

Regional climate risk assessment from climate models using probabilistic machine learning

In the format provided by the
authors and unedited

Contents

List of Supplementary Figures	4
List of Supplementary Tables	11
SI 1 Related work	13
SI 1.1 Supervised learning for super-resolution	13
SI 1.2 Empirical statistical downscaling methods	15
SI 1.3 Generative modeling approaches	15
SI 2 GenFocal identifies accurately tropical cyclones (TCs) in the North Atlantic	17
SI 2.1 Physically plausible TC tracks and structures	17
SI 2.2 Track density	17
SI 2.3 Counts and intensities of TCs	17
SI 2.4 Morphology of detected TCs tracks	19
SI 3 GenFocal models accurately multivariate spatiotemporal statistics	22
SI 3.1 Statistics of single variables	22
SI 3.2 Statistics of derived variables	23
SI 3.3 Extreme statistics of joint distributions	24
SI 3.4 Spatial correlations	25
SI 3.5 Temporal correlations	25
SI 3.6 Statistics of heat streaks	26
SI 4 Future climate risk assessment	36
SI 4.1 Changes in summer temperatures over the Western U.S.	36
SI 4.2 Changes in North Atlantic tropical cyclone activity	36
SI 5 Statistical downscaling baselines	40
SI 5.1 Bias Correction and Spatial Disaggregation (BCSD)	40
SI 5.2 Seasonal Trends and Analysis of Residuals Empirical Statistical Downscaling model (STAR-ESDM)	41
SI 6 Data	42
SI 6.1 Input datasets	42
SI 6.2 Modeled variables	42
SI 6.2.1 Debiasing	43
SI 6.2.2 Super-resolution	43
SI 6.3 Regridding	43
SI 7 Evaluation metrics	44
SI 7.1 Pointwise distribution errors	44
SI 7.1.1 Mean absolute bias (MAB)	44
SI 7.1.2 Mean Wasserstein distance (MWD)	44
SI 7.1.3 Percentile mean absolute error (MAE)	45
SI 7.2 Correlations	45
SI 7.2.1 Spatial correlation	45
SI 7.2.2 Spatial spectrum	46

SI 7.2.3	Temporal spectrum	47
SI 7.3	Tail dependence	47
SI 8	Evaluation protocol	49
SI 8.1	Derived variables	49
SI 8.2	Heat streaks	50
SI 8.3	Tropical cyclone detection	50
SI 8.3.1	Criteria	50
SI 8.3.2	TCG index	51
SI 8.3.3	Calibration	51
SI 8.3.4	Characteristics	51
SI 9	GenFocal: methodology and implementation details	53
SI 9.1	Main idea	53
SI 9.2	Setup	54
SI 9.3	Overview	54
SI 9.4	Bias correction	55
SI 9.4.1	Rectified flow	56
SI 9.4.2	Modeling details	56
SI 9.4.3	Neural architecture	57
SI 9.4.4	Hyperparameters	59
SI 9.4.5	Training, evaluation and test data	59
SI 9.4.6	Computational cost	60
SI 9.5	Super-resolution	60
SI 9.5.1	Conditional diffusion model	60
SI 9.5.2	Modeling details	61
SI 9.5.3	Sampling long temporal sequence	62
SI 9.5.4	Neural architecture	63
SI 9.5.5	Hyperparameters	65
SI 9.5.6	Training, evaluation, and test data	65
SI 9.5.7	Computational cost	65
SI 9.6	GenFocal Variants	66
SI 9.6.1	Direct Super-Resolution (SR)	66
SI 9.6.2	Quantile Mapping Super-Resolution (QMSR)	67
SI 10	Ablation studies: model selection and design choices	68
SI 10.1	Training period for the debiasing stage	68
SI 10.2	Length of the debiasing sequence	77
SI 10.3	Number of debiased variables	81
SI 10.4	Number of training steps	83
SI 10.5	Number of ensemble members	85
SI 10.6	Training data coupling in the debiasing stage	85
SI 10.7	Training period for the super-resolution stage	88
SI 10.8	Temporally coherent denoising in super-resolution	88
SI 10.9	Residual modeling in super-resolution	92
SI 11	Additional studies	95
SI 11.1	Definition of Events	95

SI 11.2	Pixel-wise statistics	95
SI 11.3	Multi-day frostbite episodes	96
SI 11.4	Tail Dependencies	96

List of Supplementary Figures

Figure SI 1	Ten meter wind speed for August 28, 2005. Daily average of Global 10 meter wind speed for the day of August 28 2005, from two ensemble members of LENS2 and ERA5, where we can observe the substantial differences between the different samples, particularly, as hurricane Katrina (a red blob next to Florida in the ERA5 sample) is absent from the climate samples.	13
Figure SI 2	Evolution of a hurricane downscaled by GenFocal. Plots of 10-meter wind speed at 60-hour intervals for a Category 1 hurricane projected by GenFocal. Colored dots track the tropical cyclone eye and its intensity in the Saffir-Simpson scale. The tropical cyclone evolves from a depression (brown) to a storm (orange) and ultimately a Category 1 hurricane (yellow).	17
Figure SI 3	TC tracks and their density. Tracks and their density for a LENS2 member in the North Atlantic in the time period 2010-2019 (a), for the same member we show a sample generated by GenFocal (b), BCSD (c), STAR-ESDM (d), SR (e) and QMSR (f). The observed tracks from the ERA5 reanalysis are shown in Main Text Fig. 2c. Note other methods (including the original LENS2) have the unrealistic concentration of TCs over Venezuela and the Pacific Coast of Mexico as well as the unphysical tracks over the Sahara desert.	18
Figure SI 4	Distribution of TC counts. Distributions of North Atlantic TC counts in the August-September-October season of 2010-2019 for the raw and downscaled LENS2 ensemble (100 members), using the different methods considered, and the ERA5 ground truth.	18
Figure SI 5	Distributions of TC intensity. Saffir-Simpson hurricane wind scale distributions for detected tropical cyclones in the North Atlantic in the August-September-October period during 2010-2019. Errors bars denote one ensemble standard deviation, computed using 100 and 800 members for LENS2 and GenFocal, respectively.	19
Figure SI 6	Distribution of decadal storm counts. Histogram of decadal storm counts produced by the LENS2 ensemble, the count distribution predicted using the tropical cyclogenesis (TCG) index, and the count distribution in the GenFocal ensembles.	19
Figure SI 7	Distribution of TC wind speeds. Distributions of the pressure-derived wind speed of tropical cyclones detected in the North Atlantic basin in the August-September-October period during 2010-2019, for LENS2 (a), GenFocal (b), BCSD (c), STAR-ESDM (d), SR (e), and QMSR (f). In addition, we also add the distribution of the pressure-derived wind speed for the reference ERA5 dataset. The confidence intervals are computed across the ensemble dimension using a Gaussian assumption (2 standard deviations).	20
Figure SI 8	Distributions of TC track length. Distributions of the track lengths of detected tropical cyclones in the North Atlantic in the August-September-October period during 2010-2019, for LENS2 (a), GenFocal (b), BCSD (c), STAR-ESDM (d), SR (e), and QMSR (f). In addition, we also add the distribution of the track lengths detected in the reference ERA5 dataset.	20

Figure SI 9	Distributions of sinuosity indices. Distributions of the sinuosity indices of the detected tropical cyclones tracks in the North Atlantic in the August-September-October period during 2010-2019, for LENS 2 (a), GenFocal (b), BCSD (c), STAR-ESDM (d), SR (e), and QMSR (f).	21
Figure SI 10	Pointwise bias of weather variables. Pointwise bias over CONUS during the summers (June-August) of the evaluation period 2010-2019 for the 2 m temperature, 10 m wind speed, mean sea-level pressure and 2 m specific humidity for GenFocal (a-d), BCSD (e-h), STAR-ESDM (i-l), QMSR (m-p), and SR (q-t).	23
Figure SI 11	Pointwise Wasserstein distance. Pointwise Wasserstein distance between marginals over CONUS during the summer (June-August) for the evaluation period 2010-2019 for the 2 m temperature, 10 m wind speed, mean sea-level pressure and 2 m specific humidity for GenFocal (a-d), BCSD (e-h), STAR-ESDM (i-l), QMSR (m-p), and SR (q-t).	24
Figure SI 12	Error of the 95th percentile. Error of the 95 th percentile over CONUS during the summer (June-August) for the evaluation period 2010-2019 for the 2 m temperature, 10 m wind speed, mean sea-level pressure and 2 m specific humidity for GenFocal (a-d), BCSD (e-h), STAR-ESDM (i-l), QMSR (m-p), and SR (q-t).	25
Figure SI 13	Error of the 99th percentile. Error of the 99 th percentile over CONUS during the summer (June-August) for the evaluation period 2010-2019 for the 2m temperature, 10m wind speed, mean sea-level pressure and 2 m specific humidity for GenFocal (a-d), BCSD (e-h), STAR-ESDM (i-l), QMSR (m-p), and SR (q-t).	26
Figure SI 14	Spatial distribution of modeling errors. Spatial distribution of modeling errors for the relative humidity over CONUS during the summer (June-August) of the evaluation period 2010-2019. Pointwise bias, Wasserstein distance, and errors of the 95 th percentile and 99 th percentile are reported for GenFocal (a-d), BCSD (e-h), STAR-ESDM (i-l), QMSR (m-p), and SR (q-t).	27
Figure SI 15	Spatial distribution of modeling errors. Spatial distribution of modeling errors for the heat index, over CONUS during the summers (June-August) of the evaluation period 2010-2019. Pointwise bias, Wasserstein distance, and errors of the 95 th percentile and 99 th percentile are reported for GenFocal (a-d), BCSD (e-h), STAR-ESDM (i-l), QMSR (m-p), and SR (q-t).	27
Figure SI 16	Tail dependence of meteorological extremes. Tail dependence of pairs of meteorological extremes over period 2010-2019 from ERA5 and bias of down-scaling methods. Tail dependence shown for high temperature and humidity (a), high temperature and low humidity (b), high temperature and high wind (c), and high wind and low humidity (d). Tail dependence biases are shown for GenFocal (e-h), BCSD (i-l), STAR-ESDM (m-p), QMSR (q-t), and SR (u-x).	28
Figure SI 17	Spatial correlation for 2-m temperature. Spatial correlation for 2 m temperature around selected populous US cities, evaluated for all snapshots at 18:00 UTC. The color scale represents the correlation coefficient relative to the city (stars) within a $\pm 4^\circ$ longitude/latitude range.	29

Figure SI 18	Spatial correlation for 10 m wind speed. Spatial correlation for 10 m wind speed around selected populous US cities, evaluated for all snapshots at all times. The color scale represents the correlation coefficient relative to the city (stars) within a $\pm 4^\circ$ longitude/latitude range.	30
Figure SI 19	Spatial correlation for near-surface specific humidity. Spatial correlation for near-surface specific humidity around selected populous US cities, evaluated for all snapshots at all times. The color scale represents the correlation coefficient relative to the city (stars) within a $\pm 4^\circ$ longitude/latitude range.	31
Figure SI 20	Spatial correlation for heat index. Spatial correlation for heat index around selected populous US cities, evaluated for all snapshots at 18:00 UTC. The color scale represents the correlation coefficient relative to the city (stars) within a $\pm 4^\circ$ longitude/latitude range.	32
Figure SI 21	Spatial power spectra density. Spatial radial power spectra density (following SI 7.2.2), including the spectral error (30), for output variables generated with GenFocal and other methods.	32
Figure SI 22	Temporal power spectra density. Temporal power spectra density (following SI 7.2.3), including the spectral error (33), for a set of selected cities in CONUS and different variables for ensembles generated with GenFocal and other methods.	33
Figure SI 23	Bias in the number of heat streaks by different lengths. Bias in the number of heat-streaks per year for caution advisory considering different lengths. We show the ground truth (ERA5)(a-d), and the pointwise errors of GenFocal (e-h), BCSD (i-l), STAR-ESDM (m-p), QMSR (q-t) and SR (u-x).	33
Figure SI 24	Bias in the number of heat streaks by different lengths. Bias in the number of heat-streaks per year for extreme caution advisory considering different lengths. We show the ground truth (ERA5)(a-d), and the pointwise errors of GenFocal (e-h), BCSD (i-l), STAR-ESDM (m-p), QMSR (q-t) and SR (u-x).	34
Figure SI 25	Bias in the number of heat streaks by different lengths. Bias in the number of heat-streaks per year for danger advisory considering different heatwave lengths. We show the ground truth (ERA5)(a-d), and the pointwise errors of GenFocal (e-h), BCSD (i-l), STAR-ESDM (m-p), QMSR (q-t) and SR (u-x).	35
Figure SI 26	Projected change in daily mean near-surface temperature. Projected change in daily mean near-surface temperature in 11 cities across the Western United States, from 2017 to 2083. Results are computed as the average over $1^\circ \times 1^\circ$ regions, and 7 summers (June-August) centered around 2020 and 2080. Boxes for BCSD, STAR-ESDM, and GenFocal show the median (50th percentile) and interquartile range (25th and 75th percentiles) of an ensemble of 8 projections, and whiskers represent the ensemble minima and maxima.	37
Figure SI 27	Projected change in daily maximum near-surface temperature. Projected change in daily maximum near-surface temperature in 11 cities across the Western United States, from 2017 to 2083. Results are computed as the average over $1^\circ \times 1^\circ$ regions, and 7 summers (June-August) centered around 2020 and 2080. Boxes for BCSD, STAR-ESDM, and GenFocal show the median (50th percentile) and interquartile range (25th and 75th percentiles) of an ensemble of 8 projections, and whiskers represent the ensemble minima and maxima.	37

Figure SI 28	Projected change in the top decile of daily maximum near-surface temperature. Projected change in the top decile of the daily maximum near-surface temperature in 11 cities across the Western United States, from 2017 to 2083. Results are computed as the average over $1^\circ \times 1^\circ$ regions, and 7 summers (June–August) centered around 2020 and 2080. Boxes for BCSO, STAR-ESDM, and GenFocal show the median (50th percentile) and interquartile range (25th and 75th percentiles) of an ensemble of 8 projections, and whiskers represent the ensemble minima and maxima.	38
Figure SI 29	Change in TC landfall frequency and intensity. Change in TC landfall frequency and intensity over the first half of the 21 st century. a. Number of TC landfalls during the August–October season of years 2010–2019. b, c. Projected change in the number of TC landfalls from 2010–2019 to 2030–2039, and 2050–2059. d–f. Number and projected change in the number of tropical storm and hurricane landfalls. g–i. Number and projected change in the number of hurricane landfalls. j–l. Median maximum pressure-derived wind speed of TC landfalls and its projected change. m–o. 90 th percentile of maximum pressure-derived wind speed of TC landfalls and its projected change. All results are computed as the average over 800 downscaled projections. Wind speed changes (k, l, n, o) are only shown if statistically significant ($p < 0.05$ in a two-sided Mann–Whitney U test) and set to zero otherwise.	39
Figure SI 30	Schematic of long trajectory sampling using parallel section denoisers.	63
Figure SI 31	Summertime bias of GenFocal over CONUS. Bias of multiple fields during summers (June–August) of 2010–2019, for GenFocal models trained using data from different reference periods.	70
Figure SI 32	Summertime Wasserstein distance over CONUS. Pointwise Wasserstein distance (see SI 7.1.2) during summers (June–August) of 2010–2019, for GenFocal models trained using data from different reference periods.	71
Figure SI 33	Summertime 99th percentile error over CONUS. Pointwise error in the 99 th percentile of multiple fields during summers (June–August) of 2010–2019, for GenFocal models trained using data from different reference periods.	72
Figure SI 34	Metrics for relative humidity over CONUS. Metrics for the relative humidity, a derived variable, over CONUS during the summer (June–August) for the evaluation period 2010–2019. We include bias, Wasserstein error, error of the 95 th and 99 th percentiles for GenFocal trained using data from different reference periods.	73
Figure SI 35	Metrics for the heat index over CONUS. Metrics for the heat index, a derived variable, over CONUS during the summer (June–August) for the evaluation period 2010–2019. We include bias, Wasserstein error, error of the 95 th and 99 th percentiles for GenFocal trained using data from different training periods.	74
Figure SI 36	Distribution of TC occurrences. Distribution of the number of TCs detected by TempestExtremes in the North Atlantic during the peak hurricane season (August–October), 2010–2019. a. Distribution of TC counts for different decadal training periods. b. Distribution of the TC counts for training data periods of varying length. TCs from all GenFocal variants were calibrated separately following SI Section SI 8.3.3.	75

Figure SI 37	Saffir-Simpson scale distribution of TCs. Distribution of the intensity of TCs detected by TempestExtremes in the North Atlantic during the peak hurricane season (August–October), 2010–2019. Results shown for GenFocal models trained on different periods, and for ERA5. Error bars denote the ensemble standard deviation. TCs from all GenFocal variants were calibrated separately following SI Section SI 8.3.3.	75
Figure SI 38	TC tracks and their density. TC tracks for a single climate projection overlaying the ensemble TC track density for the peak North Atlantic hurricane season (August–October), 2010–2019. a. LENS2. b-i. GenFocal variants trained on different periods.. f. ERA5 tracks and their density. TCs from all GenFocal variants were calibrated following SI Section SI 8.3.3.	76
Figure SI 39	Downscaled variable biases. Biases of downscaled variables, over CONUS during the summer (June-August) for the evaluation period 2010-2019 for GenFocal trained with different debiasing sequence lengths.	77
Figure SI 40	Spatial distribution of heat index errors. Spatial distribution of statistical modeling errors for the heat index, one of the derived variables, over CONUS during the summer (June-August) for the evaluation period 2010-2019. We include bias, Wasserstein error, error of the 95 th and 99 th percentiles for GenFocal trained with different debiasing sequence lengths.	78
Figure SI 41	Spatial distribution of heat streak errors. Spatial distribution of statistical modeling errors in the number of heat-streaks per year for extreme caution advisory considering different debiasing sequence lengths (in days). We show the ground truth (ERA5)(a-d), and the pointwise errors of GenFocal with different lengths of the debiased sequence.	79
Figure SI 42	TC tracks and density. Tracks and their density for a LENS2 member in the North Atlantic in the time period 2010-2019 (a), and for a downscaled sample from the same member generated using GenFocal for different debiasing sequence length (b-e). Also shown are the tracks detected in the reference ERA5 data (f).	80
Figure SI 43	Distribution of TC frequency. Distribution of the number of TCs detected by TempestExtremes in the North Atlantic for the hurricane season August-September-October during 2010-2019, using downscaled data from GenFocal models with varying debiasing sequence lengths.	80
Figure SI 44	Distribution of TC frequency. Distribution of the number of TCs detected by TempestExtremes in the North Atlantic for the hurricane season August-September-October during 2010-2019 for GenFocal trained with different number of debiasing variables.	81
Figure SI 45	Tracks and their density. Tracks and their density for a LENS2 member in the North Atlantic in the time period 2010-2020 (a), one of downscaling samples from the same member generated using GenFocal trained with different number of debiased variables sets (b-d).	82
Figure SI 46	Spatial distribution of biases. Spatial distribution of statistical biases of the downscaled variables over CONUS during the summer (June-August) for the evaluation period 2010-2019 for GenFocal trained with different number of steps.	83

Figure SI 47	Distribution of TC counts. Distribution of the number of TCs detected by TempestExtremes in the North Atlantic for the hurricane season August-September-October during 2010-2019 for GenFocal trained with different number of steps.	84
Figure SI 48	TC tracks and density. TC tracks and their density for a LENS2 member in the North Atlantic in the time period 2010-2020 (a), one of downscaling samples from the same member generated using GenFocal trained with trained with different number of steps (b-e), tracks detected using the reference ERA5 data (f).	84
Figure SI 49	Metrics of heat index. Metrics of the derived variable heat index over CONUS during the summer (June-August) for the evaluation period 2010-2019 for GenFocal trained with different number of LENS2 ensemble members.	86
Figure SI 50	Bias in extreme caution advisories. Spatial distribution of mean absolute errors in the number of extreme caution advisory streaks of different lengths per year. We show the ground truth (ERA5)(a-d), and the pointwise errors of GenFocal trained with different number of LENS2 members.	87
Figure SI 51	Spatial distribution of biases. Spatial distribution of statistical biases of the downscaled variables over CONUS during the summer (June-August) for the evaluation period 2010-2019 for GenFocal with different training data couplings.	89
Figure SI 52	Spatial distribution of different errors for the heat index. Spatial distribution of different errors for the downscaled heat index compound variables over CONUS during the summer (June-August) for the evaluation period 2010-2019 for GenFocal with different training data couplings.	90
Figure SI 53	Distribution of TC counts for different training couplings. Distribution of the number of TCs detected in the North Atlantic for the hurricane season August-October during 2010-2019 for GenFocal with different training data couplings.	91
Figure SI 54	Saffir-Simpson scale distribution of TCs. Distribution of the intensity of TCs detected in the North Atlantic during the peak hurricane season (August-October), 2010-2019. Results shown for GenFocal models trained with different couplings, for small differences (in days) of the time stamps. Error bars denote the ensemble standard deviation.	91
Figure SI 55	Saffir-Simpson scale distribution of TCs. Distribution of the intensity of TCs detected in the North Atlantic during the peak hurricane season (August-October), 2010-2019. Results shown for GenFocal models trained with different couplings, in which the day-of-the-year remains fixed and we sample from similar years. Error bars denote the ensemble standard deviation.	91
Figure SI 56	TC statistics for different super-resolution model training ranges. The same debiasing model (trained on 1980-1999) is shared between all GenFocal variants.	92
Figure SI 57	Spatially averaged bias in the number of heat streaks vs. durations for selected thresholds. GenFocal 7-day and 3-day represent variants whose super-resolution step is trained on samples of different lengths, with time-coherent sampling (1-day overlap) applied at inference time.	93

Figure SI 58	Distribution of North Atlantic TC counts for super-resolution models with varying training sample lengths and time-coherent sampling. a. super-resolution models trained with <i>7-day vs. 3-day samples</i> ; sampled with time coherence. b. super-resolution model trained with 7-day samples, <i>with vs. without time-coherent sampling</i> . c. super-resolution model trained with 3-day samples, <i>with vs. without time-coherent sampling</i>	94
Figure SI 59	Frostbite streak statistics in CONUS winter (December-January-February) season, lasting [1, 3, 5, 7] days. a-d. ERA5 reanalysis. e-h. local bias of GenFocal. i-l. local bias of BCSD. and m-p. local bias of STAR-ESDM.	98
Figure SI 60	Tail dependence of low surface temperature and high wind speed in CONUS winter (December-January-February) season. a. ERA5 reanalysis. b. GenFocal c. BCSD and d. STAR-ESDM.	98

List of Supplementary Tables

Table SI 1	Summary of downscaling methods with aligned data.	14
Table SI 2	Summary of downscaling methods for unaligned data.	15
Table SI 3	Statistical modeling errors of directly downscaled variables in marginal distributions by different methods for the summers (June-July-August) in CONUS during 2010-2019. Best highlighted in bold.	22
Table SI 4	Statistical modeling errors of derived variables by different models for the summers (June-August) in CONUS during 2010-2019.	24
Table SI 5	Meteorological fields modeled by GenFocal with their corresponding variable names in the ERA5 and LENS2 datasets. All 10 fields serve as both input and output for the debiasing model, while the super-resolution model uses the top 4 fields in CONUS and top 6 fields in the North Atlantic. Units reflect those used for model training and are converted as needed in the main text.	42
Table SI 6	Calibration scaling constant (K) for Tropical Cyclone (TC) detection. Values were chosen for best fit to TC count, track length, and lifetime in the training period (from $1/K \in \{0.1, 0.2, \dots, 0.9\}$).	51
Table SI 7	Hyperparameters for the debiasing model.	59
Table SI 8	Hyperparameters for the super-resolution model.	66
Table SI 9	Effect of training data period on the mean absolute bias, Wasserstein distance, and absolute error of the 99 th percentile for the summers (June-July-August) of 2010-2019 in CONUS. The precise definitions of the metrics are included in Section SI 7.	69
Table SI 10	Effect of debiasing sequence length on mean absolute bias, mean Wasserstein distance, and mean absolute error in the 99 th percentile for different variables during summers of 2010-2019 (June-July-August) over CONUS. The metrics are defined in SI 7.	78
Table SI 11	Effect of the number of debiased variables on mean absolute bias, mean Wasserstein distance, and mean absolute error in the 99 th percentile of the downscaled output for different variables. The metrics are defined in SI 7.	81
Table SI 12	Effect of the number of training steps on mean absolute bias, mean Wasserstein distance, and mean absolute error in the 99 th percentile for different variables. The precise definitions of the metrics are included in SI 7.	83
Table SI 13	Indices of the different LENS2 members used for training.	85
Table SI 14	Effect of the number of LENS2 members used during training on mean absolute bias, mean Wasserstein distance, and mean absolute error in the 99 th percentile for different variables. The precise definitions of the metrics are included in SI 7.	85
Table SI 15	Effect of training data coupling on the mean absolute bias, mean Wasserstein distance, and mean absolute error of the 99 th percentile. Results are shown for various coupling strategies applied to the CONUS summer seasons (June–August) from 2010 to 2019. Metric definitions are provided in section SI 7.	88
Table SI 16	Impact of time-coherent sampling (section SI 9.5.3) on the mean temporal spectra error (TSE). Columns represent variants with the super-resolution step trained on 7-day or 3-day samples, and time-coherent sampling enabled (+) or not (-) during inference.	92

Table SI 17	Effect residual modeling on mean absolute bias, mean Wasserstein distance, and mean absolute error in the 99 th percentile for different variables. The precise definitions of the metrics are included in SI 7.	93
Table SI 18	Statistical modeling errors of directly and derived downscaled variables in marginal distributions for winters (December-January-February) in CONUS (2010-2019). GenFocal consistently outperforms baselines in capturing the distribution shapes (Wasserstein) and extremes (99 th percentile MAE). Best values are in bold font. The precise definitions of the metrics are included in SI 7	96
Table SI 19	Statistical errors for the number of multi-day frostbite episodes duration evaluation period (December-January-February, CONUS 2010-2019) using mean absolute bias, mean continuous ranked probability score (CRPS), and mean spread-skill ratio (SSR). Best values are in bold font.	97

SI 1 Related work

Existing ML downscaling methods predominantly rely on temporal alignment between low- and high-resolution training data [56, 50, 59, 109, 110, 53, 103, 87], which enables treating the problem as a supervised learning task. This framework has proved effective for downscaling weather forecasts [56] and regional climate model emulation [50]. However, this supervised learning approach cannot be used to learn a downscaling model for free-running climate projections to be consistent with observation assimilated historical datasets such as ERA5 [10], since there is no temporal alignment between the coarse projections and the high-resolution data (e.g., see Fig. SI 1). GenFocal adopts an original learning framework to address this challenge of mapping between unpaired data.

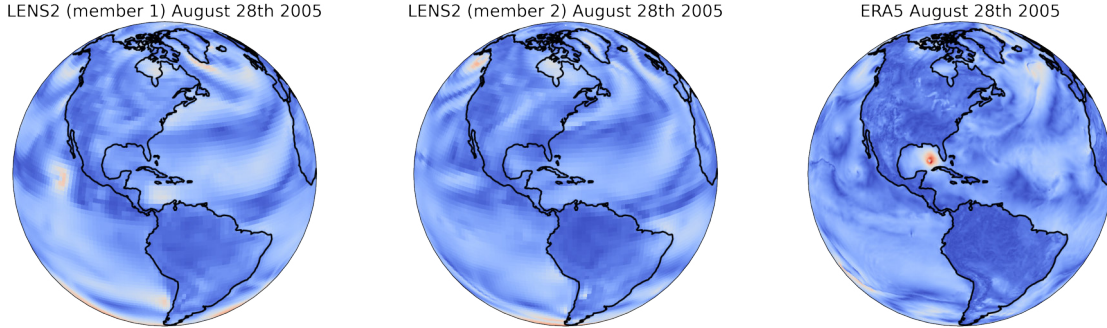


Figure SI 1: Ten meter wind speed for August 28, 2005. Daily average of Global 10 meter wind speed for the day of August 28 2005, from two ensemble members of LENS2 and ERA5, where we can observe the substantial differences between the different samples, particularly, as hurricane Katrina (a red blob next to Florida in the ERA5 sample) is absent from the climate samples.

GenFocal also stands in stark contrast to the traditional dynamical downscaling paradigm of using regional climate models (RCMs) to generate high-resolution climate data [26]. Due to its high computational cost (even at regional scale, sacrificing spatial coverage), the use of dynamical downscaling is limited to small climate-projection ensembles, thus compromising their ability to capture the risk of climate extremes [28]. Alternative statistical downscaling approaches are more efficient [70, 33] but come at the cost of highly customized designs for specific use cases [69] or fail to capture the full range of spatiotemporal correlations between meteorological fields that characterize climate [13]. Those methods are highly bespoke: methods used in hydrology [106] are markedly different from those used for tropical cyclone analysis [40]. This inflexibility limits the value they add to coarse climate projections, compared to the more flexible but expensive physics-based approaches.

In what follows, we review the existing literature on downscaling. In the interest of broad coverage, we also briefly review supervised learning methods for downscaling on temporally aligned data, namely, super-resolution, see Tables SI 1 and SI 2 for a summary.

SI 1.1 Supervised learning for super-resolution

Supervised learning is the most direct approach for downscaling aligned data, where a model learns a mapping from low- to high-resolution data using paired samples [56, 95]. Recent works using this approach are summarized in Table SI 1. These supervised methods can be broadly categorized as either deterministic or probabilistic. Most deterministic models operate under the “perfect prognosis” assumption [99], which assumes the low-resolution samples are unbiased (see [74] for an extensive

Table SI 1: Summary of downscaling methods with aligned data.

Study	Technique Used	Resolution (Low $\rightarrow$ High)	Variables Downscaled
[103]	Diffusion Model	2° (Coarsened ERA5) $\rightarrow$ 0.25° (ERA5)	2m Temperature, 10m Winds
[56]	Corrector Diffusion Model	25 km (ERA5) $\rightarrow$ 2 km (WRF Model)	2m Temperature, 10m Winds, Radar Reflectivity
[68]	Generative Diffusion Model for CAM Emulation	$\sim$ 28 km (ERA5) $\rightarrow$ 3 km (HRRR Model)	Full atmospheric state
[59]	Diffusion model	25 km (ERA5) $\rightarrow$ 5.5 km (CERRA)	Surface Temperature, Wind Speed, Geopotential Height
[87]	Video Diffusion	200 km (FV3GFS [112]) $\rightarrow$ 25 km (FV3GFS [112])	Precipitation
[109]	Generative Diffusion Model (CloudDiff)	2 km (Himawari-8 AHI [61]) $\rightarrow$ 1 km (MODIS)	Cloud related variables
[95]	Latent Diffusion Model	16 km (ERA5) $\rightarrow$ 2 km (COSMO_CLM)	2m Temperature, 10m Wind
[110]	Wavelet Diffusion Model	10 km (MRMS) $\rightarrow$ 1 km (MRMS)	Precipitation (Composite Reflectivity)
[99]	Super-Resolution CNN	$\sim$ 1.25° (MERRA-2) $\rightarrow$ 0.25° (CPC Obs.)	Daily Precipitation
[5]	CNN (DeepESD) under Perfect Prognosis	2° (ERA-Interim) $\rightarrow$ 0.5° (E-OBS)	Daily Temperature & Precipitation
[38]	ResNets (VDSR, EDSR) vs. SRCNN	2.5° (ERA5) $\rightarrow$ 0.25° (ERA5)	2m Temperature
[44]	AI-NWP (Pangu-Weather)	250 km (CMIP6) $\rightarrow$ 31 km (ERA5-like)	Full atmospheric state (focus on 2m Temperature)
[31]	Hard-Constrained Deep Learning	9 km (WRF) $\rightarrow$ 3 km (WRF)	Total Column Water, Temperature, Water Vapor, Liquid Water

review). These models seek to capture correlations between large-scale meteorological fields and local observations by training on time-aligned samples from data-assimilated simulations (e.g., ERA5) and local observational data [6, 7, 5]. The tendency of deterministic models to smoothen outputs and dampen extremes by collapsing to the conditional mean [60] has motivated the development of probabilistic methods, which typically employ generative models such as diffusion models [59, 109, 110, 53, 103] or GANs [32, 71].

Table SI 2: Summary of downscaling methods for unaligned data.

Study	Technique Used	Resolution (Low $\rightarrow$ High)	Variables Downscaled
GenFocal	Rectified Flow + Diffusion Model	1.5° (LENS2) $\rightarrow$ 0.25° (ERA5)	2m Temperature, Surface Humidity, Sea-level Pressure, 10m Winds
[30]	Nearest neighbor interpolation + Normalizing flows + CycleGAN loss	1° (NOAA20CR [16]) $\rightarrow$ 0.125° (Livneh Obs [48])	Precipitation, Daily max temperature
[10]	Filtered-weighted nearest neighbor + Diffusion Bridge	2D forced turbulent fluid and a supersaturation tracer	Vorticity, Super-saturation

One way to leverage these supervised learning approaches for climate downscaling is to enforce temporal alignment between the climate models and the high-resolution data. This can be achieved by dynamical downscaling [50, 95], or by nudging the coarse climate model of interest to follow the historical weather record [14, 19]. These approaches come with their own limitations: they require access to a well-calibrated regional climate model in the case of dynamical downscaling, and modifying the input climate model in the case of nudging climate projection. Apart from these technical barriers, both methods require running additional climate simulations that are typically computationally expensive. GenFocal overcomes the restriction of temporal alignment and addresses head-on the challenge of learning to downscale coarse climate projections to be consistent with historical weather data.

SI 1.2 Empirical statistical downscaling methods

The unpaired data assumption has been a staple of weather and climate science for decades, with many methods developed to address it [104, 66, 34, 70]. In the weather and climate literature (see [98, 12] for extensive overviews), prior knowledge can be exploited to downscale specific variables [104, 66]. Two of the most predominant methods of this type are bias correction and spatial disaggregation (BCSD) [106, 105], which combines traditional spline interpolation with a quantile matching bias correction [55] and a disaggregation step—and linear models [36]. More recent methods extend this type of methodology by adding climate signal corrections with different timescales to the climatology [34], or by adjusting historical weather analogs to follow an evolving climate [70]. The statistical simplicity of these methods comes with limitations: they do not explicitly model spatial or inter-variable correlations, which hinders their use for risk analysis of derived weather variables, such as the heat index.

SI 1.3 Generative modeling approaches

The bias between the low-resolution and the high-resolution one can be seen as a distribution mismatch from characterizing differently the same underlying system. Thus, removing the bias can be framed as an instance of unsupervised domain adaptation [29], a topic popularly studied in computer vision. Recent work has used generative models such as GANs and diffusion models to bridge the gap between two domains, usually under the umbrella of image-to-image translation or neural style transfer [11, 58, 65, 67, 79, 89, 108, 111].

Similar to ours, several recent works have adopted this view to perform debiasing. However, unlike ours, those methods increase resolution *first*, then debias. GenFocal instead debiases first, then upsample. This deliberate design is methodologically sound and avoids several pitfalls:

- Upsampling techniques such as interpolation, may incur aliasing [30, 10] and introduce compounded errors when debiasing.
- Upsampling requires the low-frequency power spectra of the two datasets match [10], precisely the problem of bias that we need to solve.

Another notable innovation is that GenFocal performs effectively temporal super-resolution with temporal coherence. In contrast, other approaches only sample snapshots and thus do not impose temporal coherence in the resulting sequences, a necessary requirement to capture accurately statistics of extended events such as tropical storms and extreme heat waves [10, 30, 101].

In practice, debiasing first at low-resolution is computationally less expensive. As GenFocal leverages diffusion models to perform super-resolution, we avoid well-known issues associated with GANs [93, 4] and normalized-flows based methods [52], which include over-smoothing, mode collapse, and large model footprints [18, 45].

Those methodological and computational advantages allow us to downscale large sequence of high-resolution (in both space and time) samples with ease. This task is beyond the capabilities of the methods mentioned above. (In our earlier work, we have compared a version of GenFocal, which outperforms those methods on flow problems that are feasible for them. For details, please see [101].)

SI 2 GenFocal identifies accurately tropical cyclones (TCs) in the North Atlantic

SI 2.1 Physically plausible TC tracks and structures

Fig. SI 2 shows the wind speed at 60-hour intervals for a Category 1 hurricane from a climate projection downscaled with GenFocal. The track shows the TC moving westerly, north of the Lesser Antilles, before recurving towards the north. The plots of 10-meter wind speed show that the strongest winds are in the right, front quadrant of the tropical cyclone. Both of these features are characteristics consistent with tropical cyclones in the North Atlantic basin.

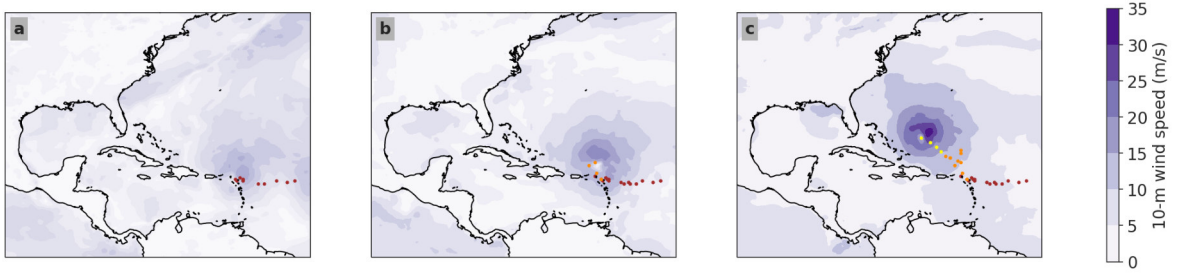


Figure SI 2: Evolution of a hurricane downscaled by GenFocal. Plots of 10-meter wind speed at 60-hour intervals for a Category 1 hurricane projected by GenFocal. Colored dots track the tropical cyclone eye and its intensity in the Saffir-Simpson scale. The tropical cyclone evolves from a depression (brown) to a storm (orange) and ultimately a Category 1 hurricane (yellow).

SI 2.2 Track density

Fig. SI 3 shows the TC tracks and densities in the original CESM2 Large Ensemble (LENS2) data and corresponding downscaled ensembles. GenFocal, shown in Fig. SI 3b, produces the most realistic TCs with a density that is remarkably close to the observed one in the ERA5 reanalysis, shown in Main Text Fig. 2c. The other models shown in Main Text Fig. 2c-f underestimate the number of TCs, overestimate TC track length, and project an unrealistic concentration of TCs over Venezuela and the Pacific Coast of Mexico. In addition, state-of-the-art (SoTA) statistical downscaling methods such as BCSD and STAR-ESDM (SI 5), as well as GenFocal variants without the generative debiasing component such as SR and QMSR (SI 9.6) predict unphysical tracks over the Sahara desert.

SI 2.3 Counts and intensities of TCs

Fig. SI 4 shows the number of detected TCs in the North Atlantic in the August-September-October season of period 2010-2019. GenFocal produces TC counts consistent with observations, in contrast to other methods, which underestimate the number of TCs. Fig. SI 5 shows the distributions of detected TC intensities. GenFocal generates distributions closely matching those in ERA5, whereas other methods tend to underestimate the number of Category 1 Hurricanes and Tropical Storms and Depressions while overestimating the number of Category 3 Hurricanes¹.

¹This bias stems from the very small pressure drop threshold (induced by the calibration procedure in SI 8.3.3) needed to calibrate other downscaling methods for optimal TC detection. The calibrated threshold pressure drop for methods

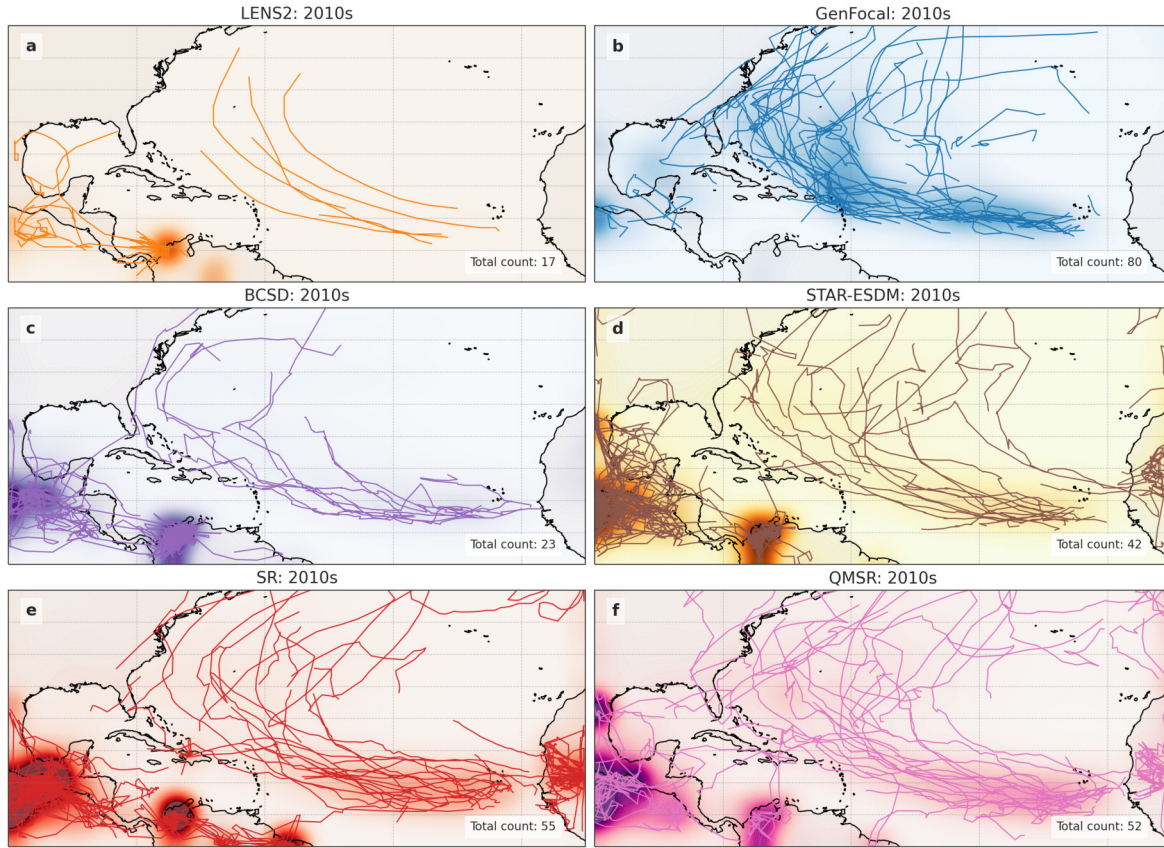


Figure SI 3: TC tracks and their density. Tracks and their density for a LENS2 member in the North Atlantic in the time period 2010-2019 (a), for the same member we show a sample generated by GenFocal (b), BCSD (c), STAR-ESDM (d), SR (e) and QMSR (f). The observed tracks from the ERA5 reanalysis are shown in Main Text Fig. 2c. Note other methods (including the original LENS2) have the unrealistic concentration of TCs over Venezuela and the Pacific Coast of Mexico as well as the unphysical tracks over the Sahara desert.

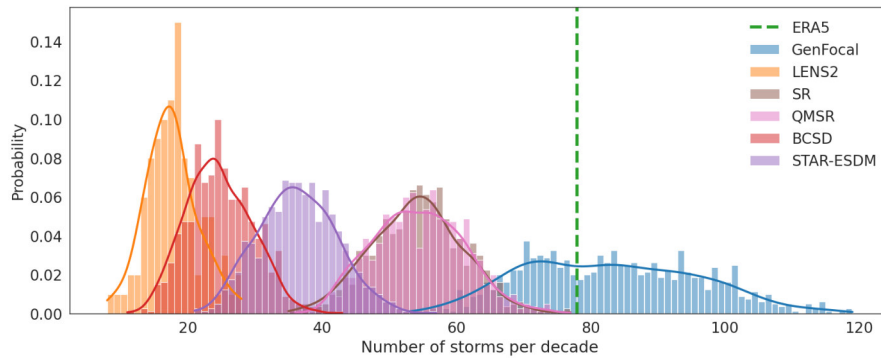


Figure SI 4: Distribution of TC counts. Distributions of North Atlantic TC counts in the August-September-October season of 2010-2019 for the raw and downscaled LENS2 ensemble (100 members), using the different methods considered, and the ERA5 ground truth.

other than GenFocal is either 20Pa or 40Pa, whereas the pressure drop for GenFocal is 120Pa. A small threshold pressure drop inflates the calibrated pressure-derived wind speed and ultimately results in higher intensity storms.

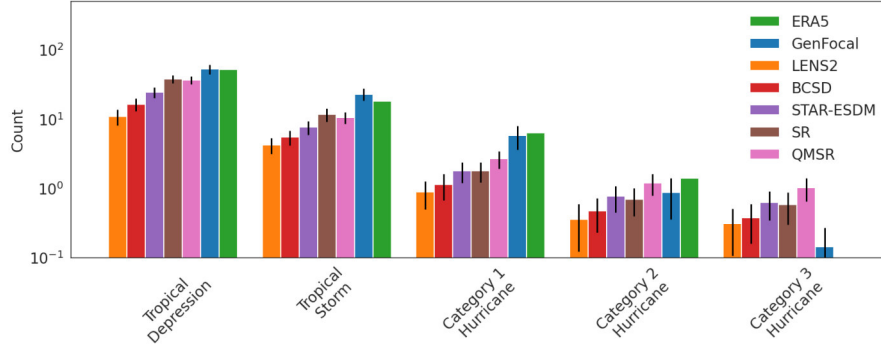


Figure SI 5: Distributions of TC intensity. Saffir-Simpson hurricane wind scale distributions for detected tropical cyclones in the North Atlantic in the August-September-October period during 2010-2019. Errors bars denote one ensemble standard deviation, computed using 100 and 800 members for LENS2 and GenFocal, respectively.

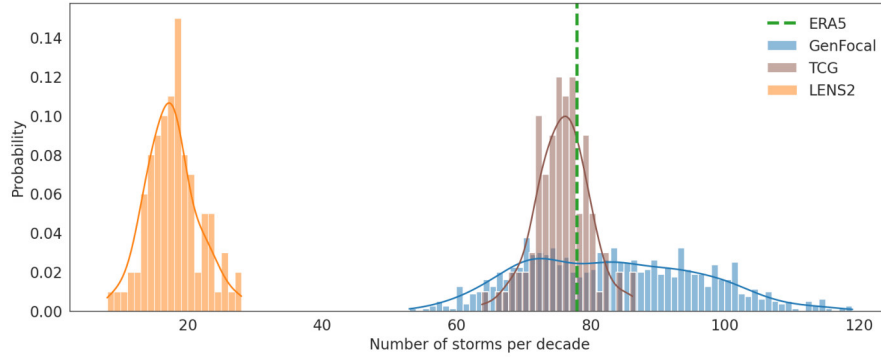


Figure SI 6: Distribution of decadal storm counts. Histogram of decadal storm counts produced by the LENS2 ensemble, the count distribution predicted using the tropical cyclogenesis (TCG) index, and the count distribution in the GenFocal ensembles.

Fig. SI 6 demonstrates that LENS2 produces significantly fewer storms in a decade than anticipated by the Tropical Cyclogenesis (TCG) index, based on large-scale patterns, while GenFocal produces decadal storm counts that are comparable to the TCG predictions and are able to generate plausible TCs whose fine details are realistic and detectable.

Fig. SI 7 shows the superior performance of GenFocal at estimating the distribution of pressure-derived wind speeds, whereas other methods tend to systematically underestimate the probability of 5-10 (m/s), and overestimate the probability of 45-60 (m/s) winds, where the observed ones in reanalysis have almost no mass. This result is consistent with Fig. SI 5.

SI 2.4 Morphology of detected TCs tracks

GenFocal accurately captures the distribution of TC track lengths, closely matching the reanalysis data (Fig. SI 8). In contrast, other methods tend to overestimate track length. Fig. SI 9 shows that GenFocal also excels at capturing the sinuosity indices of the detected tracks. The sinuosity index (SI) provides a proxy to the geometrical shapes of the tracks [91]. It is a transformation of the sinuosity, S of a storm

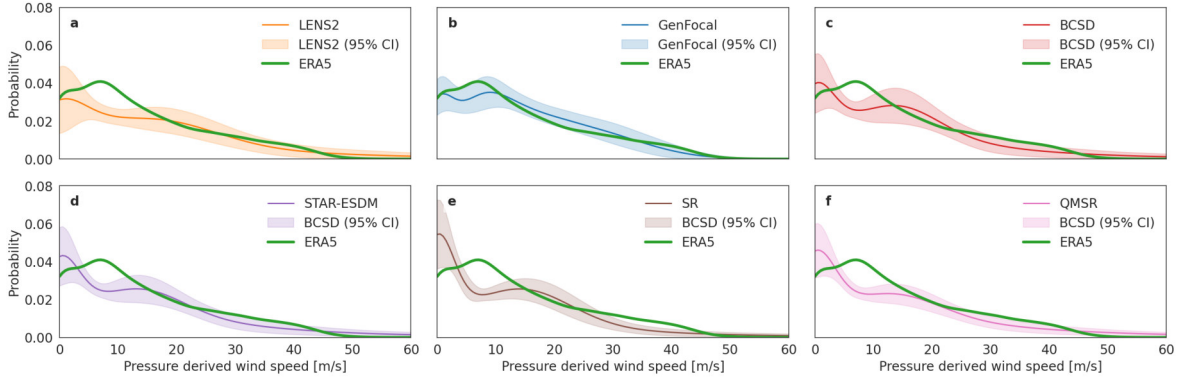


Figure SI 7: Distribution of TC wind speeds. Distributions of the pressure-derived wind speed of tropical cyclones detected in the North Atlantic basin in the August-September-October period during 2010-2019, for LENS2 (a), GenFocal (b), BCSD (c), STAR-ESDM (d), SR (e), and QMSR (f). In addition, we also add the distribution of the pressure-derived wind speed for the reference ERA5 dataset. The confidence intervals are computed across the ensemble dimension using a Gaussian assumption (2 standard deviations).

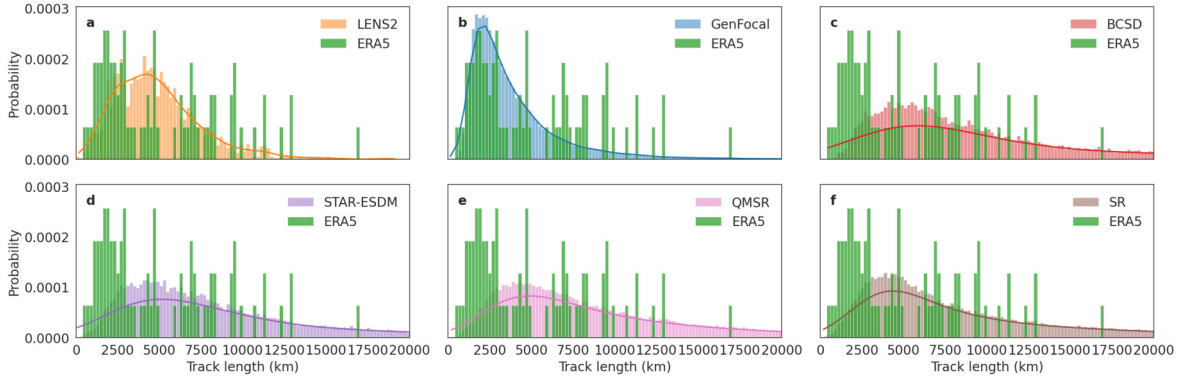


Figure SI 8: Distributions of TC track length. Distributions of the track lengths of detected tropical cyclones in the North Atlantic in the August-September-October period during 2010-2019, for LENS2 (a), GenFocal (b), BCSD (c), STAR-ESDM (d), SR (e), and QMSR (f). In addition, we also add the distribution of the track lengths detected in the reference ERA5 dataset.

path, which is defined as

$$S = \frac{l_{\text{path}}}{l_{\text{direct}}}, \quad (1)$$

where l_{path} is the total path length and l_{direct} is the direct length between the start and end points of the track. SI is defined as

$$SI = \sqrt[3]{(S-1)} \times 10 \quad (2)$$

A sinuosity index of 0 indicates a straight track, and it increases for more sinuous tracks. Fig. SI 9 shows that the tracks induced by GenFocal have a distribution similar to ERA5, whereas other methods tend to produce overly sinuous (and even erratic) tracks, as observed in Fig. SI 3.

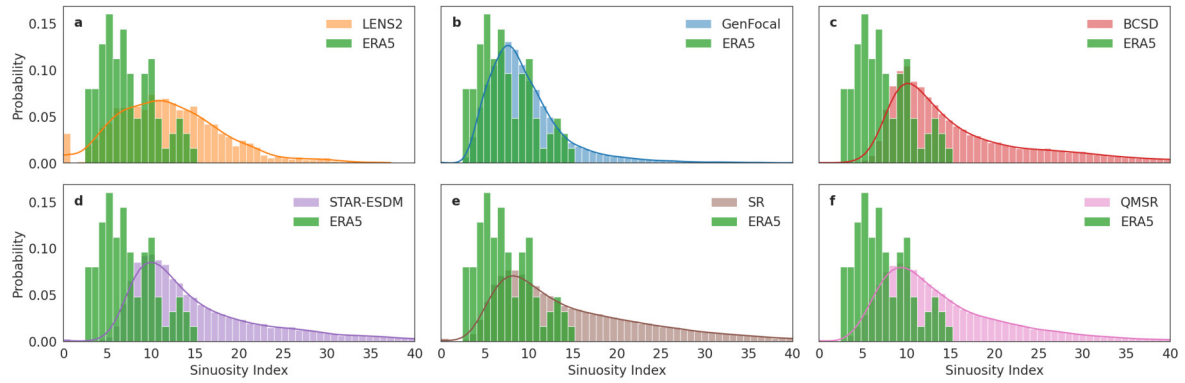


Figure SI 9: Distributions of sinuosity indices. Distributions of the sinuosity indices of the detected tropical cyclones tracks in the North Atlantic in the August-September-October period during 2010-2019, for LENS 2 (a), GenFocal (b), BCSD (c), STAR-ESDM (d), SR (e), and QMSR (f).

SI 3 GenFocal models accurately multivariate spatiotemporal statistics

We assess how well the multivariate probabilistic distributions over spatial and temporal dimensions are captured by the downscaling procedures, compared to those in ERA5 during the evaluation period (2010-2019). We focus on the summer (June-July-August) period over the Conterminous United States (CONUS) region.

We evaluate skill using the mean absolute bias (MAB) (SI 7.1.1), mean Wasserstein distance (MWD) (SI 7.1.2), and mean absolute error (MAE) in the 99th percentile (SI 7.1.3). Please refer to those sections for the definitions. Our results demonstrate that GenFocal effectively captures marginal distributions, joint distributions, distributions of derived variables (via nonlinear transformations), and their tails.

SI 3.1 Statistics of single variables

Table SI 3: Statistical modeling errors of directly downscaled variables in marginal distributions by different methods for the summers (June-July-August) in CONUS during 2010-2019. Best highlighted in bold.

Variable	GenFocal	BCSD	STAR-ESDM	QMSR	SR
Mean Absolute Bias ↓					
Temperature (K)	0.41	0.55	0.89	0.54	2.03
Wind speed (m/s)	0.19	0.12	0.15	0.15	1.83
Specific humidity (g/kg)	0.31	0.39	0.54	0.43	1.41
Sea-level pressure (Pa)	39.92	45.59	38.88	44.98	160.98
Mean Wasserstein Distance ↓					
Temperature (K)	0.47	0.61	0.93	0.59	2.08
Wind speed (m/s)	0.22	0.18	0.18	0.16	1.84
Specific humidity (g/kg)	0.36	0.42	0.56	0.46	1.47
Sea-level pressure (Pa)	52.09	47.46	41.47	46.85	162.61
Mean Absolute Error, 99 th ↓					
Temperature (K)	0.61	0.77	1.04	0.82	2.63
Wind speed (m/s)	0.48	0.39	0.50	0.41	2.43
Specific humidity (g/kg)	0.45	0.58	0.74	0.63	1.67
Sea-level pressure (Pa)	77.99	68.02	67.66	58.33	209.98

Table SI 3 compares the ability of GenFocal and other methods to capture the marginal distributions of the downscaled variables for summers (JJA) during the 2010-2019 evaluation period. GenFocal is highly competitive, often achieving the lowest errors.

While quantile mapping (QM) is statistically optimal for matching marginal distributions, it requires a subsequent super-resolution (SR) step to increase resolution. Consequently, methods like QMSR, BCSD, and STAR-ESDM show comparable performance. In contrast, SR alone performs poorly due to biases in the coarse simulation. Although GenFocal’s debiasing stage targets the joint distribution, it also frequently improves its performance on marginals.

Figs. SI 10–SI 11 show the spatial distribution of bias and Wasserstein distances (defined in SI 7.1) between the downscaled ensemble generated using different methods and ERA5 during the evaluation period. GenFocal ensembles show low bias and Wasserstein distance across all variables (Fig. SI 10), outperforming all methods in near-surface temperature and humidity. As previously noted, the absence of a debiasing step in SR leads to substantial biases and errors as shown in Fig. SI 10(q-t) and SI 11(q-t). Generally, the fields are significantly underestimated, except for humidity, which is severely overestimated in the eastern California Gulf (Mexico) and the central United States.

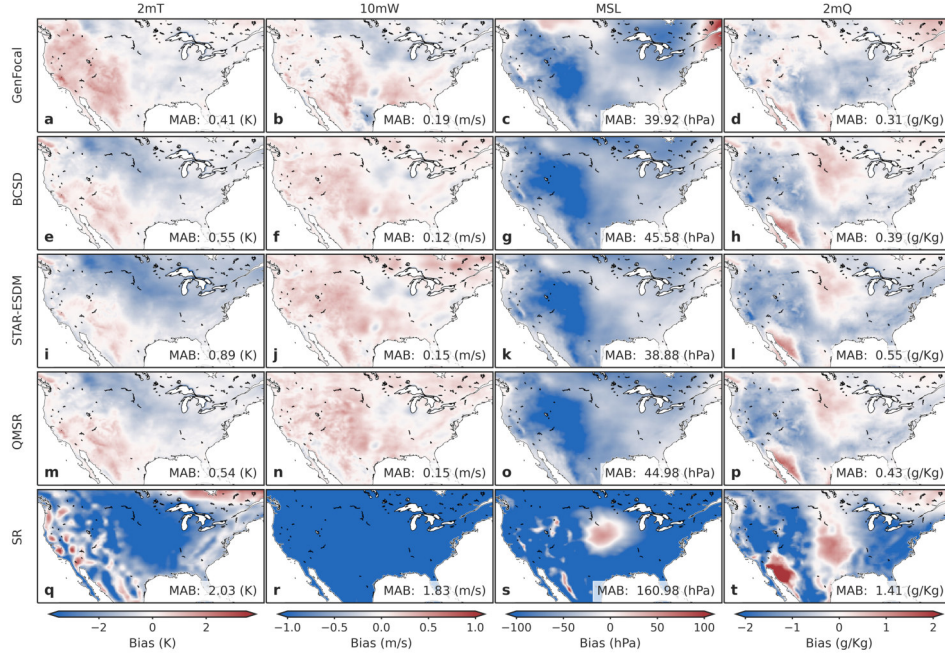


Figure SI 10: Pointwise bias of weather variables. Pointwise bias over CONUS during the summers (June-August) of the evaluation period 2010-2019 for the 2 m temperature, 10 m wind speed, mean sea-level pressure and 2 m specific humidity for GenFocal (a-d), BCSD (e-h), STAR-ESDM (i-l), QMSR (m-p), and SR (q-t).

GenFocal is also superior in recovering extreme statistics, as shown in Fig. SI 12 and Fig. SI 13 for the pixel-wise errors at the 95th and 99th percentiles for directly modeled variables, respectively.

The results in Figs. SI 10-SI 13 and Table SI 3 demonstrate that the debiasing step, through either quantile mapping (QM) or GenFocal, is crucial for obtaining statistically accurate high-resolution outputs. In contrast, super-resolution (SR) alone incurs large errors, especially in the distributional tails.

SI 3.2 Statistics of derived variables

GenFocal models explicitly the joint distribution of output variables, capturing inter-variable correlations neglected by downscaling methods that model each variable independently. We showcase the benefits of this approach by computing the statistics of derived variables that depend nonlinearly on the directly modeled variables, and comparing them to the ground truth during the evaluation period. We consider the near-surface relative humidity and the heat index (see SI 8.1 for the definition), nonlinear functions of temperature and humidity that have important effects on human health and comfort. The heat index is also used to define heat streaks in SI 8.2. The tracking errors of the statistics for the derived variables are summarized in Table SI 4, demonstrating that GenFocal substantially outperforms other methods. The spatial distributions of tracking errors are illustrated in Figs. SI 14 and SI 15. GenFocal shows substantial reductions in relative humidity bias and Wasserstein distance with respect to other methods over the Midwestern United States (Fig. SI 14). Error reductions are even more substantial and broadly distributed at the tails of the distribution. Similarly, GenFocal also reduces errors for the heat index, although less substantially.

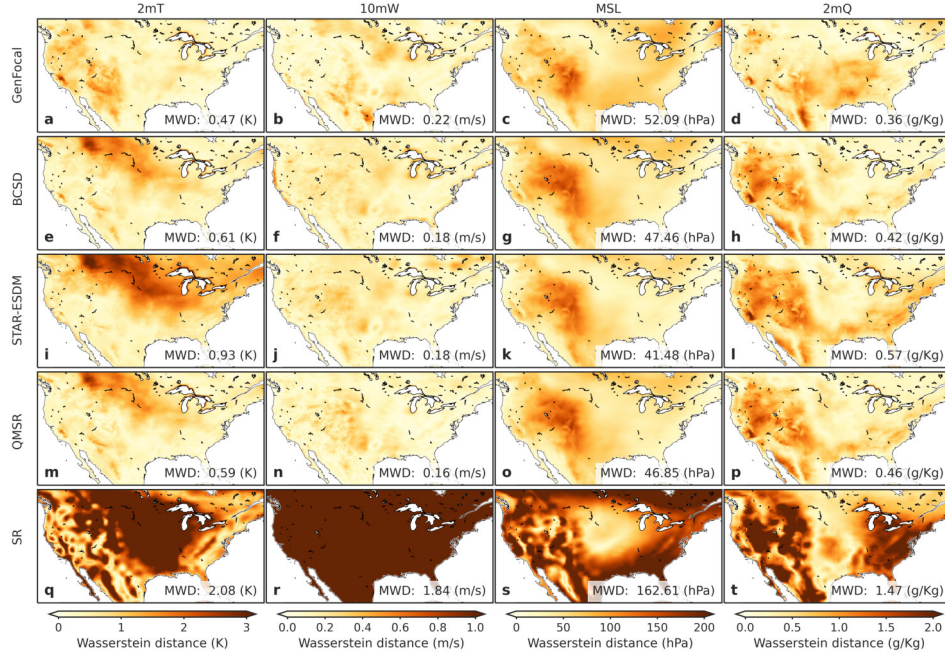


Figure SI 11: Pointwise Wasserstein distance. Pointwise Wasserstein distance between marginals over CONUS during the summer (June-August) for the evaluation period 2010-2019 for the 2 m temperature, 10 m wind speed, mean sea-level pressure and 2 m specific humidity for GenFocal (a-d), BCSD (e-h), STAR-ESDM (i-l), QMSR (m-p), and SR (q-t).

Table SI 4: Statistical modeling errors of **derived variables** by different models for the summers (June-August) in CONUS during 2010-2019.

Variable	GenFocal	BCSD	STAR-ESDM	QMSR	SR
Mean Absolute Bias ↓					
Relative humidity (%)	1.71	2.28	2.70	2.17	6.56
Heat index (K)	0.47	0.66	1.11	0.67	2.63
Mean Wasserstein Distance ↓					
Relative humidity (%)	2.10	3.54	3.77	2.51	6.81
Heat index (K)	0.53	0.72	1.14	0.72	2.69
Mean Absolute Error, 99 th ↓					
Relative humidity (%)	1.87	13.68	12.96	2.54	4.53
Heat index (K)	0.68	1.05	1.52	1.01	3.80

SI 3.3 Extreme statistics of joint distributions

In Fig. SI 16, we investigate further GenFocal’s capacity in capturing the correlation of meteorological extremes in terms of the tail dependence (see SI 7.3 for its definition). The tail dependence evaluates the probability that two variables will present extreme behavior simultaneously, which is of great importance for assessing compound risk. High temperature and humidity extremes can have important effects on human health, whereas dry hot extremes can increase agricultural losses.

GenFocal captures well the frequency of humid and hot extremes (Fig. SI 16a,e). All other methods considered tend to underestimate the co-occurrence of extremely humid and hot conditions in the U.S. Midwest (Fig. SI 16i,m,q,u). All methods show higher skill at capturing dry and hot summer extremes, with GenFocal and BCSD providing the most accurate assessment of compound risk (Fig. SI 16f,j).

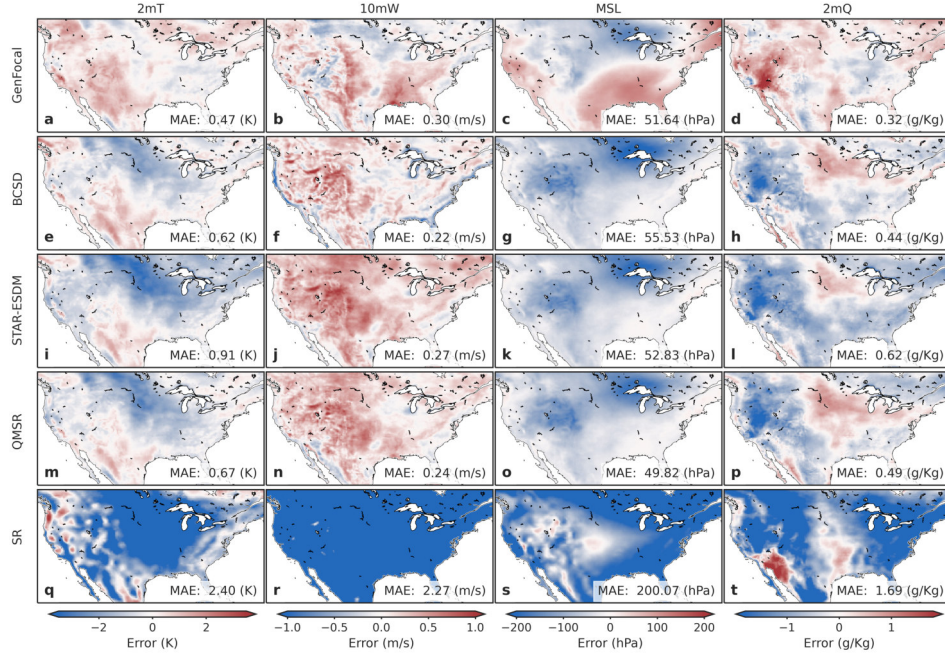


Figure SI 12: Error of the 95th percentile. Error of the 95th percentile over CONUS during the summer (June-August) for the evaluation period 2010-2019 for the 2 m temperature, 10 m wind speed, mean sea-level pressure and 2 m specific humidity for GenFocal (a-d), BCSD (e-h), STAR-ESDM (i-l), QMSR (m-p), and SR (q-t).

Results are also presented for the co-occurrence of high wind speeds and temperatures, and high wind speeds and low humidity. For both, GenFocal presents the lowest tail dependence bias with respect to the ERA5 reanalysis.

SI 3.4 Spatial correlations

GenFocal provides a more accurate representation of spatial correlations (defined in SI 7.2), as shown in Figs. SI 17-SI 20. Furthermore, the QMSR variant achieves error levels similar to GenFocal and lower than other methods, highlighting the diffusion-based super-resolution model’s advantage over random historical analogs.

Similar observations can be made from the spatial radial spectra in Fig. SI 21, where overall errors are lowest for GenFocal and QMSR, signaling the effectiveness of diffusion-based super-resolution.

SI 3.5 Temporal correlations

We also demonstrate the capacity of GenFocal in capturing the temporal statistics of the directly modeled variables. Fig. SI 22 shows the temporal power spectral density (following SI 7.2.3) of different variables for a set of different cities in CONUS during the evaluation period (summers in the 2010s). Overall, we observe that GenFocal outperforms BCSD and STAR-ESDM in the 2m temperature and specific humidity, while remaining competitive for the 10m wind speed.

As both QMSR and SR use a similar time-coherence super-resolution approach as GenFocal, they

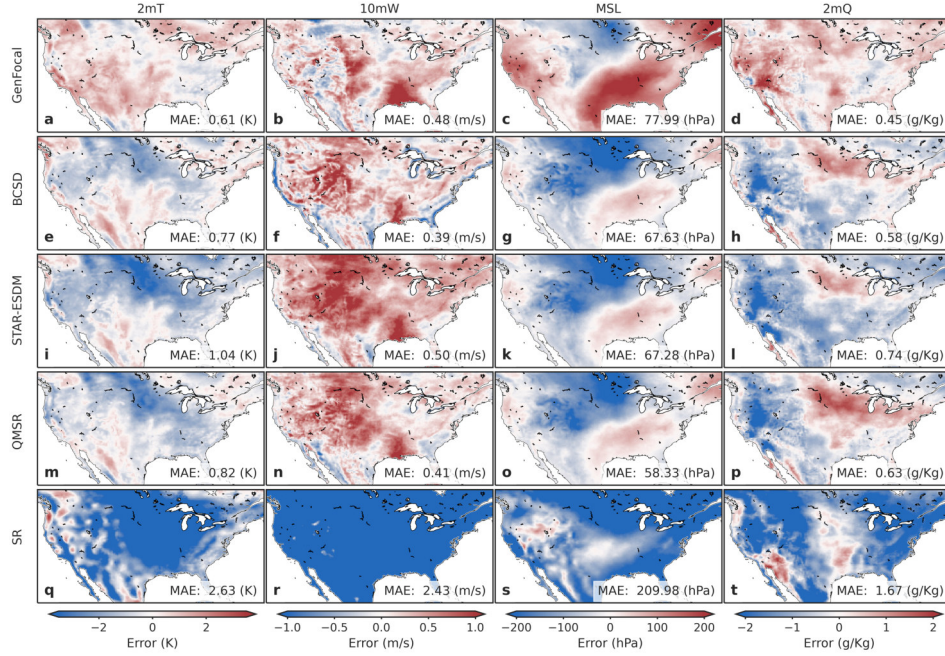


Figure SI 13: Error of the 99th percentile. Error of the 99th percentile over CONUS during the summer (June-August) for the evaluation period 2010-2019 for the 2m temperature, 10m wind speed, mean sea-level pressure and 2 m specific humidity for GenFocal (a-d), BCSD (e-h), STAR-ESDM (i-l), QMSR (m-p), and SR (q-t).

provide competitive results when compared to the disaggregation-based methods. We observe from Fig. SI 22 that QMSR also outperforms BCSD and STAR-ESDM, and in some cases is slightly better than GenFocal, while SR outperforms BCSD and STAR-ESDM in all but the 10 m wind speed case, trailing only GenFocal in all the variables.

SI 3.6 Statistics of heat streaks

Given GenFocal’s superior performance in capturing temporal coherent statistics and distributions of derived variables, we further compare the heat streaks generated by the different models including the variants of GenFocal introduced in SI 9.6. We show the biases on the number of heat-streaks under different intensities and durations. Figs. SI 23–SI 25 show the local bias in the mean number of streaks per year for the increasing intensity (from “caution” to “danger”). Each figure shows the bias for increasing duration (from 1 day to 7 days) for a fixed intensity.

In general, GenFocal outperforms other methods for a significant margin particularly as the intensity and duration increases. We observe that the geographical distribution of the bias is fairly similar among the methods that rely on QM for the debiasing, whereas GenFocal and SR present different geographical bias patterns.

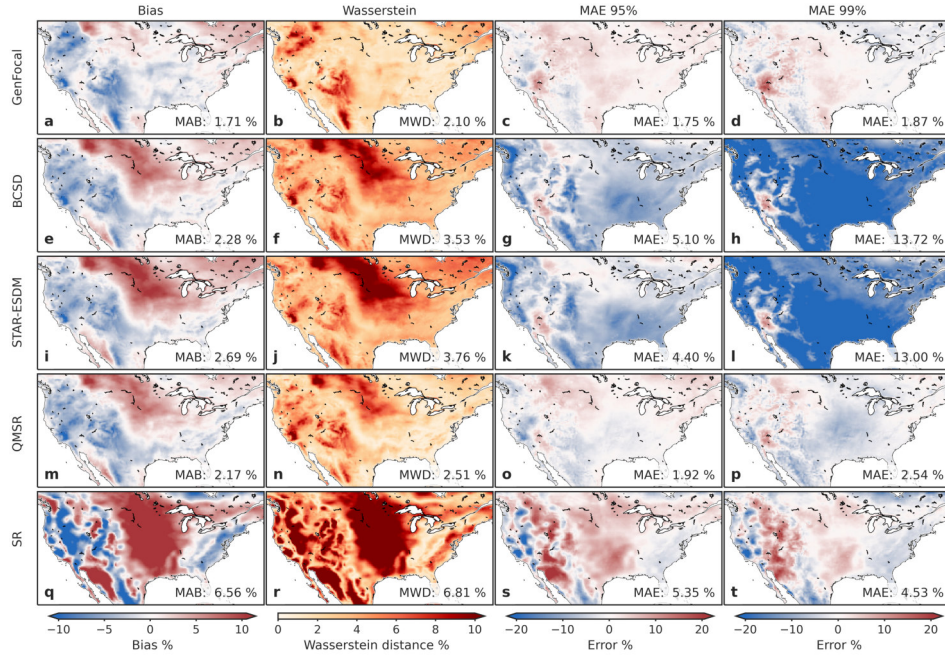


Figure SI 14: Spatial distribution of modeling errors. Spatial distribution of modeling errors for the relative humidity over CONUS during the summer (June-August) of the evaluation period 2010-2019. Pointwise bias, Wasserstein distance, and errors of the 95th percentile and 99th percentile are reported for GenFocal (a-d), BCSD (e-h), STAR-ESDM (i-l), QMSR (m-p), and SR (q-t).

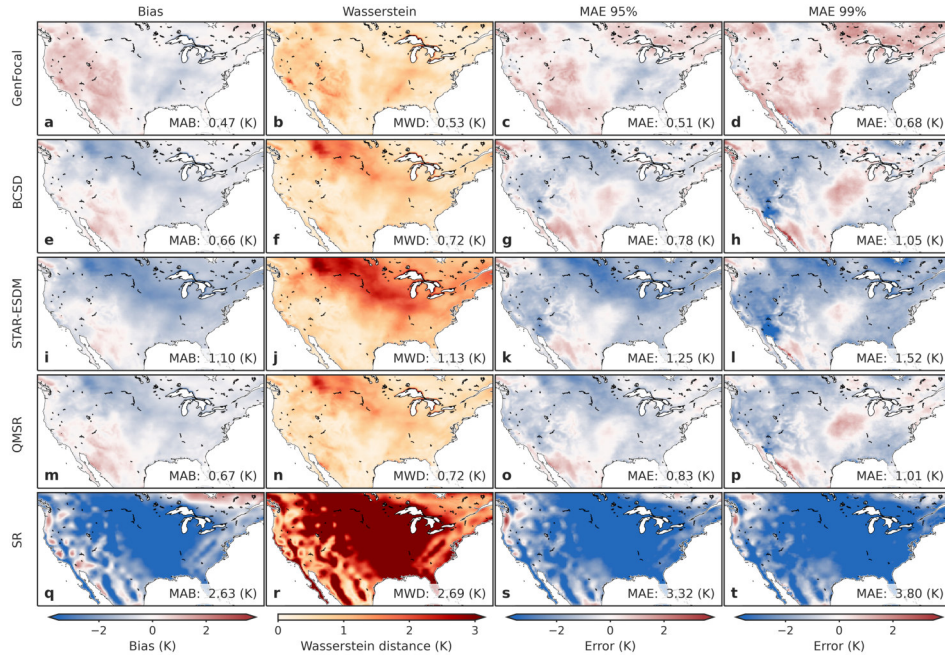


Figure SI 15: Spatial distribution of modeling errors. Spatial distribution of modeling errors for the heat index, over CONUS during the summers (June-August) of the evaluation period 2010-2019. Pointwise bias, Wasserstein distance, and errors of the 95th percentile and 99th percentile are reported for GenFocal (a-d), BCSD (e-h), STAR-ESDM (i-l), QMSR (m-p), and SR (q-t).

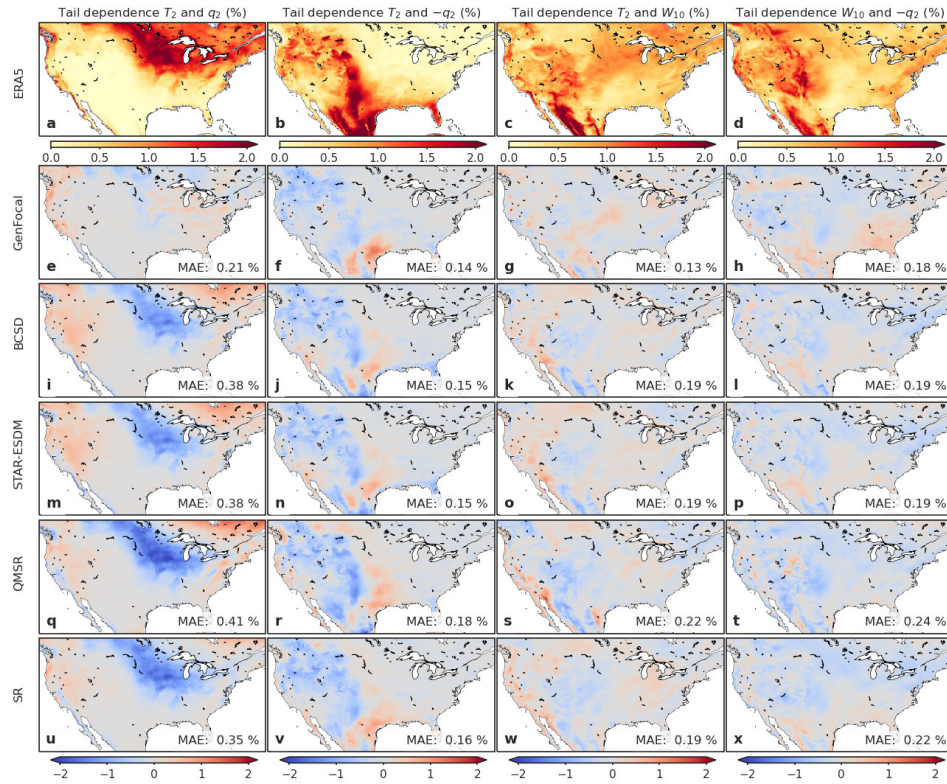


Figure SI 16: Tail dependence of meteorological extremes. Tail dependence of pairs of meteorological extremes over period 2010-2019 from ERA5 and bias of downscaling methods. Tail dependence shown for high temperature and humidity (a), high temperature and low humidity (b), high temperature and high wind (c), and high wind and low humidity (d). Tail dependence biases are shown for GenFocal (e-h), BCSD (i-l), STAR-ESDM (m-p), QMSR (q-t), and SR (u-x).

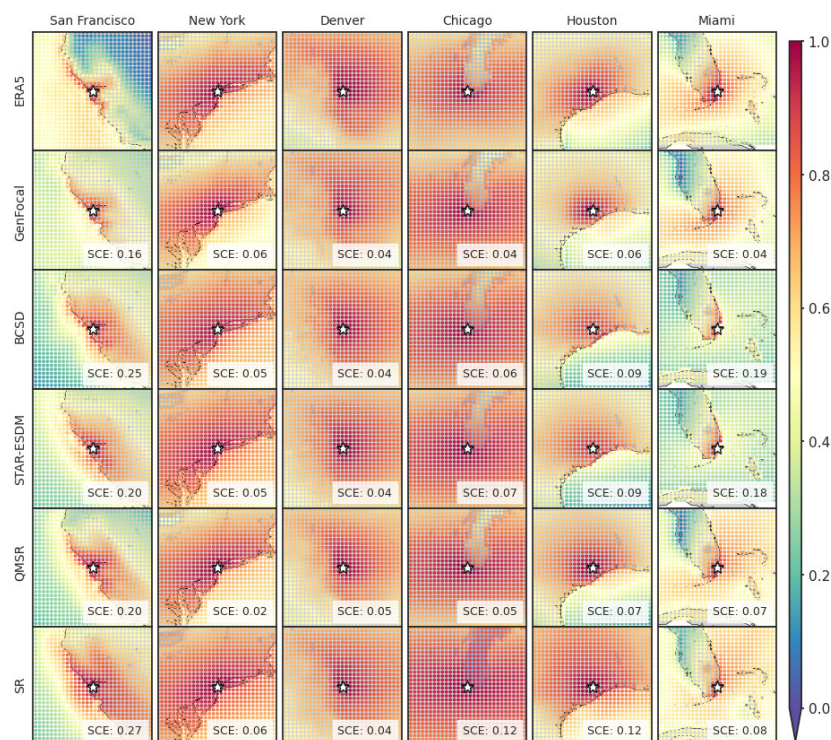


Figure SI 17: Spatial correlation for 2-m temperature. Spatial correlation for 2 m temperature around selected populous US cities, evaluated for all snapshots at 18:00 UTC. The color scale represents the correlation coefficient relative to the city (stars) within a $\pm 4^\circ$ longitude/latitude range.

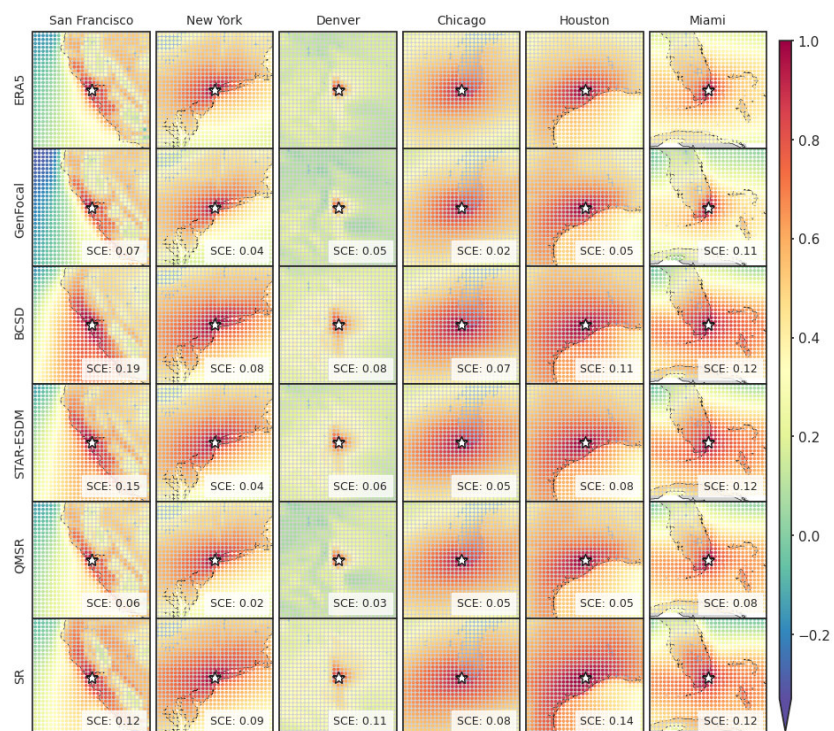


Figure SI 18: Spatial correlation for 10 m wind speed. Spatial correlation for 10 m wind speed around selected populous US cities, evaluated for all snapshots at all times. The color scale represents the correlation coefficient relative to the city (stars) within a $\pm 4^\circ$ longitude/latitude range.

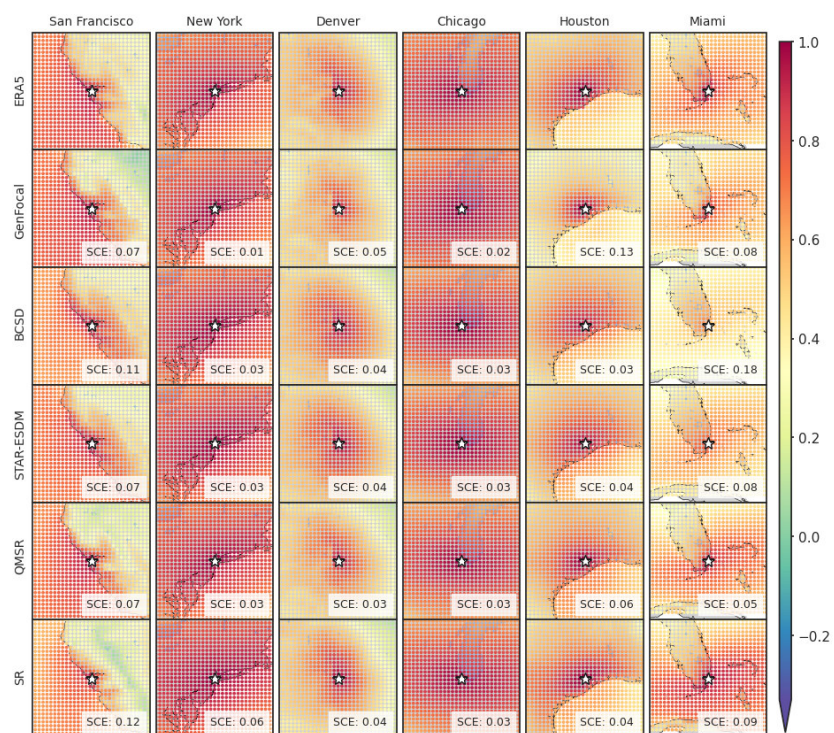


Figure SI 19: Spatial correlation for near-surface specific humidity. Spatial correlation for near-surface specific humidity around selected populous US cities, evaluated for all snapshots at all times. The color scale represents the correlation coefficient relative to the city (stars) within a $\pm 4^\circ$ longitude/latitude range.

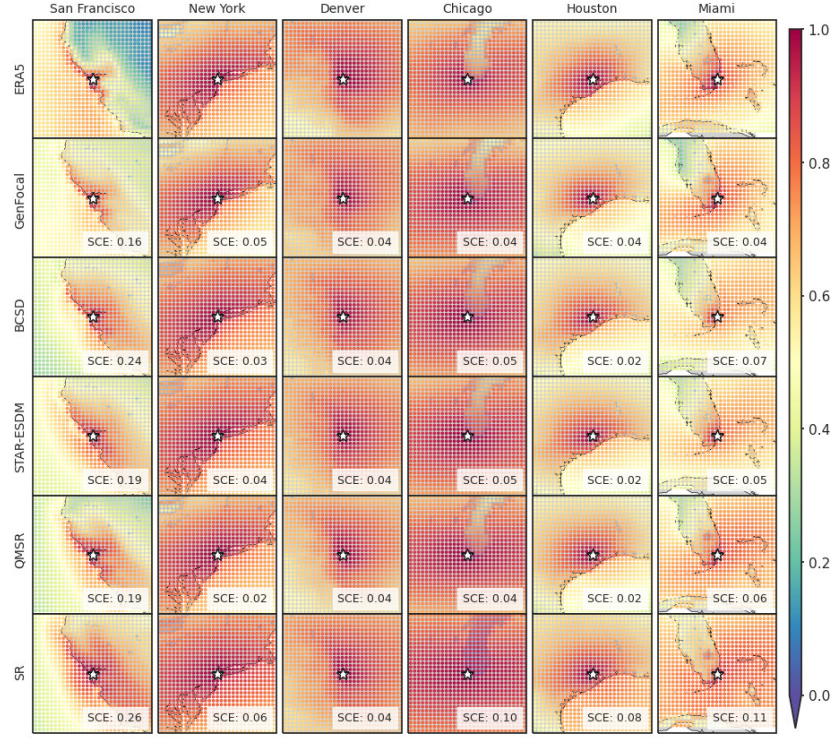


Figure SI 20: Spatial correlation for heat index. Spatial correlation for heat index around selected populous US cities, evaluated for all snapshots at 18:00 UTC. The color scale represents the correlation coefficient relative to the city (stars) within a $\pm 4^\circ$ longitude/latitude range.

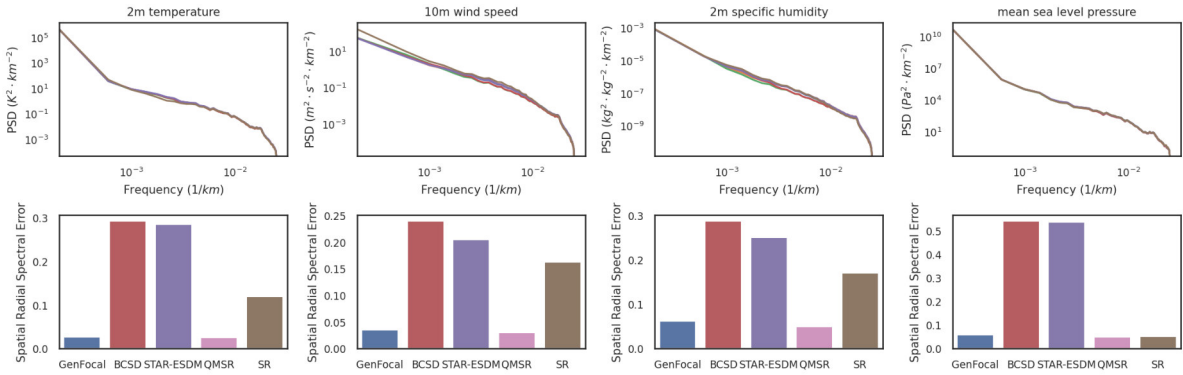


Figure SI 21: Spatial power spectra density. Spatial radial power spectra density (following SI 7.2.2), including the spectral error (30), for output variables generated with GenFocal and other methods.

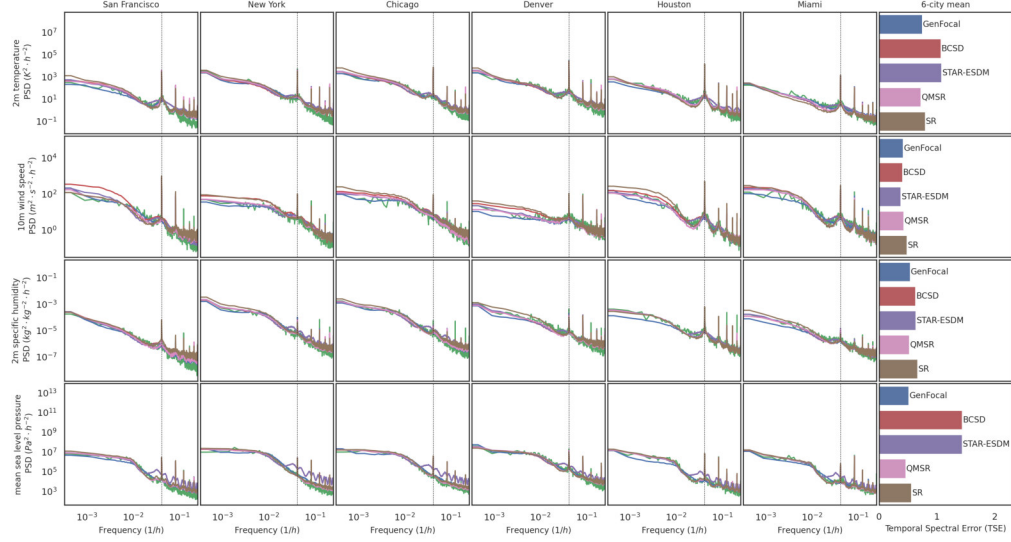


Figure SI 22: Temporal power spectra density. Temporal power spectra density (following SI 7.2.3), including the spectral error (33), for a set of selected cities in CONUS and different variables for ensembles generated with GenFocal and other methods.

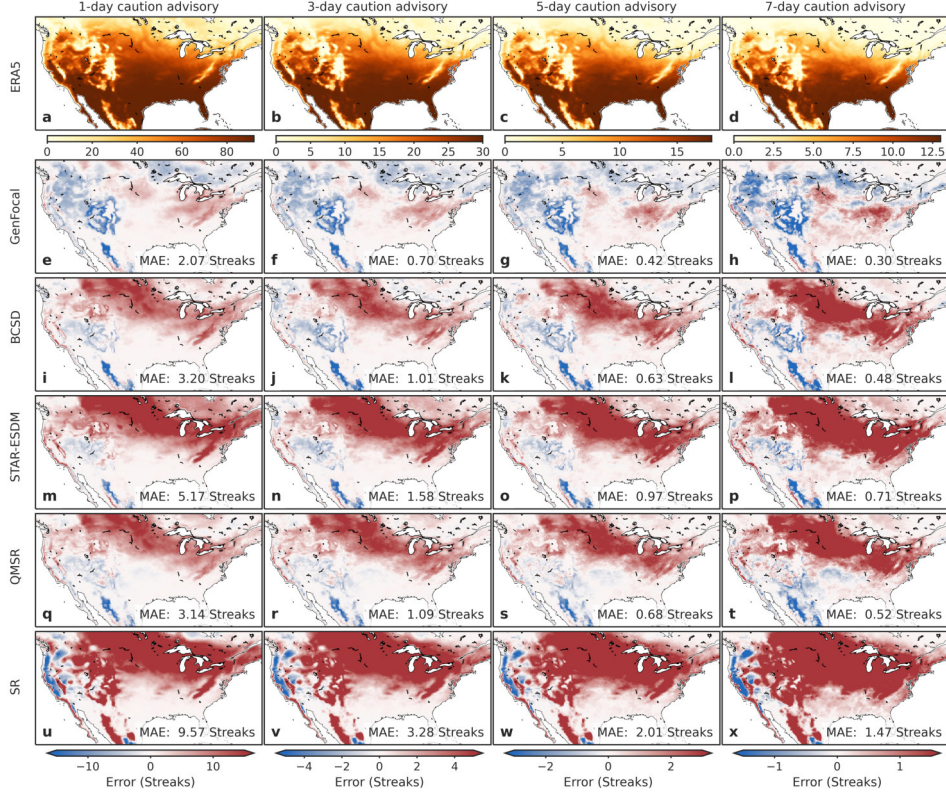


Figure SI 23: Bias in the number of heat streaks by different lengths. Bias in the number of heat-streaks per year for **caution** advisory considering different lengths. We show the ground truth (ERA5)(a-d), and the pointwise errors of GenFocal (e-h), BCSD (i-l), STAR-ESDM (m-p), QMSR (q-t) and SR (u-x).

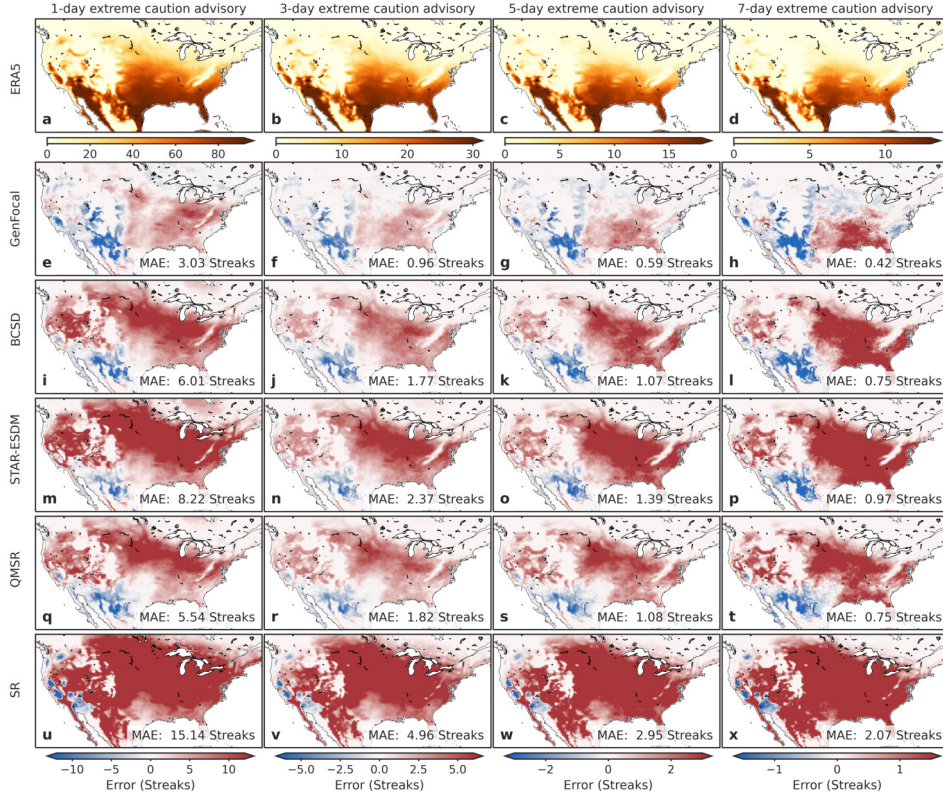


Figure SI 24: Bias in the number of heat streaks by different lengths. Bias in the number of heat-streaks per year for **extreme caution** advisory considering different lengths. We show the ground truth (ERA5)(a-d), and the pointwise errors of GenFocal (e-h), BCSD (i-l), STAR-ESDM (m-p), QMSR (q-t) and SR (u-x).

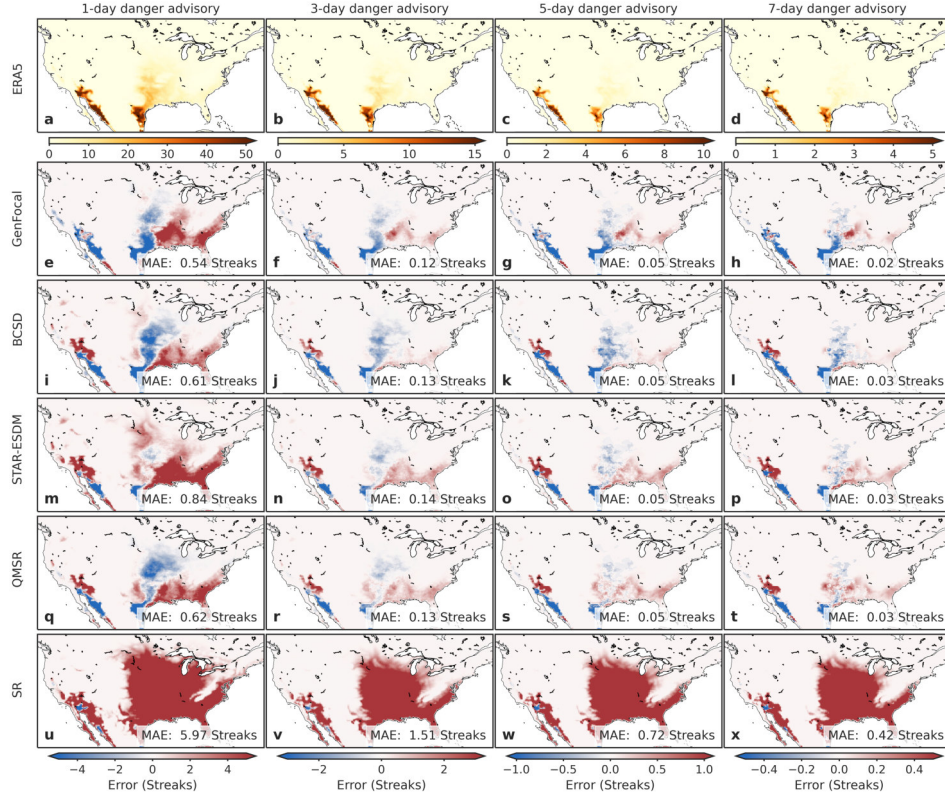


Figure SI 25: Bias in the number of heat streaks by different lengths. Bias in the number of heat-streaks per year for **danger** advisory considering different heatwave lengths. We show the ground truth (ERA5)(a-d), and the pointwise errors of GenFocal (e-h), BCSD (i-l), STAR-ESDM (m-p), QMSR (q-t) and SR (u-x).

SI 4 Future climate risk assessment

This section explores the application of GenFocal to assess future changes in regional climate risk consistent with input coarse-scale climate projections. In particular, we analyze trends in the distribution of summer near-surface temperatures over the western U.S., and changes in tropical cyclone activities in the North Atlantic basin.

SI 4.1 Changes in summer temperatures over the Western U.S.

The distribution of near-surface temperature is strongly affected by increasing atmospheric greenhouse gas concentrations. This results in significant changes in the risk of temperature extremes over time. We analyze the ability of GenFocal to project these changes at a regional scale over major cities in the western U.S., focusing on periods 2017-2023 and 2077-2083.

Since observational references do not exist for future time periods, we compare GenFocal projections to projections from the Western United States Dynamically Downscaled Dataset (WUS-D3) [73]. In particular, we evaluate the distribution changes from projections of CESM2 dynamically downscaled to 45 km and 9 km resolution using the Weather Research and Forecasting (WRF) model [82]. We denote these projections as WRF 45 km and WRF 9 km, respectively. Dynamical downscaling is performed after debiasing the CESM2 projections with respect to the ERA5 reanalysis over the historical period. Therefore, the debiasing and downscaling setup is similar to that of GenFocal. Note that WUS-D3 is generated using physics-based dynamical downscaling, an expensive operation that was restricted to the western US.

We align the grids of all projections by interpolating the GenFocal and WRF 45 km data to the WRF 9 km grid. The WRF 9 km is averaged to a similar effective scale as GenFocal by Gaussian filtering. Results are also provided for the statistical downscaling baselines BCSD and STAR-ESDM, using the same interpolation as GenFocal.

Fig. SI 26 illustrates changes in daily mean near-surface temperature in 11 cities across the western U.S. GenFocal projects large differences in temperature changes across locales, consistent with the dynamically downscaled projections. Coastal California cities like Los Angeles or San Diego experience a much slower warming rate than inland cities Portland or Salt Lake City. BCSD and STAR-ESDM fail to capture these regional differences, projecting a much more uniform warming.

GenFocal not only captures changes in daily mean temperature, but also changes in summer temperature extremes. Fig. SI 27 shows the change in daily maximum temperatures, and Fig. SI 28 illustrates changes in the top decile of daily maximum temperatures. The regional differences in these changes projected by GenFocal and WRF are also largely consistent. BCSD and STAR-ESDM, again, show a much smaller variations among regions.

SI 4.2 Changes in North Atlantic tropical cyclone activity

We assess the sensitivity of TC activity projected by GenFocal to changing environmental conditions in the North Atlantic by downscaling climate projections from the early (2010-2019) through the mid (2050-2059) 21st century under the SSP3-7.0 shared socioeconomic pathway [62]. The mid-century projection (2050-2059) roughly corresponds to the first decade surpassing a 2°C global surface temperature change since preindustrial levels, a common warming level in climate change assessments of TC activity [42].

Fig. SI 29 evaluates North Atlantic TC activity changes from 800 downscaled projections generated with

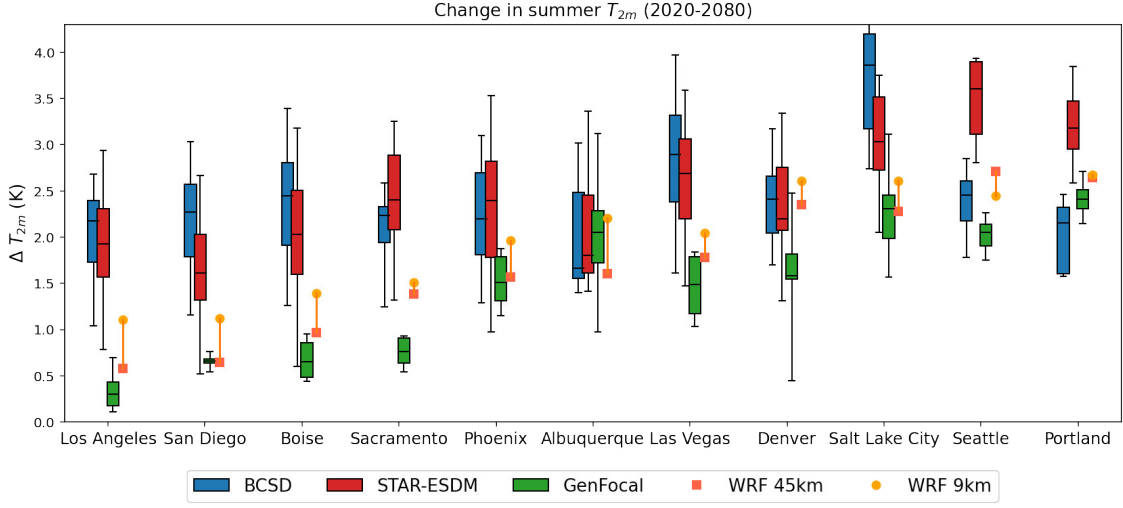


Figure SI 26: Projected change in daily mean near-surface temperature. Projected change in daily mean near-surface temperature in 11 cities across the Western United States, from 2017 to 2083. Results are computed as the average over $1^\circ \times 1^\circ$ regions, and 7 summers (June-August) centered around 2020 and 2080. Boxes for BCSD, STAR-ESDM, and GenFocal show the median (50th percentile) and interquartile range (25th and 75th percentiles) of an ensemble of 8 projections, and whiskers represent the ensemble minima and maxima.

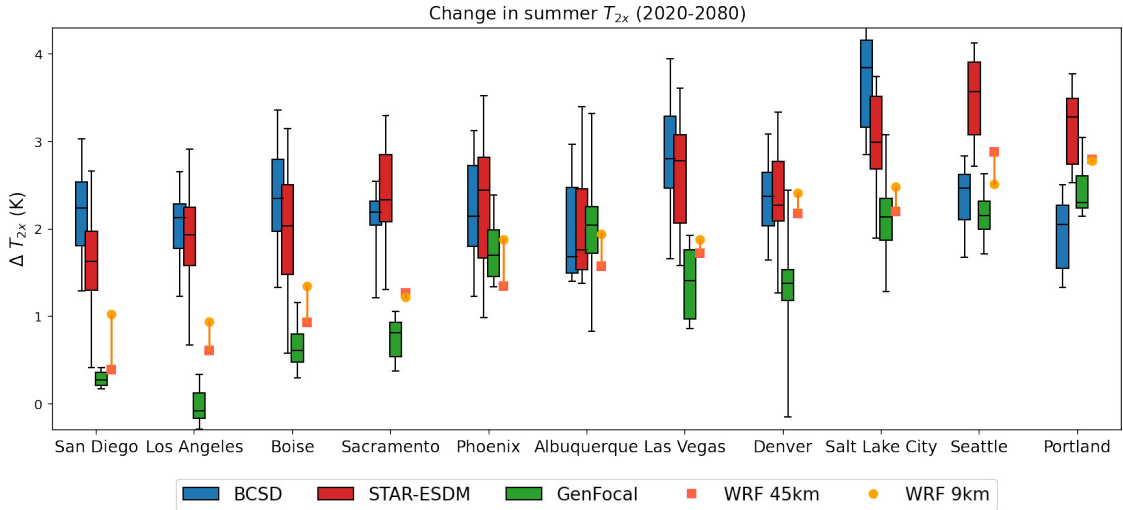


Figure SI 27: Projected change in daily maximum near-surface temperature. Projected change in daily maximum near-surface temperature in 11 cities across the Western United States, from 2017 to 2083. Results are computed as the average over $1^\circ \times 1^\circ$ regions, and 7 summers (June-August) centered around 2020 and 2080. Boxes for BCSD, STAR-ESDM, and GenFocal show the median (50th percentile) and interquartile range (25th and 75th percentiles) of an ensemble of 8 projections, and whiskers represent the ensemble minima and maxima.

GenFocal from the original 100 climate projections in the LENS2 ensemble (GenFocal samples downscale 8 trajectories per ensemble member.). Changes are evaluated for decades 2030-2039 and 2050-2059 with

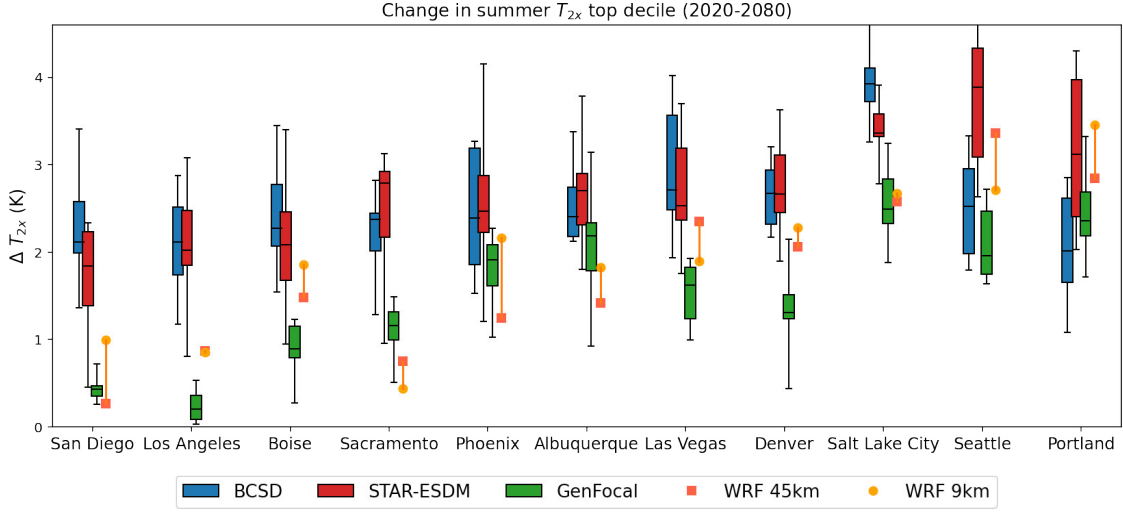


Figure SI 28: Projected change in the top decile of daily maximum near-surface temperature.

Projected change in the top decile of the daily maximum near-surface temperature in 11 cities across the Western United States, from 2017 to 2083. Results are computed as the average over $1^\circ \times 1^\circ$ regions, and 7 summers (June-August) centered around 2020 and 2080. Boxes for BCSD, STAR-ESDM, and GenFocal show the median (50th percentile) and interquartile range (25th and 75th percentiles) of an ensemble of 8 projections, and whiskers represent the ensemble minima and maxima.

respect to 2010-2019. GenFocal projects increased TC risk over the U.S. East Coast, and particularly from the Carolinas to New Jersey, due to an increase in landfall frequency (Fig. SI 29 c,f,i) and intensity (Fig. SI 29 l,o).

Elevated coastal risk is already observed in the 2030-2039 projections, but becomes more pronounced by mid-century. Increases in landfall frequency are projected both for low-intensity tropical depressions, for tropical storms, and for hurricanes; although for the latter the increased risk is limited to the Carolinas and Virginia. Increases in the TC-driven winds are also found to be significant for TCs of median and high intensity between Florida and Massachusetts. Increased coastal TC risk over a similar span of the U.S. East Coast has also been projected by [3] using the Risk Analysis Framework for Tropical Cyclones (RAFT) model. Discrepancies between GenFocal and RAFT, which forecasts reduced risk in the Northeast, may stem from RAFT's assumption of no change in the location of TC genesis, whereas other studies have projected a northward shift of TC genesis in the North Atlantic basin [25].

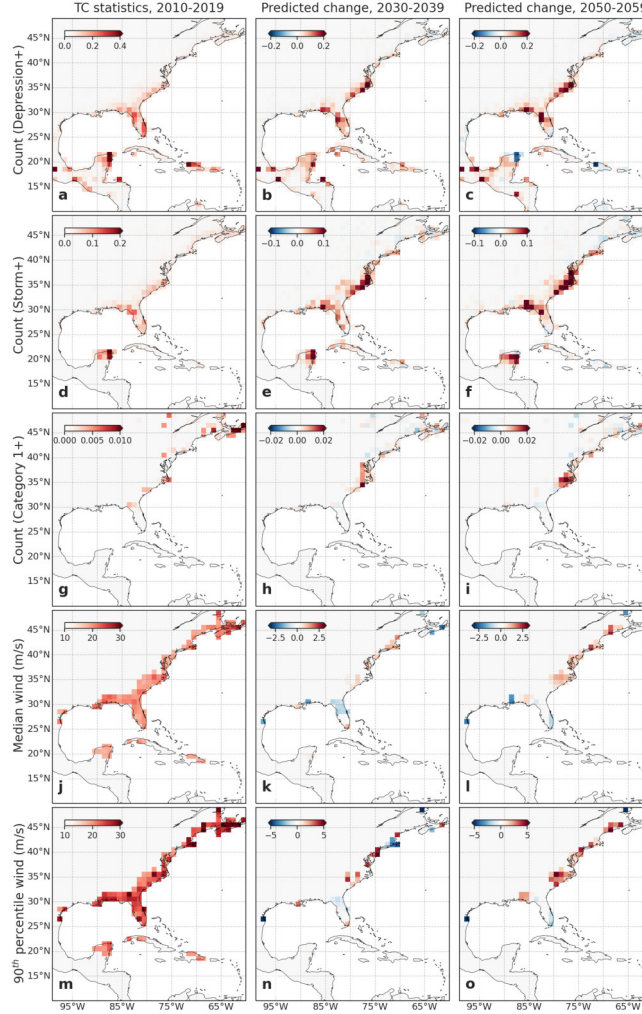


Figure SI 29: Change in TC landfall frequency and intensity. Change in TC landfall frequency and intensity over the first half of the 21th century. **a.** Number of TC landfalls during the August-October season of years 2010-2019. **b, c.** Projected change in the number of TC landfalls from 2010-2019 to 2030-2039, and 2050-2059. **d-f.** Number and projected change in the number of tropical storm and hurricane landfalls. **g-i.** Number and projected change in the number of hurricane landfalls. **j-l.** Median maximum pressure-derived wind speed of TC landfalls and its projected change. **m-o.** 90th percentile of maximum pressure-derived wind speed of TC landfalls and its projected change. All results are computed as the average over 800 downscaled projections. Wind speed changes (**k, l, n, o**) are only shown if statistically significant ($p < 0.05$ in a two-sided Mann-Whitney U test) and set to zero otherwise.

SI 5 Statistical downscaling baselines

Most existing ML-based downscaling methods are inapplicable for climate risk assessment, as they require time-aligned data which is often unavailable (Table SI 1; SI 1). We therefore compare GenFocal to two prominent statistical techniques that do not have this requirement: the widely-used Bias Correction and Spatial Disaggregation (BCSD) [106, 105, 92] and the state-of-the-art STAR-ESDM [33], recommended for the US Fifth National Climate Assessment [96]. As both rely on matching univariate marginals via quantile mapping, they fail to effectively model joint distributions.

To isolate the source of performance gains, we test two variants of GenFocal (SI 9.6). First, removing the debiasing step entirely—thus treating the model as supervised under a perfect prognosis assumption (SI 9.6.1)—produces large biases and artifacts, such as generating TCs in the Sahara (Figs. SI 3, SI 10, SI 11). Second, replacing our debiasing step with quantile mapping outperforms other baselines but still falls short of GenFocal. While this variant performs well on metrics insensitive to spatiotemporal coherence, the performance gap widens considerably on metrics that are, such as TC and heatwave statistics.

GenFocal’s superior performance in assessing risks from events with complex spatiotemporal and inter-variable dependencies—detailed in SI 2, SI 3, and SI 10—demonstrates the importance of fully probabilistic, high-dimensional modeling. We did not explore other alternatives, such as deterministic super-resolution or GAN-based models, because prior work shows they underperform for large super-resolution factors [101]. Deterministic models tend to produce overly smooth samples by collapsing to the conditional mean [60], while GANs suffer from training instability and rapid quality degradation. These limitations are particularly relevant given our high super-resolution factor of 6×6 spatially and 12 temporally, totalling $432 = 6 \times 6 \times 12$.

SI 5.1 Bias Correction and Spatial Disaggregation (BCSD)

Bias Correction and Spatial Disaggregation (BCSD) is a widely used statistical downscaling method [92, 75, 63], originally designed for applications in hydrology [107]. The method consists of three main stages: bias correction, spatial disaggregation, and temporal disaggregation.

Bias correction based on Gaussian quantile mapping. The goal of this step is to map the quantile of y to that of the coarse-grained observation data y' :

$$\tilde{y}_{\text{anom}} = \frac{y - \text{clim_mean}[y]}{\text{clim_std}[y]} \cdot \text{clim_std}[y'], \quad (3)$$

where the climatological mean and standard deviation are calculated over the BCSD training period (1961-1999). The quantiles are computed relative to member-specific climatology. Unlike GenFocal, which normalizes using the aggregated climatology of the limited set of training members (4 total), this method may favor BCSD due to better climatology estimates because of its incorporation of more training data. The competitive performance of GenFocal, despite this difference, highlights its robustness.

Spatial disaggregation. In the second stage, cubic interpolation is applied to the quantile-mapped anomaly, followed by the addition of the climatological mean of the high-resolution observations:

$$x_{\text{daily_mean}} = \text{Interp}[\tilde{y}_{\text{anom}}] + \text{clim_mean}[x]. \quad (4)$$

This step yields outputs with the desired spatial resolution, but retains the temporal resolution of the input data, which is daily.

Temporal disaggregation. The final stage involves randomly selecting a historical sample sequence from the high-resolution dataset covering the period represented by the spatially disaggregated data, in this case a day, and corresponding to the same time of the year, in this case the same day-of-year. The spatially disaggregated data is then substituted by this sequence after adjusting it to match the daily mean of the spatially disaggregated sample:

$$x_{\text{BCSD}} = x_{\text{hist_sample}} - \text{daily_mean}[x_{\text{hist_sample}}] + x_{\text{daily_mean}}. \quad (5)$$

This step ensures that the outputs achieve the target temporal resolution.

SI 5.2 Seasonal Trends and Analysis of Residuals Empirical Statistical Downscaling model (STAR-ESDM)

STAR-ESDM is a statistical downscaling method that decomposes climate fields into several components, each characterized by different timescales of variability [33]. The method relies on access to high-resolution data over a reference period, which is used to correct biases in the input data. The coarse input climate field y is modeled as

$$y = \tau_y + \text{clim_mean}[y - \tau_y] + \Delta\text{clim_mean}[y - \tau_y] + y_{\text{anom}}, \quad (6)$$

where τ_y is a third-order parametric fit of the long-term trend of the coarse climate field, $\text{clim_mean}[y - \tau_y]$ is its detrended climatological mean over the reference period, $\Delta\text{clim_mean}[y - \tau_y]$ represents the climatological mean change of the detrended field from the reference to the testing period, and y_{anom} is the resulting residual anomaly.

STAR-ESDM downscales coarse input fields by mapping each of the components of decomposition (6) to the distribution of the high-resolution reference dataset. First, the long-term trend is debiased such that its mean m_y over the reference period coincides with that of the high-resolution dataset m_x ,

$$\tilde{\tau}_x = \text{Interp}[\tau_y - m_y] + m_x. \quad (7)$$

Second, the climatological mean of the coarse field is mapped to the climatological mean of the high-resolution data, assuming that the change in climatology of the coarse data from the reference to the test period is a good approximation of the same change at high-resolution:

$$\Delta\text{clim_mean}[x - \tau_x] \approx \text{Interp}[\Delta\text{clim_mean}[y - \tau_y]]. \quad (8)$$

Finally, the coarse anomaly y_{anom} is mapped to the distribution of the high-resolution climate data, using again the climate change in the coarse data as a proxy for the climate change in the high-resolution data,

$$\tilde{x}_{\text{anom}} = \text{Interp}[y_{\text{quant}}] \cdot \text{clim_std}[x] \cdot \frac{\text{clim_std}[y - \tau_y] + \Delta\text{clim_std}[y - \tau_y]}{\text{clim_std}[y - \tau_y]}, \quad (9)$$

where $\Delta\text{clim_std}[y - \tau_y]$ is the difference in climatological standard deviation of the coarse climate data between the test period and the reference period, and the quantile of the coarse anomaly is computed with respect to the modified climatology,

$$y_{\text{quant}} = \frac{y_{\text{anom}}}{\text{clim_std}[y - \tau_y] + \Delta\text{clim_std}[y - \tau_y]}. \quad (10)$$

In equation (9), $\text{clim_std}[x]$ is the climatological standard deviation of the high-resolution dataset over the reference period. The STAR-ESDM downscaled climate field is then constructed as

$$x_{\text{STAR}} = \tilde{\tau}_x + \text{clim_mean}[x - \tau_x] + \Delta\text{clim_mean}[x - \tau_x] + \tilde{x}_{\text{anom}}. \quad (11)$$

Table SI 5: Meteorological fields modeled by GenFocal with their corresponding variable names in the ERA5 and LENS2 datasets. All 10 fields serve as both input and output for the debiasing model, while the super-resolution model uses the top 4 fields in CONUS and top 6 fields in the North Atlantic. Units reflect those used for model training and are converted as needed in the main text.

Meteorological field	Unit	ERA5 variable	LENS2 variable
Near-surface temperature	K	2m_temperature	TREFHT
Near-surface wind speed magnitude	m/s	$(u_component_of_wind^2 + v_component_of_wind^2)^{\frac{1}{2}}$	WSPDSRFV
Near-surface specific humidity	kg/kg	specific_humidity (level=1000 hPa)	QREFHT
Sea level pressure	Pa	mean_sea_level_pressure	PSL
Geopotential at 200 hPa	m	geopotential (level=200 hPa)	Z200
Geopotential at 500 hPa	m	geopotential (level=500 hPa)	Z500
U component of wind at 200 hPa	m/s	u_component_of_wind (level=200 hPa)	U200
U component of wind at 850 hPa	m/s	u_component_of_wind (level=850 hPa)	U850
V component of wind at 200 hPa	m/s	v_component_of_wind (level=200 hPa)	V200
V component of wind at 850 hPa	m/s	v_component_of_wind (level=850 hPa)	V850

SI 6 Data

SI 6.1 Input datasets

We use the Community Earth System Model Large Ensemble (LENS2) dataset [76] for our low-resolution climate dataset. LENS2 was produced using the Community Earth System Model Version 2 (CESM2), a climate model that has interactive atmospheric, land, ocean, and sea-ice models [17]. LENS2 is configured to estimate historical climate and the future climate scenario SSP3-7.0, following the CMIP6 protocol [23]. LENS2 skillfully represents the response of historical climate to external forcings [24]. LENS2 output is available from 1850-2100, with a horizontal grid spacing of 1° , and 100 simulation realizations. In this work, we use a coarse-grained version of the LENS2 ensemble at 1.5° horizontal resolution.

The ERA5 reanalysis dataset [35] is our high-resolution weather dataset. ERA5 uses a modern forecast model and data assimilation system with all available weather observations to produce an estimate of the atmospheric state. This estimate includes conditions ranging from the surface to the stratosphere. ERA5 data is available from 1940 to near present at a horizontal grid spacing of 0.25° . ERA5 estimated extremes of temperature and precipitation agree well with observations in areas where topography changes slowly [83].

SI 6.2 Modeled variables

We consider a set of four surface variables to downscale, which were chosen in order to evaluate the statistics of the spatiotemporal events of interest, namely heat-streaks and TCs.

The two-step nature of GenFocal renders it highly versatile, as the debiasing step and the super-resolution

steps are decoupled. This allows for some interesting properties, e.g., the debiasing step can be performed globally, while the super-resolution can be performed within different regions, and the meteorological fields downscaled can also be different, provided that the fields in the super-resolution step are a subset of the debiased ones.

We showcase these two properties by downscaling climate data over the North Atlantic and over CONUS (see [SI 2](#) and [SI 3](#) respectively), and by using different variables for the debiasing and the super-resolution steps. In what follows we show the variables used for each step with their names and units.

SI 6.2.1 Debiasing

As shown in [SI 10.3](#), modeling extra variables in the debiasing steps results in improved results, particularly for TC tracking (see [SI 10.3](#)). As such, we explicitly model 10 variables in the debiasing step. We include 4 surface variables: near-surface temperature, wind speed magnitude, specific humidity, and sea level pressure; and 6 variables within the mid or upper troposphere: geopotential at 200 and 500 hPa, and both components of the wind speed at 200 and 500 hPa. The official names for these variables, as documented in the datasets, are listed in [Table SI 5](#).

Although we do not super-resolve the above-surface variables, they provide extra signal for the debiasing step, as they are correlated with some of the near-surface variables.

SI 6.2.2 Super-resolution

We target four surface variables in our downscaling pipeline: near-surface temperature, wind speed magnitude, specific humidity, and sea level pressure, which constitute a subset of the debiased variables (top 4 rows in [Table SI 5](#)). In CONUS, these variables coincide with the modeled variables (both input and output). In North Atlantic, we additionally include two geopotential fields, at 200 and 500 hPa respectively, in the super-resolution model.

SI 6.3 Regridding

The ERA5 dataset is natively 0.25° and LENS2 is 1° . Here we use linear interpolation to regrid the data to 1.5° using the underlying spherical geometry of the data, instead of performing interpolation in the lat-lon coordinates. We additionally compute daily averages of the ERA5 data to match the temporal resolution of LENS2 in the debiasing process.

SI 7 Evaluation metrics

This section details the various metrics employed to assess statistical accuracy. In particular, we focus on measuring the marginals (i.e. pointwise distribution errors), such as bias, Wasserstein distance and extreme quantiles. Additionally, we incorporate metrics that account for correlations across space, time and fields.

For completeness, the trajectory in the downscaled ensemble is represented as a five-tensor:

$$x_{i,j,t,f,m}, \quad (12)$$

where the i, j indices account for the space (latitude and longitude), t for the time, f for the different fields (or variables), and m for the members in the ensemble. The reference data from ERA5 reanalysis shares the same structure but lacks the member index, and is denoted as $x_{i,j,t,f}^{\text{ref}}$.

While most metrics involve temporal aggregation over the evaluation period, the time index can sometimes be further decomposed into three components $t = (t_h, t_d, t_y)$, representing hour, day-of-the-year, and year indices. This decomposition is commonly used in climatological computations, where each sub-index is contracted differently. In this section, however, it is only applied to the computation of the tail dependence, requiring special attention to avoid evaluations dominated by the diurnal cycle.

SI 7.1 Pointwise distribution errors

The following metrics measure the distribution difference between the predicted samples concatenated into a 5-tensor x , and the reference samples concatenated into a 4-tensor x^{ref} , where $x \in \mathbb{R}^{N_{\text{lat}} \times N_{\text{lon}} \times N_t \times N_f \times N_{\text{ens}}}$ and $x^{\text{ref}} \in \mathbb{R}^{N_{\text{lat}} \times N_{\text{lon}} \times N_t \times N_f}$. Here $N_f = 4$ (or 6 when considering the derived variables in SI 8.1), N_m is 100 for LENS2 (see SI 9.4.5), and 800 for BCSO, STAR-ESDM, QMSR, SR, and GenFocal (each LENS2 member yields 8 new downscaled samples).

SI 7.1.1 Mean absolute bias (MAB)

We define the bias as the difference between the ensemble mean of the point-wise distributions

$$\text{Bias}_{i,j,f} = \frac{1}{N_t} \left[\frac{1}{N_m} \sum_{m,t} x_{i,j,t,f,m} - \sum_t x_{i,j,t,f}^{\text{ref}} \right] \quad (13)$$

where t covers the period under consideration, e.g., summer (June-July-August) during the evaluated years. The bias for different variables is plotted in Figs. SI 10, SI 14, and SI 15 over CONUS.

The mean absolute bias is defined as the spatial average of the absolute bias,

$$\text{MAB}_f = \frac{1}{N_{\text{lon}} N_{\text{lat}}} \sum_{i,j} |\text{Bias}_{i,j,f}|. \quad (14)$$

This quantity is reported in Table SI 3 for the directly modeled variables, and in Table SI 4 for the derived variables. The MAB is also reported in the annotations in Figs. SI 10, SI 14, and SI 15.

SI 7.1.2 Mean Wasserstein distance (MWD)

The Wasserstein-1 metric for each location represents the L^1 norm between the predicted and reference distributions.

Algorithmically, this metric involves constructing empirical cumulative distribution functions CDF and CDF^{ref} for the predicted and reference samples respectively. For the first we aggregate both in time and ensemble, (t and m indices), and for the second we only aggregate in time. We can write this data dependency as

$$x_{i,j,:,f,:} \rightarrow \text{CDF}_{i,j,f}(\cdot) \quad x_{i,j,:,f}^{\text{ref}} \rightarrow \text{CDF}_{i,j,f}^{\text{ref}}(\cdot), \quad (15)$$

where the m -index is aggregated for the 800 ensemble members, and the t is aggregated during the evaluation period.

Then the pointwise Wasserstein distance is computed

$$\text{WD}_{i,j,f} = \sum_{q=1} \left| \text{CDF}_{i,j,f}(x_q) - \text{CDF}_{i,j,f}^{\text{ref}}(x_q) \right| \omega_q, \quad (16)$$

where x_q are the quadrature points over which the integrand is evaluated, and are chosen to cover the union of the support for both predicted and reference distributions; and ω_q are the quadrature weights, which in this case are defined by $\omega_q := x_{q+1} - x_q$. This quantity is shown for different variables in Fig. SI 11.

The (spatially averaged) Mean Wasserstein distance (MWD) as reported in Tables SI 3 and SI 4 is then computed as:

$$\text{MWD}_f = \frac{1}{N_{\text{lon}} \cdot N_{\text{lat}}} \sum_{i,j} \text{WD}_{i,j,f}. \quad (17)$$

SI 7.1.3 Percentile mean absolute error (MAE)

This metric measures the mean absolute difference between the p^{th} percentiles of the predicted and reference samples. For each i, j coordinate and each f field, we aggregate over the member and time indices to create histograms from which the percentiles are computed. For the reference data, we only aggregate over the time index. We use `numpy.percentile` function (abbreviated to `Pctl`) with different data following

$$x_{i,j,:,f,:} \rightarrow \text{Pctl}_{i,j,f}(\cdot) \quad x_{i,j,:,f}^{\text{ref}} \rightarrow \text{Pctl}_{i,j,f}^{\text{ref}}(\cdot). \quad (18)$$

We define the pointwise percentile error of the p^{th} percentile as

$$\text{AE}_{i,j,f}(p) = \left| \text{Pctl}_{i,j,f}(p) - \text{Pctl}_{i,j,f}^{\text{ref}}(p) \right|. \quad (19)$$

This is the quantity shown in Figs. SI 12, SI 13, SI 14, and SI 15. We also consider a spatially averaged quantity for each field given by

$$\text{MAE}_f(p) = \frac{1}{N_{\text{lon}} N_{\text{lat}}} \sum_{i,j} \text{MAE}_{i,j,f}(p). \quad (20)$$

This is the quantity reported in Table SI 3.

SI 7.2 Correlations

SI 7.2.1 Spatial correlation

For a given target location given by indices i, j and a nearby location k, l , we first compute their sample means following

$$\bar{x}_{i,j,f} = \frac{1}{N_{\text{ens}} N_t} \sum_{t,m} x_{i,j,t,f,m}, \quad \text{and} \quad \bar{x}_{k,l,f} = \frac{1}{N_{\text{ens}} N_t} \sum_{t,m} x_{k,l,t,f,m}, \quad (21)$$

which allows us to compute the correlation between locations (i, j) and (k, l) as

$$\rho_{ij,kl,f} = \frac{\sum_{t,m} (x_{i,j,t,f,m} - \bar{x}_{i,j,f}) (x_{k,l,t,f,m} - \bar{x}_{k,l,f})}{\sqrt{\sum_{t,m} (x_{i,j,t,f,m} - \bar{x}_{i,j,f})^2} \sqrt{\sum_{t,m} (x_{k,l,t,f,m} - \bar{x}_{k,l,f})^2}}. \quad (22)$$

The reference correlation is computed similarly but without aggregation in the member index, i.e.,

$$\bar{x}_{i,j,f}^{\text{ref}} = \frac{1}{N_t} \sum_t x_{i,j,t,f}^{\text{ref}}, \quad (23)$$

$$\rho_{ij,kl,f}^{\text{ref}} = \frac{\sum_t (x_{i,j,t,f}^{\text{ref}} - \bar{x}_{i,j,f}^{\text{ref}}) (x_{k,l,t,f}^{\text{ref}} - \bar{x}_{k,l,f}^{\text{ref}})}{\sqrt{\sum_t (x_{i,j,t,f}^{\text{ref}} - \bar{x}_{i,j,f}^{\text{ref}})^2} \sqrt{\sum_t (x_{k,l,t,f}^{\text{ref}} - \bar{x}_{k,l,f}^{\text{ref}})^2}}. \quad (24)$$

Computing the correlation coefficient across all nearby locations within a selected range yields the correlation matrix $P_{ij,f} = \{\rho_{ij,kl,f}\}$. This matrix is shown in the plots in [SI 3.4](#) and in Main Text Fig. 3d-e. Then we compute the pointwise spatial correlation error (SCE) as

$$\text{SCE}_{ij,kl,f} = |\rho_{ij,kl,f} - \rho_{ij,kl,f}^{\text{ref}}|, \quad (25)$$

which is shown in Main Text Fig. 3(i, j, n, o, s, and t).

Finally, the SCE is then quantified using the ℓ^1 norm as a flattened vector between the predicted and reference correlation matrices:

$$\text{SCE}_{ij,f} = \|P - P_{\text{ref}}\|_{\ell^1} = \frac{1}{N_k N_l} \sum_{k,l} |\rho_{ij,kl,f} - \rho_{ij,kl,f}^{\text{ref}}|. \quad (26)$$

This is the metric shown in all the plots of [SI 3.4](#).

SI 7.2.2 Spatial spectrum

Spatial structure can be analyzed through the power spectral density (PSD). The outputs are first transformed to frequency domain via the 2-dimensional Discrete Fourier Transform (DFT):

$$x_{:,t,f,m} \rightarrow X_{t,f,m}(\cdot, \cdot), \quad (27)$$

where X denotes the Fourier coefficients. The energy of a frequency component (ξ_k, ξ_l) is given by

$$\Phi_{t,f,m}(\xi_k, \xi_l) = \frac{1}{A} |X_{t,f,m}(\xi_k, \xi_l)|^2, \quad (28)$$

where A represents the area of the region (approximated as a rectangle) over which the spectrum is computed. The 2-dimensional spectrum is converted into a 1-dimension radial spectrum by binning along radial frequency $\xi_r = \sqrt{\xi_k^2 + \xi_l^2}$ and summing the frequency components within each bin

$$\tilde{\Phi}_{t,f,m}(\xi_r) = \sum_{\sqrt{\xi_k^2 + \xi_l^2} \in [\xi_r - \Delta\xi_r, \xi_r + \Delta\xi_r]} \Phi_{t,f,m}(\xi_k, \xi_l). \quad (29)$$

The spatial radial spectral error (SRSE) between the predicted and reference spectra is computed by first averaging along time and ensemble dimension, taking the absolute difference in their logarithms and averaging across frequencies

$$\text{SRSE}_f = \frac{1}{N_{\xi_r}} \sum_{\xi_r} \left| \frac{1}{N_t N_{\text{ens}}} \sum \log \tilde{\Phi}_{t,f,m} - \frac{1}{N_t} \sum \log \tilde{\Phi}_{t,f}^{\text{ref}} \right|, \quad (30)$$

where N_{ξ_r} denotes the number of radial frequency bins. The average spectra and errors are shown in Fig. [SI 21](#).

SI 7.2.3 Temporal spectrum

Temporal correlations in the output can be similarly analyzed through the PSD. The outputs are first transformed to the frequency space via the 1-dimensional DFT in time:

$$x_{i,j,:,f,m} \rightarrow X_{i,j,f,m}(\cdot), \quad (31)$$

with corresponding energy

$$\Phi_{i,j,f,m}(\xi_s) = \frac{1}{T} |X_{i,j,f,m}(\xi_s)|^2, \quad (32)$$

where T represents the length of the time series, ξ_s is the s th frequency component. The temporal spectral error (TSE) between the predicted and reference spectra is quantified by the mean log ratio difference:

$$\text{TSE}_{i,j,f} = \frac{1}{N_{\xi_s}} \sum_{\xi_s} \left| \frac{1}{N_{\text{ens}}} \sum_m \log \Phi_{i,j,f,m}(\xi_s) - \log \Phi_{i,j,f}^{\text{ref}}(\xi_s) \right|, \quad (33)$$

where N_{ξ_s} denotes the number of frequency components considered in the temporal DFT. We aggregate the error over spatial dimensions

$$\text{TSE}_f = \frac{1}{N_{\text{lon}} N_{\text{lat}}} \sum_{i,j} \text{TSE}_{i,j,f}, \quad (34)$$

which are shown in the last column of Fig. SI 22.

SI 7.3 Tail dependence

We evaluate the correlation of extremes of climate fields f and g through the tail dependence, estimated non-parametrically following Schmidt and Stadtmüller [80]. We start by computing the percentiles for both variables

$$x_{i,j,:,f,:} \rightarrow \text{Pctl}_{i,j,f}(\cdot) \quad x_{i,j,:,g,:} \rightarrow \text{Pctl}_{i,j,g}(\cdot), \quad (35)$$

and the co-occurrence of both variables exceeding a certain percentile

$$\Lambda_{i,j,fg}(p) = \frac{100}{N_{\text{ens}} N_t \cdot p} \sum_{t,m} \mathbf{1}_{[(x_{i,j,t,f,m} > \text{Pctl}_{i,j,f}(p)) \wedge (x_{i,j,t,g,m} > \text{Pctl}_{i,j,g}(p))]}, \quad (36)$$

where $\mathbf{1}_S$ is the indicator function that evaluates to 1 or 0 depending on whether the logical expression S is true or not. Drawing upon the homogeneity property of tail copulae [80], we compute the tail dependence by averaging over a list (length N_p) of threshold percentiles evenly spaced in the range [90, 95]:

$$\tilde{\Lambda}_{i,j,fg} = \frac{1}{N_p} \sum_{p \in [90, 95]} \Lambda_{i,j,fg}(p). \quad (37)$$

The tail dependence for the reference data is computed in a similar fashion: the only difference is the exclusion of ensemble index m in (35) and (36). The tail dependence error (TDE) is taken as the absolute difference with the corresponding reference tail dependence

$$\text{TDE}_{i,j,fg} = |\tilde{\Lambda}_{i,j,fg} - \tilde{\Lambda}_{i,j,fg}^{\text{ref}}|, \quad (38)$$

and optionally aggregated over spatial dimensions

$$\text{TDE}_{fg} = \frac{1}{N_{\text{lon}} N_{\text{lat}}} \sum_{i,j} \text{TDE}_{i,j,fg}. \quad (39)$$

This metric is reported in Main Text Fig. 3 and in Fig. [SI 16](#). Note that the tail dependence for both the upper and lower extremes can be readily assessed by negating the involved variables accordingly. For instance, we may apply transformation $g \rightarrow -g$ to evaluate the dependence of high percentiles of f and low percentiles of g .

SI 8 Evaluation protocol

In this section we describe how the derived variables are computed from the GenFocal outputs defined in Sec. SI 6.2, and how spatiotemporal events of interest are defined and detected, particularly heat streaks in Sec. SI 8.2 and TCs in Sec. SI 8.3. For the latter phenomena we also describe how the detection and calibration are performed.

SI 8.1 Derived variables

Here we describe how the derived variables are calculated. In addition to the explicitly modeled variables, we utilize surface elevation, a static quantity, to convert sea level pressure to pressure at surface height.

Relative humidity. To calculate the near-surface relative humidity, we first compute the pressure at surface height z_s using the barometric formula

$$P = P_0 \cdot \left(1 + \frac{\Gamma \cdot z_s}{T_{\text{ref}}} \right)^{-\frac{g \cdot M}{R \cdot \Gamma}}, \quad (40)$$

where P_0 denotes the sea level pressure (Pa), $T_{\text{ref}} = 288.15$ is the reference surface temperature (K), Γ is the standard tropospheric lapse rate for temperature (-0.0065 K/m), M is the molar mass of air (0.02896 kg/mol), g and R are the gravitational acceleration (9.8 m/s²) and universal gas constant (8.31447 J/mol/K) respectively.

Next we compute the saturation vapor pressure using the Clausius–Clapeyron relation [20]

$$e_s(T) = P_{\text{trip}} \left(\frac{T}{T_{\text{trip}}} \right)^\alpha \exp \left[\beta_v \left(\frac{1}{T_{\text{trip}}} - \frac{1}{T} \right) \right], \quad (41)$$

where $P_{\text{trip}} = 611$ Pa, $T_{\text{trip}} = 273.15$ K, T is the temperature at 2 meters, $\beta_v = 6773.38$ K, and $\alpha = -4.98$. Finally, we compute the actual vapor pressure as

$$e = \frac{q \cdot P}{\epsilon + (1 - \epsilon) \cdot q}, \quad (42)$$

where q denotes the near-surface specific humidity in kg/kg, and $\epsilon = 0.622$. Finally, the relative humidity is expressed as the ratio of actual vapor pressure to the saturation vapor pressure. Written as a percentage:

$$RH = \frac{e}{e_s} \cdot 100. \quad (43)$$

Heat index. The heat index quantifies the perceived temperature by modeling the human body’s thermoregulatory response to combined air temperature and humidity. It was initially introduced by Steadman [88] for a range of moderate temperatures and humidities, and recently extended to all combinations of the aforementioned variables by Lu and Romps [51]. In this work, we use the full extension of Lu and Romps, which yields the heat index as the solution of a system of algebraic equations defined by the temperature and relative humidity. A numerical solution to this system, which requires using an iterative solver, is provided in Appendix A of their original work [51].

As a cautionary tale, we note that we initially followed NOAA’s heat index definition, based on a polynomial extrapolation of Steadman’s model [57]. Using this polynomial fit to estimate summer extremes in the heat index across the continental U.S. yielded unrealistically high values of the heat index over the Rockies, the Sierra Nevada, the Cascades, and the Great Lakes. This stresses the importance of using the full definition of the heat index to explore extremes, even in the current climate [77].

SI 8.2 Heat streaks

NOAA provides 4 advisory levels based on the heat index: caution, extreme caution, danger and extreme danger; triggered by heat index values exceeding 80°F, 90°F, 103°F and 125°F (300K, 305K, 312.6K, 325K), respectively.

Here, we define heat streaks as non-overlapping s -day periods where the daily maximum heat index meets or exceeds a specified advisory level HI_{advisory} on each day. We calculate the number of s -day heat streaks from a time series of daily maximum heat indices $\{HI_{\text{max},1}, \dots, HI_{\text{max},n}\}$ as follows:

1. Identify all days where $HI_{\text{max},i} \geq HI_{\text{advisory}}$. Let the indices of these days be $\{i\}_{\text{advisory}}$.
2. Count the number of non-overlapping sequences of s consecutive indices within $\{i\}_{\text{advisory}}$. This count represents the number of s -day heat streaks, denoted as H_{advisory}^h .

For a given period (e.g., 2010-2019), we compute the annual average s -day heat streak count for each heat advisory level (i.e. {caution, extreme caution, danger, extreme danger}) across all ensemble members. The error is then the mean absolute difference between the predicted and reference annual average heat streak counts:

$$\text{heat streak error} = \left| H_{\text{advisory}}^s - H_{\text{advisory, ref}}^s \right|, \quad (44)$$

where the mean $\bar{(\cdot)}$ is calculated over the ensemble members.

SI 8.3 Tropical cyclone detection

SI 8.3.1 Criteria

Tropical Cyclones (TCs) are detected using the open-source TempestExtremes [97] software package with the following criteria:

- Downscaled time slices are analyzed at 6 hour intervals (i.e. a temporal downsampling factor of 3 with respect to the GenFocal output time resolution). LENS2 time slices are analyzed at daily intervals, matching the input time resolution.
- Local minima in sea level pressure (SLP) are identified, requiring an SLP increase of at least 200 Pa within a 5.0 great circle distance (GCD). Smaller minima within a 6.0 GCD are merged.
- Wind speed must exceed 10 m/s for at least 2 days of snapshots (8 for downscaled and 2 for LENS2). The surface elevation of the minima must remain below 100 meters for the same duration.
- The minima must persist for at least 54 hours, with a maximum gap of 24 hours in the middle.
- The maximum allowable distance between points along the same path is 8.0 GCD.

We note that detecting tropical cyclones typically requires further filtering based on upper-level geopotential gap or temperature thresholds to identify the presence of warm-core structures. Such qualifications are excluded from the definition above, as our emphasis is on downscaling near-surface variables. Nonetheless, the criteria remain consistent for both predicted and reference samples, and provide a representative assessment of the associated risks.

Instances of cyclones detected above criteria are outputted as sequences of (longitude, latitude) coordinates representing the locations of the SLP minima, along with the associated SLP values.

Table SI 6: Calibration scaling constant (K) for Tropical Cyclone (TC) detection. Values were chosen for best fit to TC count, track length, and lifetime in the training period (from $1/K \in \{0.1, 0.2, \dots, 0.9\}$).

Method	Inverse scaling constant ($1/K$)
GenFocal	0.6
BCSD	0.2
Star-ESDM	0.2
QMSR	0.2
SR	0.2
LENS2	0.1

SI 8.3.2 TCG index

We can estimate how many storms we could expect in the LENS2 ensembles using the Tropical Cyclogenesis (TCG) index [94]. This index predicts the number of storms in a region as a function of the monthly means of several different variables (wind shear, low-level vorticity, relative humidity, and sea-surface temperatures).

SI 8.3.3 Calibration

Due to inherent limitations of the LENS2 input, the magnitude of SLP depressions is systematically underestimated in downscaled projections. This results in a reduced frequency of occurrence of tropical storms and hurricanes when applying TC detection algorithms directly on the downscaled data. To address this limitation, we follow the prevalent approach of calibrating the downscaled output to match the observed frequency of TCs over a reference period [22, 39]. This is achieved via a conditional affine transformation of the magnitude of SLP depressions:

$$P_0^* = \begin{cases} KP_0 + (1 - K)P_{0,\text{amb}} & \text{if } P_0 \leq P_{0,\text{amb}} \\ P_0 & \text{if } P_0 > P_{0,\text{amb}} \end{cases} \quad (45)$$

where P_0^* denotes the calibrated SLP minimum of the tropical cyclone, $K > 1$ a calibration constant and $P_{0,\text{amb}} = 1010$ hPa represents the ambient SLP.

This calibration effectively sharpens local pressure gradients by proportionally decreasing the SLP values below the ambient threshold. It enables the detection algorithm to identify weaker signals that would otherwise be missed. We perform a sensitivity analysis across a range of K values and select the value that results in the best overall match of TC statistics (count, track length and lifetime) during the training period. The same selected scaling constant is then applied for evaluation and future projections.

This calibration procedure is applied to all baselines and ablation models to establish a consistent basis for comparison. The selected K are listed in Table SI 6. Notably, GenFocal exhibits the smallest required calibration change, as indicated by a K value closest to 1.

SI 8.3.4 Characteristics

To ensure consistent interpretation throughout this work, the definitions of the TC characteristics referred to are provided below.

Count. The total number of TCs identified within a specified region and time period.

Cyclogenesis density. A geospatial quantification of the frequency of TC formation in a given region, represented by a histogram of the first point in a TC track binned to a specified spatial resolution over a particular period. To represent the overall density, we average the frequency over the ensemble and present results in a spatial map or zonal/meridional averages.

Length. The cumulative distance (in km) traversed by a single TC instance from its genesis to dissipation.

Lifetime. The total duration (in hour) for which a TC instance maintains its identity, from its genesis to dissipation.

Pressure-derived wind. Wind speed calculated directly from the detected minimum SLP, following [2].

Saffir-Simpson category. A classification scale (tropical depression, tropical storm, and category 1 to 5 hurricanes) for TC intensity based on the pressure derived wind speed [78, 81].

Sinuosity index. A measure of the curvature of a tropical cyclone track [91].

Track density. A geospatial quantification of the frequency of TC passage through a given region, represented by a histogram of detected TC centers binned to a specified spatial resolution over a particular period. To represent the overall density, we average the frequency over the ensemble and present results either in a spatial map of raw average count (e.g. Main Text Fig. 5a) or a contour plot (e.g. Main Text Fig. 2a-c).

Landfalls. Landfall locations are identified by comparing TC tracks with a quarter-degree land-sea grid. A track point is considered over land if the nearest grid cell contains more than 50% land area. For each track, the landfall location is defined as the first point that meets this criterion. The 50% threshold land area threshold filters out the small islands of the West Indies, which is why our landfall plots do not show any landfalls over the Antilles.

SI 9 GenFocal: methodology and implementation details

GenFocal is an end-to-end statistical learning approach for downscaling. In particular, it focuses on downscaling from climate simulations to reanalysis, which is a proxy to the ground-truth weather states in the past. Once learned, the downscaling operation can be applied to future climate projections so that climate impact risks can be assessed at a high-resolution in both spatial and temporal dimensions. GenFocal grounds risk assessment of future climate projection on past observations.

The design behind GenFocal addresses three important modeling challenges in downscaling from global climate simulation to observed regional weather states (using reanalysis as a proxy in this work). First, climate simulation is coarse and thus biased with respect to fine-scaled weather. Second, the two sets of data lack temporal alignment at the granularity needed for risk assessment, such as days or hours; their correspondence is, at best, decadal. Third, downscaled states need to maintain temporal coherence over extended periods (weeks or seasons), which is crucial for robustly estimating compound extreme weather events such as tropical cyclones or heat streaks.

SI 9.1 Main idea

The main idea of GenFocal is schematically illustrated in Main Text Fig. 1. To overcome the challenges of bias and lack of granular alignment, GenFocal introduces an intermediate latent variable $y' \in \mathcal{Y}'$, a sample of the low-resolution but unbiased weather-consistent state:

$$p(x|y) = \int_{\mathcal{Y}'} p(x|y')p(y'|y) dy' = p(x|C'x = y')\delta(y' = Ty), \quad (46)$$

where C' is a deterministic *known* coarse-graining map while T is a deterministic *unknown* debiasing map, forming a Dirac distribution at the bias-corrected but low-dimensional y' .

The debiasing operator T is instantiated as a rectified flow [47] to *match the distributions* of the *global* low-resolution climate and weather spaces (see Main Text Fig. 1b). The super-resolution step $p(x|y')$ employs a conditional diffusion model [85] to add fine-grained details in space and increase the temporal resolution from daily means to 2-hourly (see Main Text Fig. 1c). To model and enhance temporal coherence, GenFocal “stacks” multiple snapshots (y_s) as inputs. The super-resolution step then employs a domain decomposition technique to ensure temporal consistency across long sequences of x (see SI 9.5.3 and Fig. SI 30).

Consequently, the design philosophy of GenFocal establishes a probabilistic description of the problem as a foundational principle. This description is necessary due to the lack of spatiotemporal correspondence between climate simulations and reanalysis data, except on the coarse levels of $\mathcal{O}(100 \text{ km})$ and decades. Concretely, any downscaling approach, physical or statistical, needs to address two issues: debiasing the input data and increasing its coarse resolution. The latter is reminiscent of image and video super-resolution, which can be tackled with statistical learning approaches. The former is very challenging as traditional statistical approaches, such as postprocessing, are inadequate due to the lack of direct correspondence required for supervised learning.

GenFocal addresses both of these challenges. From a methodological standpoint, it is noteworthy that while the two statistical learning models employed by GenFocal are named differently, they share the unified underlying theme of framing generative AI as probabilistic distribution matching and density estimation for high-dimensional random variables.

SI 9.2 Setup

We formulate the statistical downscaling problem by modeling two stochastic processes, $X_t \in \mathcal{X} := \mathbb{R}^d$ and $Y_t \in \mathcal{Y} = \mathbb{R}^{d'}$ with $d > d'$, representing a high-resolution weather process and low-resolution simulated climate process [54] respectively. These processes are governed by

$$dX_t = F(X_t, t)dt, \quad (47)$$

$$dY_t = \text{GCM}(Y_t, t)dt + \sigma(Y_t, t)dW_t, \quad (48)$$

where F embodies the generally unknown high-fidelity dynamics of X_t , and the dynamics of Y_t are often parameterized by a stochastically forced GCM [62], in which the form of σ is a modelling choice. Each stochastic process² is associated with a time-dependent measure, $\mu_x(X, t)$ and $\mu_y(Y, t)$, such that $X_t \sim \mu_x(t)$ and $Y_t \sim \mu_y(t)$, each governed by their corresponding Fokker-Planck equations. We assume an *unknown* time-invariant statistical model $C: \mathcal{X} \rightarrow \mathcal{Y}$ that relates X_t and Y_t via a possibly nonlinear downsampling map. For brevity, we omit the time-dependency of the random variables X and Y in subsequent discussion.

In general, (48) is calibrated via measurement functionals to (47) using a single observed trajectory: the historical weather. The goal of statistical downscaling is to approximate the inverse of C with a downscaling map D , trained on data for $t < T$, for a finite horizon T , such that $D_{\#}\mu_y(t) \approx \mu_x(t)$ for $t > T$. Here, $D_{\#}\mu_y(t)$ denotes the push-forward measure of $\mu_y(t)$ through D , and D is assumed to be time-independent.

Note that D is necessarily a stochastic mapping. Thus, we formulate the task of identifying D as sampling from a conditional distribution [60]. We define the operator $D \times id$, where id is the identity map, such that $(D \times id)_{\#}\mu_y(t) = D_{\#}\mu_y(t) \times \mu_y(t) \approx \mu_{x,y}(t)$, where $\mu_{x,y}(t)$ is the underlying *unknown* joint distribution. Assuming this joint distribution admits a conditional decomposition, we have:

$$\mu_{x,y}(X, Y, t) \approx D_{\#}\mu_y(X, t) \times \mu_y(Y, t) = p(X | Y)\mu_y(Y, t), \quad (49)$$

where p is time-independent.

Thus far, we have cast statistical downscaling as learning to sample from $p(x | y)$, which allows us to compute statistics of interest of $D_{\#}\mu_y(t) \approx \mu_x(t)$ via Monte-Carlo methods. We rewrite $p(x | y)$ as the conditional probability distribution $p(x | C(x) = y)$. Finally, as p is assumed time-independent we model the elements $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ as random variables with marginal distributions, μ_x and μ_y where $\mu_x = \int \mu_x(X, t)dt$ and $\mu_y = \int \mu_y(Y, t)dt$. Thus, our objective is to learn to sample $p(x | C(x) = y)$ given only access to samples of the marginals X and Y .

There are two issues: we do not know C and even if C is given (approximately), it is not obvious how we can sample efficiently from $p(x | C(x) = y)$.

SI 9.3 Overview

Without any additional assumption, it is difficult to learn C from training data. GenFocal stipulates a *structural decomposition inductive prior*:

$$C = T^{-1} \circ C', \quad \text{such that} \quad (T^{-1} \circ C')_{\#}\mu_x = \mu_y, \quad (50)$$

where C consists of two components:

²For simplicity in exposition, we follow [62] where the important time-varying effects of the seasonal and diurnal cycles have been ignored, along with jump process contributions.

- *Downsampling*³ The range of $C' : \mathcal{X} \rightarrow \mathcal{Y}'$ defines an *intermediate* space $\mathcal{Y}' = \mathbb{R}^{d'}$ of low-resolution samples with measure $\mu_{Y'} := C'_\# \mu_x$ (see Main Text Fig. 1c). The key assumption is that this step only reduces resolution but does not introduce bias.
- *Biasing* The invertible biasing map $T^{-1} : \mathcal{Y}' \rightarrow \mathcal{Y}$ defines a correspondence between the two low-dimensional spaces. Conversely, T , the inverse of this biasing map, defines the map to debias: $T_\# \mu_y = \mu_{y'} = C'_\# \mu_x$ (see Main Text Fig. 1b).

Thus, downscaling, the inverse of C , becomes a sequential two-step process:

- *Bias correction*: Apply a transport map to match the probabilistic distributions such that

$$T_\# \mu_y = C'_\# \mu_x. \quad (51)$$

- *Statistical Super-resolution*: For the joint variables $X \times Y'$, approximate $p(x | C'x = y')$.

Introducing the intermediate space $\mathcal{Y}'$ is, in equivalence, to define the conditional distribution $p(x|y)$ via a latent variable, which in return leads to (46). The Dirac distribution is chosen to reflect the deterministic and invertible mapping. An extension to probabilistic mapping is possible and left for future work.

GenFocal employs two state-of-the-art generative AI techniques to build the bias correction and super-resolution maps: the bias correction step is instantiated by a conditional flow matching method [47], whereas the super-resolution step is instantiated by a conditional denoising diffusion model [84] coupled with a domain decomposition strategy [9] to upsample both in space and time and create time-coherent sequences.

SI 9.4 Bias correction

For debiasing, since the samples from $\mathcal{Y}$ and $\mathcal{Y}'$ are not aligned, we seek a map between the distributions. This is a weaker notion than sample-to-sample correspondence, which physics-based downscaling methods might be able to offer. In exchange, statistical distribution match, as shown in this work, can also be effective in debiasing yet remaining computationally advantageous.

The notion of distribution matching has a long history in applied mathematics going back to Gaspar Monge in the late 1700s and Leonid Kantorovich in the 50s, who formalized this idea, and kicked off the field of optimal transport [100]. In our context, the optimal transport framework would seek to solve the problem

$$\min_T \int c(Ty, y) d\mu_y(y) \text{ with } T_\# \mu_y = \mu_{y'} := C'_\# \mu_x, \quad (52)$$

for a cost function c measuring the cost moving ‘probabilistic mass’. Note that following this approach, the debiasing map T satisfies the constraint in (51) by construction.

Due to limitations of existing methods for solving (52) (which are briefly summarized in SI 1.3), we adopt a rectified flow approach [47], a methodology under the umbrella of generative models. Rectified flow results in a *invertible* map instantiated by the solution map of an ODE, which solves an entropy-regularized optimal transport problem [46], and it has empirically shown to be well suited for relatively large dimension (as compared to control based approaches such as neural ODE [15]), and it has a relatively low sample complexity.

³Here we suppose that the downsampling map acts both in space and in time, by using interpolation in space, and by averaging in time using a window of one day.

SI 9.4.1 Rectified flow

Rectified flow constructs the debiasing map T as the solution map of an ODE given by

$$\frac{dy}{d\tau} = v_\phi(y, \tau) \quad \text{for } \tau \in [0, 1], \quad (53)$$

whose vector field $v_\phi(x, \tau)$ is parameterized by a neural network (see SI 9.4.3 for further details). By identifying the input of the map as the initial condition, we have that $T(y) := y(\tau = 1)$. We train v_ϕ by solving

$$\ell(\phi) = \min_{\phi} \mathbb{E}_{\tau \sim \mathcal{U}[0,1]} \mathbb{E}_{(y_0, y_1) \sim \pi \in \Pi(\mu_y, \mu_{y'})} \|(y_1 - y_0) - v_\phi(y_\tau, \tau)\|^2, \quad (54)$$

where $y_\tau = \tau y_1 + (1 - \tau)y_0$. $\Pi(\mu_y, \mu_{y'})$ is the set of couplings with marginals given by the distributions from $\mathcal{Y}$ and $\mathcal{Y}'$ respectively. Once v_ϕ is learned, we debias any given y by solving (53) using the 4th-order Runge-Kutta solver.

SI 9.4.2 Modeling details

Our implementation of rectified flow introduces several custom modeling choices that are highly effective when dealing with climate data: (a) modeling in the anomaly space; (b) modeling the seasonality of climate by climatological distribution coupling; (c) modeling temporal coherence.

Let $y \in \mathcal{Y}$ denote a biased low-resolution sequence of consecutive snapshots (namely, the climate state at times $t, t + \Delta t, \dots, t + n_s \Delta t$), and $y' \in \mathcal{Y}'$ denote an unbiased low-resolution sequence, where $\mathcal{Y}'$ is the image of $\mathcal{X}$ through the linear downsampling map C' (see Main Text Fig. 1a). In our setup, the space of biased low-resolution dataset $\mathcal{Y}$ is given by a collection of 100 trajectories from the LENS2 ensemble dataset. Each trajectory, which we denote by $\mathcal{Y}^i$ (such that $\mathcal{Y} = \bigcup_i \mathcal{Y}^i$), has slightly different spatiotemporal statistics that we leverage to further extract statistical performance from our debiasing step. We characterize the statistics of each trajectory using their climatological mean and standard deviation in the training set, namely $\bar{y}^i$ and σ_y^i , which are estimated using the samples within the training range. The space of the unbiased low-resolution sequences $\mathcal{Y}'$, is given by the daily means of the ERA5 historical data regridded to 1.5° resolution. We denote the climatological mean and standard deviation of the set as $\bar{y}'$ and $\sigma_{y'}$ respectively.

To render the training more efficient, we normalize the input and output data using their *climatology* following: $\hat{y} = (y - \bar{y}^i)/\sigma_y^i$ for $y \in \mathcal{Y}^i$, and $\hat{y}' = (y' - \bar{y}')/\sigma_{y'}$. Then we seek to find the smallest deviation between the two *anomalies*.

We specialize the map T as follows. We incorporate the climatological mean and standard deviation into the vector field $v_\theta(y, \tau; \bar{y}^i, \sigma_y^i)$ and identify the solution of the revised ODE

$$\frac{d\hat{y}}{d\tau} = v_\theta(\hat{y}, \tau; \bar{y}^i, \sigma_y^i), \quad \text{for } \tau \in (0, 1), \quad (55)$$

at the terminal time as the unbiased anomaly, i.e., $\hat{y}' = \hat{y}(1)$. This is then de-normalized, resulting in $Ty = y' = \hat{y}' \odot \sigma_{y'} + \bar{y}'$, where $\odot$ is the Hadamard product.

The training loss is revised accordingly

$$\min_{\theta} \mathbb{E}_{i \in \mathbb{I}} \mathbb{E}_{\tau \sim \mathcal{U}[0,1]} \mathbb{E}_{(y_0, y_1) \sim \pi \in \Pi(\mu_{y^i}, \mu_{y'})} \|\hat{y}_0 - \hat{y}_1 - v(\hat{y}_\tau, \tau; \hat{y}^i, \sigma_y^i)\|^2, \quad (56)$$

where $\hat{y}_\tau = \tau \hat{y}_1 + (1 - \tau)\hat{y}_0$, $\Pi(\mu_{y^i}, \mu_{y'})$ is the set of couplings with climatologically aligned marginals, and $\mathbb{I}$ are the indexes of the training trajectories instantiated by the different LENS2 ensemble members.

The choice of the coupling implicitly defines the spatiotemporal structure of the bias to be rectified. We assume a seasonally-varying bias by coupling data pairs that correspond to similar time stamps (possibly up to a couple of years) for both LENS2 and ERA5 samples. Although, for simplicity in this case we use a coupling that uses the same time-stamps for both LENS2 and ERA5 samples to ensure the same climatology and to capture the correct slowly-varying drift in the distributions induced by the climate change signal. For an ablation study see section [SI 10.6](#)

Time-coherence is implicitly included in this step. At each iteration, data is extracted from a long contiguous sequence of snapshots. For example, with a batch size of 16 and a debiasing sequence length of 8, we extract $128 = 16 \times 8$ consecutive snapshots from a single LENS2 member and ERA5. That long sequence is then divided into 16 short sequences and fed to each core. We observed that choosing short sequences from the training dataset in a fully independent manner was prone to overfitting; this effect was attenuated by feeding a batch of contiguous sequences as described above. This approach also helps optimize training by reducing data loading latency, as it minimizes the number of reads from disk.

For the length of each debiasing sequence, empirically we found that 2–8 contiguous days provides good performance on the validation set. For an ablation study of the chunk size, please see Section [SI 10](#). Once the model is trained, we solve (55) using an adaptive Runge-Kutta solver, which allows us to align the simulated climate manifold to the weather manifold.

SI 9.4.3 Neural architecture

For the architecture we use a 3D U-ViT [8], with 6 levels. The input to the network are three 4-tensors, $\hat{y}$, $\bar{y}^i$, and σ_y^i ; each of dimensions $8 \times 240 \times 121 \times 10$ plus a scalar corresponding to the evolution time τ . Here the 8 corresponds to the 8 contiguous days, and the 10 channels correspond to the surface and level fields being modeled as shown in Table [SI 5](#). The output is one 4-tensor corresponding to the instant velocity of $\hat{y}_\tau$. In this case, $\bar{y}^i$, and σ_y^i are used as conditioning vectors. These variables are interpolated to the new grid, and pre-processed using a convolutional neural network, then they are concatenated to $\hat{y}$ along the channel dimension.

Resize and aggregation layers for encoding As the spatial dimensions of input tensors, 240×121 , are not easily amenable to downsampling, i.e, they are not multiples of small prime numbers, we use a resize layer at the beginning. The resize layers performs a cubic interpolation to obtain a 3-tensor of dimensions $8 \times 128 \times 10$, followed by a two-dimensional convolutional network with lat-lon boundary conditions: periodic in the longitudinal dimension (using the `jax.numpy.pad` function with `wrap` mode) and constant padding in the latitudinal dimension, which repeats the value at the end of the array (using the `jax.numpy.pad` function with `edge` mode).

For the $\hat{y}$ inputs, the convolutional network works as a dealiasing step. It has a kernel size of (7, 7), which we write as:

$$h_{\hat{y}} = \text{Conv2D}(8, 7, 1) \circ \mathcal{I}(\hat{y}), \quad (57)$$

where $\text{Conv2D}(N, k, s)$ denotes a convolutional layer with N filters, kernel size (k, k) and stride (s, s) .

The conditioning inputs, i.e., the statistics $\bar{y}^i$, and σ^i , go through a slightly different process: they are also interpolated, but they go through a shallow convolutional network composed of one two-dimensional convolutional layers followed by a normalization layer with a swish activation function, and another two-dimensional convolutional layer. Here, both convolutional layers have a kernel size (3, 3). The first has an embedding dimension of 10 as it acts as an anti-aliasing layer while the second has an embedding dimension of 128 as it seeks to project the information into the embedding space. In summary, we have

$$h_{\bar{y}^i} = \text{Conv2D}(128, 3, 1) \circ \text{Swish} \circ \text{LN} \circ \text{Conv2D}(4, 3, 1) \circ \mathcal{I}(\bar{y}^i), \quad (58)$$

$$h_{\sigma^i} = \text{Conv2D}(128, 3, 1) \circ \text{Swish} \circ \text{LN} \circ \text{Conv2D}(4, 3, 1) \circ \mathcal{I}(\sigma^i). \quad (59)$$

Then all the fields are concatenated along the channel dimension, i.e.,

$$h = \text{Concat}[h_{\tilde{y}}; h_{\tilde{y}^i}; h_{\sigma^i}], \quad (60)$$

of dimensions $8 \times 256 \times 128 \times 266$. The last dimension comes from the concatenation of $h_{\tilde{y}}$ which has channel dimension 10, together with $h_{\tilde{y}^i}$ and h_{σ^i} , which have a channel dimension of 128 each.

Spatial downsampling stack After the inputs are rescaled, projected and concatenated, we feed the composite fields to an U-ViT. For the downsampling stack we use 4 levels, at each level we downsample by a factor two in each dimension, while increasing the number of channels by a factor of two, so we only have a mild compression as we descend through the stack.

The first layer takes the output of the merge and resizing, and we perform a projection

$$h_0 = \text{Conv2D}(128, 3, 1)(h), \quad (61)$$

where h is the latent input from the encoding step. Then h_0 is successively downsampled using a convolution with stride (2, 2), and an embedding dimension of hidden_i , where i is the level of the U-Net.

$$h_{i,0}^{\text{down}} = \text{Conv2D}(\text{hidden}_i, 1, 2)(h_{i-1, n_{res}-1}), \quad (62)$$

where n_{res} is the number of resnet at each level, and hidden_i is the dimension of the hidden states for each level as given in Table SI 7. The output of the downsampled embedding is then then processed by a sequence of $n_{res} = 6$ resnet blocks following:

$$h_{i,j+1}^{\text{down}} = h_{i,j}^{\text{down}} + \text{Conv2D}(c^i, 3, 1) \circ \text{Do}(0.5) \circ \text{Swish} \circ \text{FiLM}(e) \circ \text{GN} \circ \text{Conv2D}(c^i, 3, 1) \circ \text{Swish} \circ \text{GN}(h_{i,j}^{\text{down}}), \quad (63)$$

where $c^i = \text{hidden}_i$, the number of channels at each level, Do_p is dropout layer with probability p , here j runs from 0 to $n_{res} - 1$. In addition, time embedding e , is processes with a Fourier embedding layer with a dimension of 256, which is then used in conjunction with a FiLM layer following

$$\begin{aligned} \text{FiLM}(x; \sigma_\tau) &= (1.0 + \text{Dense} \circ \text{FourierEmbed}(\sigma_\tau)) \cdot x + \text{Dense} \circ \text{FourierEmbed}(\sigma_\tau), \\ \text{FourierEmbed}(\sigma_\tau) &= \text{Dense} \circ \text{SiLU} \circ \text{Dense} \circ \text{Concat}([\cos(\alpha_k \sigma_\tau), \sin(\alpha_k \sigma_\tau)]_{k=1}^K) \end{aligned} \quad (64)$$

where α_k are non-trainable frequencies evenly spaced on a logarithmic scale between 0 and 10000, and $K = 128$. Finally, GN stands for a group normalization layer with 4 groups.

Attention Processing For the attention layers we use a ViViT-like model with 2D position encoding, axial transformer in each direction, 128 heads, the token sizes depends at which level the attention processing is performed. Also, the temporal and spatial attentions are decoupled so they can be used (or not) independently.

Spatial upsampling stack The upsampling stack takes the downsampled latent variables and sequentially increases their resolution while merging them with skip connections until the original resolution is reached. This process, within the U-ViT model, is completely different from the super-resolution stage of the framework as shown in Main Text Fig. 1, which is treated in detail in SI 9.5 The upsampling stack contains the same number of levels and residual blocks as the downsampling one. At each level, it adds the corresponding skip connection in the upsampling stack:

$$h_{i,0}^{\text{up}} = h_{i,0}^{\text{up}} + h_{i,0}^{\text{down}}, \quad (65)$$

followed by the same blocks defined in (63), followed by an upsampling block

$$h_{i-1, n_{res}-1}^{\text{up}} = \text{Conv2D}(\text{hidden}_{i-1}, 3, 1) \circ \text{channel2space} \circ \text{Conv2D}(\text{hidden}_i \cdot 2^2, 3, 1) \circ h_{i, n_{res}-1}^{\text{up}}, \quad (66)$$

where the **channel2space** operator expands the $\text{hidden}_i \cdot 2^2$ channels into $2 \times 2 \times \text{hidden}_i$ blocks locally, effectively increasing the spatial resolution by 2 in each direction.

Table SI 7: Hyperparameters for the debiasing model.

Debias architecture	
Output shape	$8 \times 240 \times 121 \times 10$
Spatial downsampling ratios	[2, 2, 2, 2, 2, 2]
Residual blocks	[6, 6, 6, 6, 6, 6]
Hidden channels	[768, 768, 768, 1024, 1280, 1536]
Axial attention layers in space	[False, False, False, False, True, True]
Axial attention layers in time	[False, False, False, True, True, True]
Trainable parameters	2,656,553,626
Training	
Device	TPU v5p, 4×4
Duration	500,000 steps
Batch size	16 (with data parallelism)
Learning rate	cosine annealed (peak= 1×10^{-4} , end= 1×10^{-7}), 1,000 linear warm-up steps
Gradient clipping	max norm = 0.6
Time sampling	$\mathcal{U}(10^{-3}, 1 - 10^{-3})$
Condition dropout (p_u)	0.5
Inference	
Device	8×Nvidia H100s
Integrator	Runge-Kutta 4th order
Solver number of steps	100

Decoding and resizing We apply a final block to the output of the upsampling stack.

$$x_{\text{out}} = \text{Conv2D}(10, 3, 1) \circ \text{SiLU} \circ \text{LayerNorm} \circ h_0^{\text{up}}. \quad (67)$$

followed by a resizing layer as the one defined in (57), with number of channels equal to the number of input fields. This operation brings back the output to the size of the input.

SI 9.4.4 Hyperparameters

Table SI 7 shows the set of hyperparameters used for the flow architecture, as well as those applied during the training and sampling phases of the rectified flow model. We also include the optimization algorithm used for minimizing (56), along with the learning rate scheduler and weighting.

SI 9.4.5 Training, evaluation and test data

We trained the debiasing stage of GenFocal using 4 LENS2 members `cmip6_1001_001`, `cmip6_1251_001`, `cmip6_1301_010`, and `smbb_1301_020`, using data from 1980 to 1999. We point out that the first three members share the same forcing using the original CMIP6 BMB protocol [76], but different initializations to sample internal variability, whereas the last one uses a smoothened version of the same forcing (see details in [76]). Debiasing is performed with respect to the coarse-grained ERA5 data for the same period.

For model selection we used the following 14 LENS2 members: `cmip6_1001_001`, `cmip6_1041_003`, `cmip6_1081_005`, `cmip6_1121_007`, `cmip6_1231_001`, `cmip6_1231_003`, `cmip6_1231_005`, `cmip6_1231_007`, `smb_1011_001`, `smbb_1301_011`, `cmip6_1281_001`, `cmip6_1301_003`, `smbb_1251_013`, and `smbb_1301_020`, using data from 2000 to 2009.

For testing we use the full 100-member LENS2 ensemble from 2010 to 2019. The full ensemble used for testing contains members with different forcings and perturbations.

SI 9.4.6 Computational cost

Training the rectified flow model took approximately three days using four TPU v5p nodes (16 cores total), with one sample per core. Each host loaded a sequence of 32 contiguous daily snapshots per iteration (four sequences of 8 consecutive snapshots), which were then distributed among the cores. For inference, each sample of 8 snapshots takes around 45 seconds to be debiased in an H100. The full debiasing step took around 9 hours to process each 140-year ensemble member on a host with 8 H100 GPUs. As the process is embarrassingly parallel, debiasing the full 100-member LENS2 ensemble for 140 years took about 9 hours using 100 nodes, each equipped with 8 H100 GPUs. Estimating each H100 costs about \$5 USD per hour at the current market rate, this debiasing step costs about \$36,000 USD and can be further reduced through engineering optimization.

SI 9.5 Super-resolution

In contrast to bias correction, super-resolution is a probabilistic supervised learning problem. The coarse-graining map C' is an operation by downsampling the ERA5 data from 2-hourly and 0.25° to daily 1.5° , thus forming a pair of aligned data sample $(y'_i = C'x_i, x_i)$. To learn the super-resolution operation, i.e., the inverse of the downsampling, we use a conditional diffusion model [84, 86], popularized by latest advances in image and video generation.

SI 9.5.1 Conditional diffusion model

In this section, we provide a brief high-level description of the generic diffusion-based generative modeling framework. While different variants exist, we mostly follow that of [41] and refer interested readers to its Appendix for a detailed explanation of the methodology.

Diffusion models are a type of generative model that work by gradually adding Gaussian noise to real data until they become indistinguishable from pure noise (forward process). The unique power of these models is their ability to reverse this process, starting from noise and progressively refining it to create new samples that resemble the original data (backward process, or sampling).

Mathematically, we describe the forward diffusion process as a stochastic differential equation (SDE)

$$dz_\tau = \sqrt{\dot{\sigma}_\tau \sigma_\tau} d\omega_\tau, \quad z_0 \sim p_{\text{data}}, \quad \tau \sim [0, 1] \quad (68)$$

where σ_τ is a prescribed noise schedule and a strictly increasing function of the diffusion time τ (note: to be distinguished from real physical time t), $\dot{\sigma}_\tau$ denotes its derivative with respect to τ , and ω_τ is the standard Wiener process. The linearity of the forward SDE implies that the distribution of z_τ is Gaussian given z_0 :

$$q(z_\tau | z_0) = \mathcal{N}(z_\tau; z_0, \sigma_\tau^2 I), \quad (69)$$

with mean z_0 and diagonal covariance $\sigma_\tau^2 I$. For $\tau = 1$, i.e. the maximum diffusion time, we impose $\sigma_{\tau=1} \gg \sigma_{\text{data}}$ such that $q(z_1 | z_0)$ can be faithfully approximated by the isotropic Gaussian $\mathcal{N}(z_1; 0, \sigma_1^2 I) := q_1$.

The main underpinning of diffusion models is that there exists a *backward SDE*, which, when integrated from $\tau = 1$ to 0, induces the same marginal distributions $p(z_\tau)$ as those from the forward SDE (68) [1, 86]:

$$dz_\tau = -2\dot{\sigma}_\tau \sigma_\tau \nabla_{z_\tau} \log p(z_\tau, \sigma_\tau) d\tau + \sqrt{2\dot{\sigma}_\tau \sigma_\tau} d\omega_\tau. \quad (70)$$

In other words, with full knowledge of (70) one can easily draw samples $z_1 \sim q_1$ to use as the initial condition and run a SDE solver to obtain the corresponding x_0 , which resembles a real sample from p_{data} . However, in (70), the term $\nabla_{z_\tau} \log p(z_\tau, \sigma_\tau)$, also known as the *score function*, is not directly known. Thus, the primary machine learning task associated with diffusion models is centered around expressing and approximating the score function with a neural network, whose parameters are learned from data. Specifically, the form of parameterization is inspired by Tweedie’s formula [21]:

$$\nabla_{z_\tau} \log p(z_\tau, \sigma_\tau) = \frac{\mathbb{E}[z_0|z_\tau] - z_\tau}{\sigma_\tau^2} \approx \frac{D_\theta(z_\tau, \sigma_\tau) - z_\tau}{\sigma_\tau^2}, \quad (71)$$

where D_θ is a *denoising* neural network that predicts the clean data sample z_0 given a noisy sample $z_\tau = z_0 + \varepsilon\sigma_\tau$ (ε is drawn from a standard Gaussian $\mathcal{N}(0, I)$). Training D_θ involves sampling both data samples z_0 and diffusion times τ , and optimizing parameters θ with respect to the mean denoising loss defined by

$$\mathcal{L}(\theta) = \mathbb{E}_{z_0 \sim p_{\text{data}}} \mathbb{E}_{\tau \sim [0, T]} [\lambda_\tau \|D_\theta(z_0 + \varepsilon\sigma_\tau, \sigma_\tau) - z_0\|^2], \quad (72)$$

where λ_τ denotes the loss weight for noise level τ . We use the weighting scheme proposed in [41] as well as the pre-conditioning strategies therein to improve training stability.

At sampling time, the score function in SDE (70) is substituted with the learned denoising network D_θ using expression (71).

In the case that an input is required, i.e. sampling from conditional distribution $p(z_\tau|y)$, the input y is incorporated by the denoiser D_θ as an additional input. Classifier-free guidance (CFG) [37] may be employed to trade off between maintaining coherence with the conditional input and increasing coverage of the target distribution. Concretely, CFG is implemented through modifying the denoising function $\tilde{D}_\theta$ at sampling time:

$$\tilde{D}_\theta = (1 + g)D_\theta(z_\tau, \sigma_\tau, y) - gD_\theta(z_\tau, \sigma_\tau, \emptyset), \quad (73)$$

where g is a scalar that controls the guidance strength (increasing g means paying more attention to y) and $\emptyset$ denotes the null conditioning input (i.e., a zero-filled tensor with the same shape as y), such that $D_\theta(x_\tau, \sigma_\tau, \emptyset)$ represents unconditional denoising. The unconditional and conditional denoisers are trained jointly using the same neural network model, by randomly dropping the conditioning input from training samples with probability p_u .

SI 9.5.2 Modeling details

We specialize the general framework of conditional distribution models to modeling weather and climate data. GenFocal has several specific components that take into consideration the unique properties of the data to facilitate learning.

We take advantage the prior knowledge that a spatially-interpolated linear mapping $\mathcal{I}(y')$ is a strong approximation to x by modeling the residual $r := x - \mathcal{I}(y')$ by using the conditional diffusion model to sample from $p(r|y')$ and add the residual back to $\mathcal{I}(y')$ as the final output of the super-resolution. Furthermore, a substantial portion of the variability in r_h is due to its strong seasonal and diurnal periodicity. To avoid learning these predictable patterns and direct the model’s focus toward capturing non-trivial interactions, we learn to sample $\tilde{r}$, the residual normalized by its climatological mean and standard deviation computed over the training dataset:

$$\tilde{r} = \frac{r - \text{clim_mean}[r]}{\text{clim_std}[r]}. \quad (74)$$

The input y' is also strongly seasonal. However, we do not remove its seasonal components and instead normalize with respect to its date-agnostic mean and standard deviation:

$$\tilde{y}' = \frac{y' - \text{mean}[y]}{\text{std}[y]}, \quad (75)$$

which ensures that the model is still able to leverage the seasonality in the input signals when deriving its output.

In summary, samples are obtained as

$$x(y'; \omega) = \mathcal{I}(y') + \text{clim_mean}[r] + \text{clim_std}[r] \cdot S(\tilde{y}'; \omega) \quad (76)$$

where $S(\tilde{y}'; \omega)$ denotes the sampling function (i.e. solving the reverse time SDE end-to-end) for $\tilde{r}$ given the normalized coarse-resolution input $\tilde{y}'$, and a noise realization ω .

SI 9.5.3 Sampling long temporal sequence

After the denoiser is trained, we may initiate a backward diffusion process by solving (70) from $\tau = 1$ to $\tau = 0$, using initial condition $z_1 \sim q_1$. We employ a first-order exponential solver, whose step formula (going from noise level σ_i to σ_{i-1}) reads

$$z_{i-1} = \frac{\sigma_{i-1}^2}{\sigma_i^2} z_i + \left(1 - \frac{\sigma_{i-1}^2}{\sigma_i^2}\right) D_\theta(z_\tau, \sigma_\tau, \tilde{y}') + \frac{\sigma_{i-1}}{\sigma_i} \sqrt{\sigma_i^2 - \sigma_{i-1}^2} \varepsilon, \quad (77)$$

where $\varepsilon \sim \mathcal{N}(0, I)$. The generated sample would be the residual for a 7-day window (i.e. model duration) corresponding to the daily mean in $\tilde{y}'$.

To generate an arbitrarily long sample trajectory with temporal coherence, we stagger multiple 7-day windows, denoted by $\{S_0, \dots, S_{M-1}\}$, such that there is a one-day overlap between neighboring windows S_j and $S_{j\pm 1}$. A separate backward diffusion process is initiated on each period S_j , leading to denoise outputs $\{d_j\}$. As such, each overlapped window has two distinct denoise outputs at every step, denoted d_j^{right} and d_{j+1}^{left} , which in general do *not* take on the same values despite the corresponding inputs z_j^{right} and z_{j+1}^{left} being the same.

To consolidate, we take the average between them, and replace the corresponding outputs on both sides to ensure that d_j is consistent between the left and right denoising windows. This in turn creates a “shock” that renders the overlapped region *less coherent* with respect to the other parts in their respective native denoising windows. However, the incoherence are expected to be insignificant under the presence of noise and more importantly, should decrease in magnitude as the backward process proceeds and the noise level reduces. At the end of denoising, one would expect a fully coherent sample across all denoising windows. A schematic for this technique is shown in Fig. SI 30.

Mathematically, the step formula in the overlapped region can be described as

$$\begin{aligned} z_{i-1,j}^{\text{right}} &= \frac{\sigma_{i-1}^2}{\sigma_i^2} z_{i,j}^{\text{right}} + \frac{1}{2} \left(1 - \frac{\sigma_{i-1}^2}{\sigma_i^2}\right) (d_{i,j}^{\text{right}} + d_{i,j+1}^{\text{left}}) \\ &\quad + \frac{\sigma_{i-1}}{\sigma_i} \sqrt{\sigma_i^2 - \sigma_{i-1}^2} \varepsilon_j^{\text{right}}. \end{aligned} \quad (78)$$

It is important to note that the random vector in the same overlapped region should be identical, i.e. $(\varepsilon_j^{\text{right}} = \varepsilon_{j+1}^{\text{left}})$.

The complete sampling procedure is described in Algorithm 1. In practice, we place each denoising window on a different TPU core so that all windows can be denoised in parallel. Consolidation of overlapping windows then takes place through collective permutation operations (`lax.ppermute` functionality in JAX), which efficiently exchanges information among cores.

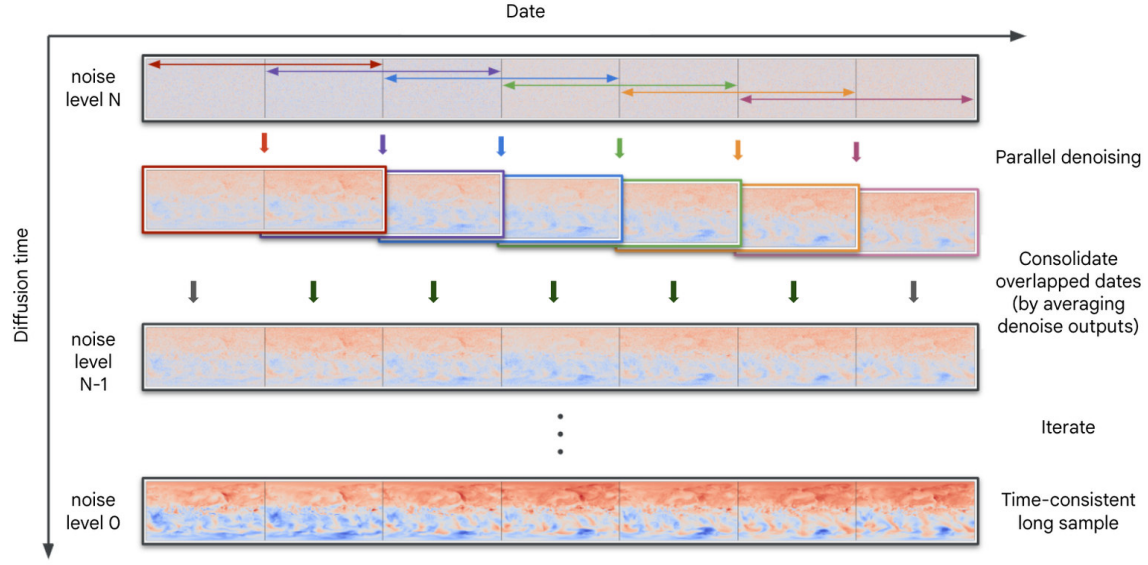


Figure SI 30: Schematic of long trajectory sampling using parallel section denoisers.

SI 9.5.4 Neural architecture

The diffusion model denoiser D_θ is implemented using a U-ViT, which consists of a downsampling and an upsampling stack, each composed of convolutional and axial attention layers. The denoiser takes as inputs noised samples z_τ , the conditioning inputs $\tilde{y}'$, and the noise level σ_τ . The output is the climatology-normalized residual sample

$$\tilde{r}_h = D_\theta(z_\tau, \sigma_\tau, \tilde{y}'). \quad (79)$$

The output samples $\tilde{r}_h$ span D_{lon} degrees in longitude, D_{lat} degrees in latitude and 7 days in time, leading to tensor shape $84 \times 4D_{\text{lon}} \times 4D_{\text{lat}} \times 4$ (quarter degree spatial and bi-hourly temporal resolutions), whose dimensions representing time, longitude, latitude and variable dimensions respectively. z_τ is a noisy version of $\tilde{r}_h$ and thus share the same size. $\tilde{y}'$ also has the same number of dimensions, but is in lower resolution with shape $7 \times D_{\text{lon}} \times D_{\text{lat}} \times 4$, while σ_τ is a scalar.

Encoding. The input $\tilde{y}'$ is merged with the noisy sample z_τ . We first apply an encoding block

$$h_{\tilde{y}'} = \text{Conv2D}(192, 3, 1) \circ \text{SiLU} \circ \text{LN} \circ \text{Conv2D}(4, 7, 1) \circ \text{Interp} \circ \tilde{y}', \quad (80)$$

which first transfers $\tilde{y}'$ to the same shape as z_τ through interpolation (cubic in space and nearest neighbor in time), followed by a layer normalization (LN), sigmoid linear unit (SiLU) activation and a spatial convolutional layer (parameters inside the brackets indicate output feature dimension, kernel size and stride respectively) that encode the input into latent features. The latent features are concatenated with z_τ in the channel dimension and projected by a convolutional layer to form the input to the subsequent downsampling stack:

$$h = \text{Conv2D}(128, 3, 1) \circ \text{Concat}([z_\tau, h_{\tilde{y}'}]). \quad (81)$$

Downsampling stack. The downsampling stack consists of a sequence of levels, each at a coarser resolution than the previous. Each level, indexed by i , further comprises a strided convolutional layer

Algorithm 1 Sampling long trajectories using overlapped denoisers. Each denoiser takes $84 \times 4D_{\text{lon}} \times 4D_{\text{lat}} \times 4$ input noise shape and generates outputs of the same shape. With 1-day overlap windows and $M = 16$ denoisers, the total trajectory shape amounts to $1164 \times 4D_{\text{lon}} \times 4D_{\text{lat}} \times 4$ (97 days).

```

1: procedure LONGTRAJECTORYSAMPLER( $D_\theta(z, \sigma, y)$ ,  $\sigma_i \in \{N, \dots, 0\}$ ,  $S_j \in \{0, \dots, M-1\}$ )
2:   sample  $z_N \sim \mathcal{N}(0, \sigma_N^2 I)$  ▷ Sample shape is the that of the overall trajectory.
3:    $\{z_{N,0}, \dots, z_{N,M-1}\} \leftarrow \text{extract}(z_N, \{S_0, \dots, S_{M-1}\})$  ▷ Each  $z_{N,j}$  is in denoiser shape.
4:   for  $i \in \{N, \dots, 1\}$  do ▷ Iterate over diffusion steps.
5:     for  $j \in \{0, \dots, M-1\}$  do
6:        $d_{i,j} \leftarrow D_\theta(z_{i,j}, \sigma_i, y_j)$  ▷ Denoise each section independently.
7:     end for
8:     for  $j \in \{0, \dots, M-1\}$  do
9:        $d_{i,j}^{\text{left}} \leftarrow (d_{i,j}^{\text{left}} + d_{i,j-1}^{\text{right}})/2$  ▷ Consolidate with left neighbor (for  $j \neq 0$ ).
10:       $d_{i,j}^{\text{right}} \leftarrow (d_{i,j}^{\text{right}} + d_{i,j+1}^{\text{left}})/2$  ▷ Consolidate with right neighbor (for  $j \neq M-1$ ).
11:    end for
12:    for  $j \in \{0, \dots, M-1\}$  do ▷ Update overlapping regions in the denoise targets.
13:       $d_{i,j} \leftarrow \text{setLeft}(d_{i,j}, d_{i,j}^{\text{left}})$ 
14:       $d_{i,j} \leftarrow \text{setRight}(d_{i,j}, d_{i,j}^{\text{right}})$ 
15:    end for
16:    sample  $\varepsilon_j \sim \mathcal{N}(0, I)$  ▷ Draw new noise for the current SDE step.
17:     $\{\varepsilon_{i,0}, \dots, \varepsilon_{i,M-1}\} \leftarrow \text{extract}(\varepsilon_j, \{S_0, \dots, S_{M-1}\})$  ▷ The same overlap region gets the same noise.
18:    for  $j \in \{0, \dots, M-1\}$  do ▷ Apply consolidated exponential denoise update.
19:       $z_{i-1,j} \leftarrow (\sigma_{i-1}^2/\sigma_i^2)z_{i,j} + (1 - \sigma_{i-1}^2/\sigma_i^2)d_{i,j} + (\sigma_{i-1}/\sigma_i)\sqrt{\sigma_i^2 - \sigma_{i-1}^2}\varepsilon_{i,j}$ 
20:    end for
21:  end for
22:   $z_0 \leftarrow \text{combine}(\{z_{0,0}, \dots, z_{0,M-1}\}, \{S_0, \dots, S_{M-1}\})$  ▷ Combines denoiser sections into a complete trajectory.
23:  return  $z_0$ 
24: end procedure

```

that applies spatial downsampling

$$h_{i,0}^{\text{down}} = \text{Conv2D}(c_i, 3, q_i) \circ h_{i-1}^{\text{down}}, \quad (82)$$

followed by 4 residual blocks defined by

$$h_{i,j}^{\text{down}} = h_{i,j-1}^{\text{down}} + \text{Conv2D}(c_i, 3, 1) \circ \text{SiLU} \circ \text{FiLM}(\sigma_\tau) \circ \text{LN} \circ \text{Conv2D}(c_i, 3, 1) \circ \text{SiLU} \circ \text{LN} \circ h_{i,j-1}^{\text{down}} \quad (83)$$

where j denotes the index of the residual block. FiLM is a linear modulation layer

$$\begin{aligned} \text{FiLM}(x; \sigma_\tau) &= (1.0 + \text{Linear} \circ \text{FourierEmbed}(\sigma_\tau)) \cdot x + \text{Linear} \circ \text{FourierEmbed}(\sigma_\tau), \\ \text{FourierEmbed}(\sigma_\tau) &= \text{Linear} \circ \text{SiLU} \circ \text{Linear} \circ [\cos(\alpha_k \sigma_\tau), \sin(\alpha_k \sigma_\tau)], \end{aligned} \quad (84)$$

where α_k are non-trainable embedding frequencies evenly spaced on a logarithmic scale.

At higher downsampling levels (corresponding to lower resolutions), we additionally apply a sequence of axial multi-head attention (MHA) layers along each dimension (both spatial and time) at the end of each block, defined by

$$h_i^{\text{down}} = h_i^{\text{down}} + \text{Linear}(c_i) \circ \text{MHA}(k) \circ \text{LayerNorm} \circ \text{PosEmbed}(k) \circ h_i^{\text{down}}, \quad (85)$$

where k denotes the axis over which attention is applied. The fact that attention is sequentially applied one dimension at a time ensures that the architecture scales favorably as the input dimensions increase.

The outputs from each block are collected and fed into the upsampling stack as skip connections, similar to the approach used in classical U-Net architectures.

Upsampling stack. The upsampling stack can be considered the mirror opposite of the downsampling stack - it contains the same number of levels and residual blocks. At each level, it first adds the corresponding skip connection in the upsampling stack:

$$h_i^{\text{up}} = h_i^{\text{up}} + h_i^{\text{down}}, \quad (86)$$

followed by the same residual and attention blocks defined in (83) and (85). At the end of the level, we apply an upsampling block defined by

$$h_i^{\text{up}} = \text{Conv2D}(c_i, 3, 1) \circ \text{channel2space} \circ \text{Conv2D}(c_i q_i^2, 3, 1) \circ h_i^{\text{up}}, \quad (87)$$

where the `channel2space` operator expands the $c_i q_i^2$ channels into $q_i \times q_i \times c_i$ blocks locally, effectively increasing the spatial resolution by q_i .

Decoding. We apply a final block to the output of the upsampling stack:

$$x_{\text{out}} = \text{Conv2D}(4, 3, 1) \circ \text{SiLU} \circ \text{LayerNorm} \circ h_0^{\text{up}}. \quad (88)$$

Preconditioning. As suggested in [41], we *precondition* D_θ by writing it in an alternative form

$$D_\theta(z_\tau, \sigma_\tau, \tilde{y}') = c_{\text{skip}}(\sigma_\tau) z_\tau + c_{\text{out}}(\sigma_\tau) F(c_{\text{in}}(\sigma_\tau) z_\tau, c_{\text{noise}}(\sigma_\tau), \tilde{y}'), \quad (89)$$

where F is the U-ViT architecture described above and

$$c_{\text{skip}} = \frac{1}{1 + \sigma_\tau^2}; \quad c_{\text{out}} = \frac{\sigma_\tau}{\sqrt{1 + \sigma_\tau^2}}; \quad c_{\text{in}} = \frac{1}{\sqrt{1 + \sigma_\tau^2}}; \quad c_{\text{noise}} = 0.25 \log \sigma_\tau, \quad (90)$$

such that the input and output of F is approximately normalized.

SI 9.5.5 Hyperparameters

Table SI 8 shows the set of hyperparameters used for the denoiser architecture, as well as those applied during the training and sampling phases of the diffusion model. We also include the optimization algorithm, learning rate scheduler and weighting for minimizing (72).

SI 9.5.6 Training, evaluation, and test data

The super-resolution stage is trained *independently of the debiasing stage*, using perfectly time-aligned ERA5 data samples at the input (1.5-degree, daily) and output (0.25-degree, bi-hourly) resolutions.

Training is conducted on continuous 7-day windows randomly selected in the training range, with each day beginning at 00:00 UTC. Spatially, the model super-resolves a rectangular patch of fixed size. Consistent with the debiasing step, data from 1960–1999 is used for training, 2000–2009 for evaluation, and 2010–2019 for testing.

SI 9.5.7 Computational cost

The diffusion model is trained on TPUv5p hosts, utilizing a total of 32 cores, which takes approximately 3 days. For sampling, 16 TPUv5e cores are employed in parallel. Each core denoises a single 7-day window, collectively generating a 97-day temporally consistent long sample⁴. Excluding JAX compile

⁴Our parallel strategy yields 97 days, calculated as: number of cores $\{16\} \times (\text{model days } \{7\} - \text{overlap } \{1\}) + \text{overlap } \{1\}$. This means increasing the number of cores effectively extends the total sample length. Alternatively, sequential sampling of 7-day windows can be performed, where sample length is independent of the number of cores and scales with inference time.

Table SI 8: Hyperparameters for the super-resolution model.

Denoiser architecture	
Output shape	$84 \times 240 \times 120 \times 4$ (CONUS); $84 \times 360 \times 180 \times 6$ (Atlantic)
Time span	7 days
Spatial downsampling ratios	[3, 2, 2, 2] (CONUS); [3, 3, 2, 2] (Atlantic)
Residual blocks	[4, 4, 4, 4]
Hidden channels	[128, 256, 384, 512]
Use attention layers	[False, False, True, True]
Trainable parameters	around 150 million (both CONUS and Atlantic)
Training	
Device	TPUv5p, $2 \times 4 \times 4$
Duration	300,000 steps
Batch size	128 (with data parallelism)
Learning rate	cosine annealed (peak= 1×10^{-4} , end= 1×10^{-7}), 1,000 linear warm-up steps
Gradient clipping	max norm = 0.6
Noise sampling	$\sigma_\tau \sim \text{LogUniform}(\text{min}=1 \times 10^{-4}, \text{max}=80)$
Noise weighting (λ_τ)	$1 + 1/\sigma_\tau^2$
Condition dropout (p_u)	0.15
Inference	
Device	TPUv5e, 4×4
Noise schedule	$\sigma_\tau = \frac{\tan(3\tau-1.5) - \tan(-1.5)}{\tan(1.5) - \tan(-1.5)} \cdot 80$, $\tau \sim [0, 1]$
SDE solver type	1st order exponential
Solver steps	$(\sigma_{\max}^{1/7} + \frac{i}{255}(\sigma_{\min}^{1/7} - \sigma_{\max}^{1/7}))^7$
CFG strength (g)	1.0
Overlap	1 day (12 time slices)
# of days coherently denoised	97 days

time, a one-time overhead that makes subsequent realizations significantly more efficient, each sample requires about 3 minutes to complete. The temporal length of the generated samples scales linearly with the number of TPU cores used, while clock time remains relatively constant. At a cost of estimated \$1.2 USD per hour of the current market rate, the super-resolution step incurs a cost of \$0.08 USD per sample day in a $[60^\circ, 30^\circ]$ region. For 100 ensemble members over 10 years (3 months per year, 8 samples per ensemble member), the estimated total inference cost is approximately \$61,440. This cost can be further reduced with accelerated sampling algorithms and other engineering optimization.

SI 9.6 GenFocal Variants

The two-stage design of GenFocal enables a “plug and play” approach for integrating different bias correction and super-resolution components. We describe two such components below, which we use as ablation studies to examine the effectiveness of our bias connection component, introduced in [SI 9.4](#).

SI 9.6.1 Direct Super-Resolution (SR)

We can examine how well a super-resolution operation, optimized on the reanalysis ERA5 can overcome the bias in the low-resolution climate data. We term this method of downscaling as SR, with the generative super-resolution described in [SI 9.5](#) being directly applied on LENS2.

SI 9.6.2 Quantile Mapping Super-Resolution (QMSR)

We have also experimented with the quantile mapping component of BCSD (described in [SI 5.1](#)), with a bit adaptation, as a debiasing procedure, followed by GenFocal’s super-resolution operation. We term this approach as QMSR. The adaptation we need is to add back the mean of the downsampled data:

$$y_{\text{qm}} = \frac{y - \text{clim_mean}[y]}{\text{clim_std}[y]} \cdot \text{clim_std}[C'x] + \text{clim_mean}[C'x]. \quad (91)$$

The resulting output y_{qm} retains the low spatial resolution and can serve as the input for our diffusion-based upsampling model. This is the “QM” baseline referred to in [Table SI 3](#).

For both variants, during the generative super-resolution steps, the inputs and outputs are respectively normalized and denormalized in the same way as described by [\(75\)](#) and [\(74\)](#), where the normalization statistics are derived from ground truth low-resolution ERA5. (For SR, experiments with input normalization using statistics of the LENS2 dataset led to worse evaluation results across almost all metrics.)

SI 10 Ablation studies: model selection and design choices

We study the sensitivity of GenFocal’ to a few design choices and implementation details. Earlier studies provide further insight into variations of diffusion-based super-resolution, see §5.1.3 in [102], and [49]. The main ablation studies and findings in this section cover the following:

- Reference periods used for training the debiasing stage (SI 10.1). Using training data from recent reference periods, which is better constrained by satellite observations, results in better models. More data improves the representation of extremes.
- Length of the debiasing sequence (SI 10.2). Longer debiasing sequences lead to improvements in most statistics.
- Number of debiased variables being modeled (SI 10.3). Debiasing 10 variables improves GenFocal’s ability to capture TC statistics compared to variants with 4 or 6 debiased variables. Since computational costs scale with the number of modeled variables, we leave the selection of an optimal variable set to future work.
- Number of training steps (SI 10.4). Additional training steps beyond 300k lead to an overestimation of the number of tropical cyclones, possibly due to overfitting.
- Number of LENS2 ensemble members used for training (SI 10.5). While LENS2 contains 100 members, we train on a small subset and evaluate on the full ensemble. Including multiple members in the training set enhances performance, though benefits saturate beyond four members.
- The data coupling strategy used to train the debiasing stage (SI 10.6). Coupling samples with similar climatologies (e.g., samples from the same day of the year, or from adjacent years) yields more statistically accurate results.
- The training period for the super-resolution stage (SI 10.7). GenFocal is largely insensitive to this change, provided enough data.
- The length of the temporal sequence and the stitching strategy in the super-resolution stage (SI 10.8). Temporal coherence through domain decomposition improves the statistics of spatio-temporal phenomena, more so for long-lived events.
- Importance of residual modeling on super-resolution (SI 10.9). Modeling the fine-grained deviations from the debiasing stage leads to improved results for most variables and metrics, compared to predicting the full high-resolution fields.

Throughout the ablation studies, the evaluation period (2010 - 2019) remains unchanged to ensure consistent comparison.

SI 10.1 Training period for the debiasing stage

We evaluate training period sensitivity using 2010–2019 CONUS summer and downscaled North Atlantic TC statistics. We consider models trained on individual decades from the 1960s to the 1990s, as well as models trained over multiple decades spanning the period 1960–2000.

Overall, the results underscore the critical importance of training data quality and diversity. Table SI 9 indicates that training exclusively on reanalysis data prior to 1979—which is significantly less constrained

Table SI 9: Effect of training data period on the mean absolute bias, Wasserstein distance, and absolute error of the 99th percentile for the summers (June-July-August) of 2010-2019 in CONUS. The precise definitions of the metrics are included in Section SI 7.

Variable	60s	70s	80s	90s	60s-90s	70s-90s	80s-90s
Mean Absolute Bias, ↓							
Temperature (K)	0.54	0.48	0.53	0.39	0.53	0.48	0.41
Wind speed (m/s)	0.23	0.24	0.19	0.16	0.17	0.19	0.19
Specific humidity (g/kg)	0.40	0.35	0.47	0.37	0.50	0.43	0.31
Sea-level pressure (Pa)	30.07	57.46	51.49	28.33	43.94	54.71	39.92
Relative humidity (%)	2.24	1.88	2.84	2.03	2.08	1.93	1.71
Heat index (K)	0.59	0.53	0.55	0.46	0.65	0.59	0.47
Mean Wasserstein Distance, ↓							
Temperature (K)	0.61	0.54	0.59	0.47	0.59	0.55	0.47
Wind speed (m/s)	0.28	0.29	0.22	0.21	0.20	0.22	0.22
Specific humidity (g/kg)	0.48	0.43	0.51	0.42	0.53	0.47	0.36
Sea-level pressure (Pa)	53.32	71.81	63.51	44.79	50.85	64.48	52.09
Relative humidity (%)	2.62	2.32	3.1	2.27	2.41	2.29	2.1
Heat index (K)	0.67	0.6	0.61	0.53	0.70	0.65	0.53
Mean Absolute Error, 99 th ↓							
Temperature (K)	1.02	0.83	0.87	0.60	0.64	0.68	0.61
Wind speed (m/s)	0.85	0.74	0.61	0.56	0.39	0.48	0.48
Specific humidity (g/kg)	0.83	0.68	0.58	0.41	0.4	0.42	0.45
Sea-level pressure (Pa)	128.36	80.61	107.09	92.12	60.31	69.89	77.99
Relative humidity (%)	2.39	2.27	2.10	1.99	1.90	1.93	1.87
Heat index (K)	1.2	0.96	0.92	0.67	0.77	0.81	0.68

by satellite observations—substantially degrades model performance. This decline is particularly pronounced in wind speed and humidity extremes. These findings align with previous studies suggesting that training AI weather models on pre-1979 reanalysis data does not lead to improved skill [43, 72].

Table SI 9 also demonstrates that leveraging more training data improves the representation of extremes. However, extending the training period to include pre-1979 data leads to higher bias and Wasserstein distances, indicating a trade-off between data diversity and quality. Consequently, the 1980–1999 period yields the best aggregate performance. These findings are further supported by Figs. SI 31–SI 33, which illustrate the geographical distribution of the bias, Wasserstein distance, and the 99th percentile, respectively.

Additional results for relative humidity and heat index are shown in Figs. SI 34 and SI 35. Notably, Fig. SI 35 reveals that models trained solely on pre-1979 data exhibit strong biases in heat index extremes at high latitudes; these biases are significantly mitigated in models trained on post-1979 data.

Data quality and diversity are also critical for the realistic representation of TCs. As shown by the TempestExtremes detection counts in Fig. SI 36, a single decade of training data is typically inadequate for learning accurate TC representations in the North Atlantic (Fig. SI 36a). Performance improves when training on high-activity decades like the 1990s [27], consistent with recent evidence that AI weather models only faithfully represent TCs when they are sufficiently prevalent in the training set [90]. This requirement is further illustrated by the example tracks in Fig. SI 38, which demonstrate the importance of data volume for capturing TC statistics.

In contrast, training on multi-decadal datasets yields realistic TC counts and tracks for all considered training periods (Fig. SI 36b, Fig. SI 38). However, models trained on early reanalysis data—unconstrained by satellite-based remote sensing—exhibit a slight underestimation of TCs across all categories (Fig. SI 37).

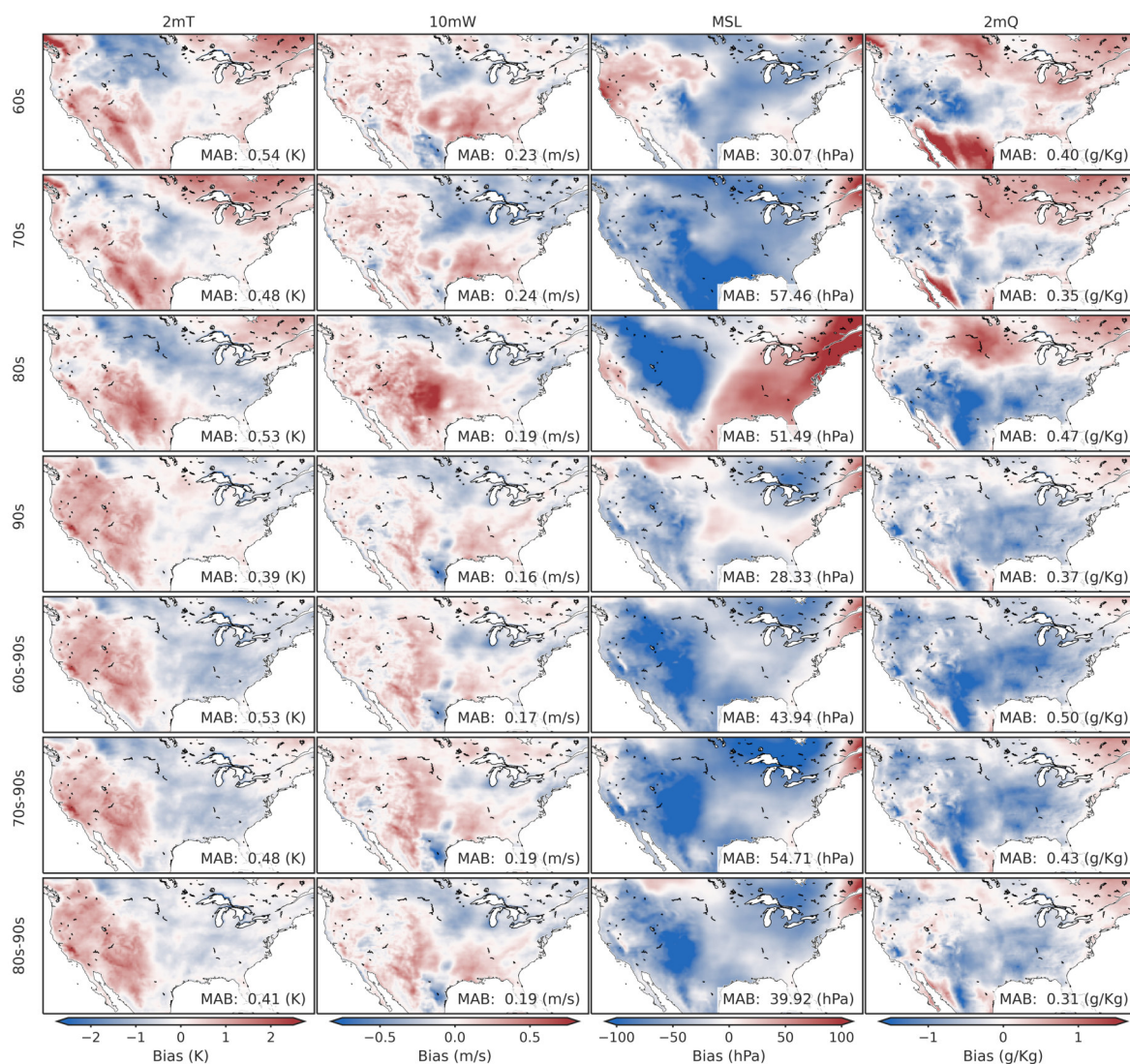


Figure SI 31: Summertime bias of GenFocal over CONUS. Bias of multiple fields during summers (June-August) of 2010-2019, for GenFocal models trained using data from different reference periods.

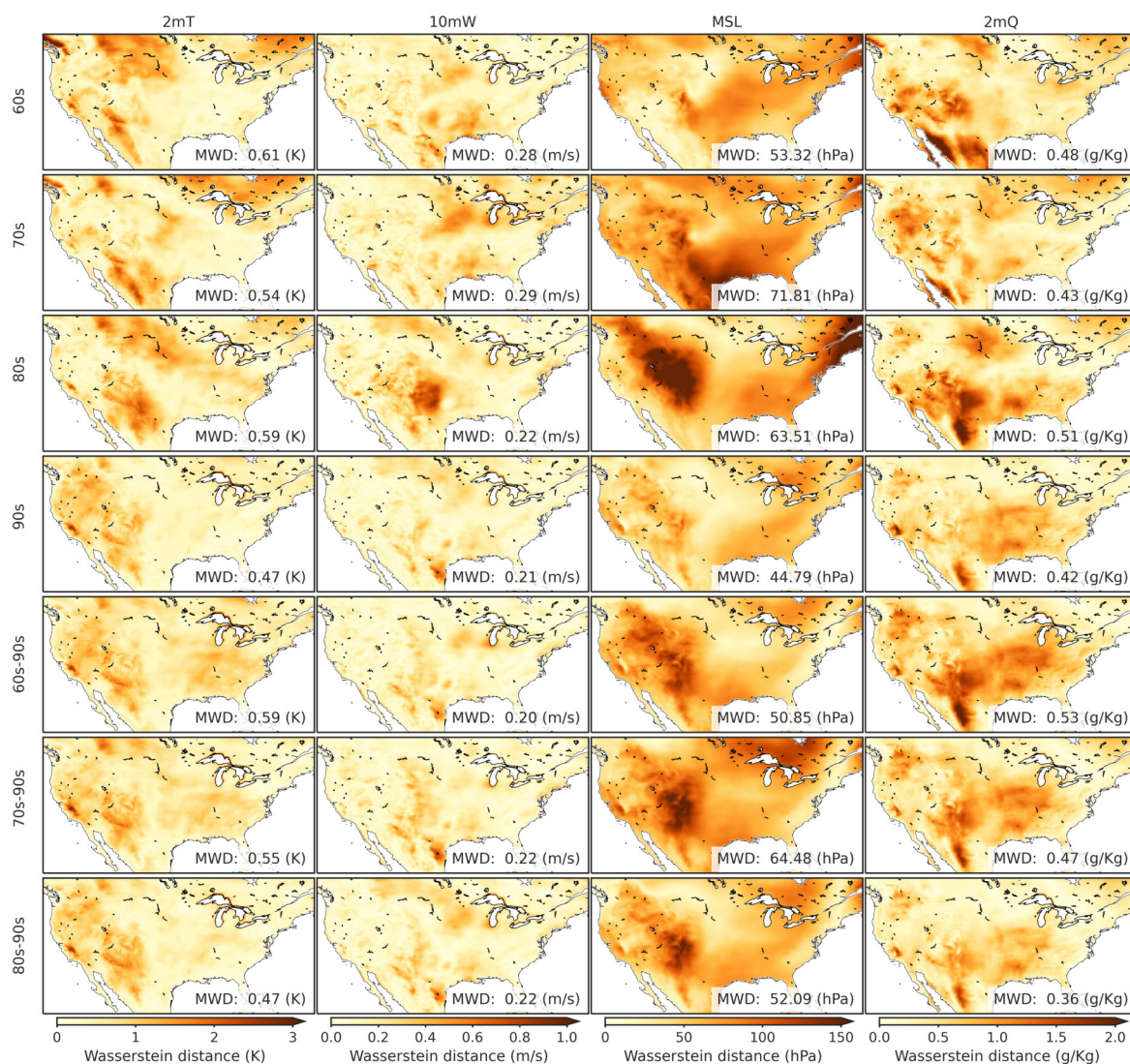


Figure SI 32: Summertime Wasserstein distance over CONUS. Pointwise Wasserstein distance (see SI 7.1.2) during summers (June-August) of 2010-2019, for GenFocal models trained using data from different reference periods.

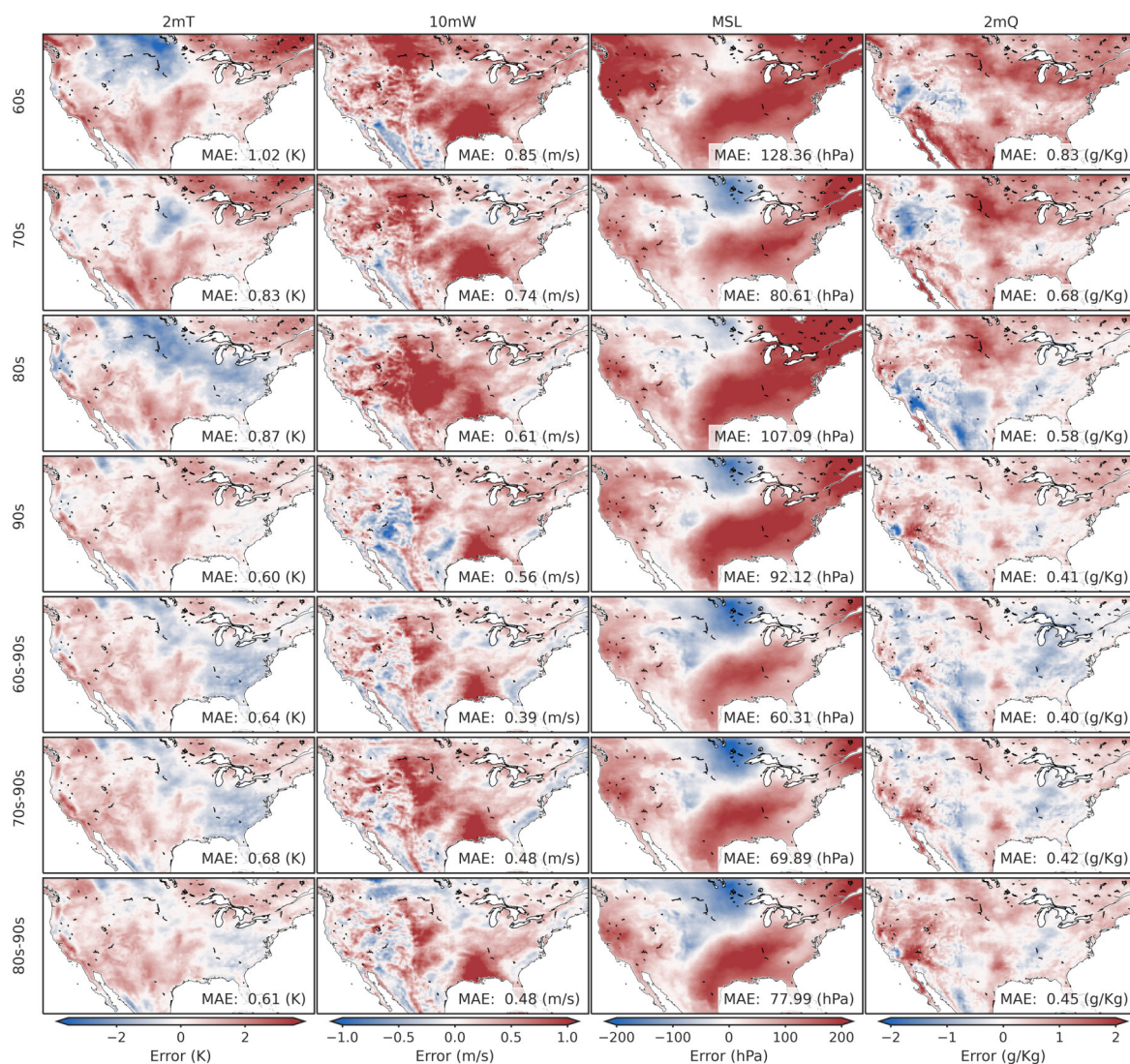


Figure SI 33: Summertime 99th percentile error over CONUS. Pointwise error in the 99th percentile of multiple fields during summers (June-August) of 2010-2019, for GenFocal models trained using data from different reference periods.

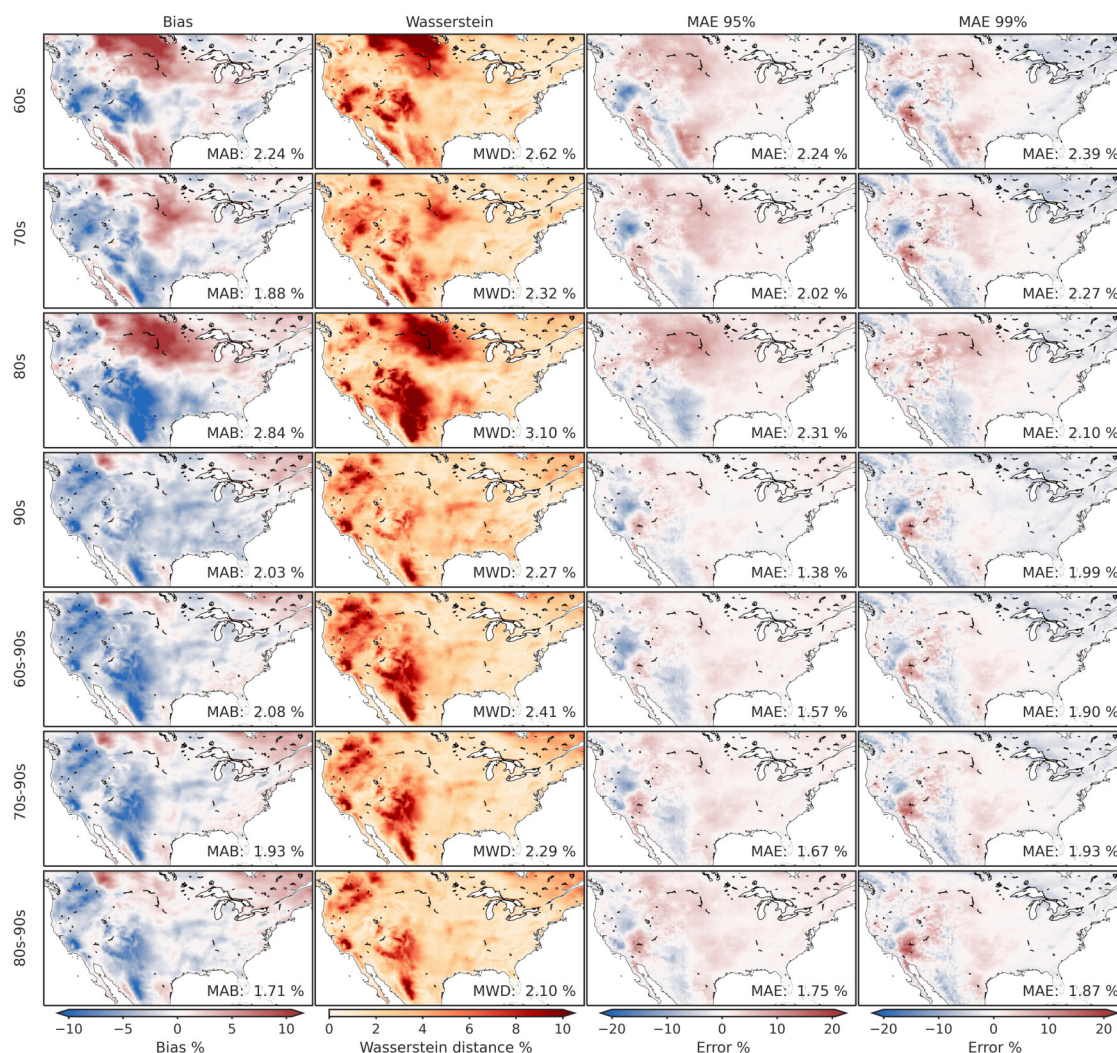


Figure SI 34: Metrics for relative humidity over CONUS. Metrics for the relative humidity, a derived variable, over CONUS during the summer (June-August) for the evaluation period 2010-2019. We include bias, Wasserstein error, error of the 95th and 99th percentiles for GenFocal trained using data from different reference periods.

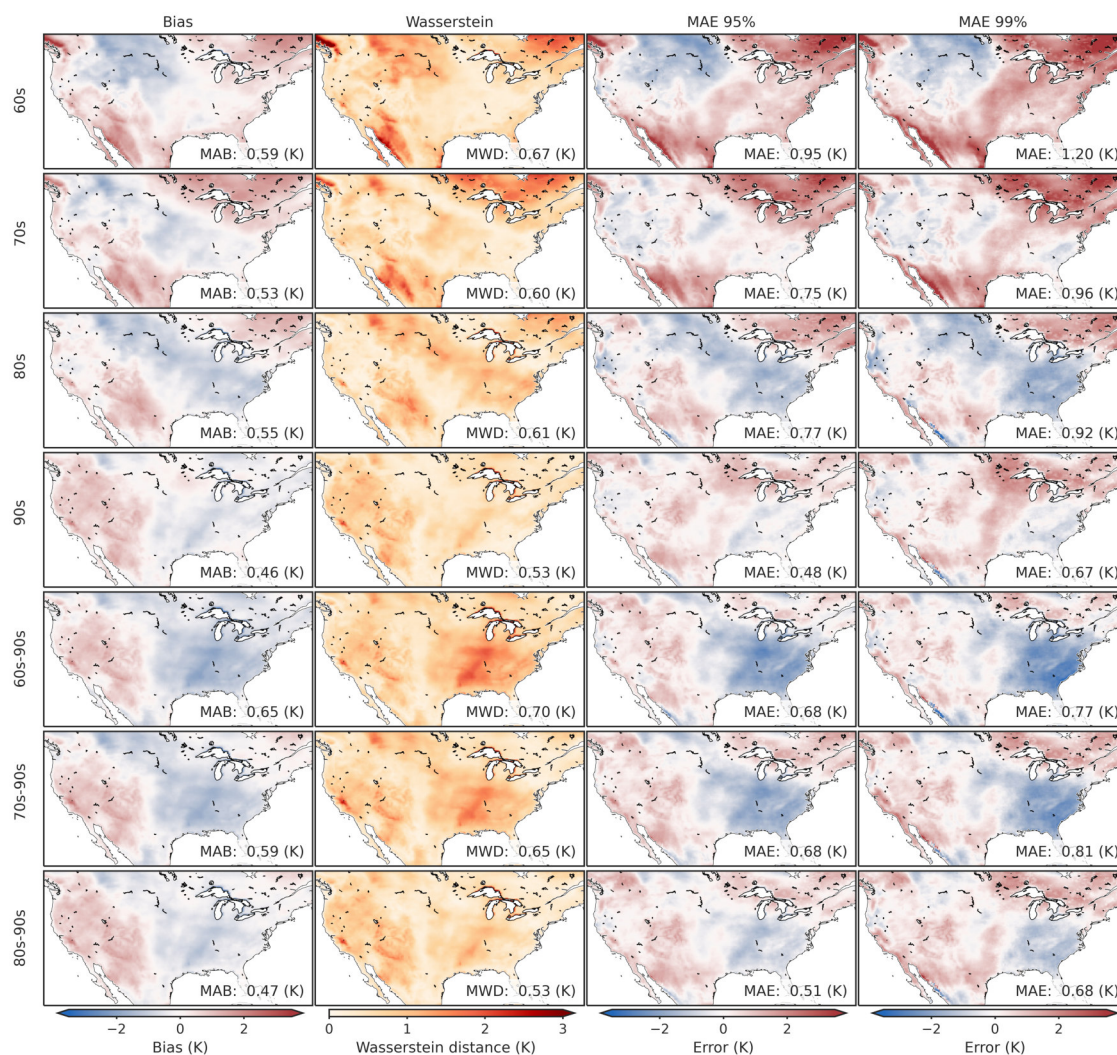


Figure SI 35: Metrics for the heat index over CONUS. Metrics for the heat index, a derived variable, over CONUS during the summer (June-August) for the evaluation period 2010-2019. We include bias, Wasserstein error, error of the 95th and 99th percentiles for GenFocal trained using data from different training periods.

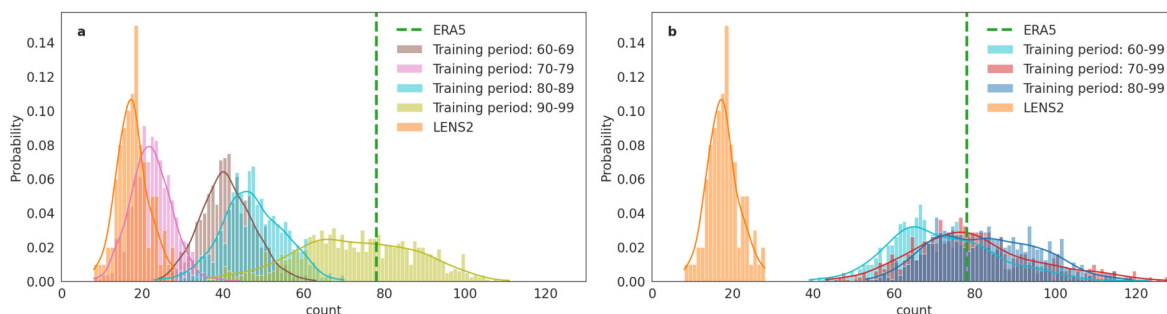


Figure SI 36: Distribution of TC occurrences. Distribution of the number of TCs detected by TempestExtremes in the North Atlantic during the peak hurricane season (August–October), 2010–2019. **a.** Distribution of TC counts for different decadal training periods. **b.** Distribution of the TC counts for training data periods of varying length. TCs from all GenFocal variants were calibrated separately following SI Section [SI 8.3.3](#).

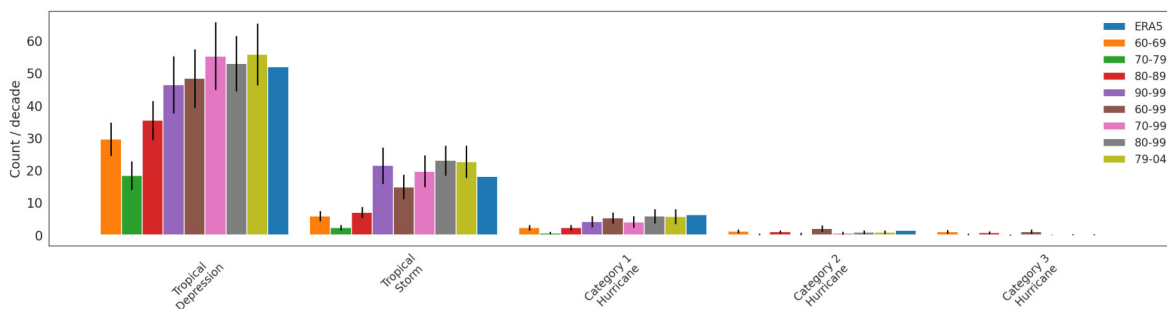


Figure SI 37: Saffir-Simpson scale distribution of TCs. Distribution of the intensity of TCs detected by TempestExtremes in the North Atlantic during the peak hurricane season (August–October), 2010–2019. Results shown for GenFocal models trained on different periods, and for ERA5. Error bars denote the ensemble standard deviation. TCs from all GenFocal variants were calibrated separately following SI Section [SI 8.3.3](#).

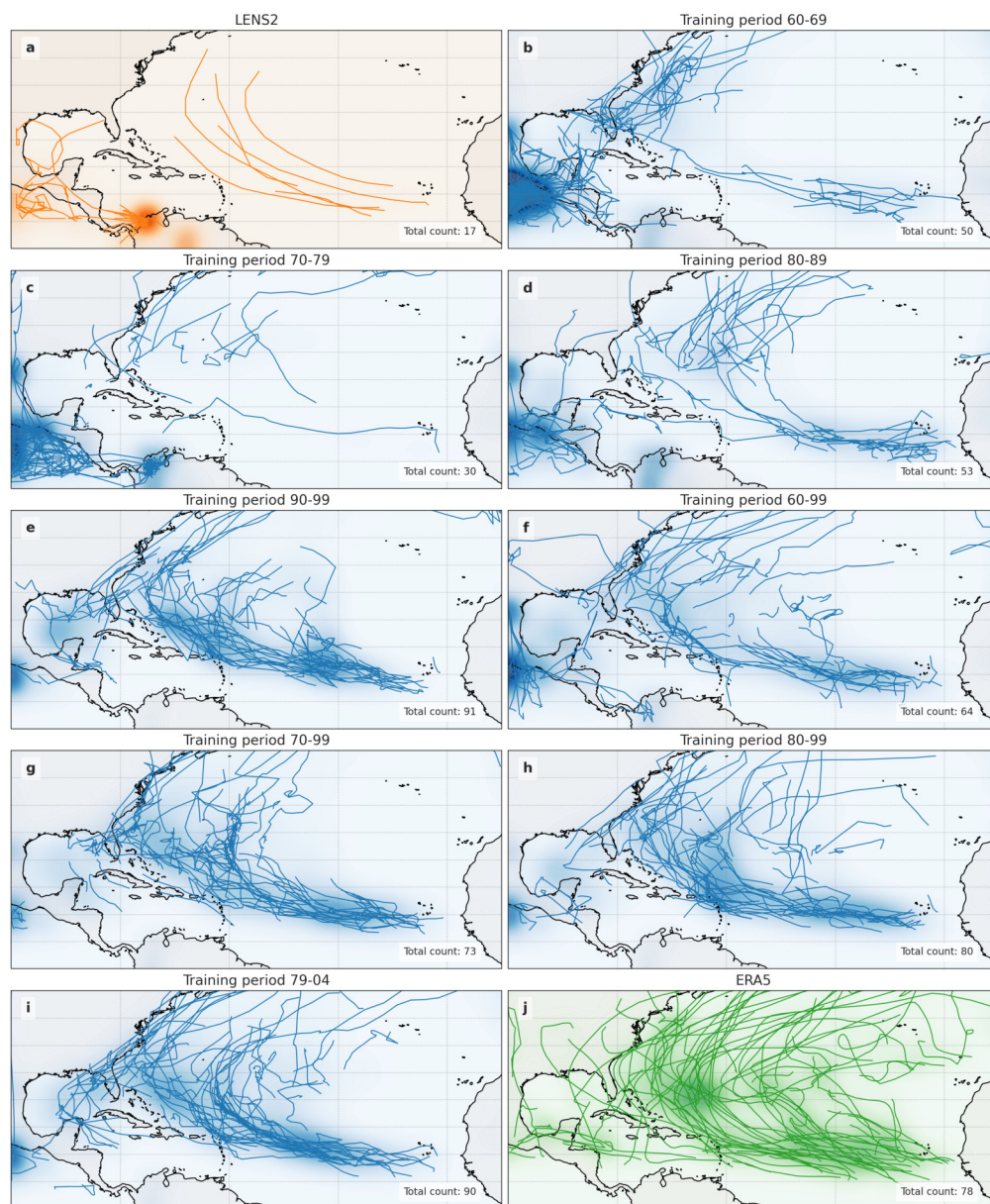


Figure SI 38: TC tracks and their density. TC tracks for a single climate projection overlaying the ensemble TC track density for the peak North Atlantic hurricane season (August–October), 2010–2019. **a.** LENS2. **b-i.** GenFocal variants trained on different periods.. **f.** ERA5 tracks and their density. TCs from all GenFocal variants were calibrated following SI Section [SI 8.3.3](#).

SI 10.2 Length of the debiasing sequence

This section shows that harnessing spatiotemporal correlations in the input data leads to a reduction in distribution matching errors, by evaluating the sensitivity of GenFocal to the number of consecutive days debiased simultaneously. We retain the same architecture and number of trainable parameters as the model reported in the main text.

Table SI 10 summarizes the statistical errors of GenFocal models with different debiasing sequence lengths. Longer debiasing sequences lead to improvements in most statistics. Fig. SI 39 shows the spatial distribution of biases for the directly modeled variables. Fig. SI 40 shows the geographical distribution of the metrics for the heat index. In both, we observe that the geographical distribution of the errors is similar across debiasing sequence lengths, with an overall error reduction for longer sequences.

Fig. SI 41 shows the bias in the projected number of extreme caution advisory periods per year, for periods of varying length. We observe that increasing the length of the debiasing sequence uniformly decreases the bias in the number of predicted heat streaks of 1 to 7 days.

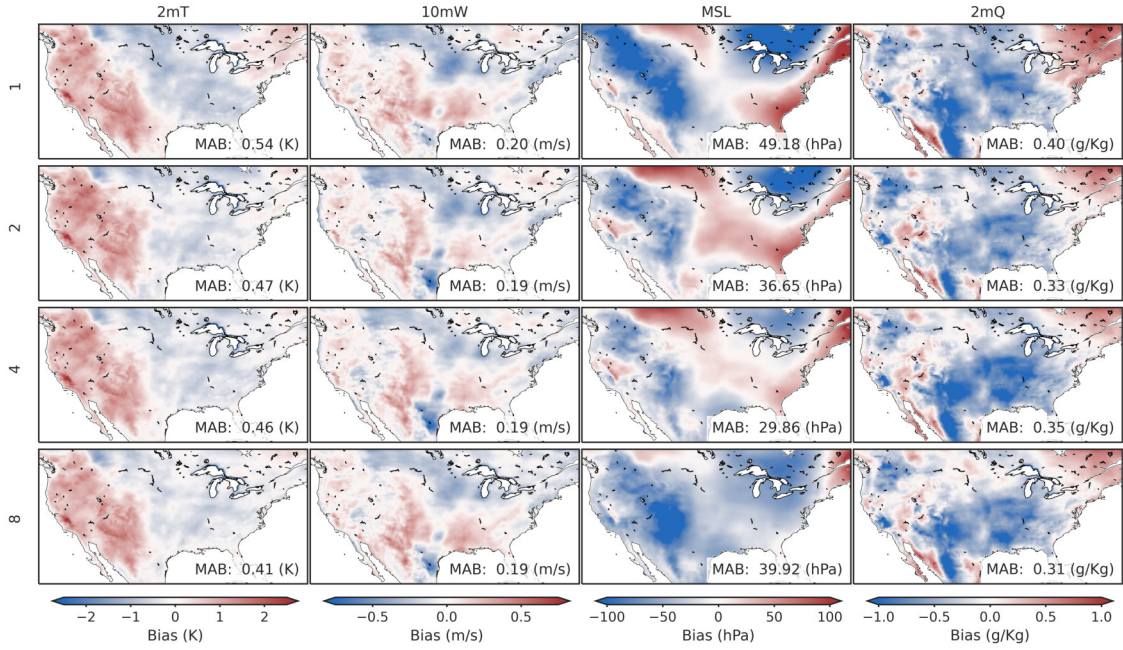


Figure SI 39: Downscaled variable biases. Biases of downscaled variables, over CONUS during the summer (June-August) for the evaluation period 2010-2019 for GenFocal trained with different debiasing sequence lengths.

Figs. SI 42 and SI 43 further show that using longer debiasing sequences also leads to tropical cyclones with more realistic trajectories in the North Atlantic basin. Furthermore, statistics of projected TCs match the observational record better.

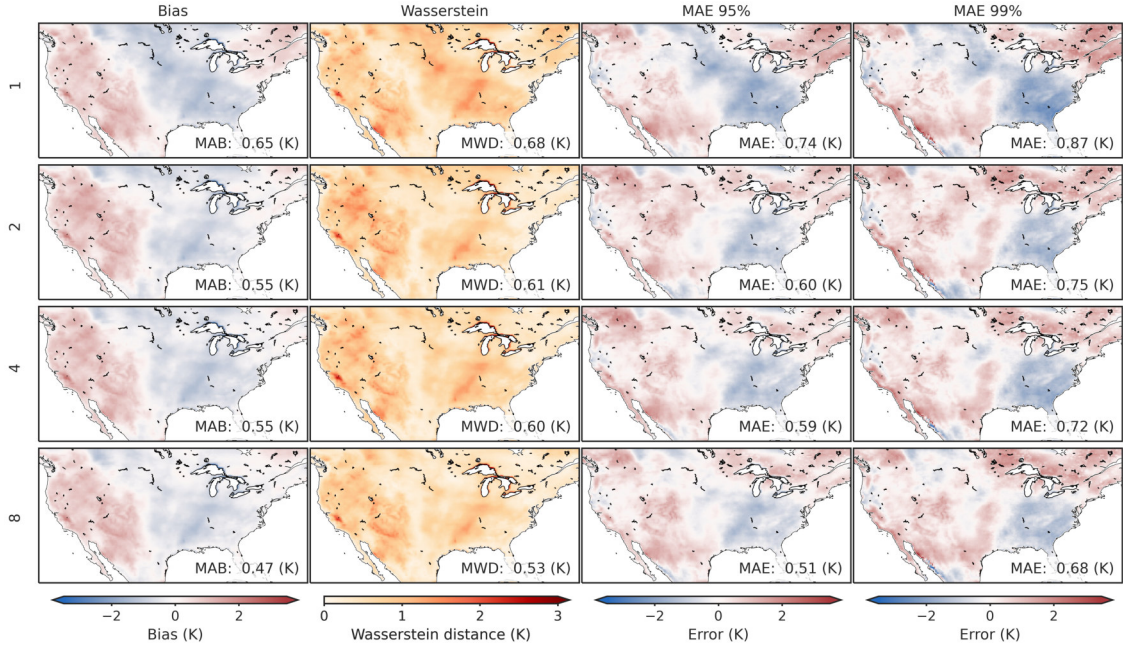


Figure SI 40: Spatial distribution of heat index errors. Spatial distribution of statistical modeling errors for the heat index, one of the derived variables, over CONUS during the summer (June-August) for the evaluation period 2010-2019. We include bias, Wasserstein error, error of the 95th and 99th percentiles for GenFocal trained with different debiasing sequence lengths.

Table SI 10: Effect of debiasing sequence length on mean absolute bias, mean Wasserstein distance, and mean absolute error in the 99th percentile for different variables during summers of 2010-2019 (June-July-August) over CONUS. The metrics are defined in SI 7.

Variable	Mean Absolute Bias ↓				Mean Wasserstein Distance ↓				Mean Absolute Error, 99 th ↓			
	1	2	4	8	1	2	4	8	1	2	4	8
Temperature (K)	0.54	0.47	0.46	0.41	0.57	0.52	0.52	0.47	0.69	0.66	0.64	0.61
Wind speed (m/s)	0.2	0.19	0.19	0.19	0.22	0.22	0.22	0.22	0.41	0.44	0.46	0.48
Specific humidity (g/kg)	0.4	0.33	0.35	0.31	0.43	0.38	0.4	0.36	0.44	0.47	0.46	0.45
Sea-level pressure (Pa)	49.18	36.65	29.86	39.92	58.08	47.72	43.36	52.09	75.57	86.13	80.84	77.99
Relative humidity (%)	1.85	1.69	1.74	1.71	2.21	2.07	2.11	2.1	1.83	1.9	1.91	1.87
Heat index (K)	0.65	0.55	0.55	0.47	0.68	0.61	0.6	0.53	0.87	0.75	0.72	0.68

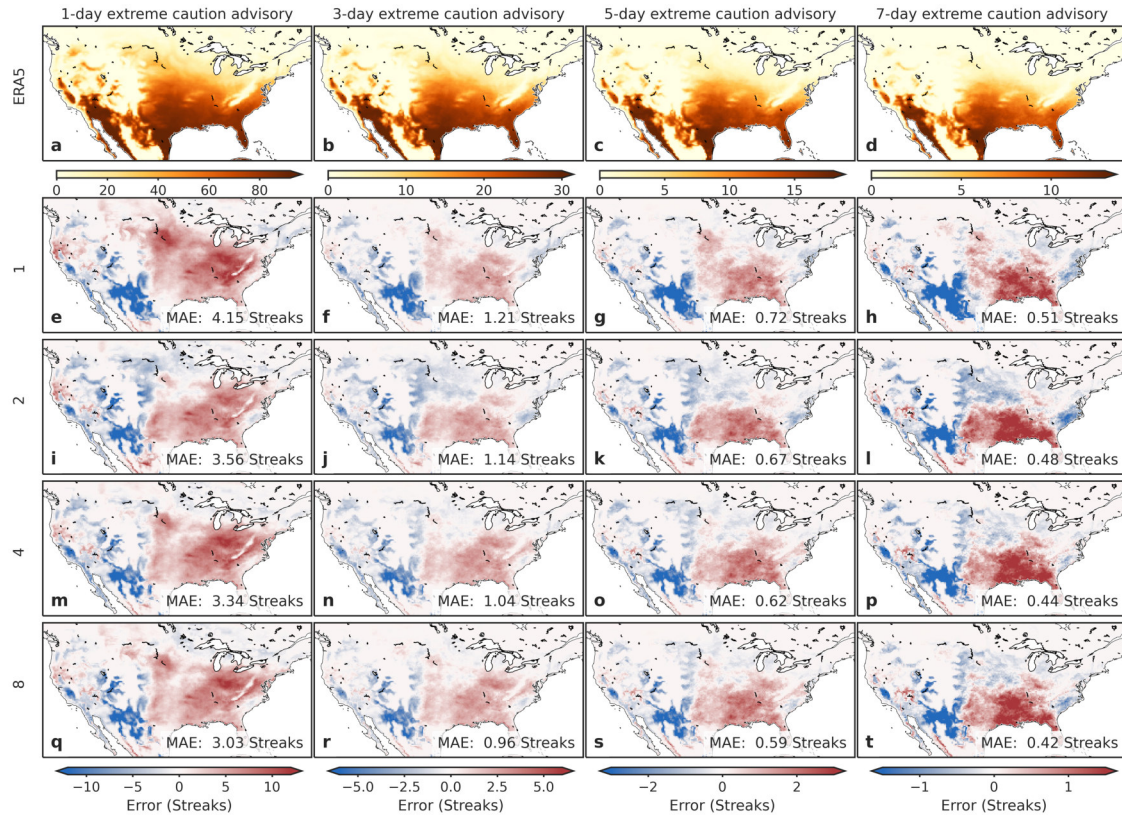


Figure SI 41: Spatial distribution of heat streak errors. Spatial distribution of statistical modeling errors in the number of heat-streaks per year for extreme caution advisory considering different debiasing sequence lengths (in days). We show the ground truth (ERA5)(a-d), and the pointwise errors of GenFocal with different lengths of the debiased sequence.

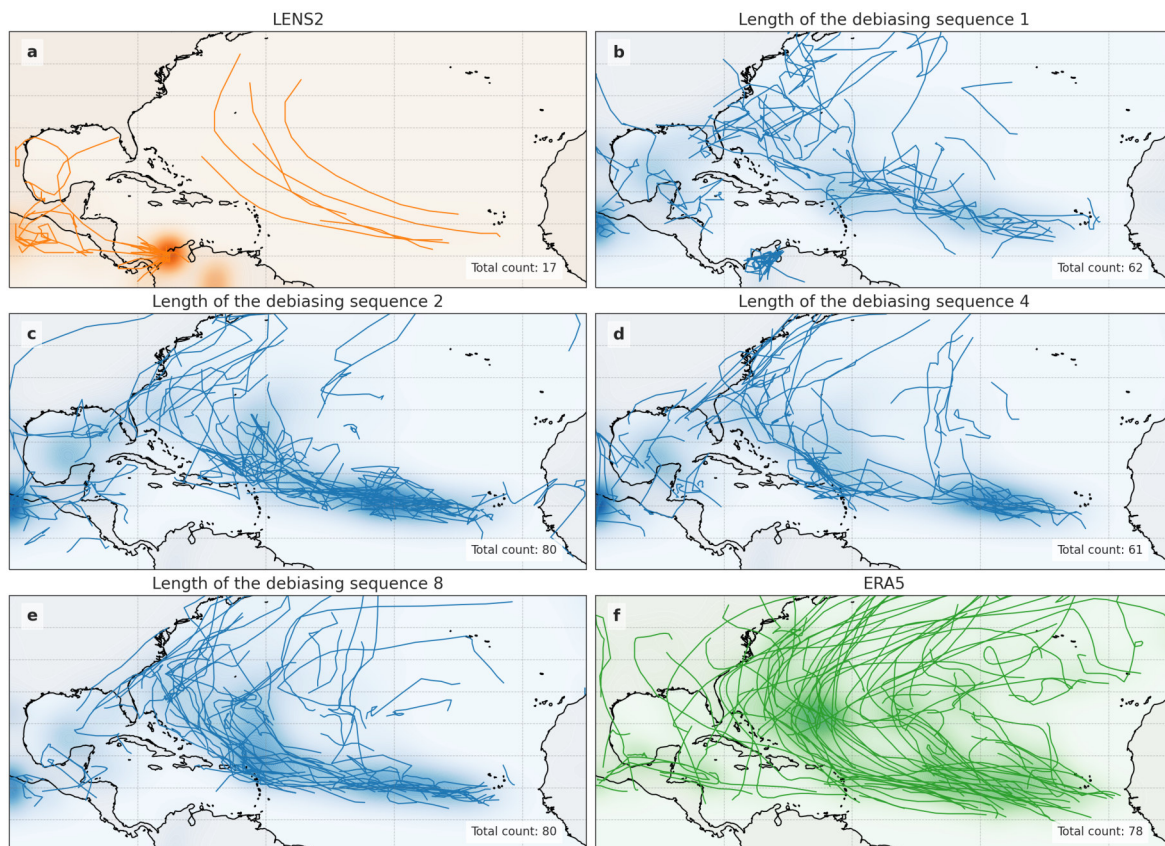


Figure SI 42: TC tracks and density. Tracks and their density for a LENS2 member in the North Atlantic in the time period 2010-2019 (a), and for a downscaled sample from the same member generated using GenFocal for different debiasing sequence length (b-e). Also shown are the tracks detected in the reference ERA5 data (f).

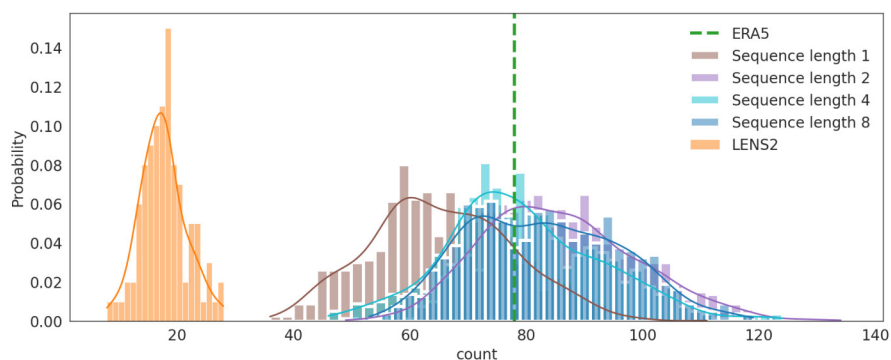


Figure SI 43: Distribution of TC frequency. Distribution of the number of TCs detected by TempestExtremes in the North Atlantic for the hurricane season August-September-October during 2010-2019, using downscaled data from GenFocal models with varying debiasing sequence lengths.

SI 10.3 Number of debiased variables

Table SI 11: Effect of the number of debiased variables on mean absolute bias, mean Wasserstein distance, and mean absolute error in the 99th percentile of the downscaled output for different variables. The metrics are defined in [SI 7](#).

Variable	Mean Absolute Bias ↓			Mean Wasserstein Distance ↓			Mean Absolute Error, 99 th ↓		
	4	6	10	4	6	10	4	6	10
Temperature (K)	0.43	0.47	0.41	0.49	0.52	0.47	0.64	0.63	0.61
Wind speed (m/s)	0.16	0.19	0.19	0.2	0.22	0.22	0.58	0.44	0.48
Specific humidity (g/kg)	0.35	0.39	0.31	0.41	0.43	0.36	0.47	0.45	0.45
Sea-level pressure (Pa)	36.82	36.67	39.92	54.14	50.22	52.09	117.07	91.22	77.99
Relative humidity (%)	1.63	1.7	1.71	2.01	2.1	2.1	1.84	1.87	1.87
Heat index (K)	0.51	0.58	0.47	0.57	0.63	0.53	0.71	0.75	0.68

We explore the sensitivity to the number of debiased variables by considering two alternative models that use 4 and 6 of the variables described in [SI 9.4.5](#), respectively. The model with 4 inputs retains the variables to be super-resolved, and the variant with 6 input variables incorporates the geopotential height at 200 and 500 hPa. In all cases, the super-resolution step only takes 4 variables as inputs.

From [Table SI 11](#) we can see that increasing the number of debiased variables leads to improvements in temperature and specific humidity, but not in wind speed or relative humidity. [Figs. SI 44](#) and [SI 45](#) show that increasing the number of debiased variables leads to more accurate TC statistics.

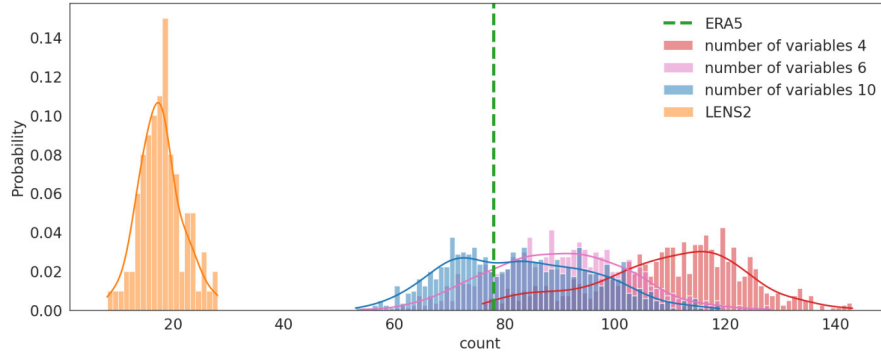


Figure SI 44: Distribution of TC frequency. Distribution of the number of TCs detected by TempestExtremes in the North Atlantic for the hurricane season August-September-October during 2010-2019 for GenFocal trained with different number of debiasing variables.

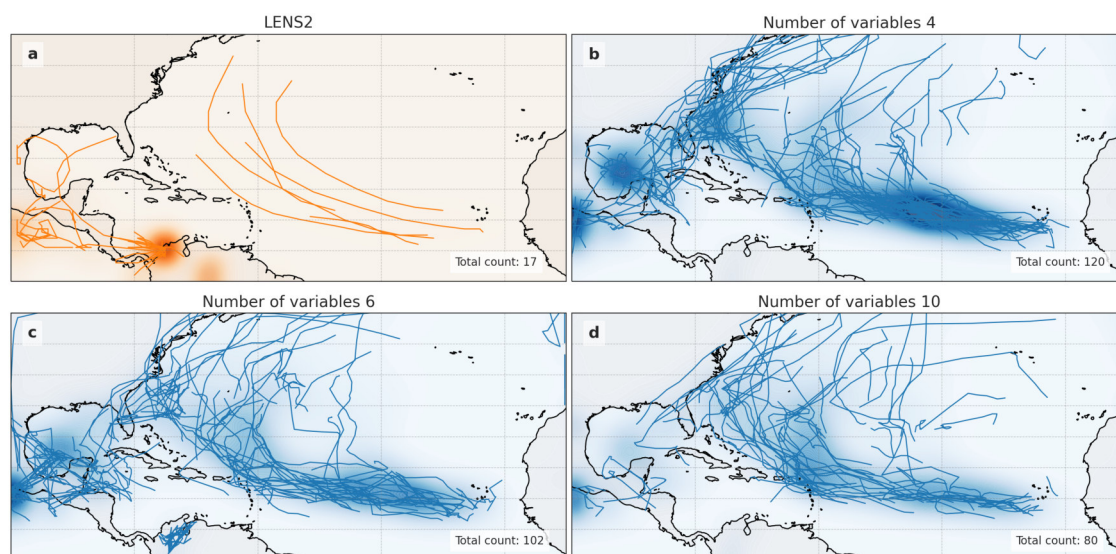


Figure SI 45: Tracks and their density. Tracks and their density for a LENS2 member in the North Atlantic in the time period 2010-2020 (a), one of downscaling samples from the same member generated using GenFocal trained with different number of debiased variables sets (b-d).

Table SI 12: Effect of the number of training steps on mean absolute bias, mean Wasserstein distance, and mean absolute error in the 99th percentile for different variables. The precise definitions of the metrics are included in SI 7.

Variable	Mean Absolute Bias ↓				Mean Wasserstein Distance ↓				Mean Absolute Error, 99 th ↓			
	300k	500k	1M	2M	300k	500k	1M	2M	300k	500k	1M	2M
Temperature (K)	0.41	0.41	0.43	0.4	0.47	0.47	0.5	0.48	0.61	0.67	0.7	0.67
Wind speed (m/s)	0.19	0.18	0.17	0.16	0.22	0.22	0.21	0.2	0.48	0.47	0.47	0.52
Specific humidity (g/kg)	0.31	0.33	0.36	0.36	0.36	0.38	0.4	0.42	0.45	0.47	0.5	0.5
Sea-level pressure (Pa)	39.92	37.07	26.6	36.74	52.09	50.83	43.63	51.67	77.99	87.53	130.13	112.24
Relative humidity (%)	1.71	1.76	1.79	1.79	2.1	2.12	2.12	2.09	1.87	1.85	1.78	1.75
Heat index (K)	0.47	0.47	0.51	0.47	0.53	0.54	0.57	0.56	0.68	0.75	0.77	0.82

SI 10.4 Number of training steps

Here, we evaluate changes in the performance of GenFocal with longer training times. Table SI 12 shows marginal improvements in the statistics of some downscaled fields over CONUS with increased training time, with the exception of the sea-level pressure, which benefits from longer training. At one million training steps we observe that some metrics start to deteriorate for some fields. Fig. SI 46 depicts the changes in the geographical distribution of biases with training time. We can observe that increasing the number of training steps does not change the distribution significantly, besides the sea-level pressure.

In contrast, longer training times deteriorate the ability of GenFocal to represent TCs, as shown in Fig. SI 47 and Fig. SI 48. GenFocal models trained for more than 300k steps tend to overestimate the frequency of tropical cyclones and reduce their track variability.

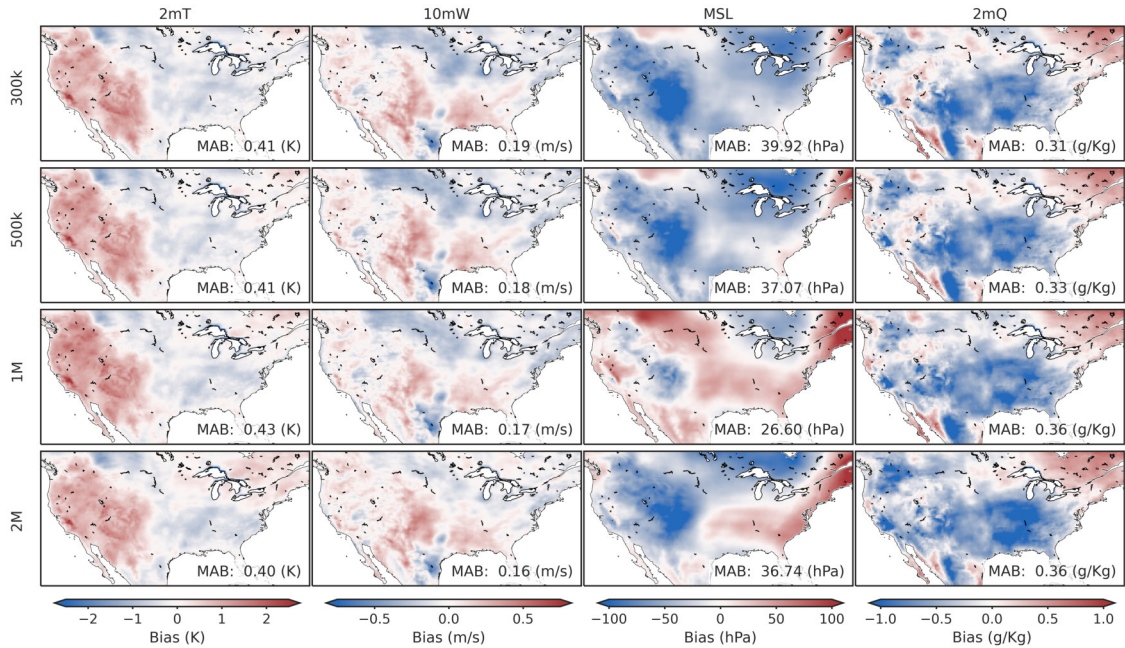


Figure SI 46: Spatial distribution of biases. Spatial distribution of statistical biases of the downscaled variables over CONUS during the summer (June-August) for the evaluation period 2010-2019 for GenFocal trained with different number of steps.

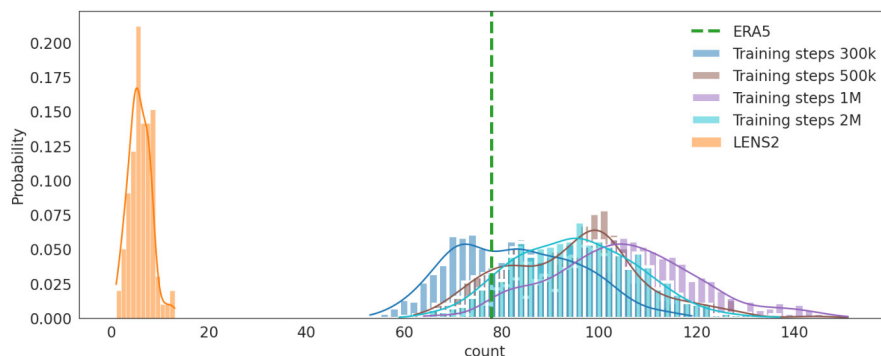


Figure SI 47: Distribution of TC counts. Distribution of the number of TCs detected by TempestExtremes in the North Atlantic for the hurricane season August-September-October during 2010-2019 for GenFocal trained with different number of steps.

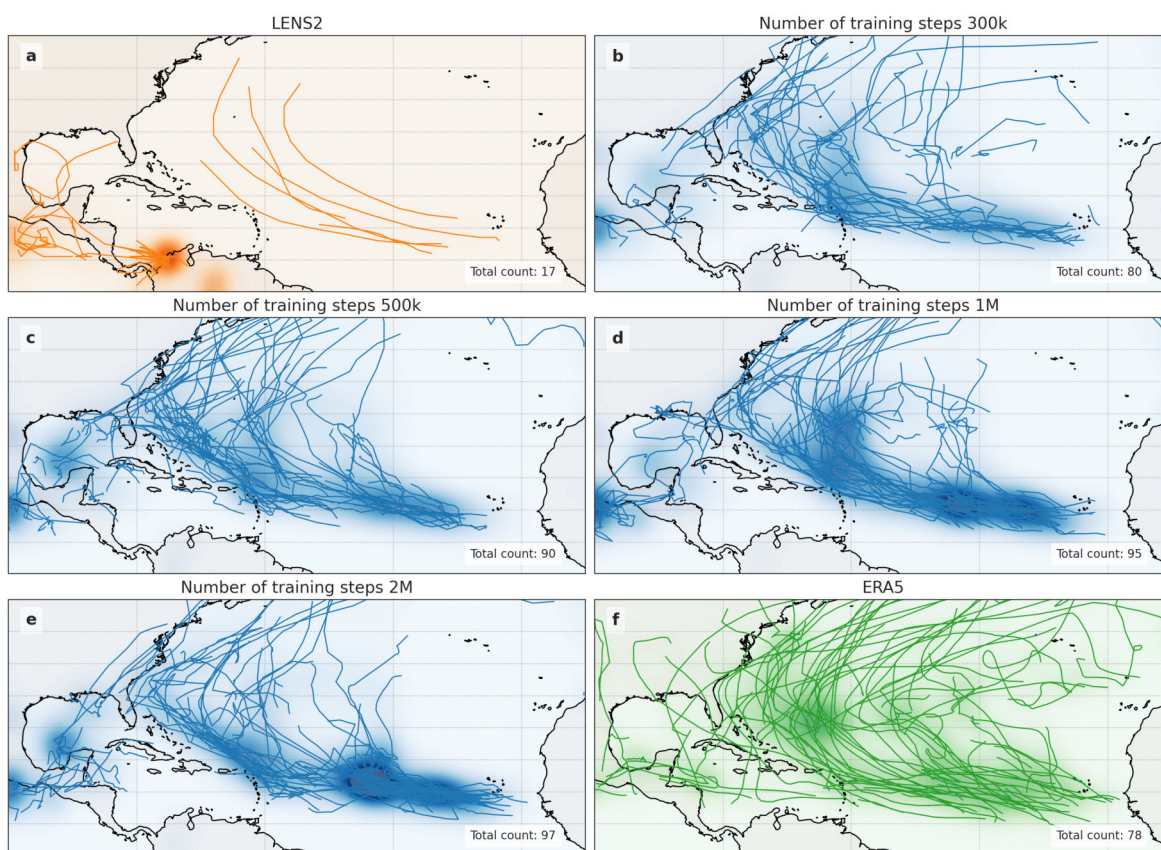


Figure SI 48: TC tracks and density. TC tracks and their density for a LENS2 member in the North Atlantic in the time period 2010-2020 (a), one of downscaling samples from the same member generated using GenFocal trained with trained with different number of steps (b-e), tracks detected using the reference ERA5 data (f).

Table SI 13: Indices of the different LENS2 members used for training.

1 member	2 members	4 members	8 members
cmip6_1001_001	cmip6_1001_001 cmip6_1251_001	cmip6_1001_001 cmip6_1251_001 cmip6_1301_010 smbb_1301_020	cmip6_1001_001 cmip6_1251_001 cmip6_1301_010 smbb_1301_020 smbb_1011_001 smbb_1301_011 cmip6_1281_001 cmip6_1301_003,

Table SI 14: Effect of the number of LENS2 members used during training on mean absolute bias, mean Wasserstein distance, and mean absolute error in the 99th percentile for different variables. The precise definitions of the metrics are included in SI 7.

Variable	Mean Absolute Bias ↓				Mean Wasserstein Distance ↓				Mean Absolute Error, 99 th ↓			
	1	2	4	8	1	2	4	8	1	2	4	8
Temperature (K)	0.5	0.47	0.41	0.39	0.55	0.52	0.47	0.45	0.84	0.61	0.61	0.75
Wind speed (m/s)	0.15	0.17	0.19	0.17	0.17	0.19	0.22	0.22	0.38	0.42	0.48	0.52
Specific humidity (g/kg)	0.37	0.36	0.31	0.27	0.4	0.4	0.36	0.35	0.57	0.4	0.45	0.62
Sea-level pressure (Pa)	41.25	33.99	39.92	60.55	46.34	42.14	52.09	72.93	62.75	63.35	77.99	84.25
Relative humidity (%)	1.78	1.55	1.71	1.76	2.06	1.88	2.1	2.16	2.75	1.82	1.87	1.96
Heat index (K)	0.64	0.6	0.47	0.4	0.69	0.65	0.53	0.47	1.24	0.85	0.68	0.83

SI 10.5 Number of ensemble members

Here we considering training the debiasing stage using 1, 2, 4 and 8 LENS ensemble members, with indices shown in Table SI 13.

The impact of this change is summarized in Table SI 14. Using more than 1 ensemble member generally improves performance. However, there is no clear improvement trend beyond using 4 members.

Fig. SI 49 shows the impact of the number of LENS ensemble members used during training on the heat index statistics. Using more members decreases the errors on average, up to 4 ensemble members. However, this behavior is not uniform, as the Wasserstein error increases in the Rockies and in the Sierra Nevada, whereas it is reduced in the East Coast.

Fig. SI 50 shows that leveraging multiple ensemble members is critical for the accurate prediction of rare extremes. Increasing the number of ensemble members from 1 to 8 decreases the error in extreme caution advisories by roughly half for periods ranging from 1 to 7 days.

SI 10.6 Training data coupling in the debiasing stage

The choice of coupling $\pi \in \Pi(\mu_y^i, \mu_y')$ in (56) is critical and primarily driven by the need to align climatologies. Specifically, samples from each dataset must be paired with consistent climatological statistics to effectively compute the unbiased anomalies transported by the flow in Eq. (3) of the Main Text.

In this section, we ablate the choice of training data coupling. In the baseline GenFocal model, the coupling enforces strict temporal alignment to ensure that both the climatology and the specific year remain consistent. This is achieved via a map based on timestamps: we sample a date within the training

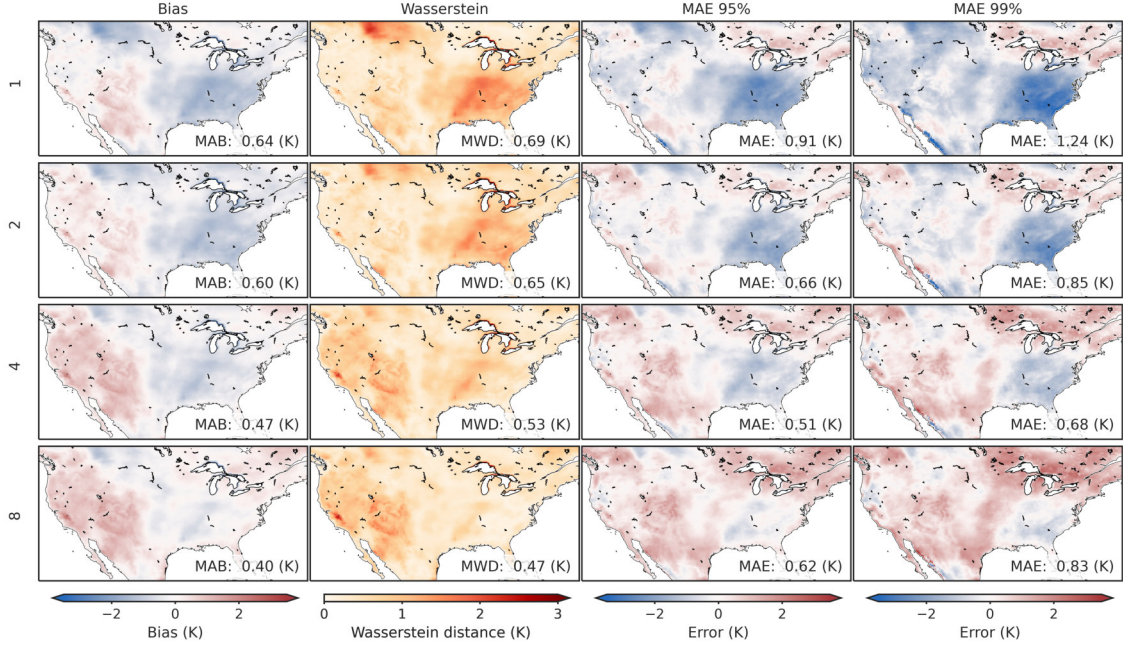


Figure SI 49: Metrics of heat index. Metrics of the derived variable heat index over CONUS during the summer (June–August) for the evaluation period 2010–2019 for GenFocal trained with different number of LENS2 ensemble members.

period and extract the corresponding samples from both datasets (sampling across different members for the LENS2 dataset). These samples are then normalized using their respective day-of-the-year climatologies. As explained in Section SI 9.4.6, the rectified flow model is trained to map the resulting anomalies from the climate simulation to the coarse-grained reanalysis.

For the ablation study, we relax the coupling by allowing the timestamps of the paired samples to differ by a defined threshold. We consider three variations:

- **Day shift ($\pm dd$):** The year matches, but the day-of-the-year may differ by up to d -number of days. In this regime, the climatologies remain relatively similar.
- **Year shift ($\pm yy$):** The day-of-the-year matches, but the years may differ by up to y -number of years.
- **Random coupling:** We use the tensor product of the marginals, $\pi = \mu_y^i \otimes \mu_y'$, effectively ignoring timestamps entirely.

To ensure a fair comparison, all ablation experiments use the *exact* same configuration, including the learning rate schedule, number of iterations, and random seed for weight initialization.

Table SI 15 summarizes the evaluation of models trained with different couplings over CONUS during the summers (June–August) of the testing period, highlighting several important trends. When the year is fixed, errors increase as the daily climatologies become misaligned. Conversely, when the day-of-the-year is aligned but the years are allowed to shift, small shifts (e.g., $\pm 1y$) can slightly improve performance, likely due to the increased diversity of training samples. However, performance regresses sharply as the year window widens further. This degradation stems from the loss of year-scale alignment; because the

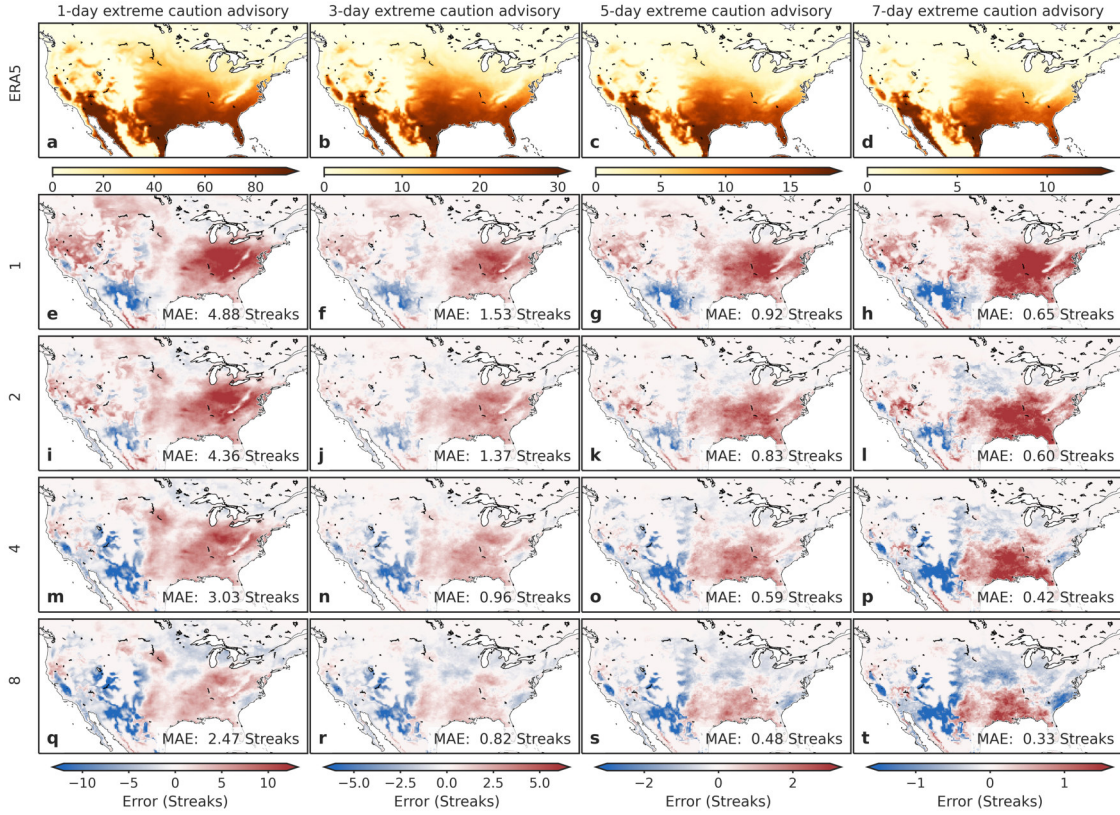


Figure SI 50: Bias in extreme caution advisories. Spatial distribution of mean absolute errors in the number of extreme caution advisory streaks of different lengths per year. We show the ground truth (ERA5)(a-d), and the pointwise errors of GenFocal trained with different number of LENS2 members.

model is trained on only 20 years of a slowly evolving distribution and must extrapolate accurately to the future, it is imperative that the network captures this time-dependent signal. Widening the coupling window blurs these subtle distributional shifts. Finally, using a random coupling (ignoring timestamps entirely) results in a sharp degradation across all metrics.

Fig. SI 51 illustrates the biases over CONUS during the summer months of 2010–2019. Increasing the misalignment in days does not drastically alter the geographical distribution of errors, with the exception of the mean sea-level pressure, which improves with small differences but deteriorates for differences beyond a week. When maintaining the same day-of-the-year but varying the years, the spatial pattern of biases remains roughly constant, though their magnitude fluctuates: decreasing for a one-year difference before rapidly increasing. In contrast, the random coupling yields a distinct spatial error distribution with significantly larger biases. A similar trend is observed in Fig. SI 52 for the heat index metrics, where the geographical distribution of errors shifts smoothly as the temporal difference (in days or years) increases. As with the direct variables, the errors for the random coupling are significantly larger.

A similar trend is observed for tropical cyclones (TCs), as shown in Fig. SI 53. Increasing the maximum day shift preserves the distribution for small shifts; however, beyond a one-week threshold, the models begin to underestimate the annual TC count. Conversely, widening the year window results in a much more rapid decline in TC frequency. In contrast, the fully random coupling severely overestimates the number of TCs. This is further corroborated by the Saffir-Simpson intensity distributions shown in

Fig. SI 54, where the response to increasing day shifts is relatively smooth, whereas year shifts induce a much sharper transition (Fig. SI 55). Finally, the completely random coupling, given by the tensor product of the marginal measures, exhibits markedly anomalous behavior, consistent with its poor performance on other metrics.

Table SI 15: Effect of training data coupling on the mean absolute bias, mean Wasserstein distance, and mean absolute error of the 99th percentile. Results are shown for various coupling strategies applied to the CONUS summer seasons (June–August) from 2010 to 2019. Metric definitions are provided in section SI 7.

Variable	± 0	$\pm 1d$	$\pm 2d$	$\pm 4d$	$\pm 8d$	$\pm 16d$	$\pm 1y$	$\pm 2y$	$\pm 4y$	random
Mean Absolute Bias, $\downarrow$										
Temperature (K)	0.41	0.58	0.59	0.62	0.66	0.56	0.4	0.49	0.56	0.8
Wind speed (m/s)	0.19	0.19	0.18	0.19	0.19	0.2	0.17	0.19	0.19	0.47
Specific humidity (g/kg)	0.31	0.41	0.44	0.46	0.5	0.48	0.25	0.30	0.31	0.47
Sea-level pressure (Pa)	39.92	30.91	35.16	35.15	42.88	55.2	50.54	53.59	49.67	143.22
Relative humidity (%)	1.71	1.73	1.8	1.8	1.88	1.92	1.77	1.97	2.00	2.82
Heat index (K)	0.47	0.69	0.71	0.73	0.77	0.64	0.41	0.52	0.61	0.88
Mean Wasserstein Distance, $\downarrow$										
Temperature (K)	0.47	0.63	0.64	0.68	0.72	0.63	0.46	0.54	0.6	0.83
Wind speed (m/s)	0.22	0.23	0.22	0.23	0.24	0.25	0.21	0.23	0.24	0.5
Specific humidity (g/kg)	0.36	0.47	0.49	0.52	0.56	0.55	0.34	0.38	0.39	0.51
Sea-level pressure (Pa)	52.09	49.53	51.22	55.44	65.59	71.86	61.41	67.08	60.36	149.25
Relative humidity (%)	2.1	2.17	2.25	2.28	2.37	2.41	2.19	2.37	2.38	3.2
Heat index (K)	0.53	0.74	0.76	0.79	0.83	0.72	0.49	0.57	0.65	0.92
Mean Absolute Error, 99 th $\downarrow$										
Temperature (K)	0.61	0.71	0.7	0.68	0.66	0.68	0.8	0.95	1.02	1.03
Wind speed (m/s)	0.48	0.5	0.47	0.53	0.60	0.62	0.45	0.48	0.49	1.22
Specific humidity (g/kg)	0.45	0.47	0.46	0.45	0.44	0.46	0.6	0.68	0.7	0.56
Sea-level pressure (Pa)	77.99	99.98	94.35	105.06	117.39	97.2	75.73	83.54	75.96	103.49
Relative humidity (%)	1.87	1.94	1.96	2.07	2.22	2.21	1.97	1.99	1.98	5.31
Heat index (K)	0.68	0.8	0.81	0.76	0.71	0.71	0.87	1.05	1.14	1.19

SI 10.7 Training period for the super-resolution stage

We observe that the performance of GenFocal is insensitive to the training period of the super-resolution component. This contrasts with the debiasing step, as illustrated in Fig. SI 56, which displays the TC statistics (including raw and categorized counts). Similar trends were observed across all other evaluation metrics. These results provide empirical evidence that the GenFocal framework effectively decomposes the problem into stationary (super-resolution) and non-stationary components.

SI 10.8 Temporally coherent denoising in super-resolution

The utility of the time-coherent sampling technique employed in the super-resolution stage (section SI 9.5.3) is directly reflected in the temporal spectra. As shown in Table SI 16, this technique leads to lower spectral errors for all variables, especially for temperature and pressure.

This technique allows GenFocal to better capture statistics of events with durations more than the training sample length. Fig. SI 57 presents the spatially averaged bias for streak counts over 1 to 3 weeks. We observe that across a wide range of thresholds and durations, GenFocal has better statistics estimation compared to BCSD and STAR-ESDM. Furthermore, the results remain insensitive to the

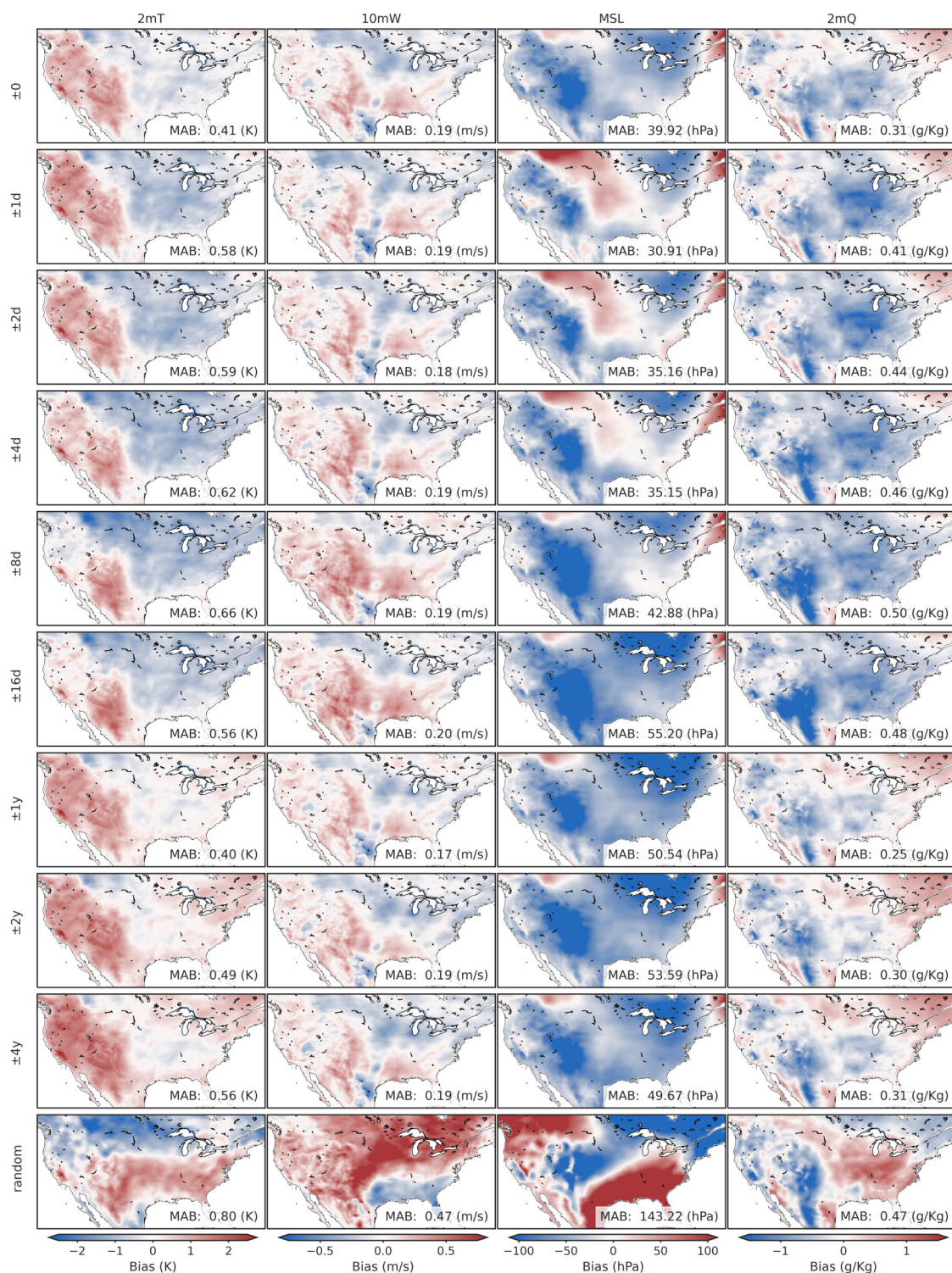


Figure SI 51: Spatial distribution of biases. Spatial distribution of statistical biases of the downscaled variables over CONUS during the summer (June-August) for the evaluation period 2010-2019 for GenFocal with different training data couplings.

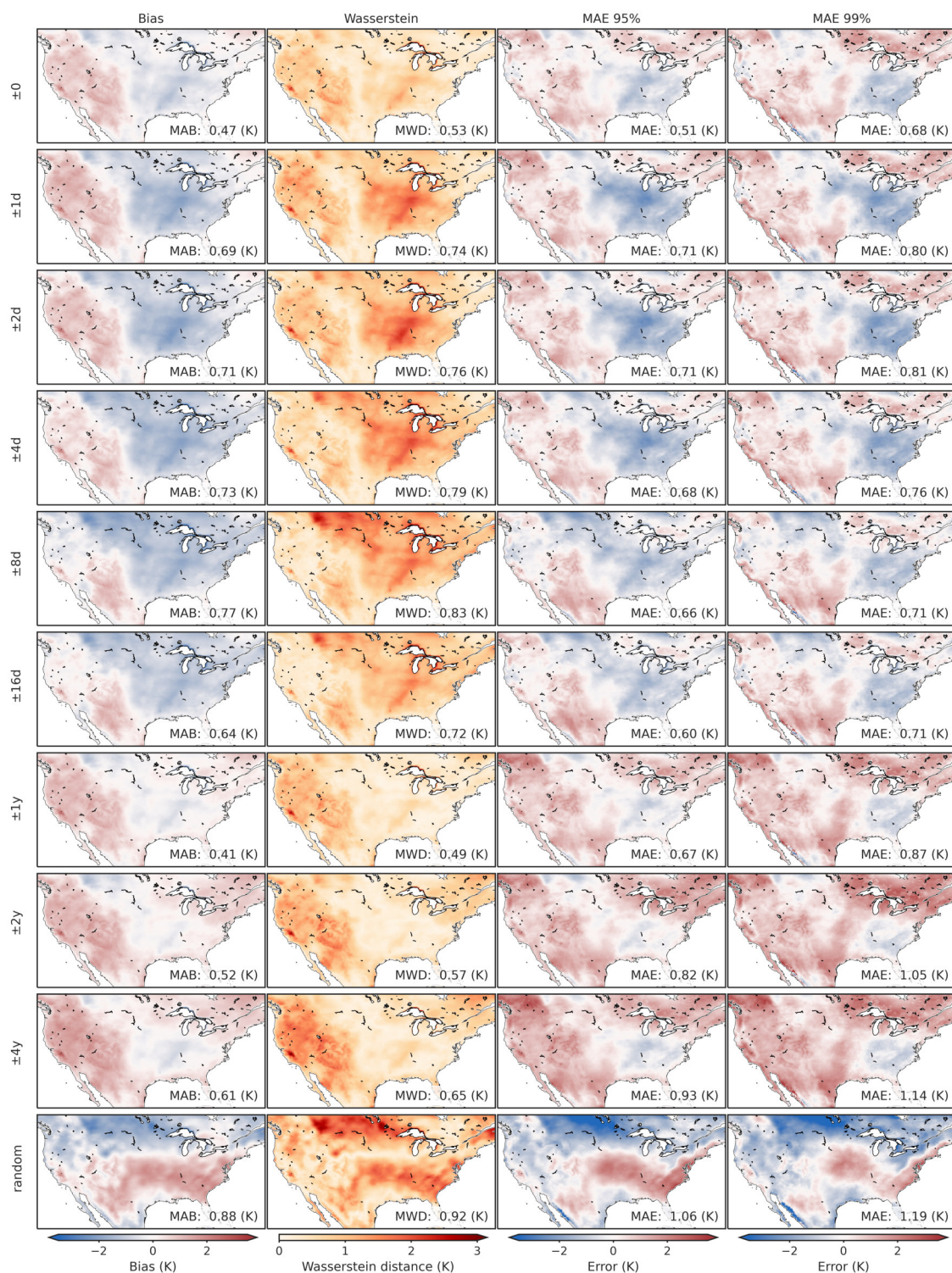


Figure SI 52: Spatial distribution of different errors for the heat index. Spatial distribution of different errors for the downscaled heat index compound variables over CONUS during the summer (June-August) for the evaluation period 2010-2019 for GenFocal with different training data couplings.

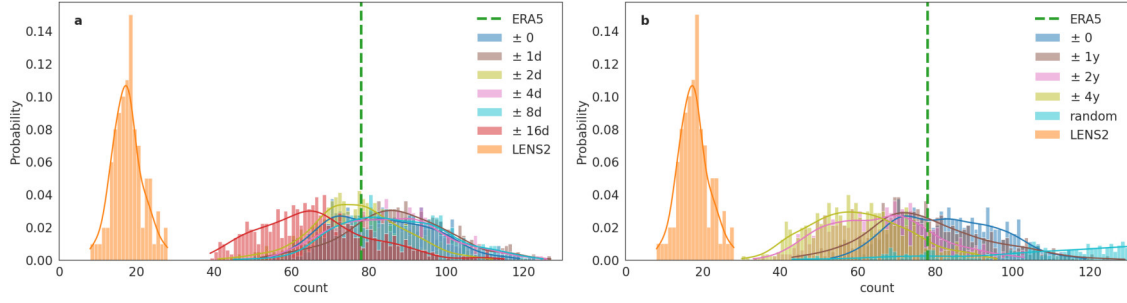


Figure SI 53: Distribution of TC counts for different training couplings. Distribution of the number of TCs detected in the North Atlantic for the hurricane season August–October during 2010–2019 for GenFocal with different training data couplings.

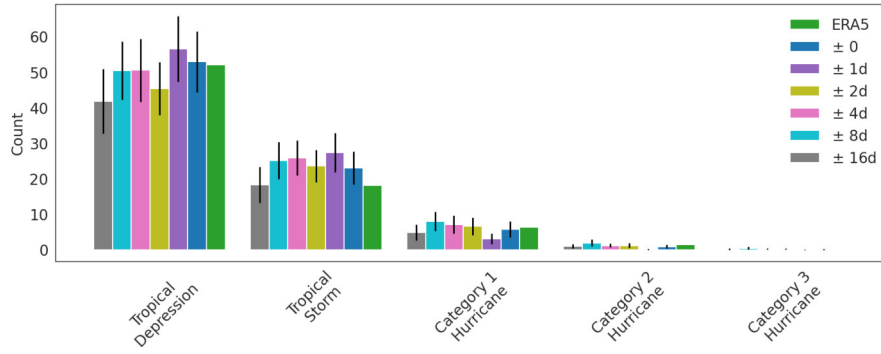


Figure SI 54: Saffir-Simpson scale distribution of TCs. Distribution of the intensity of TCs detected in the North Atlantic during the peak hurricane season (August–October), 2010–2019. Results shown for GenFocal models trained with different couplings, for small differences (in days) of the time stamps. Error bars denote the ensemble standard deviation.

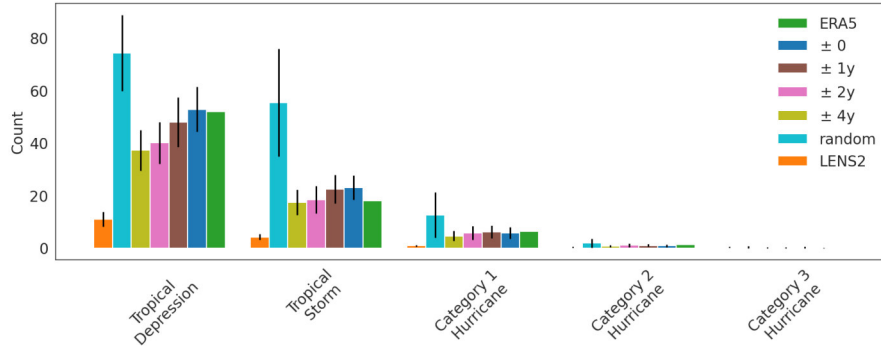


Figure SI 55: Saffir-Simpson scale distribution of TCs. Distribution of the intensity of TCs detected in the North Atlantic during the peak hurricane season (August–October), 2010–2019. Results shown for GenFocal models trained with different couplings, in which the day-of-the-year remains fixed and we sample from similar years. Error bars denote the ensemble standard deviation.

duration of the training samples for the super-resolution step (shown for 3-day and 7-day variants), demonstrating the robustness of our time-coherent sampling technique and the scalability of GenFocal

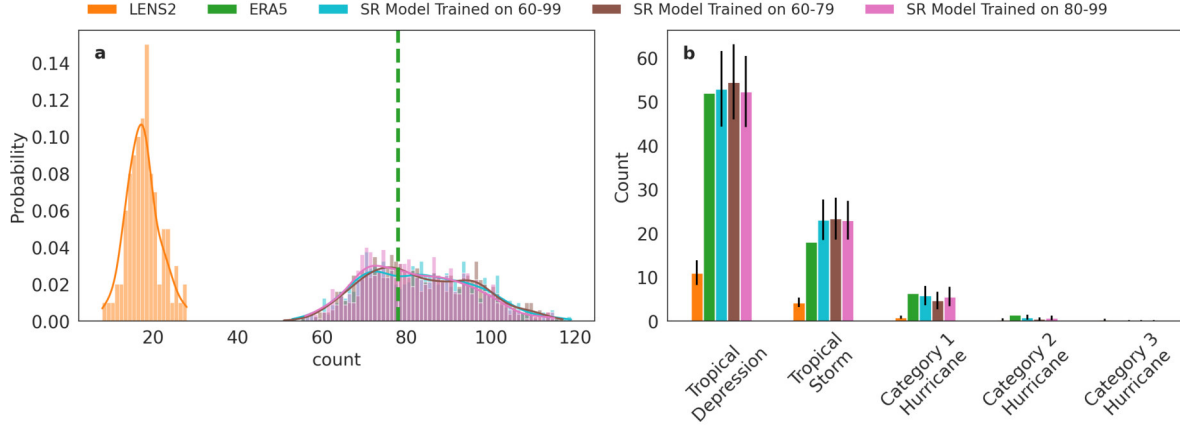


Figure SI 56: TC statistics for different super-resolution model training ranges. The same debiasing model (trained on 1980-1999) is shared between all GenFocal variants.

Table SI 16: Impact of time-coherent sampling (section SI 9.5.3) on the mean temporal spectra error (TSE). Columns represent variants with the super-resolution step trained on 7-day or 3-day samples, and time-coherent sampling enabled (+) or not (-) during inference.

	7-day +	7-day -	% change	3-day +	3-day -	% change
Temperature	0.746	0.780	-4.3	0.728	0.788	-7.6
Wind speed	0.416	0.417	-0.2	0.393	0.397	-1.0
Humidity	0.536	0.556	-3.6	0.533	0.572	-6.8
Pressure	0.513	0.566	-9.4	0.510	0.614	-16.9

in capturing the statistics of persistent, long-lasting events.

Fig. SI 58 demonstrates that time-coherent sampling results in improved TC statistics. Notably, it leads to similar distributions regardless of window size (Fig. SI 58a). Figs. SI 58b and SI 58c reveal a tendency to underestimate TCs when the technique is absent.

SI 10.9 Residual modeling in super-resolution

Table SI 17 presents the ablation between the main GenFocal model against one trained to directly predict the output, i.e. without applying the residual formulation (reference equation). We observe that modeling the residual between high-resolution target and the interpolated low-resolution input leads to significant improvements in the distribution of temperature-related variables, as well as the extreme percentiles for all variables except pressure.

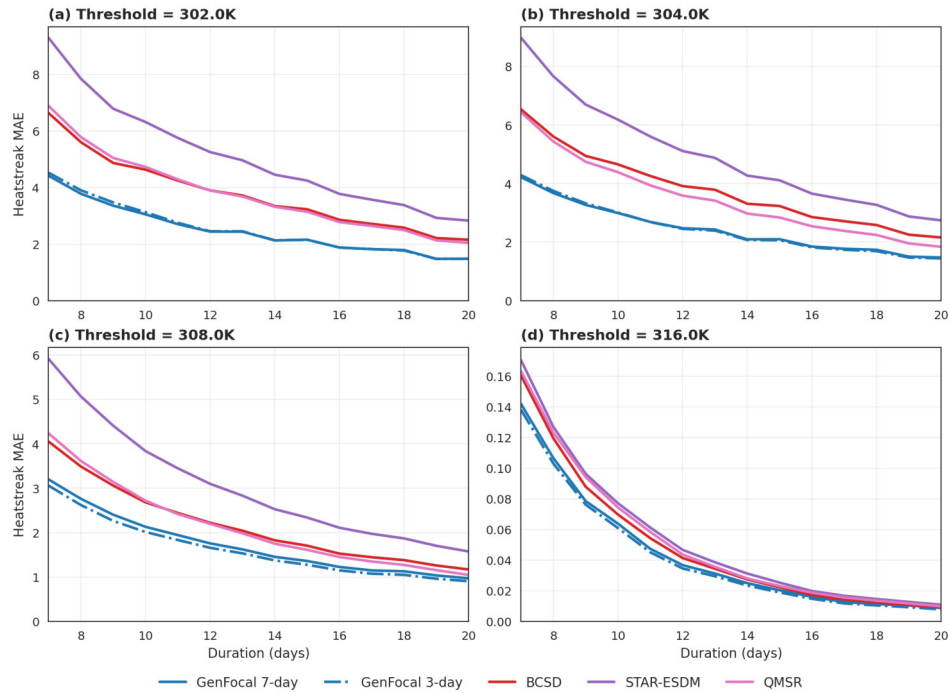


Figure SI 57: Spatially averaged bias in the number of heat streaks vs. durations for selected thresholds. GenFocal 7-day and 3-day represent variants whose super-resolution step is trained on samples of different lengths, with time-coherent sampling (1-day overlap) applied at inference time.

Table SI 17: Effect residual modeling on mean absolute bias, mean Wasserstein distance, and mean absolute error in the 99th percentile for different variables. The precise definitions of the metrics are included in SI 7.

Variable	Mean Absolute Bias ↓		Mean Wasserstein Distance ↓		Mean Absolute Error, 99 th ↓	
	residual	direct	residual	direct	residual	direct
Temperature (K)	0.41	0.73	0.47	0.75	0.61	1.16
Wind speed (m/s)	0.19	0.19	0.22	0.22	0.48	0.52
Specific humidity (g/kg)	0.31	0.29	0.36	0.37	0.45	0.75
Sea-level pressure (Pa)	39.92	46.61	52.09	56.96	77.99	73.96
Relative humidity (%)	1.71	1.98	2.10	2.16	1.87	2.08
Heat index (K)	0.47	0.79	0.53	0.82	0.68	1.37

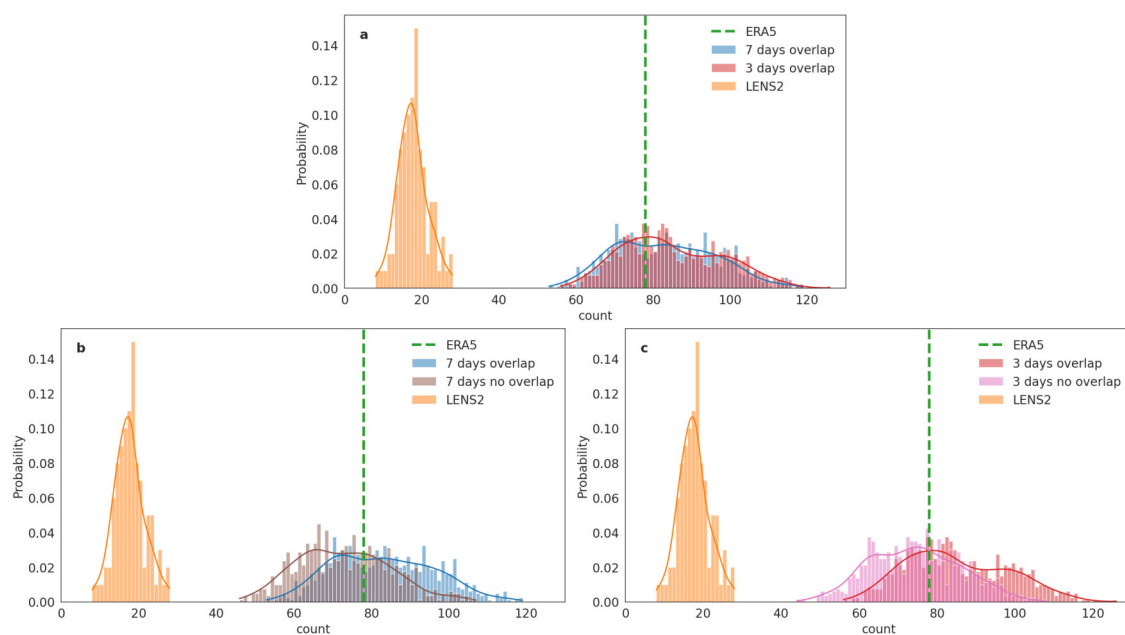


Figure SI 58: Distribution of North Atlantic TC counts for super-resolution models with varying training sample lengths and time-coherent sampling. a. super-resolution models trained with *7-day vs. 3-day samples*; sampled with time coherence. b. super-resolution model trained with 7-day samples, *with vs. without time-coherent sampling*. c. super-resolution model trained with 3-day samples, *with vs. without time-coherent sampling*.

SI 11 Additional studies

We include additional studies for environmental risks during the Northern Hemisphere winter (December, January and February) over the evaluation period. During these months, the combination of cold temperatures and strong winds can pose human safety hazards, such as hypothermia and frostbite [64]. Persistent freezing events can cause significant damage to agricultural crops and infrastructure.

To this end, we evaluate the ability of GenFocal to capture winter extremes by analyzing the tail dependence of high near-surface winds and low near-surface temperatures at a fixed time of day (12Z). We also assess the statistics of multi-day streaks of freezing daily minimum temperatures and windchill temperatures as projected by GenFocal, comparing them against the ERA5 reanalysis.

Consistent with everywhere else, we perform the model selection from the same pool of models trained for studying heat streaks in Northern Hemisphere reported in the main text and section SI 3. To save compute, we did not perform full end-to-end model selection. Using the validation data in the winters from the period of 2000-2009, we chose the model with the lowest 2m temperature bias, after the debiasing step at the coarse-level, namely without going through the super-resolution stage. We then applied the model to the test data from the period of 2010-2019, including both debiasing and super-resolution. We report metrics not only on single variables but also on derived as well as compound ones.

SI 11.1 Definition of Events

Wind Chill Temperature The wind chill temperature (WCT) is a derived meteorological index that models the rate of heat loss from exposed human skin under combined wind and temperature conditions, and adopted by the National Weather Service (NWS) and Environment Canada [64]. The WCT is defined as

$$WCT = 13.12 + 0.6215T_{air} - 11.37V^{0.16} + 0.3965T_{air}V^{0.16} \quad (92)$$

where T_{air} is the air temperature in degrees Celsius ($^{\circ}\text{C}$) and V is the wind speed at 10meter elevation in kilometers per hour (km/h).

Because WCT depends on the joint distribution of temperature and wind speed, accurate modeling requires a method that captures their inter-variable correlations.

The WCT is a critical physiological metric for public safety. It informs safety thresholds for outdoor operations, with -27°C (246.15 K) being the limit below which frostbite can occur.

SI 11.2 Pixel-wise statistics

We first evaluate the capability of GenFocal to reconstruct marginal distributions of key variables. Table SI 18 presents the statistical modeling errors for different variables during the winter months. Here GenFocal outperforms the baselines in most of the metrics. While BCSD and STAR-ESDM perform adequately on simple variables like wind speed, GenFocal demonstrates superior performance on derived variables like the wind chill temperature.

Table SI 18: Statistical modeling errors of directly and derived downscaled variables in marginal distributions for winters (December-January-February) in CONUS (2010-2019). GenFocal consistently outperforms baselines in capturing the distribution shapes (Wasserstein) and extremes (99th percentile MAE). Best values are in bold font. The precise definitions of the metrics are included in [SI 7](#)

Variable	GenFocal	BCSD	STAR-ESDM
Mean Absolute Bias ↓			
Temperature (K)	0.44	0.54	0.66
Wind speed (m/s)	0.18	0.12	0.15
Specific humidity (g/kg)	0.15	0.17	0.26
Sea-level pressure (Pa)	135.22	84.2	90.21
Windchill temperature (K)	0.53	0.70	0.86
Mean Wasserstein Distance ↓			
Temperature (K)	0.54	0.66	0.74
Wind speed (m/s)	0.20	0.22	0.21
Specific humidity (g/kg)	0.19	0.21	0.29
Sea-level pressure (Pa)	138.80	88.65	93.82
Windchill temperature (K)	0.66	0.84	0.95
Mean Absolute Error, 99 th % ↓			
Temperature (K)	1.08	1.16	1.19
Wind speed (m/s)	0.36	0.50	0.47
Specific humidity (g/kg)	0.38	0.33	0.42
Sea-level pressure (Pa)	124.20	110.82	123.94
Windchill temperature (K)	1.33	1.37	1.44

SI 11.3 Multi-day frostbite episodes

Beyond instantaneous statistics, temporal coherence is vital for assessing prolonged risk. We define a “frostbite episode” as consecutive days where the wind chill temperature remains below the critical threshold of -27°C (246.15 K). Capturing the risk of these episodes is essential for public health warnings.

Table [SI 19](#) summarizes the performance across three metrics: Mean Absolute Bias (accuracy of count), Mean Continuous Ranked Probability Score (CRPS, probabilistic accuracy), and the Spread-Skill Ratio (SSR). An SSR of 1.0 indicates perfect uncertainty calibration.

The results show that GenFocal significantly outperforms BCSD across all durations. Notably, as the threshold duration increases (from 1 to 7 days), GenFocal maintains a high SSR (0.84 for 7 days) compared to the baseline (0.74). This suggests that GenFocal is not only more accurate but also more reliable in quantifying the uncertainty of extreme, multi-day cold events.

SI 11.4 Tail Dependencies

Fig. [SI 60](#) shows that GenFocal captures the co-occurrence of cold and windy extremes better than the BCSD and STAR-ESDM baselines. In particular, the improvement is most prominent across the Western and Southern United States, where the baselines tend to overestimate and GenFocal reflects statistics much closer to that of the ERA5 reanalysis.

Table SI 19: Statistical errors for the number of multi-day frostbite episodes duration evaluation period (December-January-February, CONUS 2010-2019) using mean absolute bias, mean continuous ranked probability score (CRPS), and mean spread-skill ratio (SSR). Best values are in bold font.

Duration	GenFocal	BCSD	STAR-ESDM
	Mean Absolute Bias ↓		
1 day	1.48	1.89	1.91
3 days	0.41	0.51	0.52
5 days	0.20	0.26	0.26
7 days	0.12	0.15	0.15
Mean CRPS ↓			
1 day	0.71	1.30	1.34
3 days	0.21	0.35	0.37
5 days	0.11	0.18	0.19
7 days	0.06	0.10	0.11
Mean SSR (Ideal ≈ 1.0) ↑			
1 day	0.79	0.68	0.67
3 days	0.79	0.70	0.69
5 days	0.82	0.74	0.73
7 days	0.84	0.74	0.73

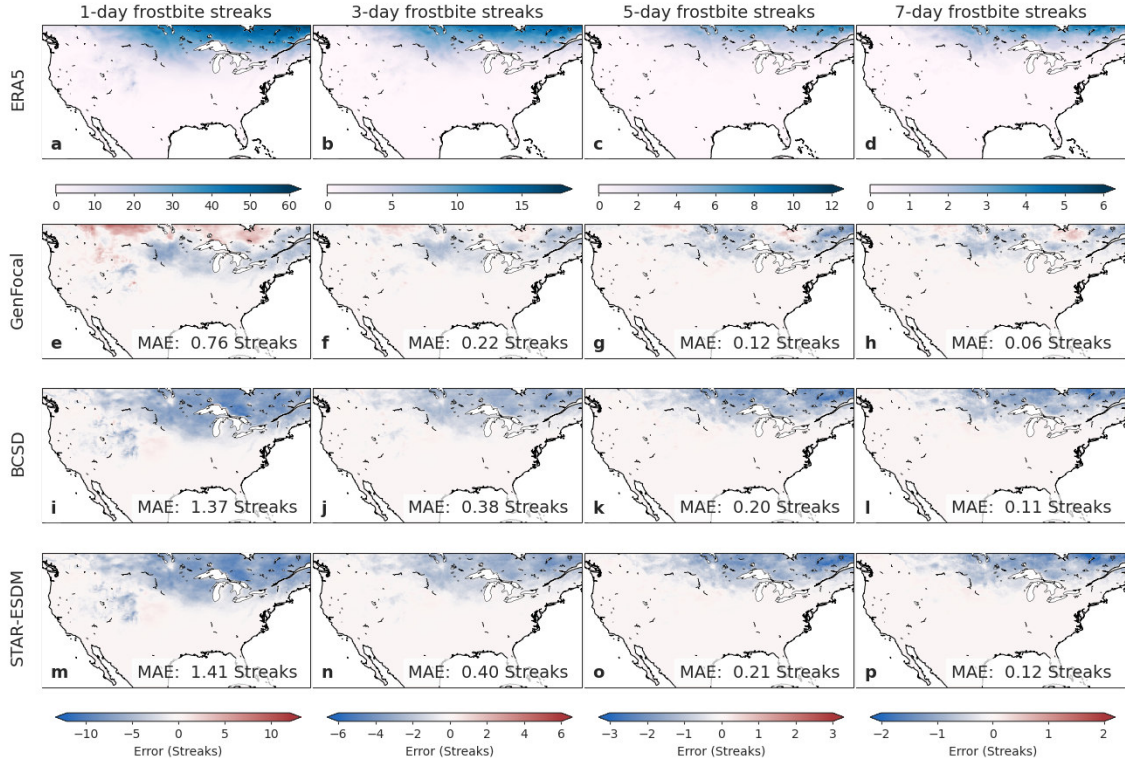


Figure SI 59: Frostbite streak statistics in CONUS winter (December-January-February) season, lasting [1, 3, 5, 7] days. a-d. ERA5 reanalysis. e-h. local bias of GenFocal. i-l. local bias of BCSD. and m-p. local bias of STAR-ESDM.

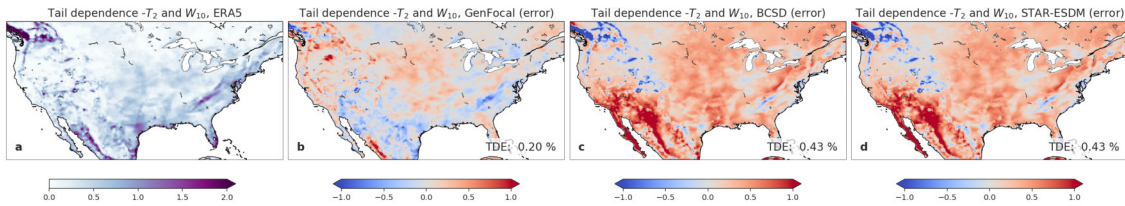


Figure SI 60: Tail dependence of low surface temperature and high wind speed in CONUS winter (December-January-February) season. a. ERA5 reanalysis. b. GenFocal c. BCSD and d. STAR-ESDM.

References

- [1] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.
- [2] Gary D Atkinson and Charles R Holliday. Tropical cyclone minimum sea level pressure/maximum sustained wind relationship for the western north pacific. *Monthly Weather Review*, 105(4):421–427, 1977.
- [3] Karthik Balaguru, Wenwei Xu, Chuan-Chieh Chang, L. Ruby Leung, David R. Judi, Samson M. Hagos, Michael F. Wehner, James P. Kossin, and Mingfang Ting. Increased U.S. coastal hurricane risk under climate change. *Science Advances*, 9(14):eadf0259, 2023.
- [4] Tristan Ballard and Gopal Erinjippurath. Contrastive learning for climate model bias correction and super-resolution. *arXiv preprint arXiv:2211.07555*, 2022.
- [5] Jorge Baño-Medina, Rodrigo Manzananas, Ezequiel Cima-devilla, Jesús Fernández, Jose González-Abad, Antonio Santiago Cofiño, and José Manuel Gutiérrez. Downscaling multi-model climate projection ensembles with deep learning (deepestd): contribution to cordex eur-44. *Geoscientific Model Development Discussions*, 2022:1–14, 2022.
- [6] Jorge Baño-Medina, Rodrigo Manzananas, and José Manuel Gutiérrez. Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, 13(4):2109–2124, 2020.
- [7] Jorge Baño-Medina, Rodrigo Manzananas, and José Manuel Gutiérrez. On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections. *Climate Dynamics*, 57(11):2941–2951, 2021.
- [8] Fan Bao, Chongxuan Li, Yue Cao, and Jun Zhu. All are worth words: a vit backbone for score-based diffusion models. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.
- [9] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.
- [10] Tobias Bischoff and Katherine Deck. Unpaired downscaling of fluid flows with diffusion bridges. *Artificial Intelligence for the Earth Systems*, 3(2):e230039, 2024.
- [11] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.
- [12] Gerd Bürger, Trevor Q Murdock, Arelia T Werner, Stephen R Sobie, and AJ Cannon. Downscaling extremes—an intercomparison of multiple statistical methods for present climate. *Journal of Climate*, 25(12):4366–4388, 2012.
- [13] Vikram Singh Chandel, Udit Bhatia, Auroop R Ganguly, and Subimal Ghosh. State-of-the-art bias correction of climate models misrepresent climate science and misinform adaptation. *Environmental Research Letters*, 19(9):094052, 2024.

- [14] Alexis-Tzianni Charalampopoulos, Shuai Zhang, Bryce Harrop, Lai-yung Ruby Leung, and Themistoklis Sapsis. Statistics of extreme events in coarse-scale climate simulations via machine learning correction operators trained on nudged datasets. *arXiv preprint arXiv:2304.02117*, 2023.
- [15] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [16] Gilbert P Compo, Jeffrey S Whitaker, Prashant D Sardeshmukh, Nobuki Matsui, Robert J Allan, Xungang Yin, Byron E Gleason, Russell S Vose, Glenn Rutledge, Pierre Bessemoulin, et al. The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, 137(654):1–28, 2011.
- [17] G Danabasoglu, J-F Lamarque, J Bacmeister, D A Bailey, A K DuVivier, J Edwards, L K Emmons, J Fasullo, R Garcia, A Gettelman, C Hannay, M M Holland, W G Large, P H Lauritzen, D M Lawrence, J T M Lenaerts, K Lindsay, W H Lipscomb, M J Mills, R Neale, K W Oleson, B Otto-Bliesner, A S Phillips, W Sacks, S Tilmes, L van Kampenhout, M Vertenstein, A Bertini, J Dennis, C Deser, C Fischer, B Fox-Kemper, J E Kay, D Kinnison, P J Kushner, V E Larson, M C Long, S Mickelson, J K Moore, E Nienhouse, L Polvani, P J Rasch, and W G Strand. The community earth system model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001916, 2020.
- [18] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
- [19] Keith W Dixon, John R Lanzante, Mary J Nath, Katharine Hayhoe, Anne Stoner, Aparna Radhakrishnan, Venkat Balaji, and Carlos F Gaitan. Evaluating the stationarity assumption in statistically downscaled climate projections: is past performance an indicator of future results? *Climatic Change*, 135(3-4):395–408, 2016.
- [20] Max Duarte, Ann S. Almgren, Kaushik Balakrishnan, John B. Bell, and David M. Romps. A numerical study of methods for moist atmospheric flows: Compressible equations. *Mon. Wea. Rev.*, 142:4269–4283, 2014.
- [21] Bradley Efron. Tweedie’s formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.
- [22] Kerry Emanuel. Response of global tropical cyclone activity to increasing CO₂: Results from downscaling CMIP6 models. *Journal of Climate*, 34(1):57 – 70, 2021.
- [23] Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific model development*, 9(5):1937–1958, 26 May 2016.
- [24] John T Fasullo, Jean-Christophe Golaz, Julie M Caron, Nan Rosenbloom, Gerald A Meehl, Warren Strand, Sasha Glanville, Samantha Stevenson, Maria Molina, Christine A Shields, Chengzhu Zhang, James Benedict, Hailong Wang, and Tony Bartoletti. An overview of the E3SM version 2 large ensemble and comparison to other E3SM and CESM large ensembles. *Earth system dynamics*, 15(2):367–386, 8 April 2024.

- [25] Andra J. Garner, Robert E. Kopp, and Benjamin P. Horton. Evolving tropical cyclone tracks in the north atlantic in a warming climate. *Earth's Future*, 9(12):e2021EF002326, 2021. e2021EF002326 2021EF002326.
- [26] Filippo Giorgi. Thirty years of regional climate modeling: Where are we and where are we going next? *Journal of Geophysical Research: Atmospheres*, 124:5696–5723, 6 2019.
- [27] Stanley B. Goldenberg, Christopher W. Landsea, Alberto M. Mestas-Núñez, and William M. Gray. The recent increase in Atlantic hurricane activity: Causes and implications. *Science*, 293(5529):474–479, 2001.
- [28] Naomi Goldenson, L Ruby Leung, Linda O Mearns, David W Pierce, Kevin A Reed, Isla R Simpson, Paul Ullrich, Will Krantz, Alex Hall, Andrew Jones, and Stefan Rahimi. Use-inspired, process-oriented GCM selection: Prioritizing models for regional dynamical downscaling. *Bulletin of the American Meteorological Society*, 104:E1619–E1629, 2023.
- [29] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2066–2073. IEEE, 2012.
- [30] Brian Groenke, Luke Madaus, and Claire Monteleoni. Climalign: Unsupervised statistical downscaling of climate variables via normalizing flows. In *Proceedings of the 10th International Conference on Climate Informatics*, CI2020, page 60–66, New York, NY, USA, 2021. Association for Computing Machinery.
- [31] Paula Harder, Alex Hernandez-Garcia, Venkatesh Ramesh, Qidong Yang, Prasanna Sattigeri, Daniela Szwarzman, Campbell Watson, and David Rolnick. Hard-constrained deep learning for climate downscaling. *Journal of Machine Learning Research*, 24(365):1–40, 2023.
- [32] Lucy Harris, Andrew TT McRae, Matthew Chantry, Peter D Dueben, and Tim N Palmer. A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, 14(10):e2022MS003120, 2022.
- [33] Katharine Hayhoe, Ian Scott-Fleming, Anne Stoner, and Donald J. Wuebbles. STAR-ESDM: A generalizable approach to generating high-resolution climate projections through signal decomposition. *Earth's Future*, 12(7):e2023EF004107, 2024.
- [34] Katharine Hayhoe, Ian Scott-Fleming, Anne Stoner, and Donald J Wuebbles. STAR-ESDM: A generalizable approach to generating high-resolution climate projections through signal decomposition. *Earth's Future*, 12(7):e2023EF004107, 2024.
- [35] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

- [36] Masoud Hessami, Philippe Gachon, Taha B.M.J. Ouarda, and André St-Hilaire. Automated regression-based statistical downscaling tool. *Environmental Modelling and Software*, 23(6):813–834, 2008.
- [37] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.
- [38] Shailesh Kumar Jha, Vivek Gupta, Priyank J Sharma, Anurag Mishra, and Saksham Joshi. Deep learning super-resolution for temperature data downscaling: a comprehensive study using residual networks. *Frontiers in Climate*, 7:1572428, 2025.
- [39] Renzhi Jing, Jianxiong Gao, Yunuo Cai, Dazhi Xi, Yinda Zhang, Yanwei Fu, Kerry Emanuel, Noah S. Diffenbaugh, and Eran Bendavid. TC-GEN: Data-driven tropical cyclone downscaling using machine learning-based high-resolution weather model. *Journal of Advances in Modeling Earth Systems*, 16(10):e2023MS004203, 2024. e2023MS004203 2023MS004203.
- [40] Renzhi Jing, Ning Lin, Kerry Emanuel, Gabriel Vecchi, and Thomas R. Knutson. A comparison of tropical cyclone projections in a high-resolution global climate model and from downscaling by statistical and statistical-deterministic methods. *Journal of Climate*, 34(23):9349 – 9364, 2021.
- [41] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.
- [42] Thomas Knutson, Suzana J. Camargo, Johnny C. L. Chan, Kerry Emanuel, Chang-Hoi Ho, James Kossin, Mrutyunjay Mohapatra, Masaki Satoh, Masato Sugi, Kevin Walsh, and Liguang Wu. Tropical cyclones and climate change assessment: Part ii: Projected response to anthropogenic warming. *Bulletin of the American Meteorological Society*, 101(3):E303 – E322, 2020.
- [43] Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632:1060–1066, 2024.
- [44] Nikolay Koldunov, Thomas Rackow, Christian Lessig, Sergey Danilov, Suvarchal K Cheedela, Dmitry Sidorenko, Irina Sandu, and Thomas Jung. Emerging ai-based weather prediction models as downscaling tools. URL <https://arxiv.org/abs/2406.17977>, 2024.
- [45] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.
- [46] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.
- [47] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.
- [48] Ben Livneh, Eric A Rosenberg, Chiyu Lin, Bart Nijssen, Vimal Mishra, Kostas M Andreadis, Edwin P Maurer, and Dennis P Lettenmaier. A long-term hydrologically based dataset of land surface fluxes and states for the conterminous united states: Update and extensions. *Journal of Climate*, 26(23):9384–9392, 2013.

- [49] Ignacio Lopez-Gomez, Zhong Yi Wan, Leonardo Zepeda-Núñez, Tapio Schneider, John Andersona, and Fei Sha. Dynamical-generative downscaling of climate model ensembles. *Proc. Natl. Acad. Sci.*, 122:e2420288122, 2025.
- [50] Ignacio Lopez-Gomez, Zhong Yi Wan, Leonardo Zepeda-Núñez, Tapio Schneider, John Anderson, and Fei Sha. Dynamical-generative downscaling of climate model ensembles. *Proceedings of the National Academy of Sciences*, 122:e2420288122, 4 2025. doi: 10.1073/pnas.2420288122.
- [51] Yi-Chuan Lu and David M. Romps. Extending the heat index. *Journal of Applied Meteorology and Climatology*, 61(10):1367 – 1383, 2022.
- [52] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. SRFlow: Learning the super-resolution space with normalizing flow. In *ECCV*, 2020.
- [53] Ran Lyu, Linhan Wang, Yanshen Sun, Hedanqiu Bai, and Chang-Tien Lu. Downscaling precipitation with bias-informed conditional diffusion model. In *2024 IEEE International Conference on Big Data (BigData)*, pages 8768–8770. IEEE, 2024.
- [54] Andrew J Majda. Challenges in climate science and contemporary applied mathematics. *Communications on Pure and Applied Mathematics*, 65(7):920–948, 2012.
- [55] Douglas Maraun. Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *Journal of Climate*, 26(6):2137 – 2143, 2013.
- [56] Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, Karthik Kashinath, Jan Kautz, and Mike Pritchard. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6:124, 2025.
- [57] National Oceanic and Atmospheric Administration (NOAA). Heat forecast tools.
- [58] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.
- [59] Fabio Merizzi, Andrea Asperti, and Stefano Colamonaco. Wind speed super-resolution and validation: from era5 to cerra via diffusion models. *Neural Computing and Applications*, 36(34):21899–21921, 2024.
- [60] Roberto Molinaro, Samuel Lanthaler, Bogdan Raonić, Tobias Rohner, Victor Armegioiu, Zhong Yi Wan, Fei Sha, Siddhartha Mishra, and Leonardo Zepeda-Núñez. Generative AI for fast and accurate statistical computation of fluids. *arXiv preprint arXiv:2409.18359*, 2024.
- [61] Arata Okuyama, Akiyoshi Andou, Kenji Date, Keita Hoasaka, Nobutaka Mori, Hidehiko Murata, Tasuku Tabata, Masaya Takahashi, Ryoko Yoshino, and Kotaro Bessho. Preliminary validation of himawari-8/ahi navigation and calibration. In *Earth Observing Systems XX*, volume 9607, pages 663–672. SPIE, 2015.
- [62] Brian C. O’Neill, Claudia Tebaldi, Detlef P. van Vuuren, Veronika Eyring, Pierre Friedlingstein, George Hurtt, Reto Knutti, Elmar Kriegler, Jean-Francois Lamarque, Jason Lowe, Gerald A. Meehl, Richard Moss, Keywan Riahi, and Benjamin M. Sanderson. The scenario model intercomparison project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, 9:3461–3482, 9 2016.

- [63] Ariel Ortiz-Bobea, Toby R Ault, Carlos M Carrillo, Robert G Chambers, and David B Lobell. Anthropogenic climate change has slowed global agricultural productivity growth. *Nature Climate Change*, 11:306–312, 2021.
- [64] Randall Osczevski and Maurice Bluestein. The new wind chill equivalent temperature chart. *Bulletin of the American Meteorological Society*, 86(10):1453 – 1458, 2005.
- [65] Baoxiang Pan, Gemma J Anderson, André Goncalves, Donald D Lucas, Céline JW Bonfils, Jiwoo Lee, Yang Tian, and Hsi-Yen Ma. Learning to correct climate projection biases. *Journal of Advances in Modeling Earth Systems*, 13(10):e2021MS002509, 2021.
- [66] Hans A Panofsky and Glenn Wilson Brier. Some applications of statistics to meteorology. (*No Title*), 1968.
- [67] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020.
- [68] Jaideep Pathak, Yair Cohen, Piyush Garg, Peter Harrington, Noah Brenowitz, Dale Durran, Morteza Mardani, Arash Vahdat, Shaoming Xu, Karthik Kashinath, et al. Kilometer-scale convection allowing model emulation using generative diffusion modeling. *arXiv preprint arXiv:2408.10958*, 2024.
- [69] David W. Pierce, Daniel R. Cayan, Daniel R. Feldman, and Mark D. Risser. Future increases in North American extreme precipitation in CMIP6 downscaled with LOCA. *Journal of Hydrometeorology*, 24(5):951 – 975, 2023.
- [70] David W. Pierce, Daniel R. Cayan, and Bridget L. Thrasher. Statistical downscaling using localized constructed analogs (LOCA). *Journal of Hydrometeorology*, 15(6):2558 – 2585, 2014.
- [71] Ilan Price and Stephan Rasp. Increasing the accuracy and resolution of precipitation forecasts using deep generative models. In *International conference on artificial intelligence and statistics*, pages 10555–10571. PMLR, 2022.
- [72] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637:84–90, 2025.
- [73] Stefan Rahimi, Lei Huang, Jesse Norris, Alex Hall, Naomi Goldenson, Will Krantz, Benjamin Bass, Chad Thackeray, Henry Lin, Di Chen, Eli Dennis, Ethan Collins, Zachary J. Lebo, Emily Slinsky, Sara Graves, Surabhi Biyani, Bowen Wang, and Stephen Cropper. An overview of the western United States dynamically downscaled dataset (WUS-D3). *Geoscientific Model Development*, 17:2265–2286, 3 2024.
- [74] Neelesh Rampal, Sanaa Hobeichi, Peter B Gibson, Jorge Baño-Medina, Gab Abramowitz, Tom Beucler, Jose González-Abad, William Chapman, Paula Harder, and José Manuel Gutiérrez. Enhancing regional climate downscaling through advances in machine learning. *Artificial Intelligence for the Earth Systems*, 3(2):230066, 2024.
- [75] Ashwin Rode, Tamma Carleton, Michael Delgado, Michael Greenstone, Trevor Houser, Solomon Hsiang, Andrew Hultgren, Amir Jina, Robert E Kopp, Kelly E McCusker, Ishan Nath, James Rising, and Jiacan Yuan. Estimating a social cost of carbon for global energy consumption. *Nature*, 598:308–314, 2021.

- [76] K. B. Rodgers, S.-S. Lee, N. Rosenbloom, A. Timmermann, G. Danabasoglu, C. Deser, J. Edwards, J.-E. Kim, I. R. Simpson, K. Stein, M. F. Stuecker, R. Yamaguchi, T. Bódai, E.-S. Chung, L. Huang, W. M. Kim, J.-F. Lamarque, D. L. Lombardozzi, W. R. Wieder, and S. G. Yeager. Ubiquity of human-induced changes in climate variability. *Earth System Dynamics*, 12(4):1393–1411, 2021.
- [77] David M Romps and Yi-Chuan Lu. Chronically underestimated: a reassessment of us heat waves using the extended heat index. *Environmental Research Letters*, 17:094017, 9 2022.
- [78] Herbert S Saffir. Hurricane wind and storm surge. *The Military Engineer*, 65(423):4–5, 1973.
- [79] Hiroshi Sasaki, Chris G. Willcocks, and Toby P. Breckon. UNIT-DDPM: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*, 2021.
- [80] Rafael Schmidt and Ulrich Stadtmüller. Non-parametric estimation of tail dependence. *Scandinavian Journal of Statistics*, 33(2):307–335, 2006.
- [81] Robert H Simpson. The hurricane disaster—potential scale. *Weatherwise*, 27(4):169–186, 1974.
- [82] WC Skamarock, JB Klemp, J Dudhia, DO Gill, Z Liu, J Berner, W Wang, JG Powers, MG Duda, D Barker, and X Huang. A description of the advanced research WRF model version 4. Technical report, NCAR, 2021.
- [83] Cornel Soci, Hans Hersbach, Adrian Simmons, Paul Poli, Bill Bell, Paul Berrisford, András Horányi, Joaquín Muñoz Sabater, Julien Nicolas, Raluca Radu, Dinand Schepers, Sebastien Villaume, Leopold Haimberger, Jack Woollen, Carlo Buontempo, and Jean-Noël Thépaut. The ERA5 global reanalysis from 1940 to 2022. *Quarterly journal of the Royal Meteorological Society. Royal Meteorological Society (Great Britain)*, 150(764):4014–4048, 1 October 2024.
- [84] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.
- [85] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.
- [86] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *CoRR*, abs/2011.13456, 2020.
- [87] Prakhhar Srivastava, Ruihan Yang, Gavin Kerrigan, Gideon Dresdner, Jeremy McGibbon, Christopher S Bretherton, and Stephan Mandt. Precipitation downscaling with spatiotemporal video diffusion. *Advances in Neural Information Processing Systems*, 37:56374–56400, 2024.
- [88] R. G. Steadman. The assessment of sultriness. Part I: A temperature-humidity index based on human physiology and clothing science. *Journal of Applied Meteorology and Climatology*, 18(7):861 – 873, 1979.
- [89] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *The Eleventh International Conference on Learning Representations*, 2023.
- [90] Y. Qiang Sun, Pedram Hassanzadeh, Mohsen Zand, Ashesh Chattopadhyay, Jonathan Weare, and Dorian S. Abbot. Can AI weather models predict out-of-distribution gray swan tropical cyclones? *Proceedings of the National Academy of Sciences*, 122(21):e2420914122, 2025.

- [91] James P Terry and Ick-Hoi Kim. Morphometric analysis of tropical storm and hurricane tracks in the north atlantic basin using a sinuosity-based approach. *International journal of climatology: a journal of the Royal Meteorological Society*, 35(6):923–934, 1 May 2015.
- [92] Bridget Thrasher, Weile Wang, Andrew Michaelis, Forrest Melton, Tsengdar Lee, and Ramakrishna Nemani. NASA global daily downscaled projections, CMIP6. *Scientific Data*, 9:262, 2022.
- [93] Chunwei Tian, Xuanyu Zhang, Jerry Chun-Wen Lin, Wangmeng Zuo, and Yanning Zhang. Generative adversarial networks for image super-resolution: A survey. *arXiv preprint arXiv:2204.13620*, 2022.
- [94] Michael K Tippett, Suzana J Camargo, and Adam H Sobel. A poisson regression index for tropical cyclone genesis and the role of large-scale vorticity in genesis. *Journal of climate*, 24(9):2335–2357, 1 May 2011.
- [95] Elena Tomasi, Gabriele Franch, and Marco Cristoforetti. Can ai be enabled to perform dynamical downscaling? a latent diffusion model to mimic kilometer-scale cosmo5. 0_clm9 simulations. *Geoscientific Model Development*, 18(6):2051–2078, 2025.
- [96] Paul A. Ullrich. Validation of LOCA2 and STAR-ESDM statistically downscaled products. Technical report, Lawrence Livermore National Laboratory (LLNL), Livermore, CA (United States), 10 2023.
- [97] Paul A Ullrich, Colin M Zarzycki, Elizabeth E McClenny, Marielle C Pinheiro, Alyssa M Stansfield, and Kevin A Reed. TempestExtremes v2. 1: A community framework for feature detection, tracking, and analysis in large datasets. *Geoscientific Model Development*, 14(8):5023–5048, 2021.
- [98] Thomas Vandal, Evan Kodra, and Auroop R Ganguly. Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation. *Theoretical and Applied Climatology*, 137:557–570, 2019.
- [99] Thomas Vandal, Evan Kodra, Sangram Ganguly, Andrew Michaelis, Ramakrishna Nemani, and Auroop R. Ganguly. DeepSD: Generating high resolution climate change projections through single image super-resolution. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’17, page 1663–1672, New York, NY, USA, 2017. Association for Computing Machinery.
- [100] Cédric Villani. *Optimal transport: old and new*. Springer Berlin Heidelberg, 2009.
- [101] Zhong Yi Wan, Ricardo Baptista, Anudhyan Boral, Yi-Fan Chen, John Anderson, Fei Sha, and Leonardo Zepeda-Núñez. Debias coarsely, sample conditionally: Statistical downscaling through optimal transport and probabilistic diffusion models. *Advances in Neural Information Processing Systems*, 36:47749–47763, 2023.
- [102] Zhong Yi Wan, Ignacio Lopez-Gomez, Robert Carver, Tapio Schneider, John Anderson, Fei Sha, and Leonardo Zepeda-Núñez. Statistical downscaling via high-dimensional distribution matching with generative models. *arXiv preprint arXiv:2412.08079*, 2024.
- [103] Robbie A Watt and Laura A Mansfield. Generative diffusion-based downscaling for climate. *arXiv preprint arXiv:2404.17752*, 2024.
- [104] R. L. Wilby, T. M. L. Wigley, D. Conway, P. D. Jones, B. C. Hewitson, J. Main, and D. S. Wilks. Statistical downscaling of general circulation model output: A comparison of methods. *Water Resources Research*, 34(11):2995–3008, 1998.

- [105] Andrew W Wood, Lai R Leung, Venkataramana Sridhar, and DP Lettenmaier. Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic change*, 62:189–216, 2004.
- [106] Andrew W Wood, Edwin P Maurer, Arun Kumar, and Dennis P Lettenmaier. Long-range experimental hydrologic forecasting for the eastern United States. *Journal of Geophysical Research: Atmospheres*, 107(D20):ACL–6, 2002.
- [107] Andrew W Wood, Edwin P Maurer, Arun Kumar, and Dennis P Lettenmaier. Long-range experimental hydrologic forecasting for the eastern United States. *Journal of Geophysical Research: Atmospheres*, 107:ACL 6–1–ACL 6–15, 10 2002.
- [108] Chen Henry Wu and Fernando De la Torre. Unifying diffusion models’ latent space, with applications to cyclediffusion and guidance. *arXiv preprint arXiv:2210.05559*, 2022.
- [109] Haixia Xiao, Feng Zhang, Lingxiao Wang, Wenwen Li, Bin Guo, and Jun Li. Clouddiff: Super-resolution ensemble retrieval of cloud properties for all day using the generative diffusion model. *arXiv preprint arXiv:2405.04483*, 2024.
- [110] Chugang Yi, Minghan Yu, Weikang Qian, Yixin Wen, and Haizhao Yang. Efficient kilometer-scale precipitation downscaling with conditional wavelet diffusion. *arXiv preprint arXiv:2507.01354*, 2025.
- [111] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *arXiv preprint arXiv:2207.06635*, 2022.
- [112] Linjiong Zhou, Shian-Jiann Lin, Jan-Huey Chen, Lucas M Harris, Xi Chen, and Shannon L Rees. Toward convective-scale prediction within the next generation global prediction system. *Bulletin of the American Meteorological Society*, 100(7):1225–1243, 2019.