What Drives General Circulation Model Biases in Precipitation Extremes? A Generalised Pareto Distribution (GPD)-Based Diagnostic of CMIP6 Tail Behaviour

Apoorva Aladalli*

* MPhil Candidate (Department of Applied Mathematics and Theoretical Physics), University of Cambridge, UK

KEYWORDS

climate change, extreme rainfall, CMIP6, GCM biases, POT GPD diagnostic

ABSTRACT

This study uses a point-scale Peaks-Over-Threshold (POT) Generalised Pareto Distribution (GPD) framework to evaluate extreme precipitation simulated by 32 CMIP6 models from 2015 to 2024. Daily rainfall exceeding high thresholds is fitted to the GPD using three estimators: Maximum Likelihood Estimation, L-moments, and Bayesian inference. Return levels for 2, 5, 10, 20, 50, and 100 year events are computed. Two climatically distinct sites are analysed, one in tropical South India and one in mid-latitude Northern Italy, with observational benchmarks from IMD4 and E-OBS datasets. Results show systematic model biases. CMIP6 models generally overestimate extreme rainfall in India and underestimate it in Italy. Biases are closely linked to convection parameterisation schemes. Zhang-McFarlane models simulate Indian monsoon extremes within about 30 percent accuracy but underpredict Italian 100-year rainfall by 50 to 65 percent. Gregory-Rowntree scheme models show mixed results. Models with simple bulk convection schemes tend to have the largest errors, often exceeding 100 percent in both regions. Notably, the finely tuned Gregory-Rowntree model KACE-1-0-G performed well in both domains, highlighting the importance of targeted calibration regardless of scheme type. Variability in threshold selection, driven by convection scheme differences and drizzle bias, affects return level estimates more than GPD shape or scale parameter variations. Differences across estimation methods are minor compared to inter-model spread, with Bayesian inference showing the most stable results. Increasing model resolution alone does not guarantee improved simulation of extreme precipitation. Overall, findings underscore that model physics, particularly convection parameterisation, is the primary determinant of extreme precipitation biases, suggesting that advancements in physical processes offer greater potential for reducing biases than resolution enhancements alone.

I. **Introduction**

Extreme weather events such as floods, heatwaves, and storms are statistically infrequent but disproportionately responsible for large-scale social, economic, and ecological disruption. A growing body of evidence, from theoretical thermodynamic reasoning based on the Clausius–Clapeyron relationship, to observational trends and climate model simulations, demonstrates that the variability, intensity, and frequency of precipitation extremes are increasing as global temperatures rise.[1] Events such as cloudbursts, flash floods, and landslides are highly localised, typically occurring at basin or sub-basin scales, and are regulated by topography, land-surface heterogeneity, and mesoscale atmospheric dynamics. Consequently, effective risk assessment and adaptation planning require high-resolution, point-scale diagnostics to inform urban drainage design, early warning systems, and resilient infrastructure planning.

In practice, decision-makers and engineers rely on climate projections from publicly available multi-model ensembles such as the Coupled Model Intercomparison Project Phase 6 (CMIP6) to assess future hydroclimatic risks.[2] Yet a central question remains: to what extent do CMIP6 models accurately reproduce the statistical characteristics of observed precipitation extremes at the spatial scales most relevant for real-world decisions?

Most model evaluation studies to date have been conducted at regional or global scales using spatially aggregated observational products or gridded reanalyses and rely on simplified statistical frameworks, particularly the block maxima method,[3] which fits annual or seasonal maxima to a Generalised Extreme Value (GEV) distribution. Likewise, the Expert Team on Climate Change Detection and Indices (ETCCDI) suite of extreme precipitation indices (for example, Rx1day, R20mm) has become a standard tool for inter-model comparison.[4] Although these methods are widely used for their simplicity and ease of interpretation, they provide limited insight into the tail of the distribution, which is essential for characterising rare, high impact events.[5] The block maxima approach retains only one value per year or season, discarding other extreme events that exceed high thresholds but do not represent the annual maximum.[6] ETCCDI indices are policy-relevant but rely on absolute thresholds (for example, the number of days with precipitation greater than 10 mm or 20 mm), which may not be climatologically meaningful across diverse regions.[7] Moreover, each index

[1] Westra, Seth, et al. "Future Changes to the Intensity and Frequency of Short-Duration Extreme Rainfall." *Reviews of Geophysics*, vol. 52, no. 3, 2013, pp. 522-555; Kharin, Viatcheslav, et al. "Changes in Temperature and Precipitation Extremes in the CMIP5 Ensemble." *Journal of Climate*, vol. 26, no. 23, 2013, pp. 8469-8482; de Vries, I, et al. "Increasing Extreme Precipitation Variability Plays a Key Role in Future Record-Shattering Event Probability." *Nature Communications*, vol. 15, 2024, 7447.

[2] Eyring, Veronika, et al. "Overview of the CMIP6 Experimental Design and Organization." *Geoscientific Model Development*, vol. 9, no. 5, 2016, pp. 1937-1958; O'Neill, Brian, et al. "The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6." *Geoscientific Model Development*, vol. 9, no. 9, 2016, pp. 3461-3482; IPCC. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report*. Cambridge University Press, 2021.

[3] Li, Chao, et al. "Changes in Annual Extremes of Daily Temperature and Precipitation in CMIP6 Models." *Journal of Climate*, vol. 34, no. 9, 2021, pp. 3441–3460; Alaya, MA Ben, et al. "An Evaluation of Block-Maximum-Based Estimation of Very Long Return Period Precipitation Extremes with a Large Ensemble Climate Simulation." *Journal of Climate*, vol. 33, no. 16, 2020, pp. 6957–6970.

[4] Alexander, L, et al. "Global Observed Changes in Daily Climate Extremes of Temperature and Precipitation." *Journal of Geophysical Research: Atmospheres*, vol. 111, no. D5, 2006, 2005JD006290; Seneviratne, Sonia, and Mathias Hauser. "Regional Climate Sensitivity of Climate Extremes in CMIP6 Versus CMIP5 Multimodel Ensembles." *Earth's Future*, vol. 8, no. 9, 2020, e2019EF001474.

[5] ibid.

[6] Coles, Stuart. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001; Katz, Richard, et al. "Statistics of Extremes in Climate Change." *Climatic Change*, vol. 100, no. 1, 2002, pp. 71-76.

[7] Donat, Markus, et al. "Updated Analyses of Temperature and Precipitation Extreme Indices Since the Beginning of the Twentieth Century: The HadEX2 Dataset." *Journal of Geophysical Research: Atmospheres*, vol. 118, no. 6, 2013, pp. 2098-2118; de Medeiros, Jeferson, et al. "Classification and Definitions of Extreme

isolates a specific attribute such as intensity, frequency, or duration and is thus sensitive to differing statistical and physical assumptions, which limits their diagnostic power for reliable local-scale assessments.

This study addresses these scale and methodological mismatches in two principal ways. First, it adopts a point-scale evaluation framework that analyses individual grid-cell level data rather than spatial aggregates, thereby aligning the spatial resolution of model evaluation with the scale of observed extremes. By selecting two climatically and dynamically distinct grid cells, one in southern India (13° N) and the other in northern Italy (44° N), the study evaluates CMIP6 model performance across fundamentally different latitudinal zones, climate regimes, and precipitation-generating mechanisms. The southern India site lies in the tropical monsoon belt and is governed by the Southwest Monsoon, which delivers over 70% of the annual rainfall between June and September. Events are typically convective, episodic, and high in intensity, and are often associated with synoptic features such as monsoon depressions, low-pressure systems, and break spells. This region is characterised by the combined influence of deep tropical convection and orographic effects, and global models continue to struggle with these processes, largely due to their coarse resolution and reliance on parameterised convection.[8] The northern Italy site lies within a mid-latitude Mediterranean climate zone, where rainfall is modulated by synoptic-scale disturbances, extratropical cyclones, and frontal systems embedded within the westerlies.[9] Seasonal peaks in precipitation in the Mediterranean and adjacent regions often occur in autumn and spring and are substantially associated with frontal uplift and extratropical cyclones, especially cold front systems. Unlike the thermodynamically unstable tropics, precipitation here has historically been linked to more stable, long duration rainfall organised by planetary-scale dynamics, indicating a sharp contrast in both the origin and structure of extremes.

Second, the study employs a Peaks-Over-Threshold (POT) approach, fitting a Generalised Pareto Distribution (GPD) to daily rainfall values exceeding a high threshold. This allows the inclusion of multiple extreme events per year and offers a more data-efficient and tail-sensitive analysis of precipitation extremes than the block maxima framework. The study compares three GPD parameter estimation methods, Maximum Likelihood Estimation, L-moments, and Bayesian inference, each of which differs in its sensitivity to sample size, distributional shape, and inferential assumptions.[10] At its core, the study interrogates systematic model bias, specifically the persistent deviations between return levels estimated from 32 CMIP6-simulated precipitation (SSP5–8.5) and those derived from observational reference datasets (IMD4 and E-OBS). Rather than attempting a full decomposition of projection uncertainty, the analysis evaluates the extent to which statistical inference methods and model structure contribute to bias in the simulation of extremes. By assessing return level sensitivity across estimation techniques and identifying models that systematically overestimate or underestimate extremes at both sites, the study characterises the magnitude and spatial structure of model–observation biases. The goal is not merely to rank models, but to understand the conditions under which some

Indices for Precipitation." *Remote Sensing of Environment*, vol. 279, 2022, 113078.

[8] Villalba-Pradas, Anahí, and Francisco Tapiador. "Empirical Values and Assumptions in the Convection Schemes of Numerical Models." *Geoscientific Model Development*, vol. 15, 2024, pp. 3447-3474; Prein, Andreas, et al. "The Future of Weather Prediction is Convection Permitting." *Bulletin of the American Meteorological Society*, vol. 96, no. 2, 2015, pp. 171-173.

[9] Rüdisühli, Stefan, et al. "Attribution of Precipitation to Cyclones and Fronts Over Europe in a Kilometer-Scale Regional Climate Simulation." *Weather and Climate Dynamics*, vol. 1, no. 2, 2020, pp. 675-699.

[10] Scarrott, Carl, and Andrew MacDonald. "A Review of Extreme Value Threshold Estimation and Uncertainty Quantification." *REVSTAT—Statistical Journal*, vol. 10, no. 1, 2012, pp. 33-60; Lazoglou, Georgia, and Christina Anagnostopoulou. "Comparing Bayesian and Classical Approaches for the Generalized Pareto Distribution in Peak Over Threshold Modeling." *Theoretical and Applied Climatology*, vol. 130, no. 3-4, 2017, pp. 1073-1086.

models outperform others in representing the upper tail of the precipitation distribution. This diagnostic insight has clear implications for improving climate model physics, especially in the treatment of convection and orography, and for enhancing the utility of model-derived information in risk-sensitive decision-making. The study contributes one component toward the broader goal of developing a statistically principled framework for benchmarking climate models and adds to the growing body of work diagnosing the capabilities and limitations of CMIP6 models in simulating precipitation extremes.

The remainder of the paper is organised as follows. Section II provides background on CMIP6 model development and the physical and numerical processes that influence simulated precipitation extremes, including convection parameterisation, resolution and time stepping, orographic representation, cloud microphysics, vertical structure, and the role of large-scale teleconnections, as well as recent advances in convection-permitting models. Section III describes the study design, data sources, and preprocessing steps, and outlines the Extreme Value Analysis framework based on a peaks-over-threshold GPD approach. Section IV presents the results, including biases in GEV parameters, threshold selection and exceedances, return level estimates, and model ranking. Section V discusses the implications of these findings, focusing on method-related biases, model physics bias, the role of resolution, and key limitations. Section VI summarises the main conclusions and highlights priorities for future work.

II. **Model Development and Theoretical Background**

A. *Convection Parameterisation and Model Bias*

Simulating precipitation extremes is strongly constrained by the parameterisation of convection. CMIP6 models, which typically operate at horizontal resolutions of about 100 to 250 km, cannot explicitly resolve convective clouds and mesoscale convective systems and therefore rely on convection parameterisation schemes.[11] These schemes vary widely across models and introduce systematic differences in the frequency, intensity, and spatial structure of simulated extremes, most notably the tendency toward a 'drizzling bias' in which models produce too frequent light rain and underestimate heavy rainfall, especially in tropical regions.[12] Different convection schemes show distinct behaviours. Zhang–McFarlane and Relaxed Arakawa–Schubert schemes, for example, tend to overproduce light rain and underrepresent heavy events, contributing to drizzle bias.[13] Gregory–Rowntree and Tiedtke-type schemes have shown regional improvements, particularly in mid-latitude frontal regimes, while newer hybrid schemes such as those used in TaiESM1 display enhanced precipitation variability and extreme rainfall skill.[14] As a result, model skill in representing extremes is highly sensitive to the choice and tuning of the convection scheme.

B. *Resolution, Time Step, and Orography*

Model skill in simulating extremes also depends on horizontal resolution, time stepping, and orographic representation. Higher horizontal resolutions of about 50 to 100 km can better capture mesoscale organisation and orographic enhancement of

[11] Arakawa, Akio. "The Cumulus Parameterization Problem: Past, Present, and Future." *Journal of Climate*, vol. 17, no. 13, 2004, pp. 2493-2525.

[12] Stephens, Graeme, et al. "Dreary State of Precipitation in Global Models." *Journal of Geophysical Research: Atmospheres*, vol. 115, no. D24, 2010, D24211; Chen, Duoying, et al. "The Convective-To-Total Precipitation Ratio and Degrees of Freedom for Convection." *Journal of Geophysical Research: Atmospheres*, vol. 126, no. 5, 2021, e2020JD034198

[13] Zhang, Mingkui, and Ying Mu. "Understanding the Role of Convective Clouds in the Hydrological Cycle." *Journal of Geophysical Research: Atmospheres*, vol. 110, 2005, D11201; Lin, Y, et al. "Changes in Extreme Precipitation Events in CMIP6 Simulations." *Journal of Climate*, vol. 26, no. 23, 2013, pp. 9469-9482.

[14] Li, Gudongze, et al. "Convection-Permitting Simulation of Persistent Precipitation Events Impacting Eastern China." *Climate Dynamics*, vol. 29, no. 7-8, 2007, pp. 723-741; Walters, David, et al. "The Met Office Unified Model Global Atmosphere 7.0/7.1 and JULES Global Land 7.0 Configurations." *Geoscientific Model Development*, vol. 12, no. 5, 2019, pp. 1909-1963; Wang, Shaocheng, et al. "Metrics and Diagnostics for the Climate Variability Timescale Spectrum." *Journal of Climate*, vol. 34, no. 9, 2021, pp. 3467-3482.

rainfall, although improved physics remains essential.[15] Shorter model time steps help to represent the life cycle of convective systems more realistically and reduce numerical artefacts in heavy precipitation.[16] The representation of orography is particularly important in regions with strong terrain gradients, where coarse grids smooth topographic features and can suppress locally intense rainfall.[17] In point-scale evaluation studies, an additional source of uncertainty arises from the spatial mismatch between observation stations and model grid cells. Even when the nearest grid point is selected, spatial offsets of several tens of kilometres are common and can markedly affect the representation of localised extremes, especially in complex terrain.[18]

C. *Microphysics, Vertical Structure, and Teleconnections*

Cloud microphysics and vertical resolution exert a strong control on simulated precipitation extremes. Many global models still use relatively simple single-moment microphysics schemes that limit raindrop size-distribution variability and evaporation processes, which in turn leads to biases in both the intensity and frequency of extremes.[19] Higher vertical resolution, particularly in the lower troposphere, improves the representation of convective development and short-duration rainfall processes.[20] Model skill is also modulated by large-scale modes of variability such as the Madden–Julian Oscillation, El Niño–Southern Oscillation, and the Quasi-Biennial Oscillation, which influence rainfall variability and extremes across the tropics and subtropics.[21] Persistent limitations in these areas contribute to structured biases in simulated extremes and help explain why models often struggle to reproduce observed tail behaviour.

D. *Advances in Convection-Permitting Models*

Convection-permitting models with grid spacing of about 1 to 4 km explicitly resolve deep convection and have been shown to markedly improve the simulation of short-duration rainfall and local precipitation extremes.[22] However, their high computational cost still limits their application to relatively small domains and shorter simulation periods, and they are not yet standard in global ensembles such as CMIP6. Interim progress has therefore focused on refinements to parameterisation schemes and moderate increases in resolution, including HighResMIP experiments, which reduce some biases in extreme rainfall but do not eliminate them.[23] Overall, current evidence indicates that biases in simulated

[15] Bador, Margot, et al. "Impact of Higher Spatial Atmospheric Resolution on Precipitation Extremes Over Land in Global Climate Models." *Journal of Geophysical Research: Atmospheres*, vol. 125, no. 13, 2020, e2019JD032184.

[16] Barrett, Andrew, et al. "One Step at a Time: How Model Time Step Significantly Affects Convection-Permitting Simulations." *Journal of Advances in Modeling Earth Systems*, vol. 11, no. 3, 2019, pp. 641-658.

[17] Pierce, David, et al. "Future Increases in North American Extreme Precipitation in CMIP6 Downscaled with LOCA." *Journal of Hydrometeorology*, vol. 24, no. 5, 2023, pp. 951-975.

[18] Haerter, Jan, et al. "Comparing Atmospheric Indices of Extreme Precipitation Events Derived From Satellite and Rain Gauge Data." *Journal of Geophysical Research*, vol. 116, 2011, D11107.

[19] Abel, S, and I Boutle. "An Improved Representation of the Raindrop Size Distribution for Single-Moment Microphysics Schemes." *Quarterly Journal of the Royal Meteorological Society*, vol. 138, no. 669, 2012, pp. 2151-2162; Morrison, Hugh, and Jan Milbrandt. "A New Double-Moment Microphysics Parameterization for Application in Cloud and Climate Models. Part I: Description." *Journal of the Atmospheric Sciences*, vol. 72, no. 1, 2015, pp. 287-311.

[20] Jenney, A, et al. "Vertical Resolution Impacts Explicit Simulation of Deep Convection." *Journal of Advances in Modeling Earth Systems*, vol. 15, no. 3, 2023, e2022MS003444.

[21] Ahn, M-S, et al. "MJO Simulation in CMIP5 Climate Models: MJO Skill Metrics and Process-Oriented Diagnosis." *Climate Dynamics*, vol. 49, no. 11-12, 2017, pp. 4023-4045; Richter, Jadwiga, et al. "Progress in Simulating the Quasi-Biennial Oscillation in CMIP Models." *Journal of Geophysical Research: Atmospheres*, vol. 125, no. 8, 2020, e2019JD032362.

[22] Kendon, Elizabeth, et al. "Do Convection-Permitting Regional Climate Models Improve Projections of Future Precipitation Change?" *Bulletin of the American Meteorological Society*, vol. 98, no. 1, 2017, pp. 79-93; Fosser, Geert, et al. "Convection-Permitting Climate Simulations Over Europe: Recent Developments in EURO-CORDEX." *Bulletin of the American Meteorological Society*, vol. 101, no. 8, 2020, pp. E1301-E1315.

[23] Prein et al. (n 8) *supra*; Haarsma, R, et al. "High Resolution Model Intercomparison Project (HighResMIP

precipitation extremes stem mainly from convective parameterisation, orographic representation, microphysics, and structural model differences, underscoring the importance of diagnosing these sources when evaluating extremes in CMIP6 model.[24]

III. **Methodology**

This study evaluates the performance of 32 CMIP6 models in simulating extreme daily precipitation at two sites: one in the Western Ghats, India (13.3°N, 75.9°E), and another in Emilia–Romagna, Italy (44°N, 12°E), for the 2015–2024 hindcast period under the SSP5–8.5 scenario. The methodology employs a point-scale peaks-over-threshold approach, fitting a GPD to daily rainfall values exceeding a high threshold. Return levels for 2-, 5-, 10-, 20-, 50-, and 100-year periods are computed using three parameter estimation methods—maximum likelihood estimation, L-moments, and Bayesian inference—to assess methodological sensitivity. Model performance is evaluated through comparative analysis of GPD parameters, threshold selection, and return level estimates against observational benchmarks.

The analysis is structured around five key methodological components: (A) Study Design and Data Sources, (B) Data Extraction and Preprocessing, (C) Extreme Value Analysis, (D) Goodness-of-Fit and Diagnostics, and (E) Bias Evaluation.

A. *Study Design and Data Sources*

The study focuses on the 2015–2024 period as a hindcast period to evaluate how well models reproduce current precipitation extremes, which is essential for testing before applying them to future risk assessment.[25] Hindcasts also inform bias correction, which depends on models capturing realistic extreme behaviour,[26] and provide insight into model deficiencies. In addition, they help identify whether biases stem from the mean state, sub-daily processes, or variability of extremes;[27] for instance, O'Gorman and Schneider[28] linked CMIP6 biases to convective scheme limitations.

All model data in this study come from the high emissions Shared Socioeconomic Pathway 5–8.5 (SSP5-8.5) scenario runs over 2015–2024. The SSP5-8.5 scenario was chosen in part because it provides available high-forcing pathways across many models, and differences between emissions scenarios remain minimal in the near term up to about 2040.[29] In other words, during 2015–2024, the climate trajectories under various SSPs are very similar, so using SSP5-8.5 does not significantly distort the comparison, and biases in hindcast simulations are primarily attributable to model physics and structural differences rather than scenario divergence.

The best-available high-resolution gridded datasets for each region were used for observed precipitation. For South India, I used the latest version of the IMD4 daily gridded rainfall dataset from the Indian Meteorological Department (IMD), which serves as the benchmark observational

v1.0) for CMIP6." *Geoscientific Model Development*, vol. 9, no. 11, 2016, pp. 4185-4208; Roberts, M, et al. "The Benefits of Global High Resolution for Climate Simulation: Process Understanding and the Enabling of Stakeholder Decisions at the Regional Scale." *Bulletin of the American Meteorological Society*, vol. 99, no. 11, 2018, pp. 2341-2359; Wehner, Michael, et al. "Evaluation of Extreme Sub-Daily Precipitation in High-Resolution Climate Models." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, no. 2190, 2021, 20190544.

[24] Stephens (n 12) *supra*; O'Gorman, Paul. "Precipitation Extremes Under Climate Change." *Current Climate Change Reports*, vol. 1, no. 2, 2015, pp. 49-59; Bador et al. (n 15) *supra*.

[25] Kharin et al. (n 1) *supra*; Wehner, Michael, et al. "Changes in Extreme Precipitation and Hydrology: A Comparison of Dynamical and Statistical Downscaling Approaches." *Journal of Hydrometeorology*, vol. 21, no. 12, 2020, pp. 2899-2913.

[26] Risser, Mark, et al. "Quantifying the Role of Internal Climate Variability in Future Climate Change: The Importance of the Ramp Function." *Climatic Change*, vol. 156, no. 3-4, 2019, pp. 605-625.

[27] Prein et al. (n 8) *supra*.

[28] O'Gorman, Paul, and Tapio Schneider. "The Physical Basis for Increases in Precipitation Extremes in Simulations of 21st-Century Climate Change." *Proceedings of the National Academy of Sciences*, vol. 106, no. 33, 2009, pp. 14773-14777.

[29] IPCC (n 2) *supra*.

dataset. IMD4 is available over the entire spatial extent of India from 1901 to the present at a resolution of 0.25° × 0.25°.[30] The dataset is constructed using an Inverse Distance Weighted (IDW) interpolation scheme, applied to daily rainfall observations from 6,955 rain gauge stations distributed across India, after stringent quality checks and homogeneity assessments.[31] In southern India, the relatively high density of rain gauge stations improves the ability of IMD4 to represent daily rainfall patterns compared to earlier products.[32] The dense station network in southern India enhances the reliability of IMD4, particularly for regional-scale precipitation features, although interpolation in complex topography such as the Western Ghats remains challenging. Given that our location of interest lies on the lee side of the Ghats, the IMD4 product is expected to offer a better representation than coarser global products, though some residual uncertainty may persist due to sharp spatial rainfall gradients in such terrain.

For Northern Italy, I used E-OBS version 30.0e, a 0.1° gridded daily precipitation dataset based on quality-controlled station data from the European Climate Assessment & Dataset network, produced using thin-plate spline interpolation.[33] E-OBS has been widely used for climate model evaluation and extreme precipitation studies in Europe,[34] and offers high reliability in northern Italy given the dense gauge network, though with known limitations in high-mountain regions.[35] The use of different grid resolutions (0.25° vs 0.1°) for India and Italy reflects that in both cases the observation grid is finer than the typical resolution of the CMIP6 models (which is on the order of 1° or coarser), so the observations are unlikely to be a limiting factor. The higher resolution of E-OBS in Europe is possible due to a very dense station network,[36] whereas IMD4's 0.25° resolution is the highest available in India given the gauge density and interpolation methods.[37] Despite the resolution difference, each dataset is of sufficient grid resolution to capture extreme daily rainfall at the point scale of interest.

All analyses use daily precipitation totals (24-hour accumulated rainfall) for consistency. Model outputs (typically reported as precipitation flux) were converted to daily total rainfall (mm day^{-1}) to match the units of the observational datasets. The target model grid cell nearest to each location coordinate was extracted to represent the model's precipitation at that site.

I chose to evaluate model rainfall at single grid points, since many hydrological applications (e.g., flood risk modelling) and bias correction schemes rely on point-scale relationships between models and observations. Evaluating model skill at the grid-point level provides the most relevant diagnostic for practical use.[38]

Since the study period spans only a decade, I adopt a stationary framework for extreme value analysis, consistent with literature[39] that suggests non-stationarity typically emerges over longer timescales.

B. *Data Extraction and Preprocessing*

[30] Pai, D, et al. "Development of a New High Spatial Resolution (0.25° × 0.25°) Long-Period (1901-2010) Daily Gridded Rainfall Dataset Over India." *Journal of the Meteorological Society of Japan*, vol. 92, no. 1, 2014, pp. 39-65.
[31] ibid.
[32] ibid.
[33] Cornes, Richard, et al. "An Ensemble Version of the E-OBS Temperature and Precipitation Datasets." *Journal of Geophysical Research: Atmospheres*, vol. 123, no. 17, 2018, pp. 9391-9409.
[34] Cornes, Richard, et al. "E-OBS: High-Resolution Gridded Mean/Max/Min Temperature, Precipitation and Sea Level Pressure for Europe & Northern Africa." *Climate Data Guide*, UCAR, 2023.
[35] Prein, Andreas, and Andreas Gobiet. "Impacts of Uncertainties in European Gridded Precipitation Observations on Regional Climate Analysis." *International Journal of Climatology*, vol. 37, no. 1, 2017, pp. 305-327.
[36] Cornes et al. (n 33) *supra*.
[37] Pai et al. (n 30) *supra*.
[38] Teutschbein, Christof, and Jörg Seibert. "Bias Correction of Regional Climate Model Simulations for Hydrological Climate-Change Impact Studies: Review and Evaluation of Different Methods." *Journal of Hydrology*, vol. 456-457, 2012, pp. 12-29.
[39] Singh, Devendra, et al. "Changing Intensification Rate of Precipitation Extremes with Temperature." *Nature Climate Change*, vol. 14, no. 5, 2024, pp. 419-425.

Daily precipitation time series for 2015–2024 were extracted from each model and observation dataset for the two target coordinates. I used nearest-neighbour interpolation to extract daily precipitation values from the climate model grids at each target location. Nearest-neighbour interpolation was prioritised over bilinear interpolation to preserve the peaks of extreme rainfall events, as large model grid cells can dampen extremes through smoothing procedures.

Data were standardised for calendar consistency, units (all converted to mm day^{-1}), and temporal alignment across the entire 10-year period. Quality control checks identified missing values, which were minimal across all datasets. All time series were visually inspected to identify and flag potential data anomalies or discontinuities.

C. *Extreme Value Analysis*

Extreme Value Analysis (EVA) was conducted using a POT approach, fitting a GPD to daily rainfall values exceeding a high threshold. The POT method allows for the inclusion of multiple extreme events per year, offering a more data-efficient and tail-sensitive analysis of precipitation extremes compared to the block maxima framework.

1. Threshold Selection

A fixed percentile threshold at the 99th percentile was applied consistently across all models and observations to ensure comparability. This threshold was validated using Mean Residual Life (MRL) plots, which provide a graphical diagnostic for assessing threshold stability. MRL plots display the mean excess of values above a threshold as a function of the threshold itself; for a valid threshold, this plot should be approximately linear.[40] Thresholds were considered appropriate if the linearity assumption held within a visually acceptable range.

2. Parameter Estimation

Three parameter estimation methods were applied to each dataset to assess methodological sensitivity:

- MLE: Standard likelihood-based approach, numerically optimised using gradient ascent methods. MLE is known for its asymptotic efficiency and is widely used in hydrological applications.
- L-moments: A robust, distribution-free method that uses linear combinations of order statistics. L-moments are particularly useful for small sample sizes and are less sensitive to outliers.[41]
- Bayesian Inference: A fully probabilistic approach that incorporates prior information about parameter values. Bayesian methods were implemented using Markov Chain Monte Carlo (MCMC) sampling with weakly informative priors. This approach provides full posterior distributions and is particularly valuable when sample sizes are limited or when parameter uncertainty needs to be quantified.[42]

Each method estimates the scale (σ) and shape (ξ) parameters of the GPD. These parameters define the tail behaviour: the shape parameter ξ controls whether the tail is bounded (ξ < 0), exponential (ξ ≈ 0), or unbounded (ξ > 0).

3. Return Level Computation

The return level z_T associated is the rainfall magnitude expected to be exceeded on average once every T years given by

$$z_T = \mu + \sigma.ln(T.\lambda) \text{ for } \xi = 0.$$

Using the fitted GPD parameters (σ – scale parameter) along with the threshold (u) and the rate of threshold exceedances (λ), return levels was calculated for return periods of 2, 5, 10, 20, 50, and 100 years for each dataset to provide a comprehensive view of extreme behaviour across multiple timescales.

[40] Coles (n 6) *supra*.

[41] Hosking, J. "L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics." *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 52, no. 1, 1990, pp. 105-124.

[42] Coles, Stuart, and Jonathan Tawn. "Modelling Extreme Multivariate Events." *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 2, 1996, pp. 377-392.

D. *Goodness-of-Fit and Diagnostics*

Model fits were evaluated using multiple diagnostic plots:

- Quantile–Quantile QQ plots: Compare quantiles of the fitted GPD distribution against empirical quantiles of the data. Points lying along the 1:1 reference line indicate a good fit.
- Probability–Probability PP plots: Compare cumulative probabilities from the fitted distribution against empirical probabilities. Similar to QQ plots, deviations from the 1:1 line indicate poor fit, particularly in the upper tail.
- Mean Residual Life MRL plots: Validate the linearity assumption required for threshold selection, as described above.

These diagnostics were generated for all models and estimation methods to identify cases where the GPD assumption may be violated or where parameter estimates were unstable.

E. *Bias Evaluation*

Model bias was quantified as the average percentage deviation of modelled return levels from observed values using the Percentage Bias (PBIAS) metric:

$$\text{PBIAS}_{avg} = \frac{100}{N} \sum_{i=1}^{N} \left(\frac{RL_i^{mod} - RL_i^{obs}}{RL_i^{obs}}\right)$$

where RL_i^{mod} is the model's estimated return level at return period I, RL_i^{obs} is the observed return level at return period I and N is the number of return levels. Positive PBIAS indicates overestimation and negative values indicate underestimation.

Model performance was interpreted as follows: PBIAS within ±10% was considered good agreement with observations, ±20% as moderate agreement, and biases larger than ±50% as indicative of poor model performance. These thresholds provide benchmarks for identifying models with systematic biases that might limit their utility for decision-making.

The comparative PBIAS across models, grouped by convection scheme and resolution, was used to identify systematic biases and assess model skill in simulating precipitation extremes. Statistical differences between models and estimation methods were evaluated visually through ranking plots and ensemble statistics.

IV. **Results**

A. *GEV Scale and Shape Parameter Biases*

I first compare the fitted GEV scale and shape parameters from each model against the observational benchmarks at the India (IMD4) and Italy (E-OBS) sites (**Figure 1**, below). Across CMIP6 models, MLE and Bayesian estimation produce very similar scale and shape values (even the sign) for a given model, whereas L-moments often diverge substantially. In the Italian (E-OBS) case, for example, as seen in Figure 1(d), most models underestimate the observed scale (≈7.8) derived using L-moments, on the other hand, MLE and Bayesian estimates show a wider spread and generally cluster around their respective, higher observational benchmarks. Only 2 of 15 models at 100 km and 2 of 17 at 250 km resolution are within ~10% of the observed scale in Italy. Higher horizontal resolution offers some improvement for India, as 3 of 15 models at 100 km (and none for 250 km) resolution meet this threshold.

Nevertheless, even at fine resolution, fewer than 25% of models achieve this level of agreement in either region. Only ACCESSCM2 estimates the scale parameter within ~10% of the observed value across all methods for Italy, while no model does so for India. 25% of models overestimate the scale in Italy and 50% do so for India; so the pattern is regionally consistent. It indicates that models tend to under-predict extreme rainfall magnitudes in Italy, but more frequently overestimate them in India.

The GEV shape parameters show analogous trends. For Italy, the observed shape is near zero, and most models (MLE / Bayesian) likewise estimate a small magnitude (positive or negative) shape, with L-moments again

producing a broader range. In Figure 1(c), it is observed that the E-OBS dataset itself yields almost identical shape estimates under MLE and Bayesian fits, indicating that this observational reference is immune to the fitting method. In contrast, at the India site, the observed shape is significantly negative (a bounded tail), and many models (MLE / Bayesian) capture a negative or near-zero shape, but L-moments fails to do so. It returns only small positive shapes for all models and even for IMD4 itself as seen in Figure 1(a).

About one-third of models simulate a negative shape for India, but L-moments produces non-negative shapes across the board. In summary, MLE and Bayesian approaches yield consistent shape estimates (agreeing in sign and magnitude) for each model, whereas L-moments cannot reproduce strongly negative shapes. Only 3–5 models per region match the observed shape within ~10% (e.g., KACE-1-0-G for India, CMCC-ESM2 for Italy) under the Bayesian method.

B. *Extreme Threshold Selection and Exceedances*

The selected thresholds matched well with the inferences from Mean Residual Life (MRL) plots in 94% of models for both regions, confirming the stability of the threshold selection across methods and models. The models' baseline extreme thresholds (e.g. the threshold fora one-in-one-year event or POT threshold) generally differ from observations. As shown in **Figure 2**, below, in the E-OBS analysis (Italy point), the observed extreme rainfall threshold is about 30.7 mm per day. Most models underestimate this value, i.e., 24 out of 32 models have lower thresholds. For example, HadGEM3-GC31-MM matches the observed threshold closely, but EC-Earth3-Veg-LR underestimates it by ≈19% despite a grid point only ≈16 km from the observation point. Geographic proximity influences threshold performance, but proximity alone does not guarantee accuracy. In both regions, several of the better-performing models, such as CMCC-ESM2 and KACE-1-0-G, sample grid points within ~50 km of the gauge.

In contrast to the Italy point, for the Indian point (observed threshold ≈33.4 mm per day), the bias reverses in that 19 of 32 models set higher thresholds than observed, and 13 are lower. KACE-1-0-G and CMCC-CM2-SR5 lie close to the observed value and HadGEM3-GC31-MM underestimates the threshold by ≈52%. Thus, E-OBS thresholds are predominantly underestimated and IMD4 thresholds are predominantly overestimated. 100 km models exhibit a slightly lower mean threshold error (8.82 mm per day) than 250 km models (9.08 mm per day) in Italy, showing a modest improvement with resolution. As shown in Figure 2, across both benchmarks, the great majority of models do not match the observed threshold. These threshold biases are carried forward into the return level estimates.

C. *Return Levels and Estimation Method Sensitivity*

Using each model's GEV fit (with the model-specific threshold), the 100-year return level (RL100) was computed. I compare the three fitting methods and also aggregate model-to model spread. Despite differences in fitted GEV parameters, the 100-year return levels (RL100) predicted by the three methods are generally of the same order for a given model, typically within 10–20% variation among MLE, Bayesian, and L-moments estimates. The average inter-method spread is ≈11–16%. In many cases, MLE and Bayesian estimates are nearly identical, while L-moments sometimes gives lower values, especially for heavy-tailed models. As shown in Figure 3, the three method curves for most models sit close together.

In worst-case outliers, ACCESSCM2 shows a method-driven RL100 variation of ≈33%. KACE-1-0-G is one of the few models performing well in both regions (E-OBS and IMD4), matching both the observed threshold and RL100 within ~10% in each case. However, method differences remain small compared to inter-model spread, which spans several hundred per cent. The ensemble mean reflects systematic biases; in E-OBS, RL100 is underestimated (~77 mm vs 100 mm

observed), and in IMD4, it is overestimated (≈145 mm vs 86 mm observed). As seen in Figure 3, these biases persist across return periods, with L-moments yielding a narrower quantile range, especially in E-OBS.

Certain models also demonstrated strong return level performance despite imperfect GPD parameter estimates. For example, HadGEM3-GC31-MM, despite having a large shape parameter bias (–121%) in E-OBS, reproduced the 100-year return level within 7.8%, likely due to accurate threshold estimation (+0.6%). Although the threshold for extremes is slightly lower in E-OBS (≈30.7 mm vs 33.4 mm for IMD4), the observed extremes in Italy are much larger in absolute terms. The maximum daily rainfall recorded in E-OBS during 2015–2024 reached 141 mm, well above the observed RL100 of ≈100 mm, and IMD4 recorded a maximum of only 78.7 mm, just slightly above its RL100 of ≈77.6 mm. Moreover, E-OBS exceeded its RL100 on two occasions in the past decade, and IMD4 exceeded its RL100 once.

D. *Model Bias and Ranking*

Finally, the models are ranked by their performance in simulating extreme precipitation. Model performance in simulating extreme precipitation return levels showed a clear dependence on the convection scheme (**Figure 4**, below). Zhang–McFarlane models (e.g. CESM2, NorESM2) widely underestimated extremes for Italy; for example, CESM2's 100-year return showed –54% bias. Similarly, NorESM2 and other ZM models showed strong underestimation at high return periods in Europe. However, only CESM2 performed well over India, with biases within ~10% across estimation methods, and other ZM models like GFDL-ESM4 and NorESM2-MM (top performer for India) overestimated Indian extremes by over 100% and 17% respectively.

For Gregory–Rowntree models, HadGEM3-GC31-MM performs reasonably well in Italy with modest RL100 underestimation (–8% to –20%), and UKESM1-0-LL overestimates by ≈30%. However, most Gregory–Rowntree models perform poorly in India, with HadGEM3-GC31-MM and ACCESSCM2 significantly underestimating and UKESM1-0-LL overestimating by up to +100%. Tiedtke-type models (EC-Earth family) were intermediate in that they generally captured large-scale features but still exhibited bias. For example, CNRM-CM6-1-HR's 100-year return level (~107 mm) was very close (+7%) to the observed 99.7mm in E-OBS, whereas other Tiedtke models (e.g. EC-Earth3) tended to under-predict the upper tail. Bulk adjustment schemes (e.g. IPSL-CM6A-LR, BCC-CSM2-MR, FGOALS-g3) showed larger biases.

In E-OBS, several bulk-scheme models had biases well outside ±20%. IPSL-CM6A-LR is a partial exception, delivering moderately accurate estimates in E-OBS (+24%) but severely overestimating extremes in IMD4 (+400%).

Across both regions, models using the Gregory–Rowntree and Tiedtke-type convection schemes generally produced lower biases, particularly in E-OBS. Zhang–McFarlane models showed strong underestimation in Europe but performed better in India. Bulk-adjustment schemes exhibited the highest biases overall, with large deviations observed in both domains. The performance varied across methods and no convection scheme delivered low-bias estimates in both regions. Importantly, KACE-1-0-G, a Gregory-Rowntree model, performed exceptionally well in both regions, ranking in the top 3 across all estimation methods.

Finally, model complexity (GCM vs ESM) was not a reliable predictor of performance. Both high-performing and poorly performing models were found among GCMs (e.g. HadGEM3-GC31-MM vs CESM2) and ESMs (e.g. UKESM1-0-LL vs NorESM2-MM). Instead, convection scheme design emerged as the dominant factor explaining model performance in simulating extreme precipitation return levels.

V. **Discussion and Limitations**

A. *Extreme Value Analysis Method Bias*

This study demonstrates that the methods used in EVA, especially in threshold selection and parameter estimation, exert an influence on return level estimates. E-OBS provides a firmer reference point when assessing model bias and return levels since it produces consistent results of GPD parameters across estimation methods, which could be attributed to a relatively dense underlying rain gauge network and high spatial resolution.[43] In contrast, the greater method sensitivity observed for IMD4 introduces additional uncertainty into the evaluation, particularly for high return periods. This highlights the importance of carefully considering the quality of the observational benchmark when interpreting bias magnitudes because an apparently large model bias against IMD4 may partly reflect uncertainties in the reference itself, whereas biases against E-OBS are likely more robust. POT offers more data points for fitting, improving parameter stability,[44] yet it is equally clear that the method is only as reliable as the data to which it is applied.[45] Any biases or limitations in the observational datasets inevitably propagate into fitted return levels, influencing not just absolute bias estimates but also the relative ranking of model performance.

Perhaps more importantly, threshold selection emerged as the dominant driver of bias in return level estimation. As noted by Coles,[46] the chosen threshold effectively defines what is deemed 'extreme'; and if poorly selected, it will skew the tail fit. In our models, it is found that even where shape and scale parameters were well aligned with observations, a bias in threshold led to substantial errors in return levels. This is particularly pronounced in convective regions like India, where coarse models tend to pre-trigger convection and hence artificially inflate the threshold (the 'drizzle bias' discussed by Stephens et al.[47]). On the other hand, in Italy's more stratiform regime, some models underestimated the threshold due to different triggering mechanisms. In both settings, the threshold effect dominated other sources of bias. This finding validates the literature's concern[48] that threshold choice remains a key uncertainty in POT approaches.

Estimator sensitivity is another issue highlighted by our results, especially when the shape parameter (ξ) approaches zero. Theoretically, as $\xi \to 0$, the GPD tail approaches an exponential form, where small changes in ξ have outsized impacts on long return period estimates.[49] In models where ξ hovered near zero, such as FGOALS-g3 over the India point ($\xi \approx -0.016$), the estimated 100-year return level (RL100) varied by up to 30mm depending on the estimator used. As Lazoglou and Anagnostopoulou[50] and de Zea Bermudez and Kotz[51] observe, likelihood-based methods such as MLE and Bayesian inference are generally more stable in this regime as opposed to L-moments which can become unreliable and fail to produce negative shape values. In my analysis, Bayesian estimates tended to be most reliable across all cases, especially where ξ was small or where sample sizes were limited. The increasing divergence between estimation methods at long return periods[52] exposes the sensitivity of tail estimates to parameter uncertainty, which is useful for diagnosing unstable model behaviour and emphasising that method

[43] Haylock, Malcolm, et al. "A European Daily High-Resolution Gridded Data Set of Surface Temperature and Precipitation for 1500-2100." *Journal of Geophysical Research*, vol. 113, D20119, 2008; Cornes (n 34) *supra*.
[44] Coles (n 6) supra; Scarrott & MacDonald (n 10) *supra*.
[45] Coles (n 6) *supra*; Beguería, Santiago, and Sergio Vicente-Serrano. "Mapping the Hazard of Extreme Rainfall by Peaks over Threshold Extreme Value Analysis and Spatial Regression Techniques." *Journal of Applied Climatology*, vol. 45, no. 1, 2006, pp. 108-124.
[46] Coles (n 6) *supra*.
[47] Stephens (n 13) *supra*.
[48] Scarrott & MacDonald (n 10) *supra*; Roth, M, et al. "Threshold Selection for Regional Peaks-Over-Threshold Data." *Journal of Applied Statistics*, vol. 43, no. 7, 2016, pp. 1291-1309.
[49] Katz, Richard. "Extremes of Daily Precipitation in Global Climate Models." *Journal of Climate*, vol. 12, no. 7, 1999, pp. 2067-2078.
[50] Lazoglou & Anagstopoulou (n 10) *supra*.
[51] de Zea Bermudez, Paolo, and Samuel Kotz. *Extreme Value Distributions: Theory and Applications*. Imperial College Press, 2010.
[52] Katz (n 49) *supra*.

choice is critical when extrapolating extremes for decision-making.

It should be noted, however, that even though methodological choices clearly affect return levels, the spread between models is substantially larger than the spread between estimation methods. This suggests that the dominant uncertainties in extreme precipitation projections arise from model physics, not statistical estimation per se. Nonetheless, good practice in EVA remains essential. Poor threshold selection or inappropriate estimators can easily confound results, as has been warned in numerous prior studies.[53] Since threshold selection often relies on judgment (e.g., 95th or 99th percentile), there is a need for more principled methods that reduce arbitrariness and promote consistency across studies.

B. *Model Physics Bias*

It is observed that model biases in extreme precipitation align strongly with convection scheme families. These biases remain consistent across estimation methods and reinforce the importance of parameterisation in shaping return-level behaviour.

Zhang–McFarlane (ZM) models such as CESM2, NorESM2-LM, and NorESM2-MM perform well in India, with RL100 biases generally within ±30%, consistent with their capacity to simulate deep convection during the monsoon. However, all three substantially underestimate extremes in Italy, with RL100 values falling short by 50–65%. This behaviour reflects known weaknesses in ZM schemes under midlatitude conditions, where premature convection stabilises the atmosphere and suppresses extreme rainfall. These findings support earlier results by Wang et al.[54] and Zhang and Mu[55], which highlight the latitude sensitivity of the ZM family.

Models using the Gregory–Rowntree (GR) scheme show more mixed results. HadGEM3-GC31-MM matches observations fairly well in E-OBS, with RL100 bias under 20%, but underestimates Indian extremes by ≈30%. UKESM1-0-LL performs less reliably, overestimating RL100 by ≈30% in Italy and by more than 80% in India. The GR approach, which delays deep convection through convergence-based triggers, captures European midlatitude extremes more effectively than tropical ones. Nonetheless, performance varies significantly within the family, and no GR model shows strong agreement across both regions.

Mass-flux lineage schemes exhibit a broader spread. CNRM-CM6-1-HR delivers the most accurate result in E-OBS, with RL100 bias under 3% and threshold error below 8 mm. In contrast, its performance deteriorates over India, where it overestimates the 100-year return level by more than 50%. CNRM-ESM2-1 shows moderate underestimation in Italy but strongly overestimates extremes in India, with a bias exceeding 65%. EC-Earth models, also part of the Tiedtke family, show mixed performance. EC-Earth3-CC performs reasonably in both domains (bias within ±20%), while EC-Earth3-Veg-LR overestimates RL100 in India by ≈23%. These inconsistencies likely arise from differences in tuning, such as entrainment rates or trigger thresholds.

Models that retain relatively simple single-moment microphysics and mass-flux convection (e.g., FGOALS-g3, BCC-CSM2-MR, IPSL-CM6A-LR) perform poorly across both regions. FGOALS-g3 underestimates RL100 by 40% in Italy and overshoots it by more than 100% in India. BCC-CSM2-MR shows extreme overestimation in India (bias over 400%) and severe underestimation in Europe. IPSLCM6A-LR, although slightly more accurate in E-OBS (bias ≈+20%), ranks among the worst

[53] Beguería, Santiago, et al. "Mapping the Hazard of Extreme Rainfall by Peaks over Threshold Extreme Value Analysis and Spatial Regression Techniques." *Journal of Applied Statistics*, vol. 43, no. 7, 2005, pp. 1291-1309.

[54] Wang, Yun, et al. "Latitudinal Differences in the Response of Temperature Extremes to Anthropogenic Forcing." *Nature Climate Change*, vol. 12, no. 5, 2022, pp. 406-412

[55] Zhang & Mu (n 13) *supra*.

performers in IMD4, with RL100 bias exceeding 300%. These models lack explicit cloud microphysics and rely on simplistic moisture-to-rain conversion rules, which leads to sharp threshold errors and unstable tail behaviour.

IPSL-CM6A-LR, though somewhat more advanced, still exhibits bulk-like behaviour. Its good performance in Italy may partly reflect both chance and the model's coarse resolution, which aligns better with the large-scale orographic and synoptic extremes (cyclones) typical of the region, but it fails in India, where capturing small-scale, intense convective bursts is critical. KACE-1-0-G demonstrates consistently strong performance across both regions, despite using a conventional Gregory–Rowntree model. Its effective tuning appears to support realistic tail behaviour in both tropical and mid-latitude regimes, which underlines the importance of targeted calibration, even within standard parameterisation frameworks. The ensemble mean is biased relative to observations but displays consistent and spatially coherent errors, making it amenable to statistical bias correction.

Overall, the structured model biases observed are fully consistent with prior literature. As Rio et al. and Emmenegger et al.[56] argued, model physics, particularly the handling of deep convection, remains the primary driver of biases in simulated precipitation extremes. My two-region comparison added the idea that these biases are systematic and region-dependent. The same model can perform well in one region (Italy) but poorly in another (India), highlighting the challenges of achieving universal skill across different climate regimes.[57] It is also worth noting the role of latitude and large-scale dynamics. Tropical extremes are dominated by deep convection, modulated by phenomena such as the Madden–Julian Oscillation (MJO) and equatorial wave[58] in contrast with midlatitude extremes that depend more on frontal systems and orographic forcing. Unsurprisingly, no single convection scheme performed equally well across both latitudes, which echoes the findings of Norris et al.[59] and Hohenegger and Bretherton.[60] Hybrid schemes, such as the one used in TaiESM1, may offer a more flexible approach, though this remains an area for future model development.

Convection-permitting models offer a promising path forward, but their high computational costs mean that for the foreseeable future, progress in CMIP-class models will depend on more sophisticated parameterisations and targeted improvements in microphysics and dynamics. Investment in these areas is likely to yield meaningful gains in extreme precipitation projections, with clear benefits for downstream risk applications.

C. *Resolution*

In this analysis, an interesting latitude-dependent pattern emerges where finer grid models tend to perform better for IMD4 (India) and some coarser-grid models still perform well for E-OBS (Italy). Finer resolution is important, particularly for convection-dominated regions, but it is not

[56] Rio, Catherine, et al. "Ongoing Breakthroughs in Convective Parameterization." *Current Climate Change Reports*, vol. 5, no. 2, 2019, pp. 95-111; Emmenegger, Laurent, et al. "Observational Constraints on Aerosol-Cloud-Convection Interactions." *Nature Reviews Earth & Environment*, vol. 4, no. 7, 2023, pp. 440-455.

[57] Zhang, Xuebin, et al. "Understanding the Causes of Climate Change Impacts on Precipitation Extremes in the Future." *Bulletin of the American Meteorological Society*, vol. 106, no. 3, 2025, pp. 423-437; Iturbide, Maialen, et al. "An Update of IPCC Climate Reference Regions for Subcontinental Analysis of Climate Model Data: Definitions and Aggregated Datasets." *Earth System Science Data*, vol. 12, no. 4, 2020, pp. 2959-2970.

[58] Mayta, V, et al. "Modulation of Intraseasonal Convective Activity by the MJO over the Central Andes." *Climate Dynamics*, vol. 62, no. 5, 2024, pp. 3475-3492.

[59] Norris, Jesse, et al. "Evaluation of the Tail of the Probability Distribution of Daily and Subdaily Precipitation in CMIP6 Models." *Journal of Climate*, vol. 34, no. 7, 2021, pp. 2657-2678

[60] Hohenegger, Cathy, and Christopher Bretherton. "Microphysics and Turbulence in Simulations of Deeper Convection." *Journal of the Atmospheric Sciences*, vol. 68, no. 8, 2011, pp. 1854-1873.

sufficient by itself, as noted by Prein et al.[61] Models like MRI-ESM2-0 demonstrate that even at finer scales, biases can persist if the convection scheme or time-stepping poorly represents storm evolution. Similarly, orographic smoothing and point–grid mismatch continue to introduce structural uncertainties.[62] Overall, these results align with Chen et al. and Bador et al.[63] in showing that increasing resolution improves the potential for realistic extremes, but model physics and temporal precision remain equally essential. Future model development could involve hybrid approaches that combine advanced physics with regionally adaptive resolution, for example, using variable-resolution grids or embedding convection-permitting nests in key regions. Such strategies may improve the global simulation of precipitation extremes without prohibitive computational costs.

Since this study evaluates extremes at the point scale, some bias arising from the distance between the model grid point and the observation site is unavoidable, particularly in regions with complex terrain or sparse station networks. Even when selecting the nearest grid point, spatial mismatches of several tens of kilometres can occur, which can influence the representation of localised extremes.

A further challenge for CMIP6 models is capturing internal climate variability, which modulates extreme rainfall on interannual to decadal timescales. At the Italy point, two events exceeding the 100-year return level occurred in the past decade, which were attributed to internal variability rather than climate change.[64] Most models struggled to reproduce this behaviour;[65] only KACE-1-0-G (for both regions) and HadGEM3-GC31-MM (for Italy) produced return level structures consistent with observed event frequency. CMIP6 models can represent mean trends in extremes but lack the ability to simulate the clustering and timing of rare, variability-driven events.

D. *Implications for Climate Model Users and Climate-Sensitive Decision-Making*

The magnitude of inter-model spread in return level estimates—spanning several hundred percent in extreme cases—raises important questions about the credibility and utility of CMIP6 ensemble predictions for decision-makers. Whilst multi-model ensembles are often presented to stakeholders as a means of quantifying uncertainty, the spread documented here reflects primarily differences in model physics rather than a principled sampling of genuine uncertainty. Citizens, water utilities, and civil protection agencies who rely on ensemble summaries (ensemble means or percentile ranges) may be misled about the true range of plausible future extremes if the ensemble is composed of models with widely divergent convection schemes and calibration choices. If biases are systematic and scheme-dependent, then claims of consensus or high confidence in ensemble predictions become contingent on the composition of the ensemble.

The literature on climate change adaptation economics indicates that uncertainty in precipitation extremes is a primary driver of infrastructure planning risk.[66] World Bank estimates place the cost of water sector adaptation globally at approximately USD 20 billion per year by 2050, and failures in climate-informed infrastructure planning can result in losses ranging from 5% to 60% in hydropower revenue alone, depending on the basin and the accuracy of precipitation projections.[67] When model-derived

[61] Prein et al. (n 8) *supra*.
[62] Barrett et al. (n 16) *supra*; Pierce et al. (n 17) *supra*.
[63] Chen, Xiuhong, et al. "Changes in Annual Extremes of Daily Temperature and Precipitation in CMIP6 Models." *Journal of Climate*, vol. 38, no. 9, 2025, pp. 3441-3460; Bador et al. (n 15) *supra*.
[64] Barnes, Clair, et al. "Limited Net Role for Climate Change in Heavy Spring Rainfall in Emilia-Romagna." Imperial College London Working Report, 2023.
[65] Wu, Yuanxin, et al. "Understanding the Contribution of Extreme Heat to the June 2021 Pacific Northwest Heat Wave." *Journal of Geophysical Research: Atmospheres*, vol. 127, no. 15, 2022, e2022JD036873.
[66] Hughes, Gordon, et al. "The Costs of Adapting to Climate Change for Infrastructure." World Bank Development and Climate Change Discussion Paper 2, 2010.
[67] ibid.

precipitation projections are significantly biased, the risk calculus underlying infrastructure investment decisions becomes distorted. Policymakers and water utilities may either fail to invest adequately in resilience measures (if models underestimate future extremes) or invest excessively in over-designed infrastructure (if models overestimate extremes), both outcomes representing economic inefficiency and misallocation of resources in constrained adaptation budgets.

Furthermore, the finding that threshold selection in extreme value analysis exerts greater control over return-level bias than variations in parameter estimation methods suggests that uncertainty in return-level projections is as much a function of methodological choices made downstream of the model output as it is a reflection of model physics.

From a decision-support perspective, these findings suggest that ensemble-based projections will be more useful to practitioners if they are stratified by convection scheme family and if scheme-specific biases are communicated alongside ensemble statistics. Rather than presenting a single multi-model mean return level to an urban planner or water utility engineer, it may be more honest and more useful to present: (i) the range of scheme-specific biases documented from evaluation against observations; (ii) the envelope of model projections annotated with scheme family and bias magnitudes; and (iii) explicit recommendations about which models are most appropriate for specific regions and applications. Such an approach would improve the scientific integrity of climate-informed adaptation planning by ensuring that decision-makers have clear insight into the physical drivers of model divergence and the likely accuracy of projections in their region of interest.

E. *Limitations*

I did not formally quantify the statistical uncertainty of my return-level estimates (e.g. via confidence or credible intervals). This was a deliberate choice to concentrate on bias structure, but it means uncertainties for any single model's 100-year event are not reported. Had I included bootstrapped confidence intervals or full Bayesian posteriors, it would have been more complete uncertainty information, but the central conclusions (regarding relative biases) would likely remain similar. In effect, by focusing on relative differences among models and methods, I have somewhat sidestepped the sampling noise issue in that the biases that persist across methods are unlikely to be purely due to random variability. Future work should incorporate formal interval estimation to check whether the observed bias magnitudes are statistically significant.

Another limitation of this study is its relatively narrow spatial and variable scope. The analysis focuses on only two locations in the tropics and mid-latitudes, without testing model skill in other important climate regimes, such as high latitudes or in the mountains, where different physical processes dominate the generation of extremes. As a result, the findings may not capture the full diversity of regional climate behaviours or model responses. Furthermore, this study evaluates precipitation extremes only using daily precipitation data, whereas other studies have used dynamical fields such as 500 hPa geopotential height, mean sea-level pressure, or wind anomalies to investigate the atmospheric conditions associated with extreme events. Including such diagnostics could offer a less siloed assessment of model skill and help attribute observed biases to circulation features or synoptic-scale drivers.

VI. **Conclusion**

This study was designed with three primary objectives: to quantify systematic biases in extreme daily precipitation simulated by 32 CMIP6 models at the point scale; to identify the dominant sources of these biases, distinguishing between contributions from statistical methodology versus model physics; and to assess the extent to which model characteristics, particularly convection

parameterisation, influence the fidelity of simulated extreme precipitation.

To address these objectives, a POT framework was employed, fitting a GPD to daily rainfall values exceeding the 99th percentile threshold. Three parameter estimation methods (Maximum Likelihood Estimation, L-moments, and Bayesian inference) were applied to quantify estimator sensitivity. Return levels for return periods ranging from 2 to 100 years were computed, and model performance was evaluated against observational benchmarks from IMD4 (India) and E-OBS (Europe) using the PBIAS metric.

The analysis revealed several key findings. First, CMIP6 models exhibit systematic and geographically divergent biases as they tend to overestimate extreme precipitation in the tropical Indian monsoon domain and underestimate it in the mid-latitude Mediterranean climate of northern Italy. Second, these biases are strongly organised by convection scheme family, with Zhang–McFarlane models performing reasonably well in monsoon regimes but substantially underestimating mid-latitude extremes, whilst Gregory–Rowntree and Tiedtke-type schemes show contrasting regional strengths and weaknesses. Third, threshold selection emerges as a critical source of uncertainty in return level estimation, with considerable variability observed across models driven by differences in convection parameterisation and drizzle bias. This variability underscores the necessity for more objective and standardised threshold selection procedures; subjective or ad hoc threshold choices introduce systematic biases that can substantially exceed variations attributable to fitted shape and scale parameters, highlighting the imperative for principled, reproducible methodologies in extreme value analysis. Finally, whilst differences among the three estimation methods are non-negligible, they remain modest relative to inter-model spread, with Bayesian inference demonstrating the most robust behaviour in edge cases where the shape parameter approaches zero.

This study contributes to the field by establishing a tailored, point-scale diagnostic framework that explicitly links CMIP6 tail behaviour to underlying model physics. By demonstrating that convection parameterisation is the primary driver of return-level bias more so than model resolution or choice of extreme value estimator the work underscores the value of physics-oriented model development over computational refinement alone. Moreover, by highlighting the dominant role of threshold selection, the analysis provides practical guidance for practitioners conducting extreme value analyses on climate model output.

On the basis of these results, the following recommendations are offered. First, efforts to improve CMIP6 skill in extreme precipitation should prioritise refinement of convection parameterisation schemes and associated microphysics and moisture treatments, particularly in convection-dominated regimes. Second, systematic and physically grounded tuning of model parameters which is evidenced by the exceptionally strong performance of KACE-1-0-G across both regions, offers a promising pathway to reducing tail biases. Third, extreme value analyses intended for impact-relevant applications should document and justify choices regarding threshold selection, percentile definition, and estimator selection, given the demonstrated sensitivity of return levels to these methodological decisions.

Future investigations could extend this diagnostic approach in several directions. Application to additional climate regions and to multi-site or spatially aggregated configurations would clarify the generality of the identified bias patterns. Integration of convection-permitting models or variable-resolution simulations, where computationally feasible, would enable systematic assessment of whether explicit convection reduces tail biases relative to parameterised approaches. Finally, formal attribution of model biases to specific parameter or process choices would support

targeted model development aimed at enhancing the utility of precipitation extremes projections for adaptation and risk-management applications.

Acknowledgements

This article is derived from my MPhil thesis at the University of Cambridge, conducted under the supervision of Dr Edoardo Borgomeo and in partnership with Eoliann SRL, Italy. I extend my sincere gratitude to Dr Borgomeo for his invaluable guidance and support, and to Robin Castellani of Eoliann for his help with data analysis.



Figure 1: Comparison of GEV Shape Parameter Estimates for Each Model under Three Methods (MLE, L-moments, Bayesian) at 100km (Left) and 250km (Right) Nominal Resolutions. *Notes: Dashed lines indicate the shape and scale value from observation data (IMD4 – Fig 1(a), (b), E-OBS – Fig 1(c), (d)) for each method. Models are arranged in increasing distance from the model data point to observation data point. MLE and Bayesian generally produce similar shape values (often of the same sign), while L-moments shows greater divergence and fails to capture negative shapes, tending to stay positive.*

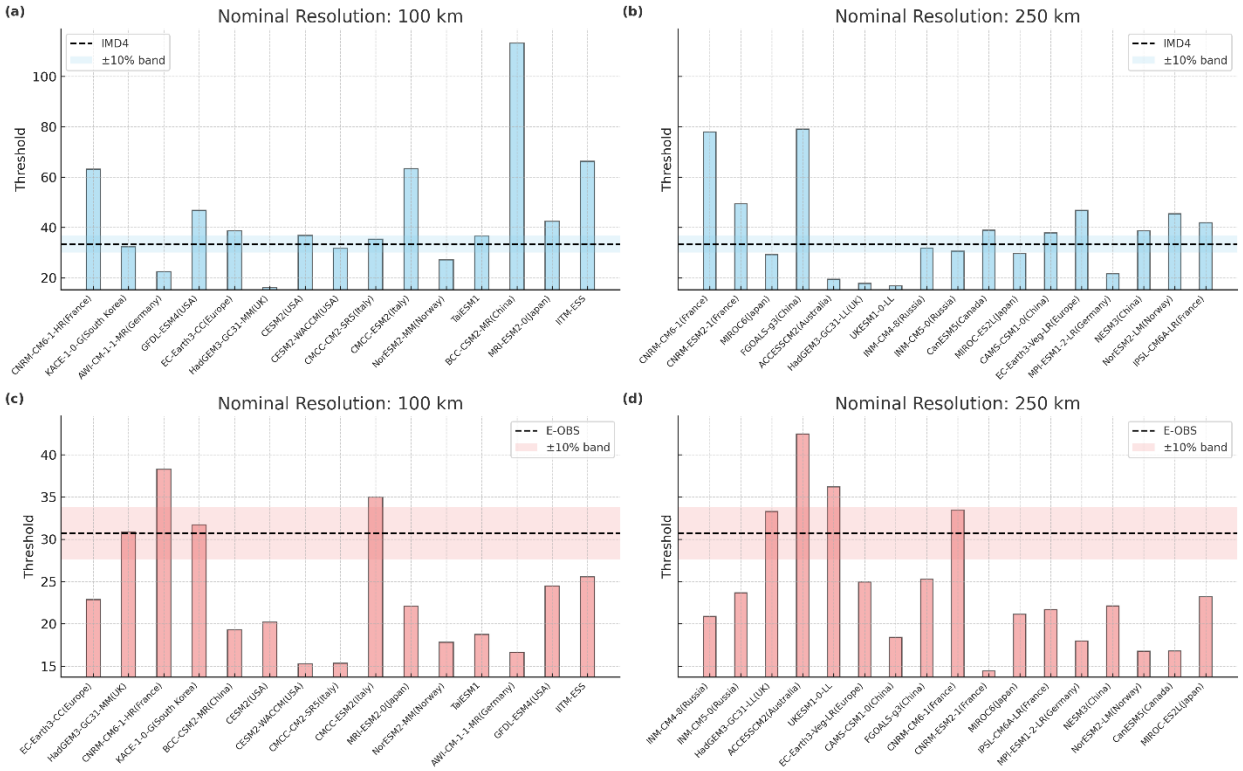


Figure 2: Comparison of 99th Percentile Threshold Value for Each Model at 100km and 250km Resolution. *Notes: Horizontal dashed lines indicate the threshold for an observation dataset (IMD4- Fig 2(a), 2(b), E-OBS- Fig 2(c),4.2(d)) and the shaded ~10% band provides a visual guide to identify models that closely match the observed value. Models are arranged in increasing distance from model data point to observation data point. Threshold value for IMD4 is often more than for observed data and vice versa for E-OBS case.*

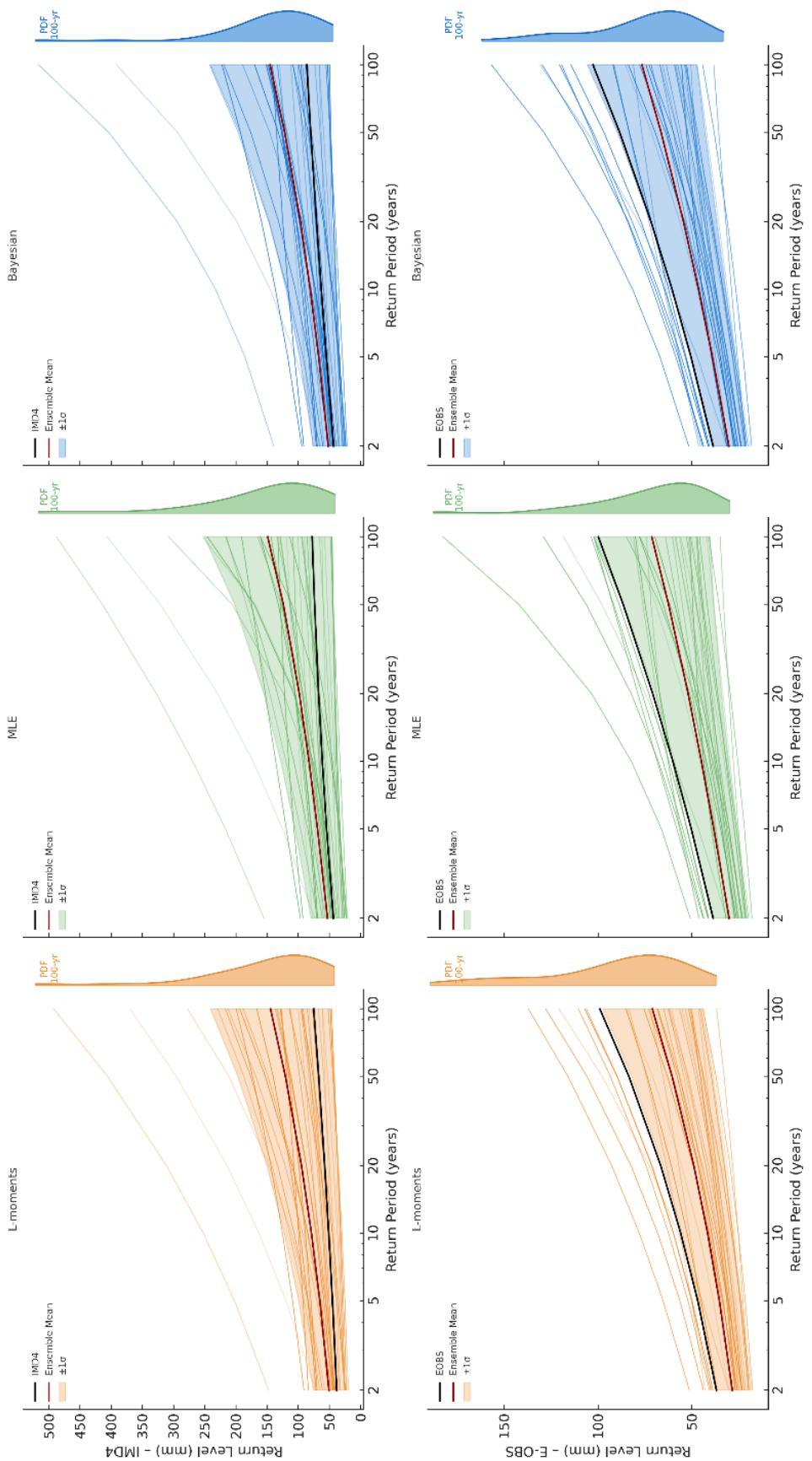


Figure 3: Return Level Estimates for Different Return Periods Using Three Parameter Estimation Methods—L-moments (a, d), Maximum Likelihood Estimation (b, e), and Bayesian Inference (c, f) for IMD4 (Top Row) and E-OBS (Bottom Row). Notes: Each panel shows return levels from individual CMIP6 models (light lines), the ensemble mean (red), and observational values (black). Shaded bands around the ensemble mean represent ~1 standard deviation. On the right of each plot is a probability density function (PDF) of the 100-year return level across models, showing ensemble spread. Return periods are shown on a log scale.

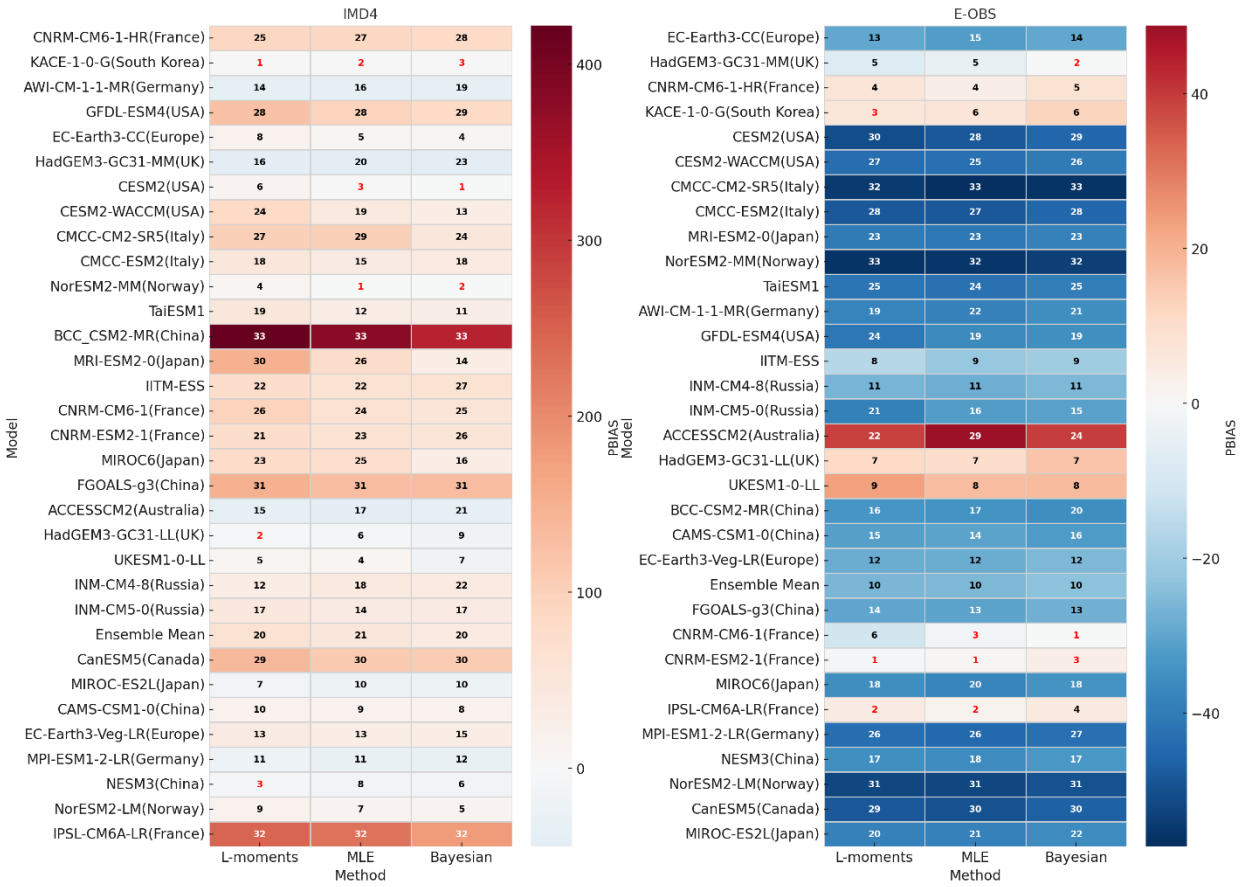


Figure 4: Model Biases in Simulating Daily Precipitation Extremes at India Site (Fig 4(a)) and Italy Site (Fig 4(b)). Notes: CMIP6 models ranked by the absolute value of Mean PBIAS (%), including the ensemble mean, grouped by resolution (100km/250km) and ordered by distance. Top three ranks for each site are written in red. Models mostly show positive bias (red shade) for the India site and mostly negative bias for the Italy site (blue shade).