Response to the Referees – EXSoDOS 1.0: downscaling of weather extremes shifts for ensemble climate projections using ground-based measurements, reanalysis and stochastic modelling

General Introduction and Summary of Major Revisions

We thank the three referees for their thorough and constructive reviews. In response, we improve the clarity, robustness, and positioning of the manuscript by (i) strengthening the quantitative evaluation of distributions and extremes, (ii) expanding the climate-change assessment with explicit tables summarizing changes in means and tails, (iii) clarifying and tabulating the methodological assumptions underlying EXSoDOS, and (iv) sharpening the novelty claim relative to existing downscaling approaches. We further (v) extend the validation scores and (vi) add robustness tests based on multi-period calibration and cross validation, (vii) clarify computational cost and data requirements, and (viii) expand the discussion of limitations, including stationarity assumptions, return period use limits, and the lack of spatial or multivariate dependence.

Below, you find the point-by-point responses to the referee's comments.

Referee 1

Referee comment: Overall, the paper is well written, and the statistical technique developed appears robust. However, there are several issues that need to be addressed before I can recommend this manuscript for publication.

Response: We thank the referee for the constructive comments, which helps us to improve the manuscript.

Referee comment: The authors have chosen 5 sites to evaluate EXSoDOS and the manuscript clearly mentions why these cases were selected. However, in my opinion only five cases globally may be insufficient to fully assess the robustness of the approach. I recommend expanding the analysis to include 20–30 stations, with representation from additional regions. Specifically, it would be valuable to incorporate sites from North and South America, as well as Australia. Including a few stations located on islands and near the Southern Ocean would further strengthen the geographic diversity of the dataset. Within India, the current focus on Puri (a coastal city) could be complemented by selecting sites from different geographical settings, such as an inland city like Delhi or a location near the Himalayas.

Response: We agree that five stations are insufficient for a comprehensive global robustness assessment. The primary objective of this manuscript is to demonstrate the EXSoDOS methodology, including calibration, validation, and application on single-station applications, rather than providing an exhaustive global benchmark. Adding more stations from multiple regions while keeping the single-station focus would make the manuscript too extensive. Yet, to further illustrate transferability, we added an additional use case in the Appendix, where precipitation is downscaled for seven randomly selected for USA with at least 60 years of observations. This additional example suggests that the workflow and performance are reproducible across regions with different climatic regimes. Nevertheless, we emphasize that for any new application, local validation remains indispensable because predictor–predictand relationships and data quality are location dependent.

Proposed manuscript text (Sect. 2.6):
"The selected use cases are intended to demonstrate the methodology and validation workflow rather than to provide an exhaustive global evaluation. For any new application, local validation remains essential because data quality and predictor–predictand relationships are location dependent."

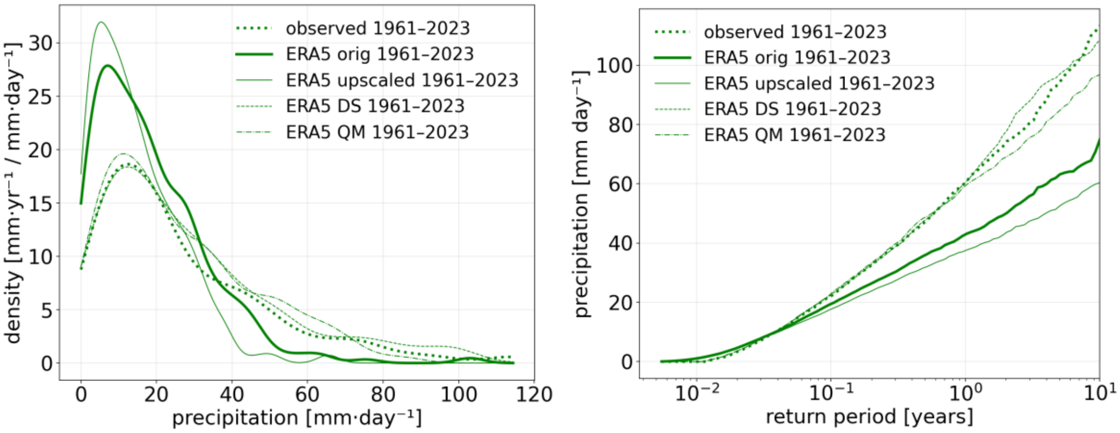


Figure S1. Idem as Fig. 4 and 5, but for combine results of 8 stations in USA, , namely USC00141593 (lat=39.5722, lon=-97.2836), USC00445050 (lat=38.0422, lon=-78.0061), USC00021664 (lat=32.0061, lon=-109.357, USC00130157 (lat=42.7536, lon=-92.8022), USC00250640 (lat=40.1306, lon=-99.8278), USC00410404 (32.1633, -95.83), USC00475808 (44.5378, -90.535).

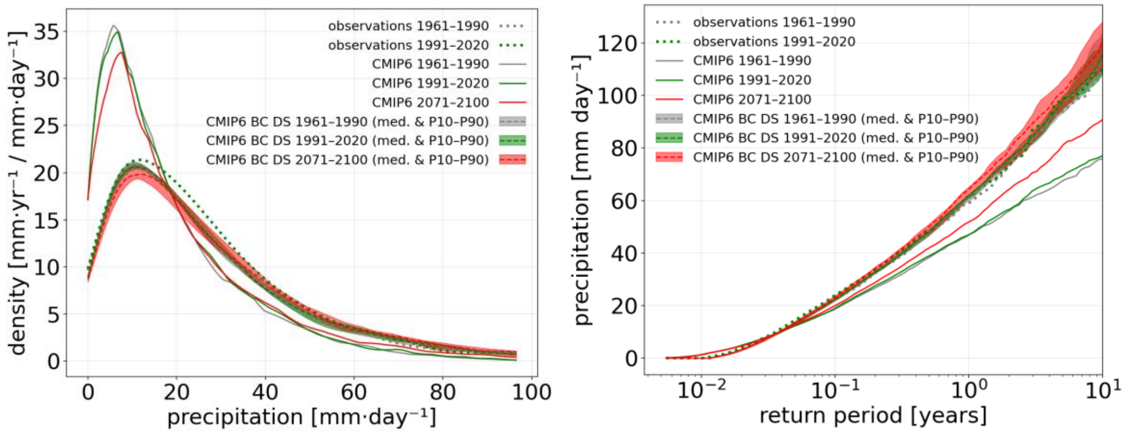


Figure S2. Idem as Fig. 6 and 7, but for combine results of 8 stations in USA.

Referee comment: Section 3.1: The current description lacks sufficient detail. While the authors state that there are differences between the four columns in Figure 3, these distinctions are difficult to visualize due to the way the figure is presented. I recommend enhancing the analysis by including one or two additional metrics and discussing the potential limitations of the Perkins distribution overlap score. (...)
Response: We thank the referee for this constructive comment. In response, we have substantially extended the quantitative validation in Sect. 3.1. In addition to the Perkins distribution overlap score, we now report a suite of complementary diagnostics that explicitly target both central tendencies and tail behaviour. These include the mean, standard deviation, 95th percentile, 1-year return level, annual number of dry days, quantile-based loss metrics, and Kolmogorov–Smirnov statistics.

This extended set of metrics allows a more objective and transparent comparison between observations, ERA5, and the different downscaling configurations, and makes differences that are visually subtle in Fig. 3 quantitatively explicit. We further added a discussion of the limitations of the Perkins overlap score, in particular its insensitivity to compensating errors and its limited ability to diagnose discrepancies in the distribution tails. The revised analysis and the new Table 3 directly address these issues.

Proposed manuscript text (Sect. 3.1):
" While the Perkins distribution overlap score provides an integrated measure of similarity between two probability density functions, it is inherently insensitive to compensating errors and offers limited insight into discrepancies in the distribution tails. Therefore, we complement the Perkins score with additional diagnostics, including quantile-based loss metrics, Kolmogorov–Smirnov statistics, and explicit indicators of extremes such as high percentiles, return levels, and dry-day frequencies. Together, these metrics provide a more complete and objective assessment of both the central tendencies and the extreme behaviour of the downscaled variables."

	Mean	Std	P95	1y return	Dry days [d/yr]	Perkins	QLoss Δ (ratio)	KS stat (p)
observed	3.3 (+0.0)	5.2 (+0.0)	21.4 (+0.0)	69.7 (+0.0)	281.6 (+0.0)	1	0.000 (1.000)	0.000 (1.00)
ERA5 orig	2.9 (-0.4)	3.5 (-1.7)	13.4 (-8.0)	38.4 (-31.3)	218.2 (-63.5)	0.53	0.146 (1.086)	0.355 (0.00)
ERA5 upscaled	2.8 (-0.5)	3.3 (-1.9)	12.1 (-9.3)	37.2 (-32.5)	206.1 (-75.5)	0.487	0.200 (1.118)	0.457 (0.00)
ERA5 DS	3.3 (-0.0)	5.2 (+0.0)	21.8 (+0.4)	64.8 (-4.8)	283.8 (+2.2)	0.953	0.000 (1.000)	0.010 (0.68)
ERA5 QM	3.1 (-0.2)	4.9 (-0.3)	20.5 (-0.9)	61.8 (-7.8)	282.3 (+0.7)	0.905	0.001 (1.001)	0.007 (0.92)

Table 3: Validation metrics for precipitation distributions from observations (Observed), original ERA5 (ERA5 orig), upscaled ERA5 (ERA5 upscaled), fully correlated stochastic downscaling (ERA5 DS), and quantile-mapping-only downscaling (ERA5 QM). For the mean, standard deviation (Std), 95th percentile (P95), 1-year return level (1y return), and annual number of dry days, absolute values are reported with deviations from observations in brackets. We further report the Perkins overlap score, the quantile loss difference (with ratio in brackets), and the Kolmogorov–Smirnov statistic with the corresponding p-value.

Referee comment: (…) Similarly, in Section 3.2, I suggest incorporating relevant statistical metrics (perhaps in the form of a table) to quantitatively assess aspects such as ‘increase’, ‘decrease’, ‘less extreme’, and ‘shift of distribution’
Response: To make all statements on increases, decreases, and distributional shifts fully quantitative, we add a new table (Sect. 3.2) summarizing annual precipitation, dry days, standard deviation, 95th percentile, and 1year return levels for observations, original CMIP6 output, and biascorrected and downscaled CMIP6 projections. Median values and 10–90% ensemble ranges are reported.

Dataset	Annual precip [mm/yr]	Dry days [d/yr]	Std	P95	1y return
Obs 1961–1990	1075	272.0	8.66	19.3	62.4
Obs 1991–2020	1173	265	8.86	21.7	62.7
CMIP6 1961–1990	861 (650–1106)	104 (36–178)	5.14 (4.45–7.52)	11.0 (6.9–13.4)	40.5 (26.9–72.1)
CMIP6 1991–2020	894 (688–1185)	103 (36–176)	5.72 (4.20–8.61)	11.2 (7.3–15.0)	43.9 (27.8–79.6)
CMIP6 2071–2100	920 (897–1348)	104 (27–179)	8.31 (6.23–12.12)	11.3 (9.34–18.0)	78.2 (51.8–103.5)

CMIP6 BC DS 1961–1990	1228 (1200–1281)	267 (265–270)	9.64 (9.35–9.95)	21.3 (20.9–22.0)	73.4 (70.7–76.9)
CMIP6 BC DS 1991–2020	1305 (1247–1316)	267 (266–270)	10.22 (9.76–10.42)	22.7 (21.7–23.2)	75.3 (74.0–77.9)
CMIP6 BC DS 2071–2100	1400 (1110–1602)	271 (267–278)	11.47 (9.85–13.31)	24.9 (20.1–28.5)	85.6 (77.1–100.9)

Table 4: Annual precipitation, dry days, standard deviation, percentile 95 value and 1-year return value of observed and modelled time series. We include the original CMIP6 climate projections (SSP585) including models listed in Tab. 1, and the bias-corrected and downscaled CMIP6 models (CMIP6 BC DS). Observations are also included as comparison. Results are shown for two historical time frames 1961–1990 and 1991–2020, and for one future timeframe 2071–2100. For the climate projections, we show median and percentile 10–90 ranges of the model ensemble.

Referee comment: The statistical framework relies on several assumptions that should be listed explicitly.

Response: We summarize the key methodological assumptions underlying EXSoDOS as an additional subsection 2.7. The key assumptions include (i) stationarity of predictor–predictand correlations, (ii) stochastic representation of subgrid variability, (iii) realism of bias-adjusted coarse scale predictors (perfect prognosis assumption), (iv) sufficient record length of observations to represent extremes up to ~10year return periods, and (v) independence across stations and variables. These assumptions are now presented in tabular form in the appendix.

Assumption	Description / rationale	Implications / how we check it
A1. Perfect-prognosis (PP) assumption	Predictors (after bias-adjustment) are assumed to be physically meaningful and transferable betweenreanalysis and GCMs (Maraun, 2016).	We mitigate PP violations by (i) bias-adjusting predictors with quantile-delta mapping (Cannon et al., 2015) per month and (ii) validating the full predictor→predictand chain against station observations.
A2. Stationary predictor–predictand dependence	The month- and category-dependent correlation ρ between normalized predictor and predictand is assumed approximately invariant under climate change (Sect. 2.2.3).	We test stability by calibrating on independent periods (1961–1990 vs 1991–2020) and even/odd-year splits; correlations and resulting tails remain within sampling uncertainty.
A3. Stochastic residual variability	Unexplained sub-grid variability is treated as a stochastic residual r drawn from a standard normal and mapped back through the empirical CDF of observations (Eq. 10–11).	This implies that tail behavior beyond the observational record cannot be guaranteed; we therefore restrict interpretation to return periods supported by record length and report

		ensemble/sampling uncertainty.
A4. Distributional representativeness of station record	Station observations are assumed long and complete enough to represent climatological distributions and extremes over 30-year windows (WMO, 2017).	We specify minimum record-length and completeness criteria, and we limit validation plots to return periods supported by sample size (Sect. 2.4).
A5. QDM bias-adjustment stationarity	Quantile-delta mapping assumes percentile-dependent model bias is approximately stationary while climate change signals in quantiles are preserved (Cannon et al., 2015).	We discuss known limitations (e.g., trend inflation when mixing scales) and keep bias adjustment on the model grid before station downscaling (Maraun, 2013).
A6. Single-site / single-variable application	EXSoDOS is applied independently per station and variable and therefore does not enforce spatial or cross-variable coherence.	We explicitly state this limitation and provide guidance on when multi-site or multivariate methods (e.g., copulas / correlation-preserving transforms; Switanek et al., 2022) are required (Sect. 4).

Table S1: Key assumptions of EXSoDOS stochastic downscaling

Minor comments

Referee comment: L22: “…generally underrepresented in climate projections.” Add 2–3 references.

Response: We add references supporting the underrepresentation of extremes in coarse-resolution climate models, for example: IPCC (2021, WGI), Sillmann et al. (2013), and Fischer & Knutti (2015). Proposed manuscript change (Introduction): “…point-scale events like heavy precipitation, heavy wind, and extreme heat are not explicitly represented in climate projections (IPCC, 2021; Sillmann et al., 2013; Fischer and Knutti, 2015).”

We note that we now state ‘not explicitly represented’ instead of ‘underrepresented’, in line with the public comments of Prof. Benestad who correctly points out that locally (point-scaled) measured and grid-scale averaged values are different by nature.

Referee comment: L38: “finer resolution grids”. It might be good to mention the resolution.

Response: We now specify typical resolutions explicitly. Proposed manuscript change (Introduction): “…statistical downscaling methods have been developed to finer-resolution grids (e.g., ~1 km for CHELSA-W5E5; Karger et al., 2023) and point observation locations, whereas typical GCM resolutions are ~1–3°.”

Referee comment: L53: “…normalized and correlated with each other.” Kindly add one sentence with a short explanation.

Response: We add clarification in Sect. 2.2: “we map predictor and predictand values to standard normal space using their empirical CDFs (i.e., apply the probability integral transform and then the inverse normal CDF), estimate correlation in that space, and then map sampled values back through the inverse empirical CDF of observations.”

Referee comment: L66: “EXSoDOS runs quickly…” Please add a metric here.

Response: We add explicit runtime metrics and clarify what is included. Proposed manuscript change (Introduction): “For a single station, a full downscaling run for one scenario typically takes ~5–10 seconds, and a 10-member ensemble completes in <1 minute on a modern CPU, excluding one-time data download and gridded bias-adjustment preprocessing. Downloading of source data (ERA5, CMIP6 climate models, station data) can take longer (network dependent) and QDM bias-adjustment (incl. ERA5 grid upscaling) on continental grids can take hours, which is performed once per model/grid.”

Referee comment: L87: Is there any reason for the 50:50 split? Citation to a previous literature might be helpful.

Response: We add the following reasoning to the manuscript: “For both calibration and validation, we require that the overall distribution and extremes to be represented on a climatological timescale. Over a 60-year period we have an equivalent of 30 year data for both, which is in line with WMO climatology assessment standards of the World Meteorological Organization (WMO, 2017).”

Referee comment: L140: How sensitive are the results to the choice for 2 additional months.

Response: Below, we show the sensitivity to the model with respect to the number of months used for calibration. The top panels show the results from original seasonal (3-monthly) calibration, and the bottom panels show the results of using single months calibration. The single months calibration tends to enlarge the climate trends for intensities between 40-60mm/day, but dampen climate trends in the high extremes with return periods > than 5 years. We argue that the results using seasonal (3-monthly) calibration are more robust since it uses a larger sample size to determine distributions and correlations between predictors and predictands, while still providing samples representative for the time of the year.

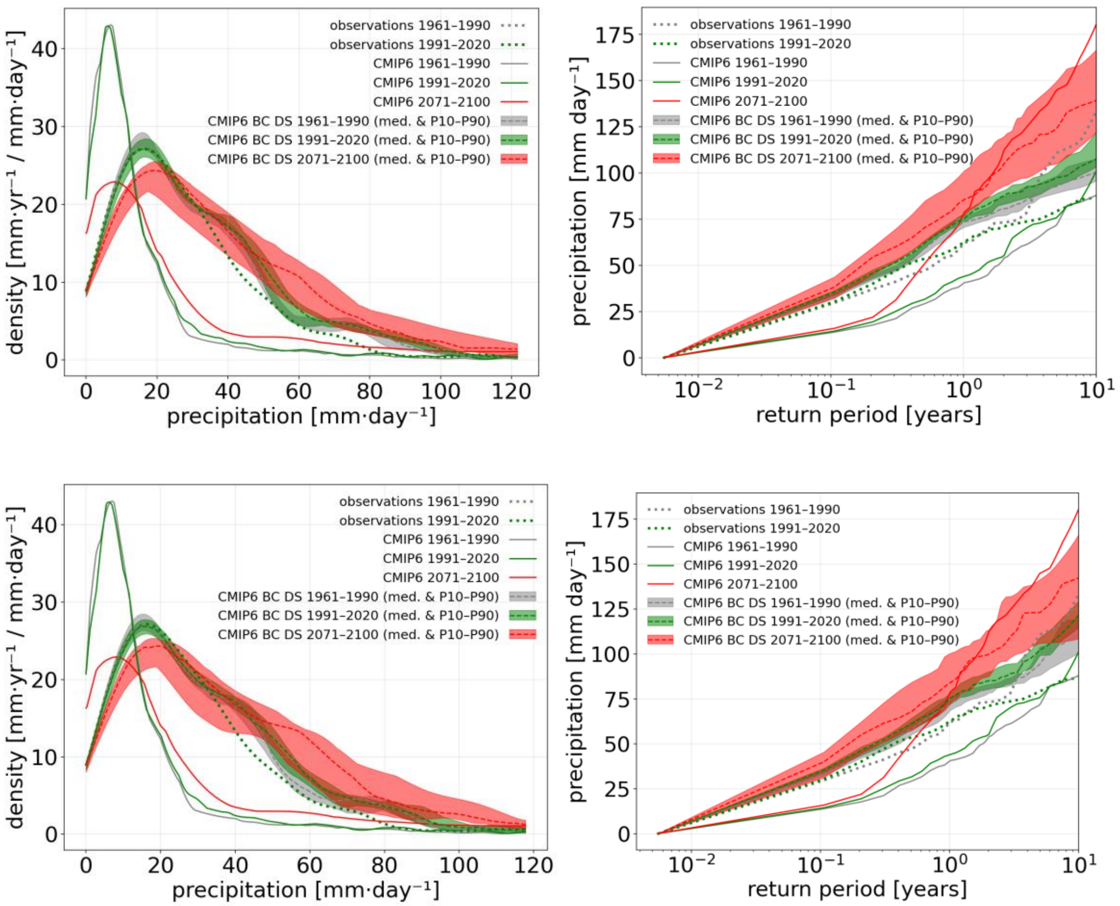


Figure: upper panels show the distribution (left) and return periods (right) from the original seasonal calibration, and lower panels show the results using single-month calibration.

Proposed manuscript text (Sect. 2.2.1, end of first paragraph): “Calibration is performed per calendar month. To increase sample size while keeping seasonality, we include data from the adjacent months (±1 month), yielding an effective 3-month seasonal window. A sensitivity test using single-month windows showed less robust tail estimates due to reduced calibration sample size per month, whereas the 3-month window provides a larger sample size and smoother transitions between months (see Appendix Fig. R1).”

Referee comment: L149: Are there existing literatures which support the choice of an exponential profile?

Response: To the author’s best knowledge, exponential (or other non-linear) profiles haven’t been considered to create predictor categories or to employ quantile delta mapping bias correction. We add to the manuscript that: “We’ve introduced an exponential profile to represent the large variation in the tails of the distribution.”

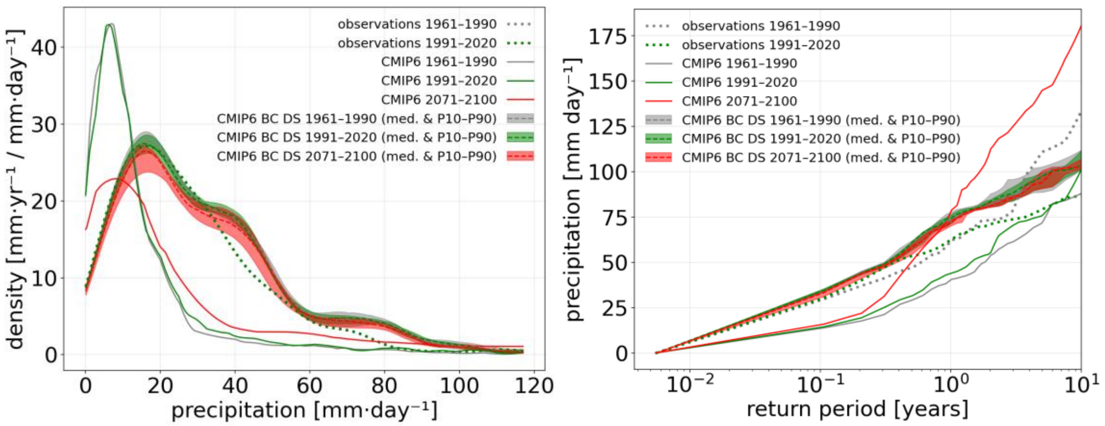
Referee comment: L192: How sensitive are the results to the choice of 'c'? Can the authors include a figure?

Response: Thanks for this remark. While the intension was to better relate predictands with their predictors, we realize that it leads to inconsistencies. For example, the rescaled variable y' ~ y/x for x >> c will correlate differently with x compared to y' ~ y for x<<c. Therefore, we remove scaling of the predictand in the revised manuscript. Therefore, sensitivity to this parameter becomes obsolete.

Referee comment: L213: I would like to see a figure/chart before and after detrending.

Response: We thank the referee for this interesting question. We suppose that it was referred to 'retrending' mentioned on L213 instead of 'detrending'.

Below, you find the results for precipitation where predictors have been detrended, but no retrending was done on the final output predictands. It shows that the overall climate change signals are much less pronounced before retrending than after retrending, and a smaller ensemble spread on the distribution and return levels were found.



Referee comment: L305-320, 325-330: Kindly add a few more citations for each case.

We make sure that each case have two references as follows:

"…many native vegetation (crop) species require the occurrence of freezing temperatures to ensure proper dormancy release and phenological development of many perennial plant species (e.g. **Evans et al., 2014; IPCC, 2021**)."

"For Spangdalhem (Germany) also in Europe, wind speed variability and extremes directly affect wind energy yield and structural loads on turbines (**Tobin et al., 2015; Pryor et al., 2012).**"

"...one of the global hotspots of extreme heat, particularly in the Middle East and South Caucasus regions (**Perkins-Kirkpatrick and Lewis, 2020; Wouters et al., 2022**)."

"Coastal India is highly vulnerable to extreme heat stress due to the combined effects of high temperature and humidity, with documented impacts on mortality and labour productivity (**Im et al., 2017; Raymond et al., 2020**)."

"As rainfall is a dichotomous variable, its inter-annual variability in time and space is difficult to assess. In Mali—and particularly in the Sikasso region—rainfall is a key driver of agricultural production, which remains a major source of livelihood for the local population. Variations in rainfall occurrence and intensity directly affect crop yields and water availability and can exacerbate vulnerability to climate hazards such as droughts and floods. Rainfall variability and the increasing relevance of heavy-rainfall extremes across the Sahel have been widely documented, with important implications for agriculture and flood risk (**Panthou et al., 2014; Sanogo et al., 2015**)"

Referee comment: Figure 3: Needs a legend.

Response: Thank you for your suggestion. We will add a legend for the revised manuscript.

Referee comment: Figures 3, 5, 6: I recommend using Kelvin as the unit of temperature.

Response: We prefer to stick to degrees Celsius, which is the unit for which weather station data is provided by the archives and weather institutes.

Referee comment: Figure 5: Check units of 'wind speed' and 'precipitation'

Response: We checked and confirmed the validity of the wind speed and precipitation units and confirmed that they are correct.

Referee comment: Figure 6, 7: Kindly consider using one common legend for the entire figure.

Response: Thanks for the suggestion. We include one legend for all the panels together in the revised manuscript.

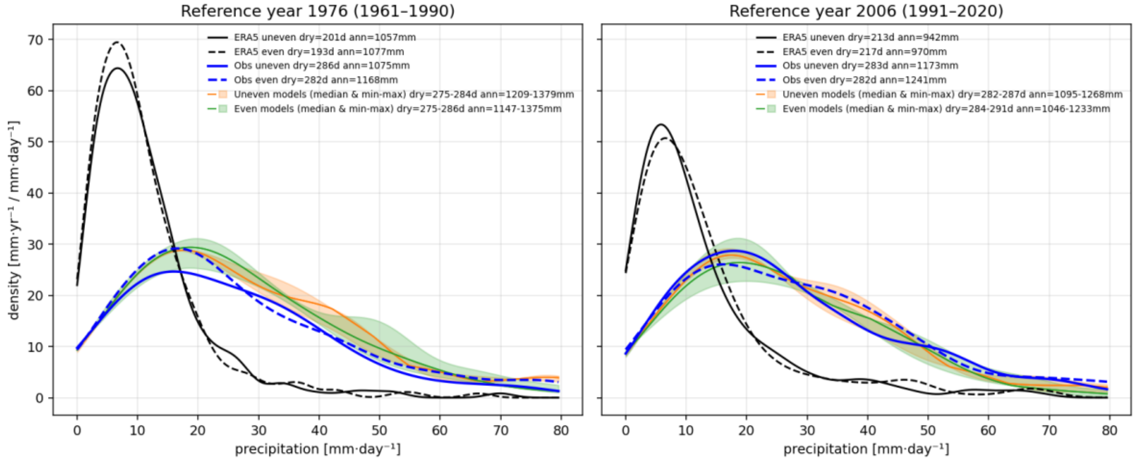
Referee 2

Referee comment: This manuscript proposes a new stochastic model to improve the prediction accuracy of climate extremes. The methodology is innovative, and the manuscript is well-written and easy to follow. However, I believe the following issues should be addressed before publication on Geoscientific Model Development.

Response: We thank the reviewer for their constructive comments, which we address below.

Referee comment: As the authors mentioned in Lines 213-215, the statistical relationship was assumed to remain unchanged under climate change. To test the robustness of the relationship, could the authors conduct cross-validation (e.g. 5-fold or leave-one-out) to evaluate the stability and generalizability of the statistical relationship?

Response: We provide robustness tests for precipitation using three calibration periods (1961–1990, 1991–2020, and the full record 1961–2023), even/odd year splits, resulting in 6 simulations per case. The spread across simulations is related to, and of the same order as, the variability over different sampling in the measurements (cfr. variability of observations across time period and even/odd years). Such a spread needs to be taken into account in climate change assessments. We include these results as an Appendix.



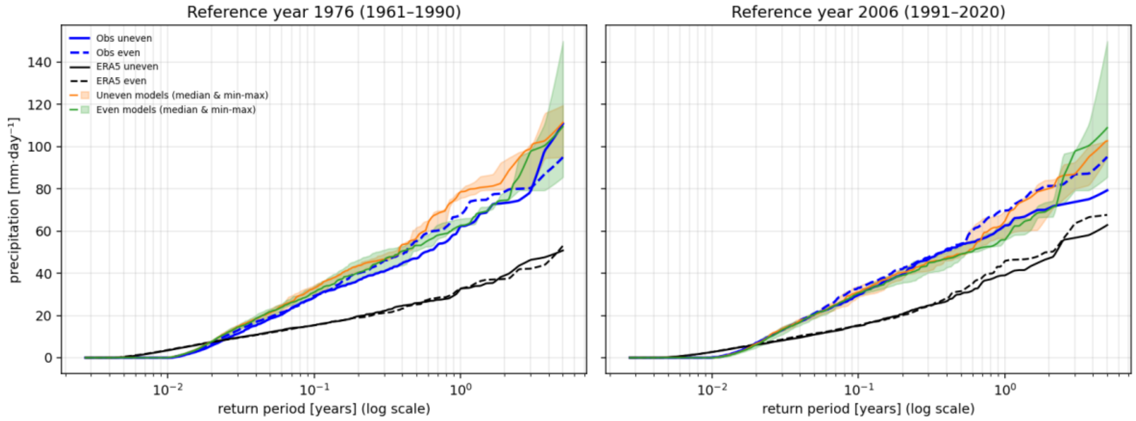


Figure: distributions (upper panels) and return levels (lower panels) for observations (Obs), ERA5 (ERA5), and downscaled results from models with the differerent calibration sets (models). We show results for even (even) and odd (uneven) years. Left panels show results for 1971-1990 and right panels show results for 1991-2020.

Station selection and landsurface characteristics

Referee comment: In addition, I am concerned about the representativeness of the five stations selected (Section 2.6.1). Why the five cases were selected to show the general utility of the model?

Response: We opted to focus on the description of the model and to demonstrate with the five cases its procedure for applications including the calibration, validation and validation. We chose five stations, rather arbitrarily, to demonstrate the utility for each of the variables. The locations are arbitrary in diverse locations in the world. The five cases are all linked to a particular challenge to climate change, which we further elaborate with 2 reference per case, see our reply to the first referee. Given these 5 cases, one can now employ the same model and procedure to any station elsewhere. To further show the applicability/transferability of EXSoDOS to other stations, we make a sixth case in the appendix which show results for 8 stations with 60-year data that are randomly chosen for the US, see below.

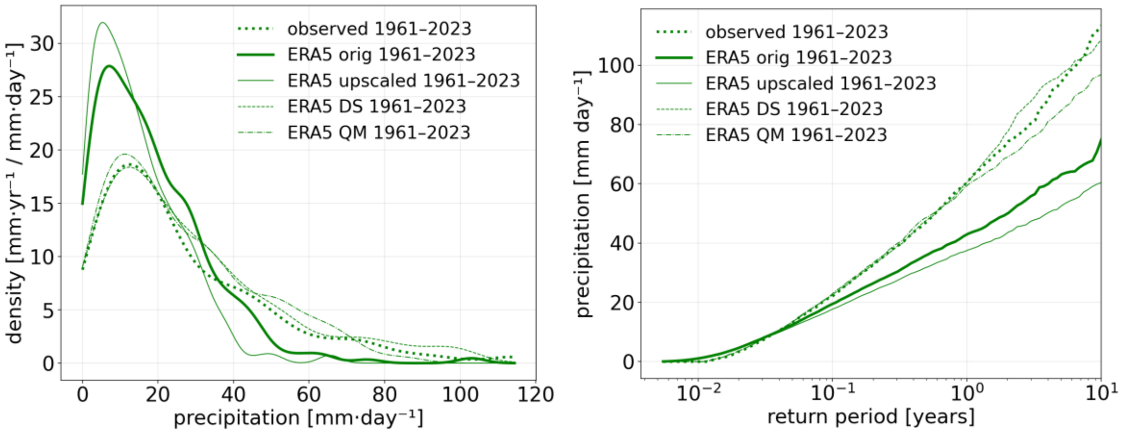


Figure S1. Idem as Fig. 4 and 5, but for combine results of 8 stations in USA.

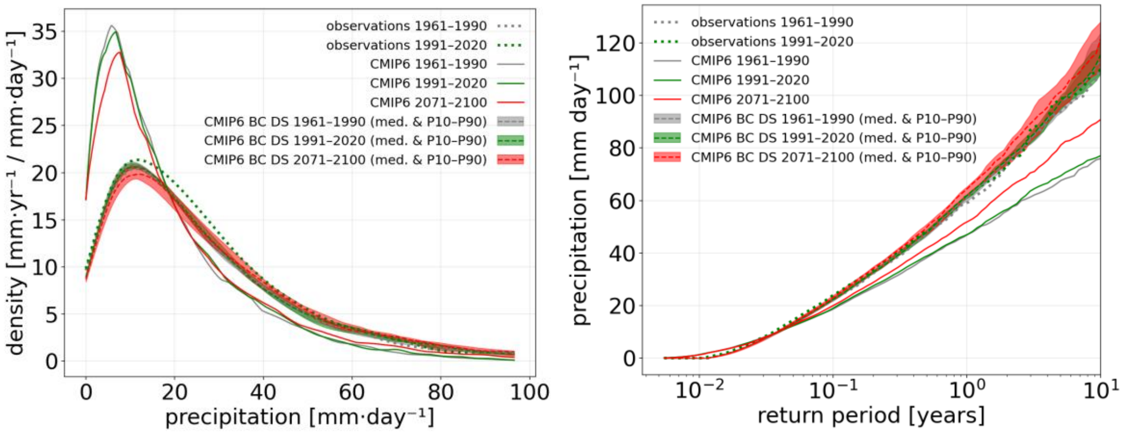


Figure S2. Idem as Fig. 6 and 7, but for combine results of 8 stations in USA, namely USC00141593 (lat=39.5722, lon=-97.2836), USC00445050 (lat=38.0422, lon=-78.0061), USC00021664 (lat=32.0061, lon=-109.357, USC00130157 (lat=42.7536, lon=-92.8022), USC00250640 (lat=40.1306, lon=-99.8278), USC00410404 (32.1633, -95.83), USC00475808 (44.5378, -90.535).

Referee comment: What are the Plant Functional Types of these stations? Understanding the PFTs could help assess whether the model performance is influenced by land-surface characteristics.

Response: We agree with the reviewer that land-surface characteristics including dominant vegetation/land cover can influence station-scale variability and extremes, and may therefore affect EXSoDOS performance. EXSoDOS implicitly captures local land-surface effects through the observed predictand record used for calibration/denormalization, but does not include explicit land-surface predictors. Reporting plant functional types is outside the scope of this paper, hence we do not include it in the manuscript. In the revised manuscript we add (Sect. 4, Limitations) a short statement that performance may depend on site characteristics, and more generally on general environmental and climatic context (e.g., adding boundary-layer stability or circulation indicators). We also clarify that extending the predictor set is a promising avenue to better represent regime-dependent land–atmosphere interactions.

Referee 3

Referee comment: This manuscript introduces EXSoDOS, a stochastic point-scale downscaling framework for representing changes in weather extremes however, several methodological assumptions, validation choices, and claims of novelty need strengthening before publication.

Response: We thank the referee for the constructive comments, which helps us to further sharpen assumptions, validation, and novelty.

Referee comment: The paper positions EXSoDOS as novel, yet its design is close to prior stochastic “perfect-prog” approaches (analog/binning, weather generators, recent station-scale frameworks). Please sharpen the novelty claim: Clarify what is conceptually new beyond the two-stage design and the pre-scaling trick; (…)

Response: We agree that the downscaling algorithm itself is similar to previous methods. EXSoDOS is positioned as novel in its explicit focus on evaluating and assessing shifts in extremes across past and future climatological periods for multiple types of variables. This was achieved by its consistent use of globally available datasets with hyper-climatological temporal coverage (ie., >>30 years) including ERA5, CMIP6 and the station data, and its non-parametric treatment of distributions. To the authors’ best knowledge, downscaling time series that can evaluate both past (ie., between 1961-1990 and 1991-2020) and future climatological shifts (ie., towards the end of the 21st century) in weather extremes haven’t been done before. In contrast to previous methods, the method that allows seasonality and including data retrieval with global coverage, the downscaling and validation is transferable to any weather station, even in locations as shown for Sikasso in Mali for which data availability is scarce but where high-quality long-term observation records are available. The method uses observed CDF instead of fitting them to particular distributions, which avoids additional assumptions, hence it allows to simulate different types of variables (Tmin,Tmax,Tmean, precip, wind and heat stress temperature).

We will highlight these novelty aspects more in the introduction and method section of the manuscript.

In the introduction:

“*To fill this gap, we present EXSoDOS, a stochastic downscaling framework designed explicitly to evaluate shifts in local weather extremes across past and future climatological periods. While the underlying stochastic downscaling approach builds on established perfect-prognosis concepts, the novelty of EXSoDOS lies in its end-to-end design for consistent assessment of extreme-event variability and return levels across multiple decades, variables, and climate states. This is achieved through the combined use of globally available long-term datasets (station observations, ERA5 reanalysis, and CMIP6 ensembles), a non-parametric treatment of distributions, and a workflow that allows direct comparison of observed, reconstructed, and projected extremes.*”

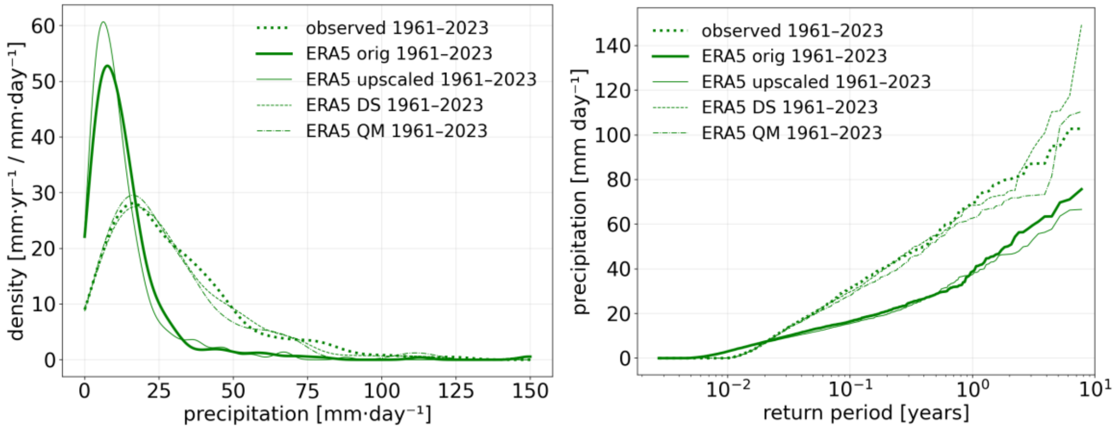
Section 2.1: "*A key design criterion of EXSoDOS is the consistent use of datasets with hyper-climatological temporal coverage (≫30 years), which enables robust estimation of variability and extremes as well as their shifts between climatological periods. By relying exclusively on globally available datasets (ERA5, CMIP6, and station observations), the framework is fully transferable to any station location where sufficiently long records exist, including data-sparse regions such as Mali.*"

Section 2.2, first paragraph, highlighting that it builds further upon previous approaches:

"The stochastic downscaling strategy follows the general philosophy of perfect-prognosis and analog-based approaches previously proposed in the literature (e.g. Volosciuk et al., 2017; Switanek et al., 2022). EXSoDOS does not introduce a fundamentally new class of stochastic models; instead, it adapts and extends these approaches within a unified framework that is explicitly designed for multi-decadal extreme-event assessment across historical and future climates."

Referee comment: (…) Add quantitative benchmarks against at least one strong baseline on the same stations/periods (e.g., BCSD-type with weather generator/analog; a Switanek-style variant without pre-scaling). (…)

Response: We add head-to-head comparison with standard quantile mapping to the distributions (Fig. 4) and return levels (Fig. 5), hence, without correlation sampling. The results for precipitation look as follows:



Referee comment: (…) Report distribution overlap, CRPS, and high-quantile loss (p ≥ 0.95).

Response: We further extend the Perkins score table (Table 2) with additional statistics and scores (also in line with the answer to comments to referee1):

	Mean	Std	P95	1y return	Dry days [d/yr]	Perkins	QLoss Δ (ratio)	KS stat (p)
observed	3.3 (+0.0)	5.2 (+0.0)	21.4 (+0.0)	69.7 (+0.0)	281.6 (+0.0)	1	0.000 (1.000)	0.000 (1.00)
ERA5 orig	2.9 (-0.4)	3.5 (-1.7)	13.4 (-8.0)	38.4 (-31.3)	218.2 (-63.5)	0.53	0.146 (1.086)	0.355 (0.00)
ERA5 upscaled	2.8 (-0.5)	3.3 (-1.9)	12.1 (-9.3)	37.2 (-32.5)	206.1 (-75.5)	0.487	0.200 (1.118)	0.457 (0.00)
ERA5 DS	3.3 (-0.0)	5.2 (+0.0)	21.8 (+0.4)	64.8 (-4.8)	283.8 (+2.2)	0.953	0.000 (1.000)	0.010 (0.68)
ERA5 QM	3.1 (-0.2)	4.9 (-0.3)	20.5 (-0.9)	61.8 (-7.8)	282.3 (+0.7)	0.905	0.001 (1.001)	0.007 (0.92)

Table 3: Validation metrics for precipitation distributions from observations (Observed), original ERA5 (ERA5 orig), upscaled ERA5 (ERA5 upscaled), fully correlated stochastic downscaling (ERA5 DS), and quantile-mapping-only downscaling (ERA5 QM). For the mean, standard deviation (Std), 95th percentile (P95), 1-year return level (1y return), and annual number of dry days, absolute values are reported with deviations from observations in brackets. We further report the Perkins overlap score, the quantile loss difference (with ratio in brackets), and the Kolmogorov–Smirnov statistic with the corresponding p-value.

While the results from to standard quantile mapping (ERA5 QM) show similar scores and distributions to the full correlated sampling (ERA5 DS), the latter one is preferred, since the former could lead to inflation of trends in extremes (Maraun, 2013). Inflation by standard quantile mapping is also suggested by our sensitivity results comparing results over different time frames 1961-1990 and 1991-2020, as discussed further below.

Referee comment: You assume correlations are static under climate change and manage range shifts via detrending (temperature) and QDM. Please: Test correlation stability across eras (1961–1990 vs 1991–2020); Discuss how non-stationarity would bias tail estimates and provide indicators to detect failure; Consider a time-varying or regime/season-partitioned correlation experiment.

Response: We test correlation stability across historical periods by calibrating the model for different periods. The first one is the full 63-year period (1961-2024) as already done, and then two additional model calibrations for the non-overlapping 30-year periods 1961-–1990 and 1991–2020. These 3 models are calibrated on uneven years and then applied to generate time series for even years. Conversely, 3 additional models are calibrated on even years to generate uneven years (‘uneven models’).

In the table below, predictor-–predictand correlations for the lowest precipitation category are reported for the different calibration sets.

	may	**jun**	**jul**	**aug**	**sep**	**oct**
1961-1990 even model	0.28	0.2	0.2	0.2	0.32	0.39
1961-1990 uneven model	0.29	0.22	0.2	0.23	0.39	0.4
1991-2020 even model	0.28	0.21	0.21	0.17	0.33	0.45
1991-2020 uneven model	0.32	0.25	0.17	0.17	0.32	0.44
1961-2023 even model	0.31	0.17	0.2	0.18	0.33	0.43
1961-2023 uneven years (reference)	0.3	0.22	0.21	0.2	0.34	0.42

Table S1: predictor-–predictand correlations for the lowest precipitation category in different months in the rainy months of Sikasso.

Correlations appear stable among the different calibrations for the different months, hence suggests invariance under climate change. However, these changes may still lead to different results in the extremes. In addition, differences also result from differences in cumulative distribution functions of the predictand used for the calibration. To test the sensitivity to the changes in the correlations and predictand CDFs, we generated time series for each of these model sets. The results in distribution and return levels can be found in the respective figures below.

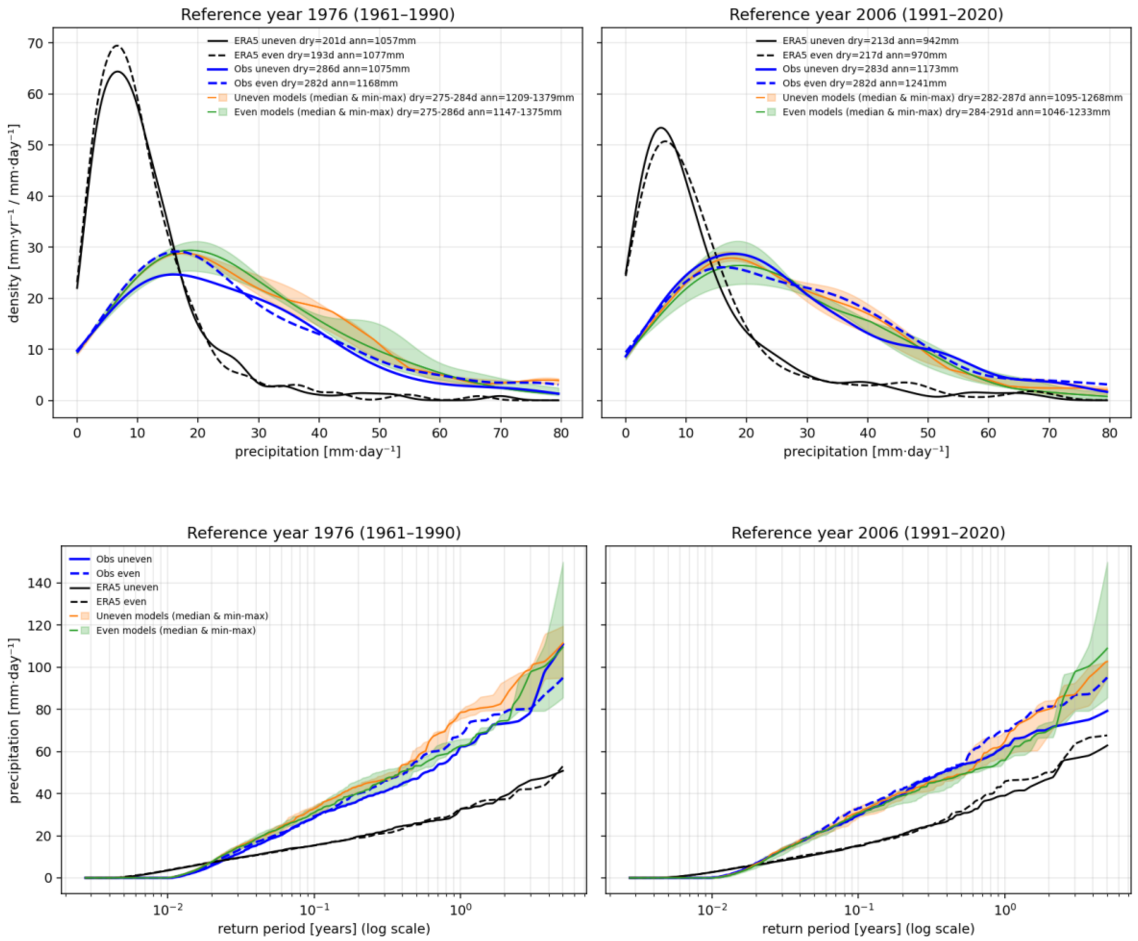


Figure: distributions (upper panels) and return levels (lower panels) for observations (Obs), ERA5 (ERA5), and downscaled results from models with the differerent calibration sets (models). We show results for even (even) and odd (uneven) years. Left panels show results for 1971-1990 and right panels show results for 1991-2020.

So we find that variations in distributions of model output (even/odd; different calibration sets) are in accordance to variations in the distributions of observation samples (even vs. odd;991-2020 vs 1961-1990). Differences are most pronounced in the tails—and the different calibration sets lead to a variability of comparable magnitude. In the revised manuscript, these robustness results are included in the Appendix (new Appendix B: cross-validation and multi-period calibration), where we provide: (i) the full set of distribution and return-level plots for the 6 simulations (3 calibration periods × even/odd), and (ii) a short summary in the main text (Sect. 3.1) stating that model uncertainty is comparable to sampling variability in the observations.

Referee comment: Validation up to ~5–10 years is reasonable given record length, but many users need 20–50-year guidance. Either demonstrate an EVT coupling (e.g., GPD) for tails or state explicit use-limits and uncertainty when extrapolating.

Response: We now explicitly state these user-limitations and uncertainty considerations in the revised manuscript in two places: (1) Sect. 2.4, where we explain that validation plots are restricted to return periods supported by record length and the calibration/validation split; and (2) Sect. 4, where we note that return periods beyond ~10 years require either substantially longer observational records or an explicit extreme-value extrapolation step (e.g., GPD/GEV), which is outside the scope of EXSoDOS v1.0. We also emphasize that users should interpret changes in the far tail as conditional on both model (GCM) uncertainty and methodological choices (e.g., detrending/retrending; see response to related comment further below).

Referee comment: EXSoDOS is applied per station and per predictor, which does not preserve spatial dependence or compound hazards (e.g., hot-humid heat stress, rain-wind). Please expand the limitations and—ideally—include a small proof-of-concept using a copula or simple correlation-preserving extension. Provide guidance on when multi-site methods are required.

Thank you for raising this important remark. EXSoDOS is indeed only applied per station and per predictor, which is applicable for single-site long-term records especially in data-scarce regions. If one needs coherent time series for multiple sites and/or multiple variables, one should introduce spatial and/or inter-variable dependence. This can be achieved with an extension that accounts for correlations among different predictors and predictands at different sites, ie., transforming (normalized) variables to an approximately independent space using the estimated correlation matrix (or a Gaussian copula), applying EXSoDOS per component, and transforming back to correlated space. This strategy is conceptually similar to the multi-site approach in Switanek et al. (2022). We add this discussion as a limitation and outlook in Sect. 4, including guidance that multi-site methods are required whenever impacts depend on spatially aggregated hazards (e.g., catchment rainfall, regional wind farms) or compound events (e.g., rain–wind, hot–humid heat stress).

Referee comment: Beyond literature review, please include at least one head-to-head comparison with an established method (e.g., standard quantile mapping, ISIMIP approach).

Response: Thanks for this suggestion. As shown in the results above, we include a comparison with a standard quantile mapping approach in the results. The method with EXSoDOS correlated sampling is preferred since it avoids inflation of trends on the predictand by the predictors (Maraun, 2013), as already mentioned in the text at L.130. We now illustrate inflation with the results for quantile mapping as shown in the figures below (in analogy to the results for the full correlated sampling above).

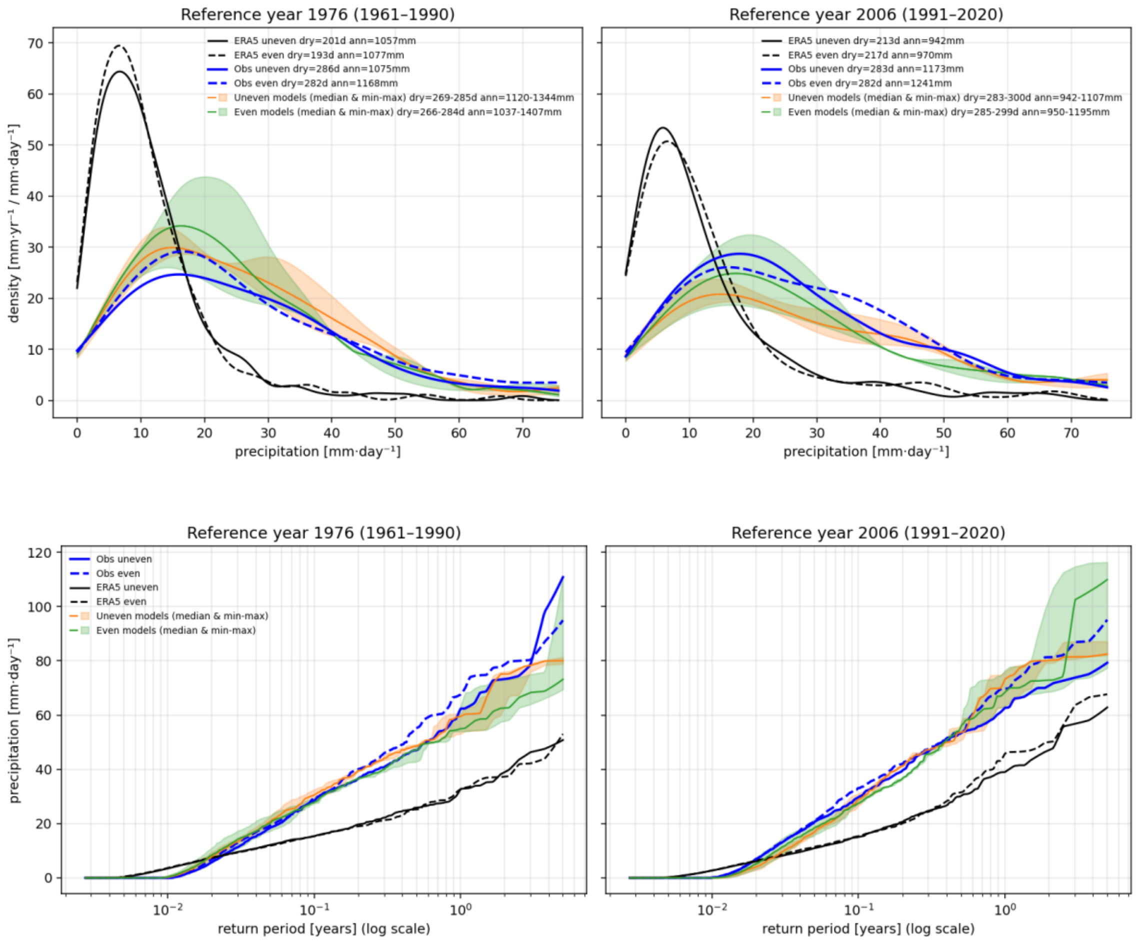


Figure: idem as figure above, but with quantile mapping instead of correlated sampling.

We find an increase in return levels in the tails (return period >= 1 year) of the predictor (ERA5 even and uneven) between the different time frames 1961-1990 and 1991-2020 (black lines). The increase in predictor return levels leads to an increase in modelled predictand return levels when using the quantile mapping approach (green and orange lines and spreads). However, return levels in the tails (return period >= 1 year) remain stable according to local observations for even years, and a decrease was found for uneven years. This discrepancy suggests inflation of outcomes by the quantile mapping approach, and such inflation is avoided when applying correlated sampling.

We also evaluate the effect of standard quantile mapping versus correlated downscaling on future climate projections. See figures and table below. It is found that climate change signals towards the end of the century are more pronounced with the quantile-mapping approach, especially in the tails, which may be due to inflation of trends. Hence, the method of statistical downscaling largely affects the outcomes, and needs to be accounted for in climate change assessments.

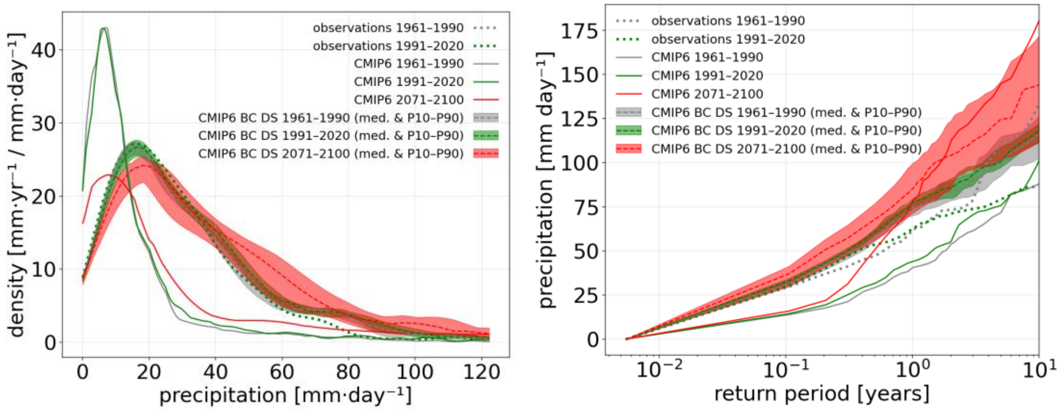


Figure: Fig. 6 and 7 for precipitation

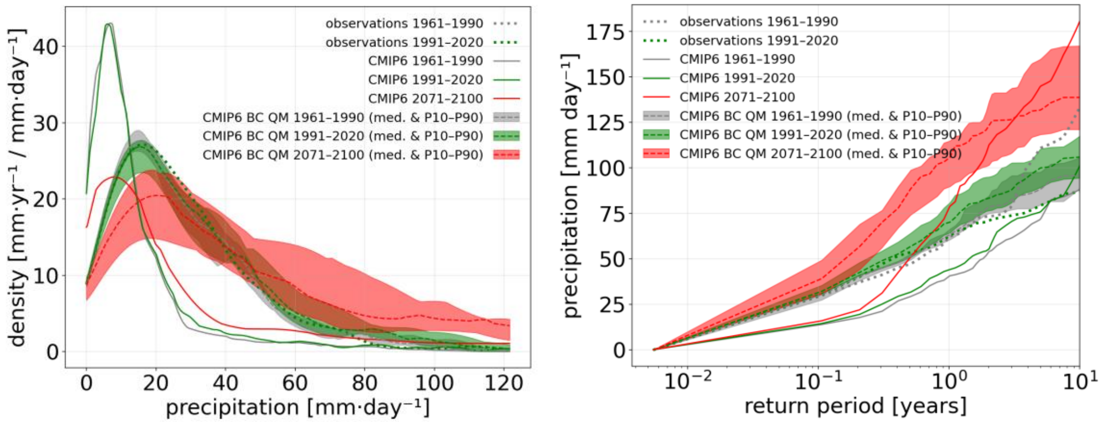


Figure: Idem as Fig. 6, 7 for precipitation but using quantile mapping for downscaling towards station level (QM).

Dataset	Annual precip [mm/yr]	Dry days [d/yr]	Std	P95	1y return
Obs 1961–1990	1075	272.0	8.66	19.3	62.4
Obs 1991–2020	1173	265	8.86	21.7	62.7
CMIP6 1961–1990	861 (650–1106)	104 (36–178)	5.14 (4.45–7.52)	11.0 (6.9–13.4)	40.5 (26.9–72.1)
CMIP6 1991–2020	894 (688–1185)	103 (36–176)	5.72 (4.20–8.61)	11.2 (7.3–15.0)	43.9 (27.8–79.6)
CMIP6 2071–2100	920 (897–1348)	104 (27–179)	8.31 (6.23–12.12)	11.3 (9.34–18.0)	78.2 (51.8–103.5)
CMIP6 BC DS 1961–1990	1228 (1200–1281)	267 (265–270)	9.64 (9.35–9.95)	21.3 (20.9–22.0)	73.4 (70.7–76.9)

CMIP6 BC DS 1991–2020	1305 (1247–1316)	267 (266–270)	10.22 (9.76–10.42)	22.7 (21.7–23.2)	75.3 (74.0–77.9)
CMIP6 BC DS 2071–2100	1400 (1110–1602)	271 (267–278)	11.47 (9.85–13.31)	24.9 (20.1–28.5)	85.6 (77.1–100.9)

Table 4: Annual precipitation, dry days, standard deviation, percentile 95 value and 1-year return value of observed and modelled time series. We include the original CMIP6 climate projections (SSP585) including models listed in Tab. 1, and the bias-corrected and downscaled CMIP6 models (CMIP6_BC_DS). Observations are also included as comparison. Results are shown for two historical time frames 1961–1990 and 1991–2020 (green), and for one future timeframe 2071–2100. For the climate projections, we show median and percentile 10–90 ranges of the model ensemble.

Dataset	**Annual precip [mm/yr]**	**Dry days [d/yr]**	**Std**	**P95**	**1y return**
CMIP6 BC QM 1961–1990	1124 (1029–1158)	267 (263–271)	8.72 (8.35–9.05)	19.5 (18.2–20.5)	63.0 (60.4–66.1)
CMIP6 BC QM 1991–2020	1182 (1140–1224)	268 (264–273)	9.40 (9.04–10.16)	20.6 (19.6–21.3)	70.1 (63.9–80.0)
CMIP6 BC QM 2071–2100	1340 (1196–1549)	275 (270–289)	13.22 (10.81–13.64)	22.5 (19.7–28.1)	105.4 (89.3–114.8)

Table. Idem as table above, but for quantile mapping.

Refereree comment: …and discuss when EXSoDOS is preferable relative to CORDEX or CHELSA-W5E5 station-scale products.

Response: We include the following text in the conclusion: 'EXSoDOS should be considered when long-term station weather station data is available and when representation extremes distribution (ie., tails) at point-scale is important to evaluate past and future climate change. In other cases, one should use existing state-of-the-art archives like CORDEX (Coppola et al., 2015) or CHELSA-W5E5 (Karger et al., 2023) enabling grid-scale climate reconstruction and projections down to 1km resolution.'

Referee comment; Specify minimum record length and completeness thresholds for station data, missing-data handling, and whether calibration is seasonal vs annual. This will aid reproducibility and user adoption.

Response: We add the following text to section 2.1: "The minimum record length needs to be 60 years, for which one can address 2 x 30 years of records to evaluate the model performance to capture past shifts in weather extremes. 30 years comes from the global standard from the World Meteorological Organization (WMO 2017). One should only use records with observational records covering >=90% of the sample period. We do not infill gaps: missing values are treated as NaN and excluded from empirical CDF correlation construction."

As already stated in the text, the bias-correction of the predictor, and the calibration is done on seasonal basis in which each month is bias-corrected / calibrated with the month before, the month itself and the month after, see L140-143 (calibration of stochastic model) L264-265 (bias adjustment of predictor).

Minor Comments

Referee comment: Figures: Panels (e.g., Figs 6–7) are visually dense; consider small multiples per variable, clearer legends, and consistent units and station names (e.g., Spangdahlem vs Dahlem). Where bands are shown, label them explicitly as ensemble percentile envelopes (10–90%) to avoid confusion with confidence intervals.

Response: We make the following changes to the figures (see example figures for precipitation above):

- remove the statistics from the figures, and put them in separate tables (extended with additional statistics for allowing for more quantitative assessment)
- explicitly mention the 10th-90th percentile envelopes.
- adapt the font size of legends and other elements.

Referee comment: Abbreviations: Spell out at first use (QDM, ERA5).

Response: We spell out QDM on first occurrence in the revised manuscript.

Referee comment: Text quality: Fix typos (e.g., "envionmental" → environmental; "algorythm" → algorithm) and standardize capitalization/diacritics.

Response: We correct these typos and increase the overall text quality in the revision

Referee comment: Data coverage: When citing a GHCN "≥30-year" requirement, specify the QC filters applied.

Response: we suggest the following QC filters at revision, namely:

Daily-minimum temperature: We will explicitly state that we use GHCN-Daily quality control by excluding values carrying non-empty quality flags and removing physically implausible values (e.g., Tmin < −90°C or > 60°C). We require sufficient completeness per 30-year window (≥90% of days present).

Daily-maximum temperature: Same QC as for Tmin (exclude non-empty quality flags; remove physically implausible values, e.g., Tmax < −90°C or > 70°C; require ≥80% completeness per 30-year window; no long-gap infilling). We also ensure internal consistency by removing days where Tmax < Tmin when both are available.

Wind speed: exclude values with non-empty quality flags, remove negative values, and require ≥90% completeness. We additionally note that wind extremes are sensitive to instrumentation and exposure changes; thus, station metadata changes (when available) should be checked and the shift assessment interpreted cautiously.

Specific Questions for the Authors

Referee comment: Data-sparse regions: How does performance degrade with shorter or lower-quality station records? Any guidance on minimum data length by variable?

Response: To perform climate-change analyses as presented in this study, one should include 2 x 30 years of data to allow evaluation of past shifts under climate change, with at least 90% of data coverage. One may consider shorter time periods, but then uncertainty on high extremes will increase drastically.

Referee comment: Sub-daily extremes: Can EXSoDOS be adapted to sub-daily metrics (e.g., hourly precipitation) and what changes would be required?

Response: Yes, it can be adapted by relating daily/hourly predictor variables to hourly predictands. Long-term hourly observations are required in that case.

Referee comment: Circulation shifts: How would systematic circulation changes (e.g., jet latitude, monsoon onset) alter predictor–predictand correlations, and can your framework detect/adapt to such shifts?

Response: You are right that circulation types can influence the predictor-to-predictand correlation! This is also suggested by the changes in correlation among the different months (see table S1 shown above) over which typical circulation types change. As such, the EXSoDOS model takes into account the seasonal-dependent correlations and CDF functions. However, these correlations and CDF functions are still considered static per month, and changes in circulation types are only considered with respect to any changes in the distributions of the coarse predictor variable. More precise circulation-dependent assessment under climate change can be done by calibrating the model for different circulation types and explicitly taking the weather type shifts into account from ERA5/GCMs, or by taking into account more predictor variables (eg., pressure, ABL stability parameters…) that link to different weather types. Machine learning algorithms (eg., neural networks) could help to overarch the complexity of the statistical relationships.

Referee comment: Computational cost: How does runtime compare to alternative downscaling methods (e.g., analog generators, quantile-mapping ensembles) for N stations and M GCMs?

Response: The runtime of EXSoDOS only takes a 10 seconds for one simulation on a modern computer. In the latest version, all processing is done as vectorize numpy/xarray operations instead pandas operations. This allows us to process multiple stations at once, as was done for 8 stations in US. The computational overhead by including additional stations is neglegible, hence the method has potential to be upscaled towards continental and global assessments. It should be noted that most of the computing time is spent on fetching (days) and biascorrecting global datasets with the quantile delta mapping (hours), but this needs to be done only once for larger continental areas. Furthermore, we find that quantile mapping is slightly faster since it doesn't require correlated random sampling. We didn't compare computational cost with other methods like weather generators, but we expect the computational cost woule be similar.

Additional changes

In line with the public comments of Prof. Benestad. We will mention statistical downcaling of probability distributions (eg., Benestad et al., 2025)

Additional references

Benestad, R. E., Parding, K. M., and Dobler, A.: Downscaling the probability of heavy rainfall over the Nordic countries, Hydrol. Earth Syst. Sci., 29, 45–65, https://doi.org/10.5194/hess-29-45-2025, 2025.

Fischer, E. M., and Knutti, R.: Anthropogenic contribution to global occurrence of heavy-precipitation and high-temperature extremes, Nature Climate Change, 5, 560–564, https://doi.org/10.1038/nclimate2617, 2015.

IPCC: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, 2021.

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., and Bronaugh, D.: Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate, Journal of Geophysical Research: Atmospheres, 118, 1716–1733, https://doi.org/10.1002/jgrd.50203, 2013.

Evans, A.M., Hsu, C.-Y., & Sheng, X. (2014). *Vernalization and the chilling requirement to exit bud dormancy: shared or separate regulation?* Frontiers in Plant Science

Pryor, S. C., Barthelmie, R. J., & Kjellström, E. (2012). Climate change impacts on wind energy: A review. *Renewable and Sustainable Energy Reviews*, 16, 430–437.

Tobin, I., Greuell, W., Jerez, S., Ludwig, F., Vautard, R., Van Vliet, M. T. H., & Breón, F.-M. (2015). Climate change impacts on the power generation potential of a European mid-century wind farms scenario. *Environmental Research Letters*, 10, 044013.

Perkins-Kirkpatrick, S. E., & Lewis, S. C. (2020). Increasing trends in regional heatwaves. *Nature Communications*, 11, 3357.

Im, E.-S., Pal, J. S., & Eltahir, E. A. B. (2017). Deadly heat waves projected in the densely populated agricultural regions of South Asia. *Science Advances*, 3, e1603322.

Raymond, C., Matthews, T., & Horton, R. M. (2020). The emergence of heat and humidity too severe for human tolerance. *Science Advances*, **6**, eaaw1838.

Panthou, G., Vischel, T., & Lebel, T. (2014). Recent trends in the regime of extreme rainfall in the Central Sahel. *International Journal of Climatology, 34*(15), 3998–4006.

Sanogo, S., Fink, A. H., Omotosho, J. A., Ba, A., Redl, R., & Ermert, V. (2015). Spatio-temporal characteristics of the recent rainfall recovery in West Africa. International Journal of Climatology, 35(15), 4589–4605.

WMO (2017). WMO Guidelines on the Calculation of Climate Normals (WMO-No. 1203). Authoritative methods, definitions, and implementation details