\mathbf{x}$, transforming them to have standard uniform margins $\mathbf{u} \sim \mathcal{U}(0,1)$.

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 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 3 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.

Distribution of the training data margins

From a machine learning perspective, transforming the margins to a 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 Huster et al. (2021) and Wiese et al. (2019) demonstrated that a GAN with a light-tailed latent space ~~cannot~~ struggles to learn the tail behaviour of a heavy-tailed distribution, leading to underestimation of the tails. Huster et al. (2021) 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.

Drawing on these results, we hypothesise that what matters is not the 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~~ certain desirable properties (Heffernan and Tawn, 2004; Keef et al., 2009; Quinn et al., 2019). We verify this hypothesis by repeating the experiments of Huster et al. (2021) in the Supplementary Material and find that we can match the results of Huster et al. (2021) using a standard GAN on light tail-transformed data. In Sect. 3, 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.

2.3 Generative model training and sampling

Figure 3 shows the workflow for training the deep generative model and sampling synthetic multi-hazard footprints. The transformed multi-hazard footprints from Fig. 2 are rescaled to take values in the range $(0,1)$ using a custom 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. 2 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.

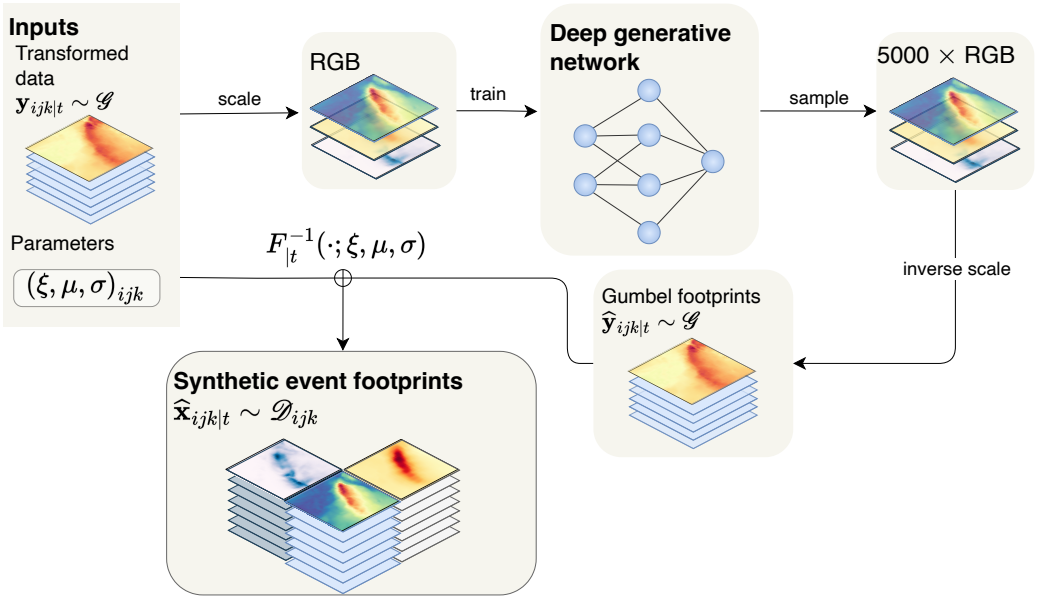


Figure 3. Schematic of the workflow for training the deep generative model. The transformed hazard footprints from Fig. 2 are rescaled to take values in the range $(0, 1)$ using a return period-based scaling method (Eq. 2) and converted to three-channel RGB images. A deep generative network is trained 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 Eq. (2). The inverse of the probability integral transform (Eq. 1) converts the synthetic footprints back to the scale of the original (deseasonalised) data.

Rescaling footprints be in (0, 1)

Depending on the distribution chosen for training $\mathcal{G}$, the transformed multi-hazard footprints (~~$\mathbf{y} \sim \mathcal{G}$~~with marginals $y \sim \mathcal{G}$) may require ~~a final,~~ an additional rescaling step to ensure they lie in the range $(0, 1)$. ~~The~~ A standard rescaling approach is min–max scaling, which maps the largest value of a dataset to 1 and the 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.~~ is our goal.

An alternative approach that allows us to specify a sensible maximum range for the ~~dataset, without specifying a physical range subjectively~~generated data, using only the training data, is to ~~rescale~~ set the maximum range using quantiles of the ~~distribution we chose for training. We can specify quantiles that correspond to~~ training distribution $\mathcal{G}$, which we can specify to correspond specific return periods~~, for example, for a~~. For example, we can set it so that the maximum possible value our method can produce, at any point on the domain, corresponds to an intensity that is on average exceeded once every R events at that point (or similarly, once every $R' = \lambda R$ years, with λ the number of events per year). For the distribution $\mathcal{G}$~~and return period~~, the minimum and maximum values associated with R ~~, the return level is $y_R = F_{\mathcal{G}}^{-1}(1 - 1/R)$~~are $F_{\mathcal{G}}^{-1}(1/R)$ and

$F_{\mathcal{G}}^{-1}(1-1/R)$, respectively. 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.~~ While this will result in the same maximum and minimum values across the $\mathcal{G}$-distributed training data, it will map back to different values in physical space.

After rescaling, the weather fields are stacked such that each multivariate footprint is now a three-channel tensor. Finally, it is converted to an RGB image, ready to feed into a generative model.

Generative adversarial network

In theory, this framework is agnostic to the choice of deep generative model, meaning common models like variational auto-encoders (Kingma and Welling, 2013), diffusion models (Ho et al., 2020), or flow-based models (Kobyzev et al., 2021) 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 (Liu et al., 2021) and StyleGAN2-ADA (Karras et al., 2020), which use a self-supervised discriminator and ~~differentiable augmentation~~ adaptive discriminator augmentation (ADA), respectively, to prevent overfitting. ~~Differentiable augmentation~~ 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~~additional differentiable augmentation (DA), StyleGAN2-ADA successfully learned the distribution of a dataset of only 100 images from scratch (Karras et al., 2020; Zhao et al., 2020). In Section 3, we will use StyleGAN2-ADA with differentiable ~~augmentationsto~~ augmentations (StyleGAN2-ADA+DA) to train a generative model on a set of multi-hazard footprints.

2.4 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 Coles et al. (1999, 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 Coles et al. (1999, 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 Heffernan and Tawn (2004) 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~~ by Heffernan and Tawn (2004, 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 multivariate case, where we can model the distribution of n other variables $\mathbf{j} = 1, ..., n$, 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 are not well characterised by existing event sets (Meiler et al., 2022). We model three variables that determine the impact of tropical cyclones: wind speed, precipitation, and atmospheric pressure at sea-level, which ~~determines~~ affects the elevation of storm surges.

3.1 Data

Our dataset consists of hourly gridded weather data from the ERA5 reanalysis product from 1940 to 2022 (Hersbach et al., 2023). We extract the northerly and easterly components of 10 m wind speeds (ms^{-1}), total precipitation (m), and atmospheric pressure at sea level (Pa) over the region and calculate wind speed as the ℓ_2-norm of the northerly and easterly components of the 10 m wind speeds.

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 $s_{|t}(\mathbf{x})$, which computes monthly medians for each margin and subtracts them, yielding a time series of climatological anomalies. Storm events are identified using the daily maximum 10 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 events.

Figure 4 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% of events). 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.

To 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 $h_{2|t}(\mathbf{x}) = \sum_{|t}(\mathbf{x}_2)$, extracting the cumulative precipitation over the events duration; and for sea level pressure $h_{3|t}(\mathbf{x}) = \min_{|t}(\mathbf{x}_3)$, extracting the lowest sea-level pressure over the event's duration.

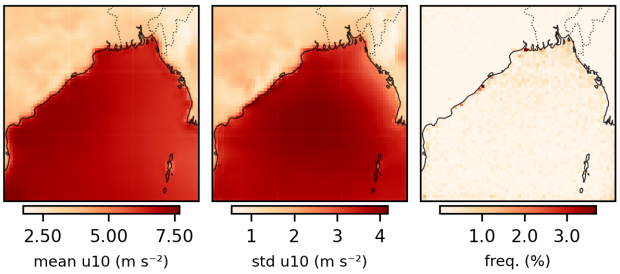


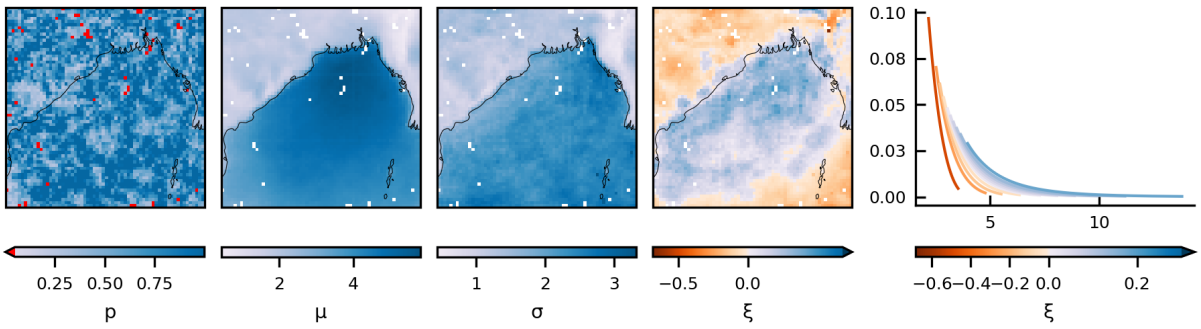
Figure 4. Map of grid cells that triggered storm events, based on being the site of the maximum wind speed over the domain on a given day.

3.3 Marginal transformations

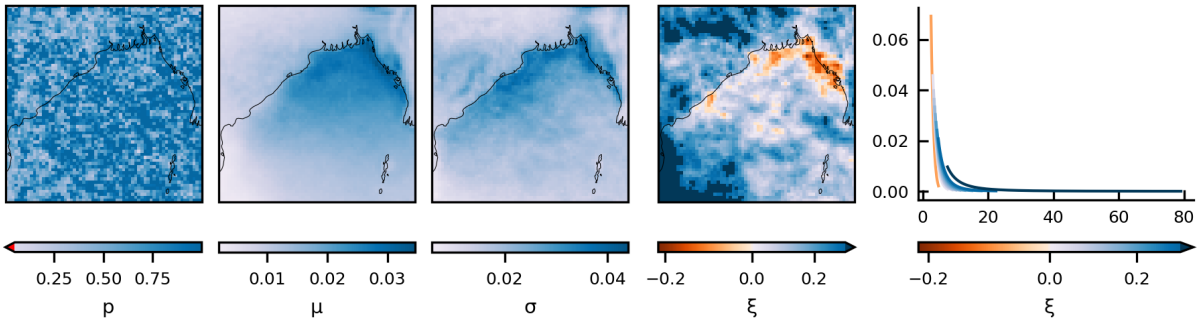
We fit the semiparametric distribution Eq. (1) to the margins of the wind speed, precipitation, and (negated) sea-level pressure footprints and in line with Heffernan and Tawn (2004), we use a generalised Pareto distribution (GPD) as the parametric tail model. Threshold selection for the GPD for each of the 12,000 margins is done using the *ForwardStop* method of Bader et al. (2018). 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 5(a)–5(c) show the fitted parameters for each of the wind, precipitation, and sea-level pressure event sets. For precipitation, the Bader et al. (2018) method successfully selected a threshold for all margins, while for wind speed and sea level pressure, it failed to select thresholds for 54 and 35 margins, respectively~~, so these~~. The rejected margins were fitted using entirely empirical distributions. These pixels are shown as red/white pixels in the fitted parameter plots.

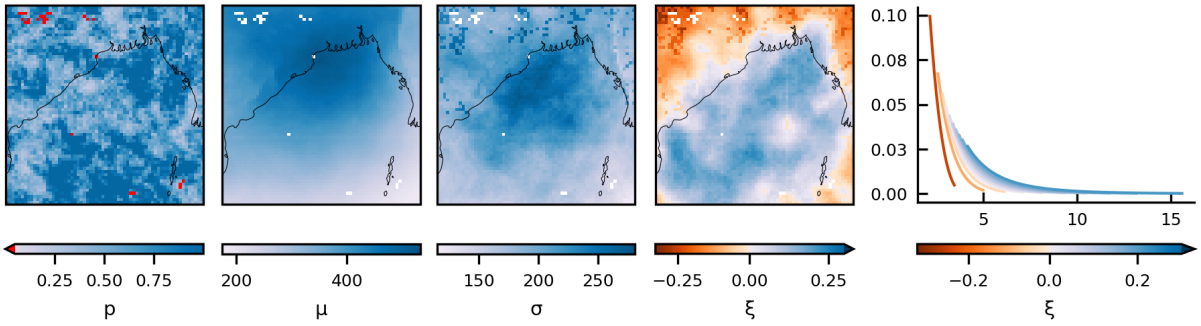
Figure 5 shows that the GPD parameters vary relatively smoothly across the domain for all three variables, with ~~all variables exhibiting larger scales and variability near the~~ a marked difference between onshore and offshore regions. The shape parameter estimates show the highest variation, with wind and sea-level pressure assigned predominantly negative shape parameters over land, suggesting the existence of an upper bound for wind speeds and a lower bound for sea-level pressure in these regions. Similarly, precipitation is assigned some negative shape parameters along the northern coastline. ~~It is~~ The existence of upper bounds for wind and rainfall have been much-contested in the literature (Harris, 2005; Serinaldi and Kilsby, 2014); however, and such conclusions should be treated with caution. Additionally, a large patch of very heavy-tailed precipitation is visible in the southwestern corner of the domain. While beyond the scope of this study, 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 (Davison et al., 2012; Huser and Wadsworth, 2022, pp. 167; pp. 2). ~~Additionally, the shape parameter estimates for all variables are~~



(a) Wind speeds at 10 m



(b) Total precipitation



(c) Sea-level 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.

~~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 (Harris, 2005). Work by Harris (2009) 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.~~

Applying the semiparametric distribution functions Eq. (1) with the parameters shown in Fig. 5, the transformed variables $\mathbf{u} = \tilde{F}(\mathbf{x})$ have a standard uniform distribution. But as discussed in Sect. 2, 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 Boulaguiem et al. (2022)'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~~ a deep learning model, we use the return period-based rescaling method described in Sect. 2. We ~~choose~~ use a 1-in-$R = 10,000$ ~~years as the maximum return level any marginal value can reach . This was~~ event return level to specify the maximum value any generated marginal can reach (choosing event-based rather than year-based scaling for simplicity). This is converted into a maximum value ~~for rescaling for~~ of $1 - 1/R$ for the uniform distribution ~~$1 - 1/R$ and~~ and $F_{\mathcal{G}}^{-1}(1 - 1/R)$ for the Gaussian and Gumbel distributions~~using $F^{-1}(1 - 1/R)$, where $F_{\mathcal{G}}^{-1}$ is the quantile function of the distribution. For the data that has not been transformed, the~~ . The lower bounds are calculated analogously, *i.e.*, using a minimum value of $1/R$ for the uniform distribution and $F_{\mathcal{G}}^{-1}(1/R)$ for the Gaussian and Gumbel distributions. As a heuristic, the no-transform 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.~~ events—approximating the ratio of the R-event to N-event maximum under Gumbel tails. While we note this is inexact, it is sufficient for our comparative purposes.

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 (Arjovsky et al., 2017; Gulrajani et al., 2017) on the 1,249-event dataset, we found that the GAN became biased towards the more common, less spatially coherent storms. To address this, we used the GAN that has been specifically developed for small datasets, the StyleGAN2-ADA model, with additional differentiable augmentation (Karras et al., 2020; Zhao et al., 2020), which can produce good results on as few as 100 training samples.

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 150 most extreme storms for 2013 epochs (*i.e.*, so that it had sampled images 300,000 times). This took approximately four hours on an ~~nVIDIA~~ NVIDIA GeForce GTX 1080 Ti GPU. 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

Visual appearance

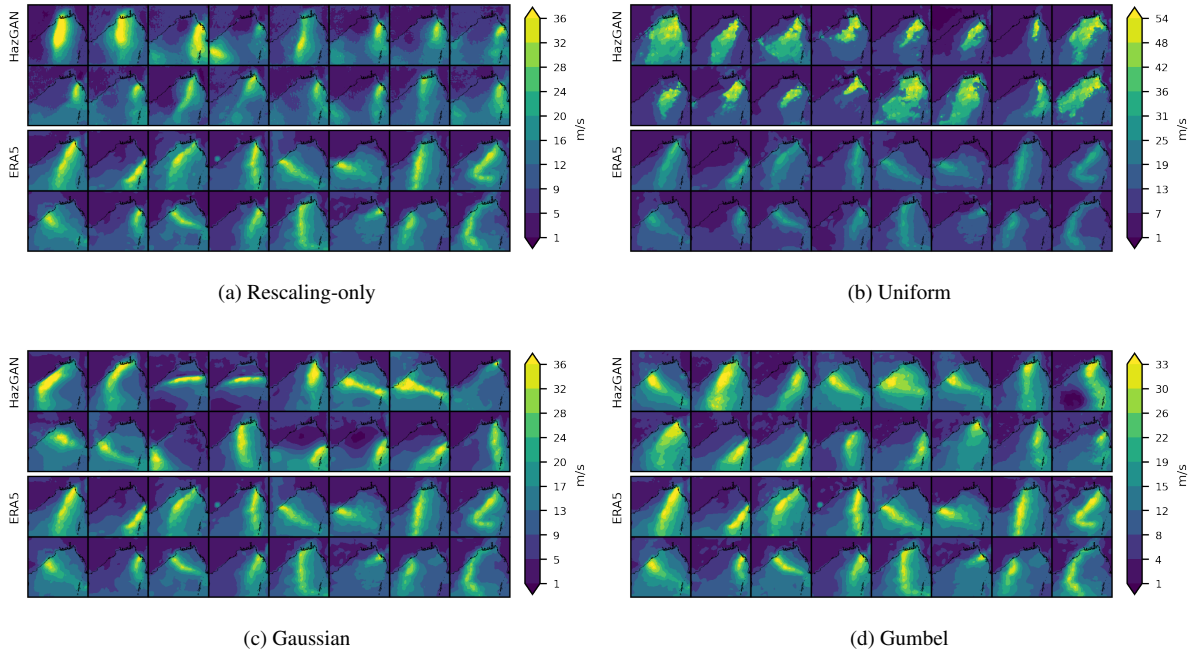
Ranking scores according to the severity function~~$r_{|ijk}(\mathbf{x})$,~~, as defined by the domain-maximum wind speed $r_{|ijk}(\mathbf{x}) = \max_{k=0|i,j}(\mathbf{x})$, Fig. 6 compares the 16 most severe storm footprints derived from the ERA5 dataset (top rows) with the most extreme footprints generated using the GAN (bottom rows) 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 provided in the Supplementary Information.

~~The~~ Qualitatively, 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.

Marginal distribution fits

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. (3), 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.

To assess how well each model captures the overall distribution of pixel values, Figure 7 shows the 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.



(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 for a GAN trained on (a) margins rescaled to [0,1] in the usual way, (b) margins transformed to uniform, (c) margins transformed to Gaussian, and (d) margins transformed to Gumbel.

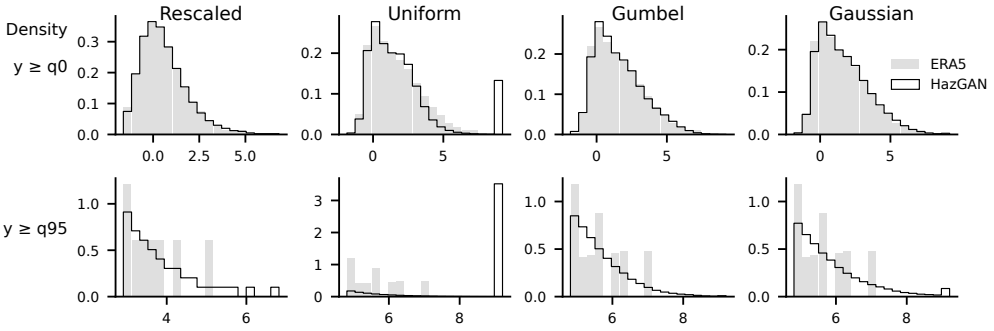
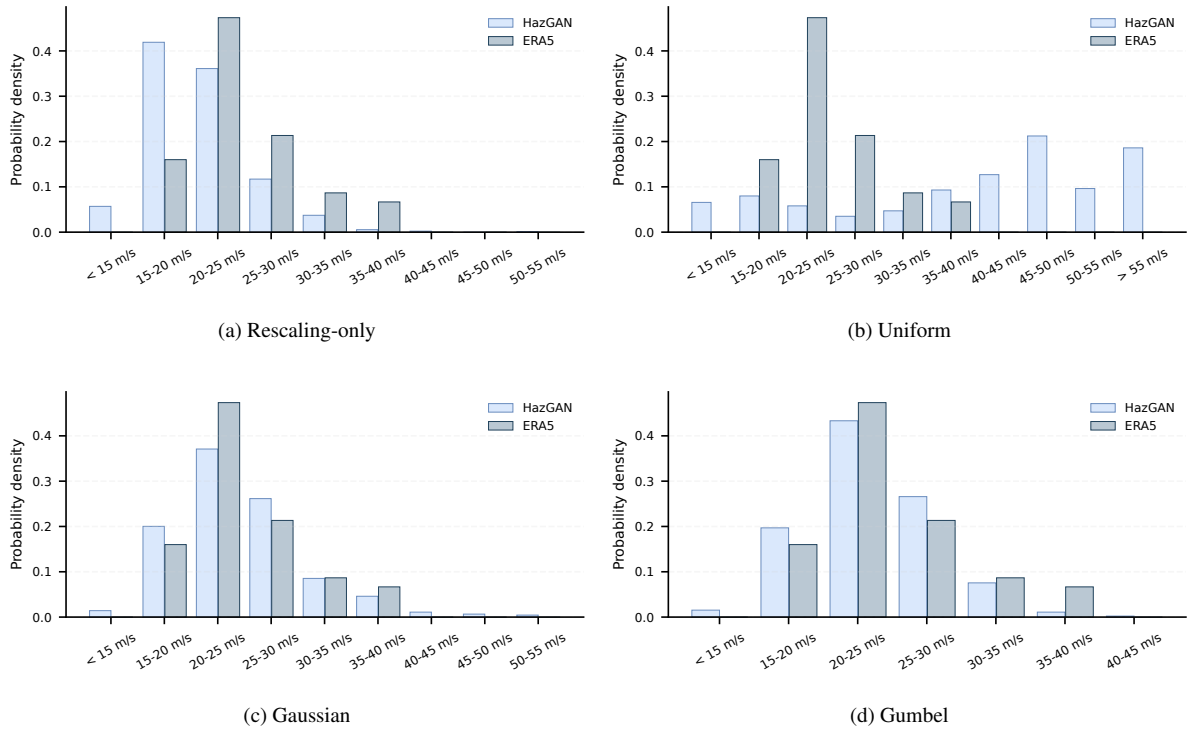


Figure 7. Flattened distribution of wind speed pixels for the data generated using each model. All variables ~~have been~~ transformed to Gumbel to enable cross-comparison.

Storm event distribution fits

To assess how well the GAN captures the storm ~~intensity distribution, as defined by the domain-maximum wind speed, we calculate~~ severity distribution, we again calculate the severity $r_{|ijk}(\mathbf{x})$ for each storm in the training and generated data and



(a) Rescaling-only

(b) Uniform

(c) Gaussian

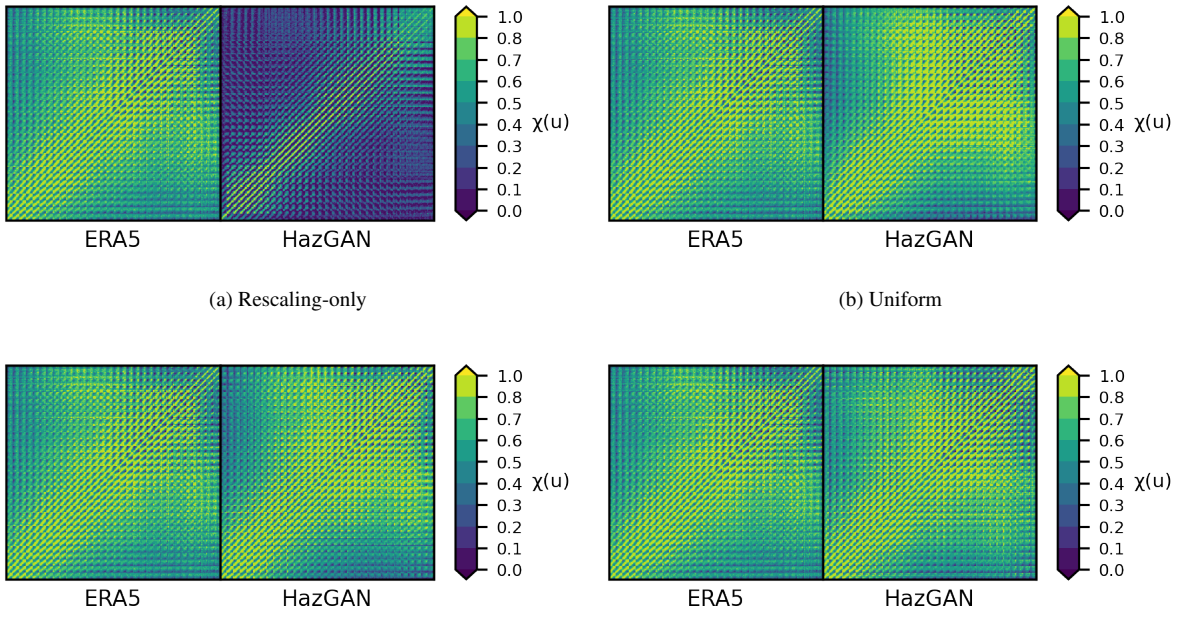
(d) Gumbel

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.

compare their distributions in Figure 8. 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).

Spatial dependence structures

To ~~check~~ assess how the model is learning the spatial dependence structures, we estimate the tail dependence coefficients $\hat{\chi}(u)$ (choosing $u = 0.8$ based on initial data exploration) between all pairs of pixels across the domain, generating for each variable a 4096×4096 matrix for each of the four generated datasets (Fig. 9). 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 extremal correlation matrices. The rescaled model performs worst, with a correlation of 0.380 (MAE = 0.309) between the ERA5 and GAN-generated correlation fields for wind speed, while the uniform, Gaussian, and Gumbel models perform much better, with correlations of 0.839 (MAE = 0.088), 0.837 (MAE = 0.089), and 0.857 (MAE = 0.083), respectively.



(a) Rescaling-only

(b) Uniform

(c) Gaussian

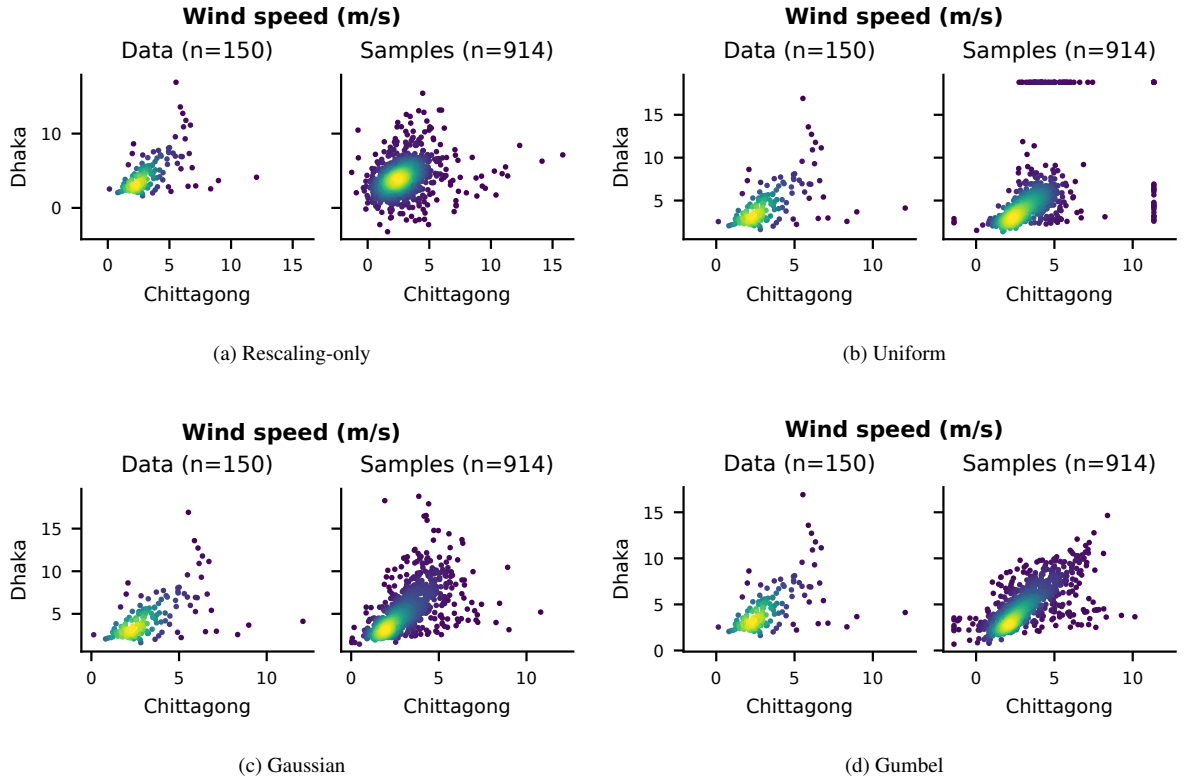
(d) Gumbel

Figure 9. Pairwise spatial extremal correlation estimates at $\hat{\chi}(0.8)$ for 10 m wind speed anomalies during storms across the Bay of Bengal.

For a more detailed view of bivariate spatial relationships, Fig. 10 shows scatter plots of 10 m wind speed anomalies between Chittagong and Dhaka, two cities in Bangladesh. The rescaling-only model produces a noticeably simpler, more ellipsoidal dependence structure than is seen in the ERA5 data, 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.

Multivariate dependence structures

Figure 11 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 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 = 0.104), 0.666 (MAE = 0.089) and 0.705 (MAE = 0.08), respectively. The rescaling-only and uniform-trained model performs much worse, with correlation of 0.193 (MAE = 0.117). ~~These~~ In all cases, the generated data had a tendency to overestimate the strength of the inter-variable dependence, with the mean overestimation (across all pairs of variables and locations) of 0.292 (uniform), 0.256 (Gaussian), 0.205 (Gumbel), and 0.241 (rescaling-only). The Gumbel-trained model performs best here, with the lowest mean error and highest correlation with the training data, while the uniform-trained model performs worst. We observe that these scores



(a) Rescaling-only

(b) Uniform

(c) Gaussian

(d) Gumbel

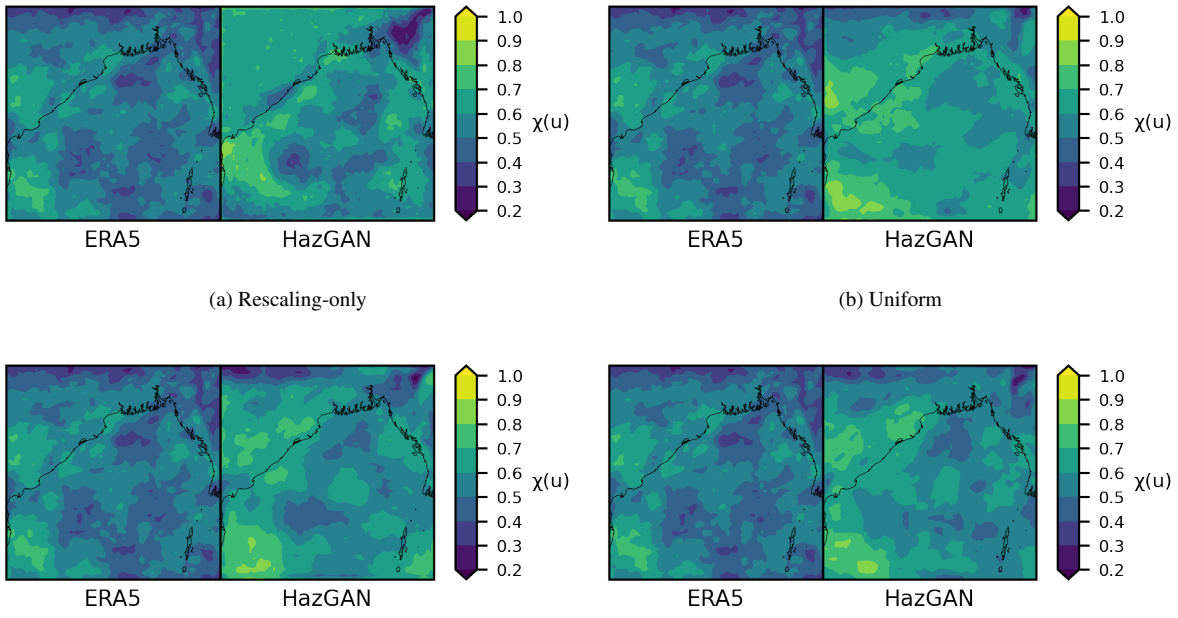
Figure 10. Scatter plots comparing the bivariate distribution of 10 m wind speed anomalies between the cities of Dhaka and Chittagong in Bangladesh.

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 Heffernan and Tawn (2004) model

We choose the Gaussian model as the best-performing model and compare its performance to the Heffernan and Tawn (2004) conditional exceedance model, which is provided in R's *texmex* package (Southworth et al., 2024). In this package, the *mexDependence* function integrates the GPD fitting of Eq. (3.2) and the fitting of the dependence model Heffernan and Tawn (2004, Eq. 4.1) into a single function, but our data has already been transformed to a uniform scale using Eq. (1), 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 1,000 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 $\hat{\chi}(0.8)$



(a) Rescaling-only

(b) Uniform

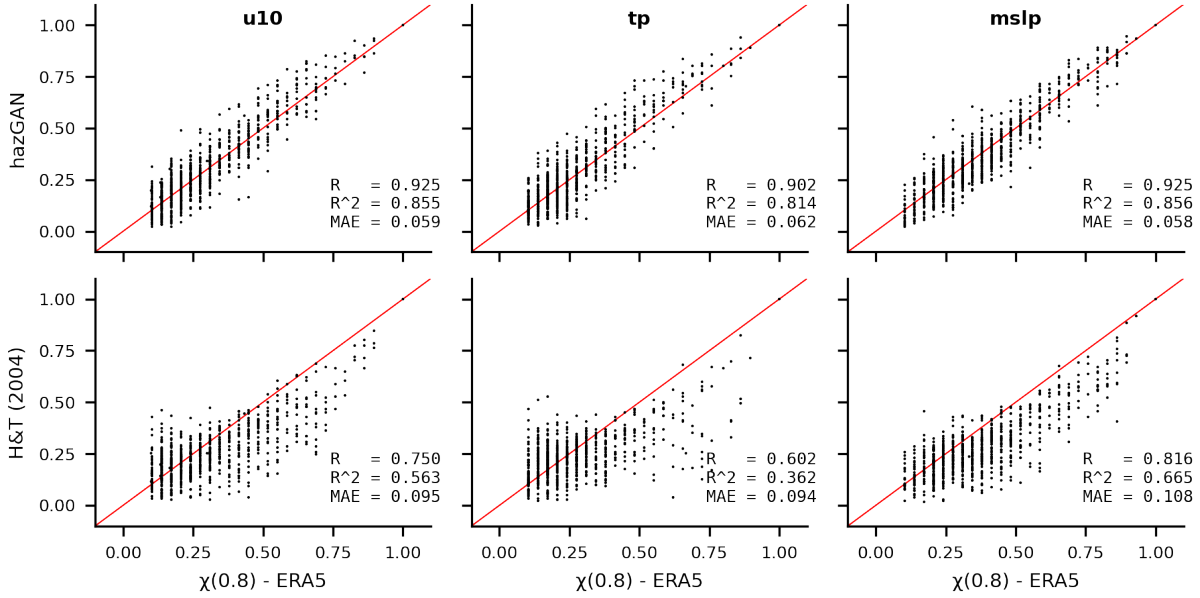
(c) Gaussian

(d) Gumbel

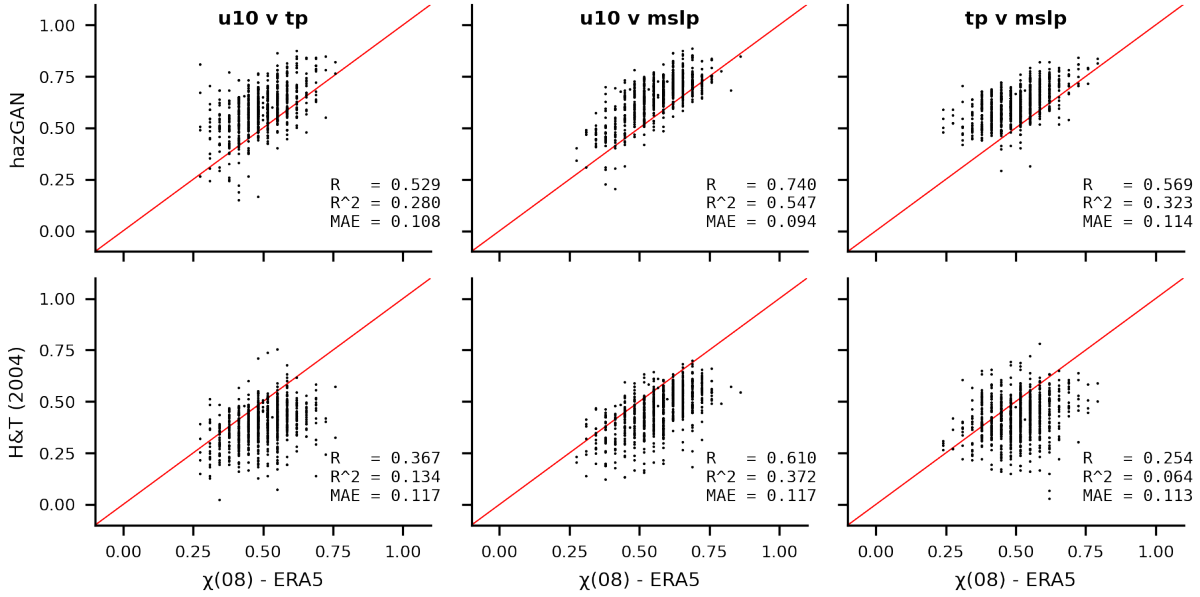
Figure 11. Inter-variable extremal correlation estimates $\hat{\chi}(0.8)$ for 10 m wind speed and total precipitation over the Bay of Bengal.

between the generated and training data. Figure 12(a) plots the distribution of spatial extremal correlations for each weather variable and 12(b) plots the inter-variable extremal correlations for all combinations of wind speed, precipitation, and sea-level pressure.

Both models ~~perform less in well in modelling~~ struggle more to model the inter-variable dependence structure than the spatial structure, suggesting that the inter-variable dependence structure is indeed ~~a more challenging joint distribution~~ more challenging to learn. Nonetheless, the GAN achieved higher correlations and lower MAE compared to the training data than the Heffernan and Tawn (2004) 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 Heffernan and Tawn (2004) model is learning 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 Heffernan and Tawn (2004) 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 (de Fondeville and Davison, 2018), which are designed to model ~~the full spatial dependence structure of extremes~~extremal spatial fields. However, we consider this beyond the scope of the current work, which aims only to demonstrate the effectiveness of the hazGAN approach.



(a) Rescaling-only



(b) Uniform

Figure 12. Comparison of how hazGAN and the Heffernan and Tawn (2004) model capture the spatial (top panel) and multivariate (bottom panel) dependence structures of the ERA5 training data. The x-axes show $\hat{\chi}(0.8)$ estimates for 1,000 randomly sampled pairs of variables in the training data and the y-axes show the $\hat{\chi}(0.8)$ estimates from the same variables in the synthetic data.

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

To illustrate an application in risk analysis, we use the event set generated by the Gaussian-based model to estimate risk to mangrove forests from storms 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.

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 (Mo et al., 2023; Bunting et al., 2022). The model predicts the probability that a mangrove patch is damaged, conditional on local winds and precipitation. A patch is defined as ~~"damaged"~~ 'damaged' if it experiences a drop in enhanced vegetation index (EVI) exceeding 20% in the aftermath of a storm. Further details of the 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. 3 as input to the mangrove 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~~ assumption implicit when return period hazard maps are treated like true events, while the independence assumption is implicit when regional risks are modelled separately and the results aggregated (Metin et al., 2020). 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~~neglected.

4.2 Results

Figure 13 shows the risk profile for widespread mangrove damages over the Bay of Bengal, plotting the expected area of mangrove forest damaged against event return period. The return period is calculated as a function of the total mangrove 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 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 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 2309.12 km^2 (23%) and a 100-year storm damages 2891.09 km^2 (29%). A

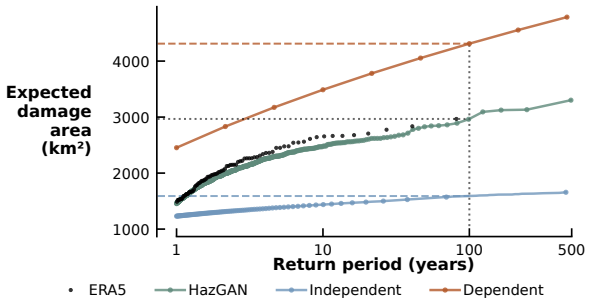


Figure 13. Risk profiles showing the expected total area of mangrove forest damaged by storms of different return periods for ERA5 and GAN-generated storm footprints. Also shown are risk profiles for the two simplifying assumptions of total dependence or independence across the domain, where a total 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 (Bunting et al., 2022), 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.

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

Method	Deviation	~~5 yr~~ 5-yr	~~10 yr~~ 10-yr	~~25 yr~~ 25-yr	~~50 yr~~ 50-yr	~~100 yr~~ 100-yr	MAE
HazGAN	Expected damage area (km^2)	-147.40	-172.92	-154.79	-8.85	-5.92	54.05
	Return period (years)	-0.11	-0.11	-2.48	8.64	17.28	0.02
Independent	Expected damage area (km^2)	-1066.51	-1218.01	-1277.82	-1289.83	-1376.78	484.37
	Return period (years)	-0.10	-0.19	-2.12	9.34	18.69	0.01
Dependent	Expected damage area (km^2)	722.18	832.63	1005.84	1217.94	1344.31	1062.16
	Return period (years)	-0.45	-0.19	-5.64	5.65	18.46	0.64

100-year storm under the total dependence assumption predicts damage to 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 33% of mangrove forest in the region (3301.93 km^2); the dependence assumption-generated data predicts damage to 48% of mangrove forests (4785.14 km^2); and the independence assumption-generated data predicts damage to only 16% of the mangrove forests (1660.71km^2).

Table 1 reveals that while the GAN-generated dataset appears to underestimate total expected damages with a mean absolute error of 54 km^2 across all return periods (reaching up to 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 484 km^2) and the dependent assumption dataset's systematic overestimation (mean absolute error of 1062 km^2).

To visualise the qualitative difference between modelling the dependence structure and assuming total or no dependence, Fig. 14 shows, for the ERA5 and synthetic datasets, a sample corresponding to a 1-in-75 year return period. ~~Realistic events~~

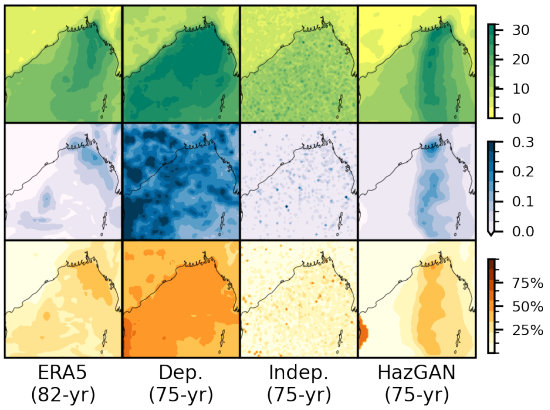


Figure 14. Footprints for events with approximately 1-in-75 year return period mangrove damages for the ERA5 and hazGAN-generated datasets, as well as samples from the two simplifying assumptions of total dependence or independence between all variables.

~~have clustering in the extremes~~ The extreme winds under realistic events are spatially clustered, while the dependent assumption ~~(hazard maps) distributes~~ distributed extreme winds evenly across the entire region~~and results~~ , resulting in far more widespread and unrealistic disruption estimates. The independence assumption shows no spatial coherence and so underestimates the impacts of spatially coherent events. By construction, the independence assumption exhibits no spatial coherence, hence underestimating the impact of spatially coherent events.

5 Discussion

~~In this paper, we have 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 Huster et al. (2021), Boulaguiem et al. (2022) and the statistical theory of multivariate extremes (Davison et al., 2012). 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~~

5.1 Main findings

A key result of this work was the insight that training a GAN (with Gaussian latents) on data transformed to have 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~~ margins resulted in improved representation of both marginal and joint tail behaviour when compared to GANs trained on rescaled data or uniform-transformed margins. The uniform transformation performed worst overall, likely because it compressed data in the tails, reducing sensitivity to marginal extremal behaviour. The rescaled transformation also performed poorly, which we hypothesise is due to the model needing to use more of

its capacity to learn the marginal distributions, which were rescaled but not otherwise standardised. The Gumbel and Gaussian transformations showed the best performance. 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 ~~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~~ StyleGAN2-ADA model used; however, benchmarking against the Heffernan and Tawn (2004) model showed a similar pattern, indicating that the poorer performance on multivariate dependence structures was due to the increased complexity of the task. Overall, the GAN trained on Gaussian margins achieved slightly better scores than the Heffernan and Tawn (2004) model for capturing the extremal dependence structure. This result is encouraging, particularly because the hazGAN model learns the full joint distribution of all 12,228 variables and locations simultaneously, whereas the Heffernan and Tawn (2004) model learns pairwise relationships separately so does not capture higher-order dependencies. In the future, it would be interesting to compare hazGAN to more recent models for spatials extremes, such as r-Pareto processes (de Fondeville and Davison, 2018). However, we consider this beyond the scope of the current work, which aims only to demonstrate the effectiveness of the hazGAN approach.

The method developed in this paper is suitable for generating the multivariate and spatially compounding varieties of compound hazards (Zscheischler et al., 2020), but does not address preconditioned or temporally compounding hazards. While the temporal problem is generally beyond the scope of this work, it may be possible to represent some types of preconditioning effects by simply using a longer temporal window for the temporal aggregation function $h_{k|t}(\mathbf{x})$. Letting, for example, one field represent 30-day cumulative rainfall preceding a storm. Otherwise, extending this method to capture either cascading events or the temporal dynamics within storms is more challenging and would require a significant extension of the method, involving a deep generative model that can handle an additional time dimension and new methods to extract events and standardise the margins.

5.2 Limitations, data quality, and marginal extremes

As training data, we used the ERA5 hourly gridded reanalysis product from 1940 to 2022 (Hersbach et al., 2023). ERA5 data was chosen as it is the most comprehensive and widely used reanalysis product currently available. However, like all reanalysis products, ERA5 has known biases and uncertainties. The 0.25° horizontal resolution, for example, cannot resolve many small-scale phenomena such as convective, turbulent, and dissipative processes, leading to parametrisations of varying quality (Steptoe and Economou, 2021; Alkhalidi et al., 2025), an underrepresentation of extremes, and omission of fine-scale details. Any 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 regional reanalysis or observational products—*e.g.*, IMDAA for the Indian Ocean (Rani et al., 2021) or HadUK for the United Kingdom (Hollis et al., 2019)—or to apply bias-correction and downscaling methods before training, although this would introduce additional uncertainties and complexities. Furthermore, this method could in theory be applied to data of any topology, provided a suitable deep generative model is used; an interesting avenue for future work would be to explore the applicability of the ~~well-known~~

~~conditional exceedance model of Heffernan and Tawn (2004). 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~~method to non-gridded datasets, such as point or graph data, which would enable the direct simulation of hazards over stations, river networks, or infrastructure networks.

The most critical limitation of the work so far is that it has only been tested on a historical dataset of 82 years, or 150 storms. This meant that validation was limited to comparisons with the training data and overfitting could not be properly assessed. Future work should focus on further validation of the method by testing it on large, controlled synthetic datasets of spatial extremes in order to fully assess the ability of the model to extrapolate far beyond the training data and capture pre-specified extremal dependence structures, tail behaviour, and check for overfitting (*e.g.* de Fondeville and Davison, 2018; Huser and Wadsworth, 202
.

~~In Sect. 2, 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.~~ Additionally, parametric fits to the extremes of small datasets of meteorological variables are known have high variability and biases (Harris, 2005; Serinaldi and Kilsby, 2014) , meaning there is signficant uncertainty in the fitted marginal distribution parameters. 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 most extremes in the training data. However, negative shape parameters were obtained for many onshore wind and sea-level pressure variables, indicating that they have upper-bounded tails. In previous studies, these negative GPD shape parameters for wind and rainfall have been demonstrated to be an artefact of mixed climates or finite sample sizes for the tails of distributions which converge to their final asymptotic form very slowly (Harris, 2005; Serinaldi and Kilsby, 2014), requiring very large samples for accurate estimation of the shape parameter. In future work, some of these effects might be mitigated by fitting a nonstationary model over the domain, allowing us to effectively 'borrow strength' between pixels and avoid mixed climates by using much higher thresholds (Davison et al., 2012; Huser and Wadsworth, 2022), although this requires further investigation.

5.3 Methods for deseasonalisation and event extraction

In the application to storms in the Bay of Bengal in Sect. 3, 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 (Dunlop, 2008) or the fire weather index for fire potential (Thompson et al., 2025; van Wagner et al., 1974).

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. 4), 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, but this has not been rigorously checked. 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, Cook et al., 2003; Cook, 2014; Zhang et al., 2018).

~~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. 5). 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 (Davison et al., 2012; Huser and Wadsworth, 2022, 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 (Harris, 2005). In this case a subasymptotic model such as, for example, XIMIS (Harris, 2009) may be the most appropriate approach to modelling the margins and we leave this as an interesting avenue for future work.~~

~~In~~

5.4 StyleGAN2 and computational resources

In Section 3, we used a StyleGAN2-ADA with differentiable augmentation to train a generative model on a set of 150 multi-hazard footprints (Karras et al., 2020; Zhao et al., 2020). We chose this model for its demonstrated ability to learn from small datasets, making it ~~appropriate~~ useful for modelling historical climata data~~, which generally contains fewer than 100 years of data (*e.g.*, Hersbach et al., 2023). It would however be interesting to investigate the applicability of alternative data-efficient deep generative models in this framework.~~

~~For training data,we used the ERA5 hourly gridded reanalysis from 1940 to 2022 (Hersbach et al., 2023). ERA5 data was chosen as it is the most comprehensive and widely used reanalysis product currently available. However, like all reanalysis products, ERA5 has known biases and uncertainties. The 0.25° horizontal resolution, for example, cannot resolve many small-scale phenomena such as convective, turbulent, and dissipative processes, leading to parametrisations of varying quality~~

(Steptoe and Economou, 2021; Alkhalidi et al., 2025), 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 (Rani et al., 2021) or HadUK for the United Kingdom (Hollis et al., 2019)—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 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.

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 Heffernan and Tawn (2004) model showed a similar pattern, indicating that the poorer performance on multivariate dependence structures was indeed due to this being a more challenging task.

Overall, the GAN trained on Gaussian margins achieved slightly better scores than the Heffernan and Tawn (2004) 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 Heffernan and Tawn (2004) 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 Heffernan and Tawn (2004) model fitted many separate models, each with less data. In the future, it learned the dependence structure between 150 samples of 12,228 variables in approximately four hours on a single NVIDIA GeForce GTX 1080 Ti GPU, and generated 914 new samples in under 30 minutes. There is potential to scale this up further: Zhao et al. (2020) demonstrated good results on 100-sample datasets of 256×256 RGB images (196,608 variables) and Karras et al. (2020) trained on datasets of a few thousand 1024×1024 RGB images (over 3 million variables), although at this point, fitting millions of generalised Pareto distributions becomes a computational bottleneck. While StyleGAN2-ADA is a powerful model, it was developed for image generation and may not be the optimal choice for modelling meteorological data. It would be interesting to compare the GAN to more recently developed spatial models for extremes, such as r-Pareto processes (de Fondeville and Davison, 2018), 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 approachinvestigate the applicability of alternative data-efficient deep generative models in this framework.

In terms of reprodicibility: , despite the use of a random seed, StyleGAN2-ADA has irreducible stochasticity, so . So while results will be similarbut , they are 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 methoddifferent models or training methods.

6 Conclusion

In this manuscriptarticle, 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 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.

References

Abdelmoaty, H. M., Papalexiou, S. M., Mamalakis, A., Singh, S., Coia, V., Hairabedian, M., Szeftel, P., and Grover, P.: Generative Adversarial Networks for Downscaling Hourly Precipitation in the Canadian Prairies, Journal of Geophysical Research: Machine Learning and Computation, 2, e2025JH000678, https://doi.org/10.1029/2025JH000678, 2025.

Alkhalidi, M., Al-Dabbous, A., Al-Dabbous, S., and Alzaid, D.: Evaluating the accuracy of the ERA5 model in predicting wind speeds across coastal and offshore regions, Journal of Marine Science and Engineering, 13, 149, https://doi.org/10.3390/jmse13010149, 2025.

Arjovsky, M., Chintala, S., and Bottou, L.: Wasserstein generative adversarial networks, in: International Conference on Machine Learning, pp. 214–223, PMLR, 2017.

Bader, B. and Yan, J.: eva: Extreme Value Analysis with Goodness-of-Fit Testing, 2020.

Bader, B., Yan, J., and Zhang, X.: Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate, Annals of Applied Statistics, https://doi.org/10.1214/17-AOAS1092, 2018.

Bailey, R., Benton, T., Challinor, A., Elliott, J., Gustafson, D., Hiller, B., Jones, A., Jahn, M., Kent, C., Lewis, K., et al.: Extreme ~~Weather and Resilience of the Global Food System, Final project report from the uk–us taskforce~~ weather and resilience of the global food system (2015). Final Project Report from the UK–US Taskforce on extreme weather and global food system resilience, The Global Food Security Programme, ~~UK,~~ 2015.

Bates, P. D., Savage, J., Wing, O., Quinn, N., Sampson, C., Neal, J., and Smith, A.: A climate-conditioned catastrophe risk model for UK flooding, Natural Hazards and Earth System Sciences, 23, 891–908, ~~2023.~~

Neal, and Smith]fathom ~~Bates, P. D., Savage, J., Wing, O., Quinn, N., Sampson, C., Neal, J., and Smith, A.: A climate-conditioned catastrophe risk model for UK flooding, Natural Hazards and Earth System Sciences, 23, 891–908,~~ https://doi.org/10.5194/nhess-23-891-2023, ~~2023.~~ 2023.

Becher, O., Pant, R., Verschuur, J., Mandal, A., Paltan, H., Lawless, M., Raven, E., and Hall, J.: A ~~Multi-Hazard Risk Framework to Stress-Test Water Supply Systems to Climate-Related Disruptions~~multi-hazard risk framework to stress-test water supply systems to climate-related disruptions, Earth's Future, 11, ~~e2022EF002946~~https://doi.org/10.1029/2022EF002946, 2023.

~~Bhatia, K., Vecchi, G., Murakami, H., Underwood, S., and Kossin, J.: Projected response of tropical cyclone intensity and intensification in a global climate model, Journal of climate, 31, 8281–8303, 2018.~~

Bhatia, S., Jain, A., and Hooi, B.: ExGAN: Adversarial generation of extreme samples, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 6750–6758, https://doi.org/10.1609/aaai.v35i8.16834, 2021.

Boulaguiem, Y., Zscheischler, J., Vignotto, E., van der Wiel, K., and Engelke, S.: Modeling and simulating spatial extremes by combining extreme value theory with generative adversarial networks, Environmental Data Science, 1, e5, https://doi.org/10.1017/eds.2022.4, 2022.

Bunting, P., Rosenqvist, A., Hilarides, L., Lucas, R., Thomas, T., Tadono, T., Worthington, T., Spalding, M., Murray, N., and Rebelo, L.-M.: Global Mangrove Extent Change 1996–2020: Global Mangrove Watch Version 3.0, Remote Sensing, https://doi.org/doi.org/10.3390/rs14153657, 2022.

Coles, S., Heffernan, J. E., and Tawn, J. A.: Dependence ~~Measures for Extreme Value Analyses~~measures for extreme value analyses, Extremes, 2, 339–365, https://doi.org/10.1023/A:1009963131610, 1999.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P.: An Introduction to Statistical Modeling of Extreme Values, vol. 208, Springer, 2001.

Cook, N.: Towards better estimation of extreme winds, Journal of Wind Engineering and Industrial Aerodynamics, 9, 295–323, 1982.

Cook, N. J.: Consolidation of analysis methods for sub-annual extreme wind speeds, Meteorological Applications, 21, 403–414, https://doi.org/10.1002/met.1355, 2014.

Cook, N. J., Harris, R. I., and Whiting, R.: Extreme wind speeds in mixed climates revisited, Journal of Wind Engineering and Industrial Aerodynamics, 91, 403–422, https://doi.org/10.1016/S0167-6105(02)00397-5, 2003.

Cross, D., Doyle, L., Dunning, P., Evans, D., Foster, N., Haseldine, L., MacDonald, A., Nix, B., Oldham, P., Smith, H., Styles, K., and Whitwham, C.: Global Flood Model: Technical Report, Tech. rep., JBA Risk Management Limited, 2020.

Davison, A. C., Padoan, S. A., and Ribatet, M.: ~~Statistical Modeling of Spatial Extremes~~Statistical modeling of spatial extremes, Statistical Science, 27, 161 – 186, https://doi.org/10.1214/11-STS376, 2012.

Davison, A. C., Huser, R., and Thibaud, E.: Spatial extremes, in: Handbook of Environmental and Ecological Statistics, edited by Gelfand, A. E., Fuentes, M., Hoeting, J. A., and Smith, R. L., CRC Press, 2019.

de Fondeville, R. and Davison, A. C.: High-dimensional peaks-over-threshold inference, Biometrika, 105, 575–592, 2018.

de Fondeville, R. and Davison, A. C.: Functional peaks-over-threshold analysis, Journal of the Royal Statistical Society Series B: Statistical Methodology, 84, 1392–1422, https://doi.org/10.1093/biomet/asy026, 2022.

Dunlop, S.: A dictionary of weather, OUP Oxford, 2008.

Engelke, S. and Ivanovs, J.: Sparse structures for multivariate extremes, Annual Review of Statistics and Its Application, 8, 241–270, https://doi.org/10.1146/annurev-statistics-040620-041554, 2021.

Faraway, J., Marsaglia, G., Marsaglia, J., and Baddeley, A.: goftest: Classical Goodness-of-Fit Tests for Univariate Distributions, https://CRAN.R-project.org/package=goftest, r package version 1.2-3, 2021.

Gaupp, F., Hall, J., Mitchell, D., and Dadson, S.: Increasing risks of multiple breadbasket failure under 1.5 and 2°C global warming, Agricultural Systems, 175, 34–45, https://doi.org/10.1016/j.agsy.2019.05.010, 2019.

Gilleland, E. and Katz, R. W.: extRemes 2.0: An Extreme Value Analysis Package in R, Journal of Statistical Software, 72, 1–39, https://doi.org/10.18637/jss.v072.i08, 2016.

Girard, S., Gobet, E., and Pachebat, J.: HTGAN: Heavy-tail GAN for multivariate dependent extremes via latent-dimensional control, International Journal of Computer Mathematics, pp. 1–41, https://inria.hal.science/hal-04700084, 2025.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.: Generative ~~adversarial nets,~~ Adversarial Nets, in: Advances in Neural Information Processing Systems, vol. 27, 2014.

Guillod, B. P., Jones, R. G., Dadson, S. J., Coxon, G., Bussi, G., Freer, J., Kay, A. L., Massey, N. R., Sparrow, S. N., Wallom, D. C., et al.: A large set of potential past, present and future hydro-meteorological time series for the UK, Hydrology and Earth System Sciences, 22, 611–634, https://doi.org/10.5194/hess-22-611-2018, 2018.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C.: Improved training of ~~wasserstein gans,~~ Wasserstein GANs, in: Advances in Neural Information Processing Systems, vol. 30, 2017.

Harris, I.: Generalised Pareto methods for wind extremes. Useful tool or mathematical mirage?, Journal of Wind Engineering and Industrial Aerodynamics, 93, 341–360, https://doi.org/10.1016/j.jweia.2005.02.004, 2005.

Harris, R. I.: XIMIS, a penultimate extreme value method suitable for all types of wind climate, Journal of Wind Engineering and Industrial Aerodynamics, 97, 271–286, https://doi.org/10.1016/j.jweia.2009.06.011, 2009.

Heffernan, J. E. and Tawn, J. A.: A conditional approach for multivariate extreme values (with discussion), Journal of the Royal Statistical Society Series B: Statistical Methodology, 66, 497–546, https://doi.org/10.1111/j.1467-9868.2004.02050.x, 2004.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on single levels from 1940 to present, https://doi.org/10.24381/cds.adbb2d47, (Accessed on 10-09-2023), 2023.

Ho, J., Jain, A., and Abbeel, P.: Denoising diffusion probabilistic models, in: Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851, https://arxiv.org/abs/2006.11239, 2020.

Hollis, D., McCarthy, M., Kendon, M., Legg, T., and Simpson, I.: HadUK-Grid—A new UK dataset of gridded climate observations, Geoscience data journal, 6, 151–159, 2019.

Hunt, K. M. R. and Bloomfield, H. C.: Landfalling tropical cyclones significantly reduce Bangladesh's energy security, EGUsphere, 2025, 1–20, https://doi.org/10.5194/egusphere-2025-4474, 2025.

Huser, R. and Wadsworth, J. L.: Advances in statistical modeling of spatial extremes, Wiley Interdisciplinary Reviews: Computational Statistics, 14, e1537, https://doi.org/10.1002/wics.1537, 2022.

Huser, R., Opitz, T., and Wadsworth, J. L.: Modeling of spatial extremes in environmental data science: Time to move away from max-stable processes, Environmental Data Science, 4, e3, https://doi.org/10.1017/eds.2024.54, 2025.

Huster, T., Cohen, J., Lin, Z., Chan, K., Kamhoua, C., Leslie, N. O., Chiang, C.-Y. J., and Sekar, V.: Pareto GAN: Extending the representational power of GANs to heavy-tailed distributions, in: International Conference on Machine Learning, pp. 4523–4532, PMLR, 2021.

Islam, T. and Peterson, R. E.: Climatology of landfalling tropical cyclones in Bangladesh 1877–2003, Natural Hazards, 48, 115–135, https://doi.org/10.1007/s11069-008-9252-4, 2009.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T.: Training generative adversarial networks with limited data, Advances in ~~neural information processing systems~~Neural Information Processing Systems, 33, 12 104–12 114, 2020.

Keef, C., Tawn, J., and Svensson, C.: Spatial dependence in extreme river flows and precipitation in Great Britain, Journal of Hydrology, 378, 240–252, https://doi.org/10.1016/j.jhydrol.2009.09.026, 2009.

Kingma, D. P. and Welling, M.: ~~Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114~~Auto-Encoding Variational Bayes, https://arxiv.org/abs/1312.6114, 2013.

Kobyzev, I., Prince, S. J., and Brubaker, M. A.: Normalizing flows: An introduction and review of current methods, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, pp. 3964–3979, https://arxiv.org/abs/1908.09257, 2021.

Lamb, R., Keef, C., Tawn, J., Laeger, S., Meadowcroft, I., Surendran, S., Dunning, P., and Batstone, C.: A new method to assess the risk of local and widespread flooding on rivers and coasts, Journal of Flood Risk Management, 3, 323–336, https://doi.org/10.1111/j.1753-318X.2010.01081.x, 2010.

Lamb, R., Garside, P., Pant, R., and Hall, J. W.: A probabilistic model of the economic risk to Britain's railway network from bridge scour during floods, Risk analysis, 39, 2457–2478, https://doi.org/10.1111/risa.13370, 2019.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4681–4690, 2017.

Lhaut, S., Rootzén, H., and Segers, J.: Wasserstein–Aitchison GAN for angular measures of multivariate extremes, https://arxiv.org/abs/2504.21438, 2025.

Liu, B., Zhu, Y., Song, K., and Elgammal, A.: Towards ~~Faster and Stabilized GANTraining for High-fidelity Few-shot Image Synthesis~~faster and stabilized GAN training for high-fidelity few-shot image synthesis., in: ~~iclr,~~ International Conference on Learning Representations, https://openreview.net/forum?id=1Fqg133qRaI, 2021.

Ljung, G. M. and Box, G. E.: On a measure of lack of fit in time series models, Biometrika, 65, 297–303, https://doi.org/10.1093/biomet/65.2.297, 1978.

Lloyd's: RDS 2025: Realistic Disaster Scenarios – Scenario Specification, Tech. Rep. EM 518 v1.0, Lloyd's Exposure Management, 2025.

Meiler, S., Vogt, T., Bloemendaal, N., Ciullo, A., Lee, C.-Y., Camargo, S. J., Emanuel, K., and Bresch, D. N.: Intercomparison of regional loss estimates from global synthetic tropical cyclone models, Nature Communications, 13, 6156, https://doi.org/10.1038/s41467-022-33918-1, 2022.

Metin, A. D., Dung, N. V., Schröter, K., Vorogushyn, S., Guse, B., Kreibich, H., and Merz, B.: The role of spatial dependence for large-scale flood risk estimation, Natural Hazards and Earth System Sciences, 20, 967–979, https://doi.org/10.5194/nhess-20-967-2020, 2020.

Mo, Y., Simard, M., and Hall, J. W.: Tropical cyclone risk to global mangrove ecosystems: potential future regional shifts, Frontiers in Ecology and the Environment, 21, 269–274, https://doi.org/10.1002/fee.2650, 2023.

Nelson, R.: An Introduction to Copulas, Springer, New York, 2006.

Obakrim, S., Benoit, L., and Allard, D.: A multivariate and space-time stochastic weather generator using a latent Gaussian framework, Stochastic Environmental Research and Risk Assessment, 39, 3677–3701, https://doi.org/10.1007/s00477-024-02897-8, 2025.

Papalexiou, S. M., Serinaldi, F., and Porcu, E.: Advancing Space-Time Simulation of Random Fields: From Storms to Cyclones and Beyond, Water Resources Research, 57, e2020WR029 466, https://doi.org/10.1029/2020WR029466, 2021.

Quinn, N., Bates, P. D., Neal, J., Smith, A., Wing, O., Sampson, C., Smith, J., and Heffernan, J.: The spatial dependence of flood hazard and risk in the United States, Water Resources Research, 55, 1890–1911, https://doi.org/10.1029/2018WR024205, 2019.

Radford, A., Metz, L., and Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks, ~~arXiv preprint arXiv:1511.06434~~https://arxiv.org/abs/1511.06434, 2015.

Rani, S. I., Arulalan, T., George, J. P., Rajagopal, E., Renshaw, R., Maycock, A., Barker, D. M., and Rajeevan, M.: IMDAA: High-resolution satellite-era reanalysis for the Indian monsoon region, Journal of Climate, 34, 5109–5133, https://doi.org/10.1175/JCLI-D-20-0412.1, 2021.

Schwarzwald, K. and Geil, K.: xagg: A Python package to aggregate gridded data onto polygons, Journal of Open Source Software, 9, 7239, https://doi.org/10.21105/joss.07239, 2024.

Serinaldi, F. and Kilsby, C. G.: Rainfall extremes: Toward reconciliation after the battle of distributions, Water Resources Research, 50, 336–352, https://doi.org/10.1002/2013WR014211, 2014.

Serinaldi, F., Bárdossy, A., and Kilsby, C. G.: Upper tail dependence in rainfall extremes: would we know it if we saw it?, Stochastic Environmental Research and Risk Assessment, 29, 1211–1233, https://doi.org/doi.org/10.1007/s00477-014-0946-8, 2015.

Southworth, H., Heffernan, J. E., and Metcalfe, P. D.: texmex: Statistical modelling of extreme values, r package version 2.4.9, 2024.

Speight, L., Hall, J., and Kilsby, C.: A multi-scale framework for flood risk analysis at spatially distributed locations, Journal of Flood Risk Management, 10, 124–137, https://doi.org/10.1111/jfr3.12175, 2017.

Stengel, K., Glaws, A., Hettinger, D., and King, R. N.: Adversarial super-resolution of climatological wind and solar data, Proceedings of the National Academy of Sciences, 117, 16 805–16 815, https://doi.org/10.1073/pnas.1918964117, 2020.

Steptoe, H. and Economou, T.: Extreme wind return periods from tropical cyclones in Bangladesh: insights from a high-resolution convection-permitting numerical model, Natural Hazards and Earth System Sciences, 21, 1313–1322, https://doi.org/10.5194/nhess-21-1313-2021, 2021.

Thompson, V., Mitchell, D., Melia, N., Bloomfield, H., Dunstone, N., and Kay, G.: Detecting rising wildfire risks for South East England, Climate Resilience and Sustainability, 4, e70 002, https://doi.org/10.1002/cli2.70002, 2025.

UNDRR: The Sendai Framework Terminology on Disaster Risk Reduction: Hazard, https://www.undrr.org/terminology/hazard, ~~accessed~~[Accessed: 9 February 2026], 2017.

van Wagner, C. E. et al.: Structure of the Canadian forest fire weather index, vol. 1333, Environment Canada, Forestry Service Ottawa, ON, Canada, 1974.

Wiese, M., Knobloch, R., and Korn, R.: Copula & marginal flows: Disentangling the marginal from its joint, ~~arXiv preprint arXiv:1907.03361~~https://arxiv.org/abs/1907.03361, 2019.

Wilks, D. S. and Wilby, R. L.: The weather generation game: a review of stochastic weather models, Progress in Physical Geography, 23, 329–357, https://doi.org/10.1177/030913339902300302, 1999.

Wilson, T., Tan, P.-N., and Luo, L.: DeepGPD: A Deep Learning Approach for Modeling Geospatio-Temporal Extreme Events, Proceedings of the AAAI Conference on Artificial Intelligence, 36, 4245–4253, https://doi.org/10.1609/aaai.v36i4.20344, 2022.

Yin, J., Gentine, P., Slater, L., Gu, L., Pokhrel, Y., Hanasaki, N., Guo, S., Xiong, L., and Schlenker, W.: Future socio-ecosystem productivity threatened by compound drought–heatwave events, Nature Sustainability, 6, 259–272, https://doi.org/10.1038/s41893-022-01024-1, 2023.

Zhang, S., Solari, G., Yang, Q., and Repetto, M. P.: Extreme wind speed distribution in a mixed wind climate, Journal of Wind Engineering and Industrial Aerodynamics, 176, 239–253, https://doi.org/10.1016/j.jweia.2018.03.019, 2018.

Zhao, S., Liu, Z., Lin, J., Zhu, J.-Y., and Han, S.: Differentiable augmentation for data-efficient ~~gan training, Advances in neural information processing systems,~~ GAN training, in: Advances in Neural Information Processing Systems, vol. 33, pp. 7559–7570, 2020.

Zscheischler, J., Westra, S., Van Den Hurk, B. J., Seneviratne, S. I., Ward, P. J., Pitman, A., AghaKouchak, A., Bresch, D. N., Leonard, M., Wahl, T., et al.: Future climate risk from compound events, Nature Climate Change, 8, 469–477, https://doi.org/10.1038/s41558-018-0156-3, 2018.

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M. D., et al.: A typology of compound weather and climate events, Nature Reviews Earth & Environment, 1, 333–347, https://doi.org/10.1038/s43017-020-0060-z, 2020.