Response to the Referees with revision – 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 once again for their thorough and constructive reviews.

We now provide a revised version of the manuscript. This is according to our answers to the referee's comments that we have sent on december 21nd 2025. Herewith, we have updated our answers that explicitly refer to the revised manuscript, and to facilitate the review process, we include all our answers from the previous document. So the referees do not have to consider the previous document.

In the revised manuscript, we improved the clarity, robustness, and positioning of the manuscript by (i) strengthened the quantitative evaluation of distributions and extremes, (ii) expanded the climate-change assessment with explicit tables summarizing changes in means and tails, (iii) clarifed and tabulating the methodological assumptions underlying EXSoDOS, and (iv) sharpened the novelty claim relative to existing downscaling approaches. We further (v) extended the validation scores and (vi) added robustness tests based on multi-period calibration and cross validation, (vii) clarified computational cost and data requirements, and (viii) expanded 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 Supplementary, 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

Under methods – use cases (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.*"

Under conclusions:

"*The model is only evaluated for 5 sites, hence new applications require additional validation with local data. Nevertheless, the EXSoDOS, including its validation and application, is transferable to any region in the world. We further exemplify transferability by showing results for precipitation for 7 random stations across USA in the supplementary material, see resp. Figs. S1 (validation) and S1 (projection).*"

The following figures have been included as supplementary:

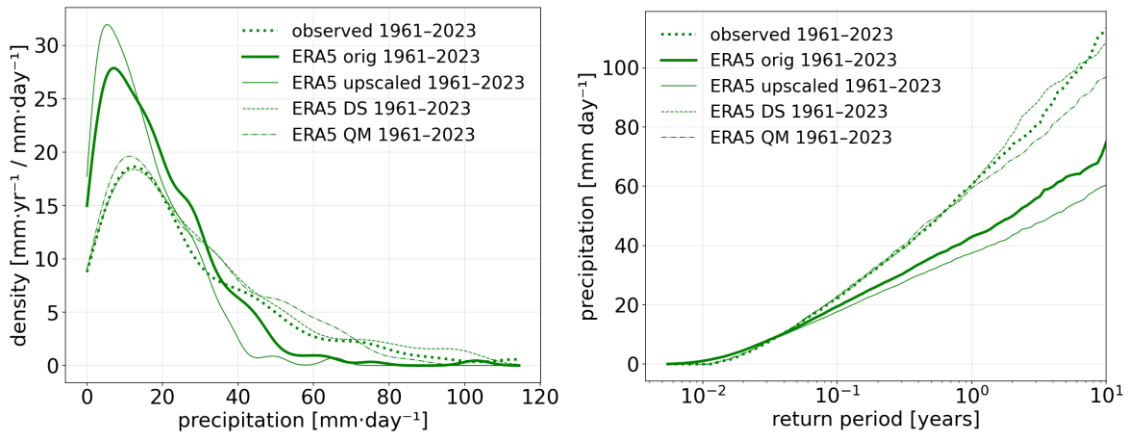


Figure S1. Idem as Fig. 4 and 5, but for combined results of 7 stations in USA, , namely USC00141593 (lat=39.5722, lon=-97.2836), USC00445050 (lat=38.0422, lon=-78.0061), USC00021664 (lat=32.0061, lon=-109.357, USC00130157 (lat=42.7536, lon=-92.8022), USC00250640 (lat=40.1306, lon=-99.8278), USC00410404 (32.1633, -95.83), USC00475808 (44.5378, -90.535).

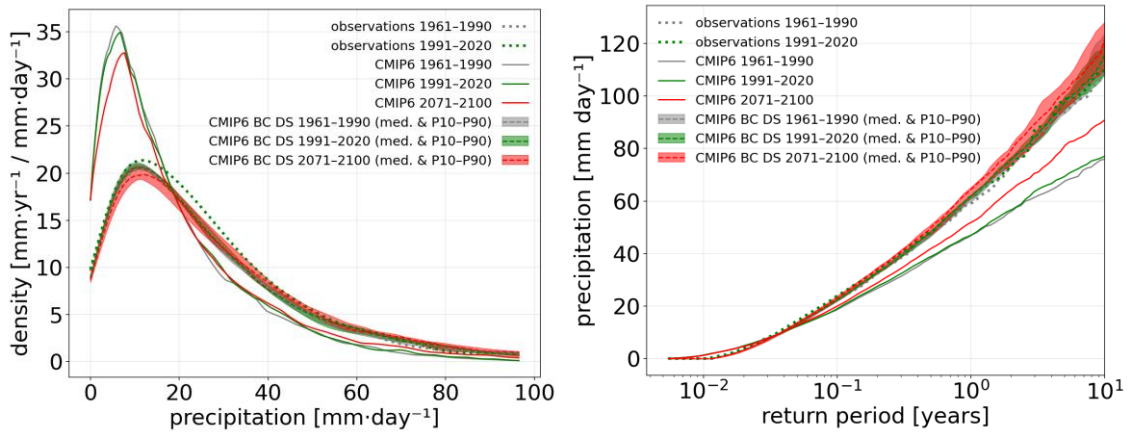


Figure S2. Idem as Fig. 6 and 7, but for combined results of 8 stations in USA.

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 behavior. 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 is included as Table 2 in the manuscript. For precipitation of Sikasso, this table looks as follows:

	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 2 (precipitation only, see manuscript for other variables): 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.

This extended set of metrics allows us for 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 also discuss 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 and added the listing of the other scores in the methods section as follows (Sect. 2.4):

"*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 distribution diagnostics, including quantile-based loss metrics, Kolmogorov–Smirnov statistics, and explicit indicators of extremes such as high percentiles, return levels (value as a function of return period, see below), and dry-day frequencies. These metrics are represented in a table and provide a comprehensive and objective assessment of both the central tendencies and the extreme behavior of the downscaled variables.*"

Given this new score table, we modified section 3.1 as follows:

"*… However, the annual cycles of ERA5 underrepresent the variability over years, as highlighted by the 10-to-90 percentile range of each day of the year in Fig. 3.* ***This is also clear from underestimated values in standard deviation, 95th percentiles and 1-year return levels (Tab. 2), especially for precipitation in Sikasso (Mali) where also the number of dry days is underestimated.*** *The underrepresentation is more subtle for the temperature variables, such as daily minimum temperature for Uccle in Belgium and the daily maximum temperature for Baku (Azerbaijan), and for heat stress temperature for Puri (India). The underrepresentation of extremes is more pronounced when ERA5 is upscaled to a coarser resolution of global climate models (1° resolution), as extreme values are averaged over larger grid sizes.*

In contrast, all results for ERA5 downscaled to the station level match better the variability of the annual cycle (Fig. 3) and distribution statistics (Tab. 2) compared to the original ERA5 and distribution statistics. Herewith, also the overall bias is removed for each variable. The improved representation of extremes by downscaling is also clear from density distributions (Fig. 4) and return periods (Fig. 5). *Distributions of precipitation after downscaling are also better matching the observations (thin dashed lines versus thick dashed lines in Fig. 4 than without downscaling of ERA5 (thick line versus thin dashed line), since they get shifted to more high extremes,* ***while also the number of dry days becomes larger and better matching observations (Tab. 2). At the same time, the downscaling reduces the overall bias in average daily precipitation for Sikasso with a downscaled value of 3.3 mm matching the observed value, whereas it is underestimated by the ERA5 value (2.9 mm). Other variables, particularly daily minimum and maximum temperature, wind speed and heat stress, also show a better match with the observed distribution after downscaling with larger spread (Fig. 4). Besides better statistics on mean and variability, the better match of the distributions after downscaling is confirmed with better score values for Perkins overlap score, the quantile loss difference and the Kolmogorov–Smirnov statistic, see Tab. 2.****"*

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 added a new table under Sect. 3.2 summarizing annual averages, dry days (for precipitation) and heat days (for heat stress), standard deviation, 95th percentile, and 1year return levels for observations, original CMIP6 output, and bias-corrected and downscaled CMIP6 projections. Median values and 10–90% ensemble ranges are also reported. The table looks as follows for daily precipitation in Sikasso (full table including other variables can be found as Tab. 3 in the manuscript):

Dataset	Window	Average	Std	P95	1y return	Dry days [d/yr]
observations	1961–1990	3.20	8.99	20.20	67.60	281.84
observations	1991–2020	3.40	9.62	22.49	69.70	281.52
CMIP6	1961–1990	2.18 (1.78–3.06)	5.14 (4.43–8.61)	9.94 (6.86–13.67)	42.39 (28.90–82.14)	229.97 (220.50–250.67)
CMIP6	1991–2020	2.35 (1.85–3.48)	5.72 (4.57–9.85)	10.86 (7.21–15.53)	47.24 (31.40–93.95)	230.83 (219.03–249.55)
CMIP6	2071–2100	2.52 (2.46–3.88)	7.89 (5.90–12.65)	11.25 (9.31–18.21)	69.75 (46.81–114.54)	235.57 (215.90–247.13)
CMIP6 BC DS	1961–1990	3.20 (3.12–3.30)	9.24 (8.82–9.30)	21.10 (20.56–22.09)	64.36 (61.90–67.14)	284.97 (282.65–285.78)
CMIP6 BC DS	1991–2020	3.40 (3.36–3.50)	9.72 (9.54–10.30)	22.54 (22.04–23.16)	69.43 (64.91–73.14)	284.48 (282.90–286.80)
CMIP6 BC DS	2071–2100	3.92 (3.28–4.26)	11.68 (10.04–12.87)	25.67 (21.93–28.59)	82.20 (70.97–90.45)	285.28 (282.83–289.37)

Table 3 (only precipitation): Annual mean, standard deviation, percentile 95 value (P95) and 1-year return level (1y return), and annual dry days 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.

We have revised section 3.2 so that it refers to the numbers in Table 3, see r. 418–451.

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

Response: We added the following text under section 2:

"The key assumptions of EXSoDOS include (see also Tab. S1)

1. *realism of bias-adjusted coarse scale predictors (perfect prognosis assumption),*
2. *stationarity of predictor–predictand correlations,*
3. *stochastic representation of subgrid variability,*
4. *sufficient record length of observations to represent extremes up to ~10year return periods,*
5. *QDM bias-adjustment stationarity (Cannon et al., 2015) and*
6. *independence across stations and variables."*

And we added the following table in the supplementary text (Tab. S1):

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 between reanalysis 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); 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	Quantile-delta mapping assumes percentile-dependent model bias is	We discuss known limitations (e.g., trend inflation when mixing scales) and keep bias

stationarity	approximately stationary while climate change signals in quantiles are preserved (Cannon et al., 2015).	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): "

…events like heavy precipitation, heavy wind, extreme heat (stress) and cold spells as observed by (point-scale) weather stations are not explicitly represented in climate projections. This results from a scale mismatch between coarse spatial resolution of GCMs and point-scale observations for which the scale of the extreme events are too small to be resolved (IPCC, 2021; Fischer and Knutti, 2015; Sillmann et al., 2013)."

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):

"*… On the other hand, statistical downscaling methods have been developed (Maraun, 2016) to finer grids of 0.5° (eg., ISIMIP3BASD by Lange, 2019) and 1km resolution (eg., CHELSA-W5E5 by Karger et al., 2023) and point observation locations (Switanek et al., 2022), whereas 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:

"*This is done by mapping 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), acquiring correlation in that space, and then map back sampled values 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 including a reference to WMO: "…*We used a 50:50 split, because we require that the overall distribution and extremes to be represented on a climatological timescale in both calibration and validation. 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.

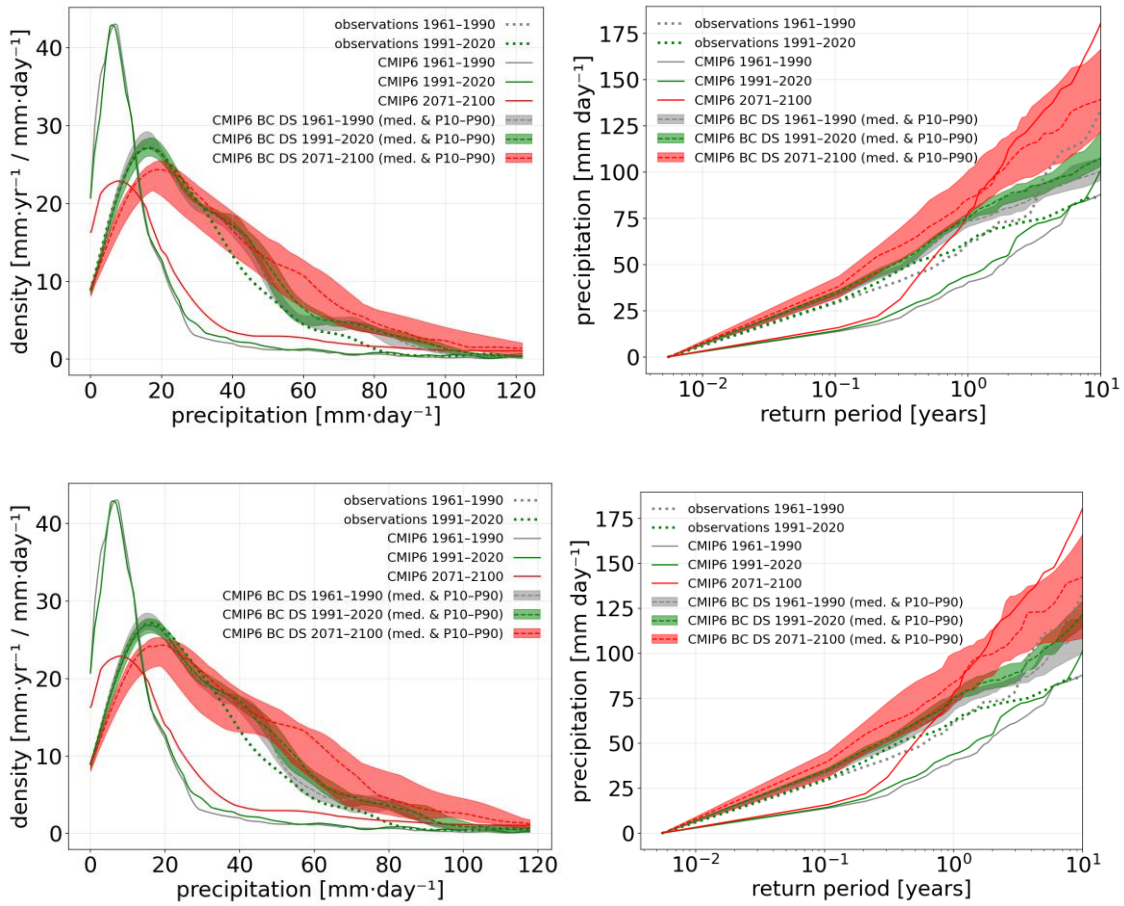


Figure: upper panels show the distribution (left) and return periods (right) from the original seasonal calibration (idem as lower right panels of resp. Fig. 6 and 7), and lower panels show the results using single-month calibration.

Proposed manuscript text (Sect. 2.2.1, end of first paragraph): "*Calibration is performed per calendar month. To increase sample size, and providing smooth transition between months while preserving seasonal representativeness, we include data from the adjacent months (±1 month), yielding an effective 3-month seasonal window. A sensitivity test using single-month windows showed sligthly less robust tail estimates due to reduced calibration sample size per month (not shown).*"

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 use an exponential profile for quantiles of category borders which provides more categories in the tails****, hence to cover the large variation in the tails of the distributions.***"

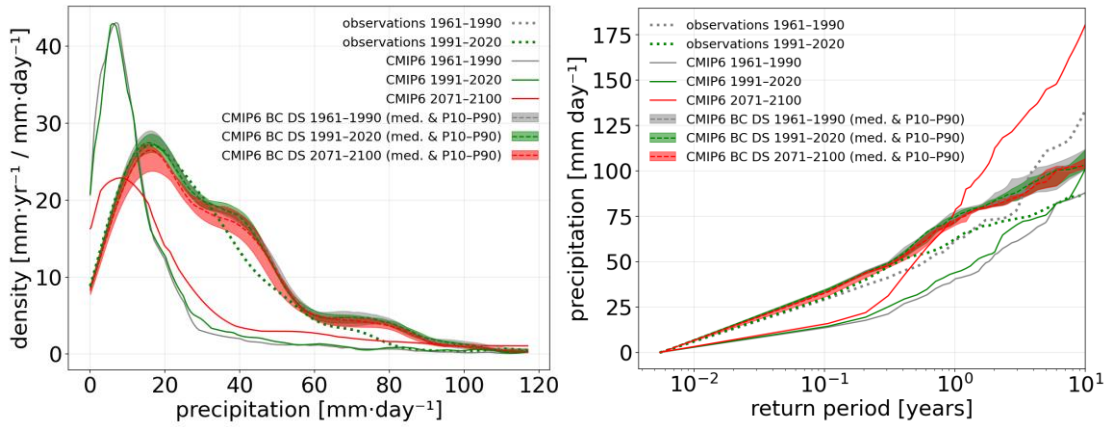
Referee comment: L192: How sensitive are the results to the choice of ‘c’? Can the authors include a figure?

Response: Thanks for this remark. While the intention 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 removed scaling of the predictand in the revised manuscript, and the 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 then 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 at least 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.* ***Brunner 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., 2016; Devis et al., 2013, Pryor et al., 2010).***"

"*…one of the global hotspots of extreme heat, particularly in the Middle East and South Caucasus regions (****eg., Fig. 1D of Wouters et al., 2022; Perkins-Kirkpatrick and Lewis, 2020;****).*"

"*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 (****eg., Fig. 1D of Wouters et al., 2022; Im et al., 2017; Raymond et al., 2020****).*"

"*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 (****Sanogo et al., 2015; Panthou et al., 2014****)*"

Referee comment: Figure 3: Needs a legend.

Response: Thank you for your suggestion. We added a legend in the upper left panel of Fig. 3.

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 now include one legend for all the panels together in the right corners of the respective figures in the revised manuscript, see Figs. 4–7.

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 daily precipitation in Sikasso using three calibration periods (1961–1990, 1991–2020, and the full record 1961–2023), even/odd year splits resulting in 6 models. We used these time blocks to maximize the size of both the calibration and validation data, because we require that the overall distribution and extremes to be represented on a climatological timescale in both calibration and validation (see also reply on comment to referee 1 regarding 50:50 split). We apply these models for the two time frames 1961–1990 and 1991-2020, see Figure S3 below. The spread across simulations is related to, and of the same order as, the variability over different sampling sets of the observations (1961–1990 vs. 1991-2020; even vs. odd years). This shows that the statistical relationships are robust under climate change. However, such a spread needs to be taken into account in climate change assessments.

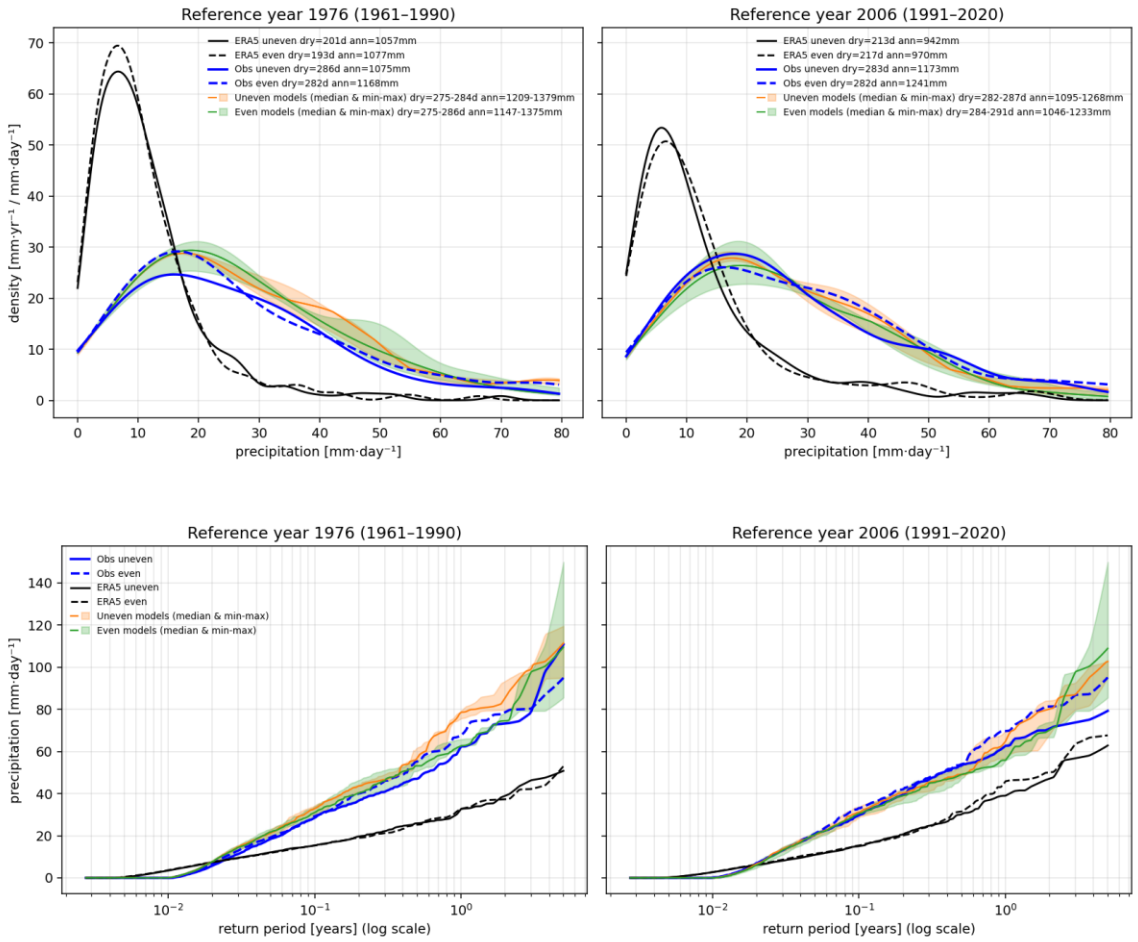


Figure S3: distributions (upper panels) and return levels (lower panels) for observations (Obs), ERA5 (ERA5), and downscaled results from models with the different calibration sets (models). We show results for even (even) and odd (uneven) years. Left panels show results for 1971-1990 and right panels show results for 1991-2020.

The results (and figure) above are included in the supporting information (also in line with a reply to a similar comment of referee 3): "***Text S1: Correlation stability and stationarity of the statistical relationship under climate change and its effect on model results***"

We refer to these results in the method section as follows:

"*Correlation stability and stationarity of the statistical relationship under climate change and its effect on model results are assessed for daily precipitation in Sikasso in Text S1, Tab. S2 and Fig. S3. It was found that the spread across simulations using 6 different calibration sets (even and odd years; 1961–2023, 1961–1990, 1991–2020) is of the same order as observation sampling sets (even vs odd years; 1961–1990 vs 1991–2020). The underlying correlation coefficient where also found to be stable. This indicates that*

statistical relationships are robust under climate change. Nevertheless, uncertainties arising from different calibration sets needs to be kept in mind in climate change assessments."

We also briefly mention the sensitivity tests in the conclusion section:

"… ***One should interpret distribution shifts, especially those in far tails, conditional on the uncertainty related to calibration sets and representativeness/stability of correlation between coarse-scale predictor and point-scale predictand under climate change (see sensitivity analysis in Text S1)****, methodological choices (e.g., detrending/retrending), and the realism of the shifts of the climate predictors at the coarse scale provided by the global climate models. …"*

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 application. 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. The five cases do not guarantee representiveness or valididity to any station in the world. Instead, as also pointed out in our reply to referee 1, one should always calibrate and validate the model again when employed for any other station in the world. Nevertheless the 5 cases are a sufficient set of examples to be able to 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 supplementary which show results for 8 stations with 60-year data that are randomly chosen for the US, see below.

Proposed manuscript text:

Under methods – use cases (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."

Under conclusions:

"*The model is only evaluated for 5 sites, hence new applications require additional validation with local data. Nevertheless, the EXSoDOS, including its validation and application, is transferable to any region in the world. We further exemplify transferability by showing results for precipitation for 7 random stations across USA in the supporting information, see resp. Figs. S1 (validation) and S2 (projection).*"

The following figures have been included as supplementary:

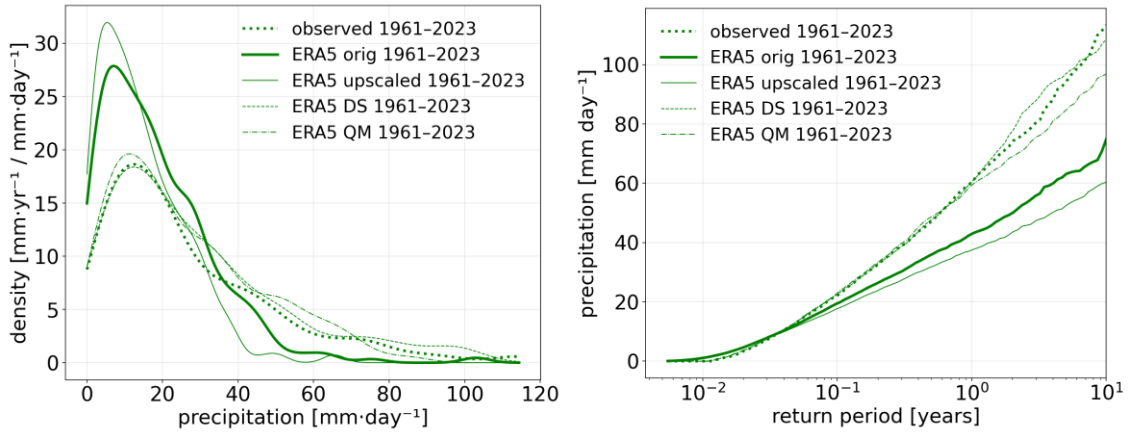


Figure S1. Idem as Fig. 4 and 5, but for combined results of 7 stations in USA, , namely USC00141593 (lat=39.5722, lon=-97.2836), USC00445050 (lat=38.0422, lon=-78.0061), USC00021664 (lat=32.0061, lon=-109.357, USC00130157 (lat=42.7536, lon=-92.8022), USC00250640 (lat=40.1306, lon=-99.8278), USC00410404 (32.1633, -95.83), USC00475808 (44.5378, -90.535).

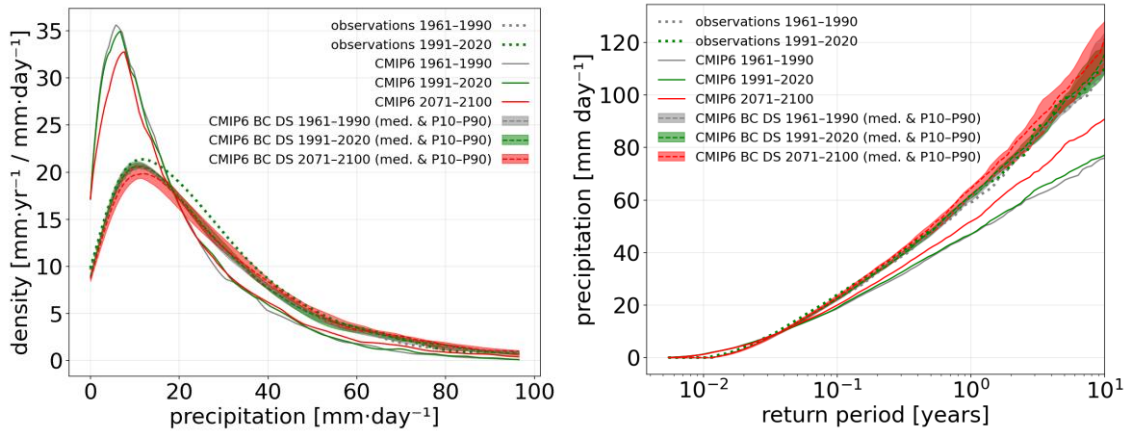


Figure S2. Idem as Fig. 6 and 7, but for combined results of 8 stations in USA.

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 and plant functional types 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. In the revised manuscript we add (under Sect. 4, conclusions) a short statement that performance may depend on site characteristics, and more generally on general environmental parameters (plant functional types, soil, urban environment) and climatic indicators (e.g., boundary-layer stability or circulation indicators…). We also clarify that extending the predictor set with these characteristics is a promising avenue to better represent regime-dependent land–atmosphere interactions. Reporting plant functional types (alongside other environmental parameters) is outside the scope of this paper, hence we do not include it in the manuscript.

We modify the text in the conclusion section as follows:

"*EXSoDOS also considers only one predictor and one predictand for one location at a time, which implies several limitations.* ***At first, the model doesn't explicitly capture its dependences on multiple local environmental (plant functional types, soil, urban environment…) and climatic parameters (e.g., boundary-layer stability or circulation indicators). As such, the complexity of physical processes may be underrepresented by the model.***

…

The stochastic model could be improved in several ways. ***… At second. extending the predictor set further with environmental and climatic parameters (mentioned above) 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., 60 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 shifts for both past (ie., between 1961-1990 and 1991-2020) and future climatological periods (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 highlight these novelty aspects more in the introduction and method section of the manuscript.

In the introduction:

“While its stochastic downscaling approach builds on established perfect-prognosis concepts, the novelty of EXSoDOS lies in its ability to evaluate shifts in local weather extremes across past and future climatological periods for different variables. This is achieved by the combined use of (1) globally available long-term datasets including weather station observations, ERA5 (ECMWF Re-Analysis 5; Hersbach et al., 2020), and the GCM (global climate model) ensemble from CMIP6 (Coupled Model Intercomparison Project Phase 6; Eyring et al., 2016}, (2) a non-parametric treatment of distributions, and a (3) workflow that allows direct comparison of observed, reconstructed, and projected

extremes. Its end-to-end design makes consistent assessment of extreme-event variability and return levels across multiple decades, and variables for any climate region possible."

Section 2.1: "*A key design feature of EXSoDOS is the consistent use of datasets with hyper-climatological temporal coverage (>60 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 a sufficiently long record of observations exists, including data-sparse regions such as Mali.*"

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

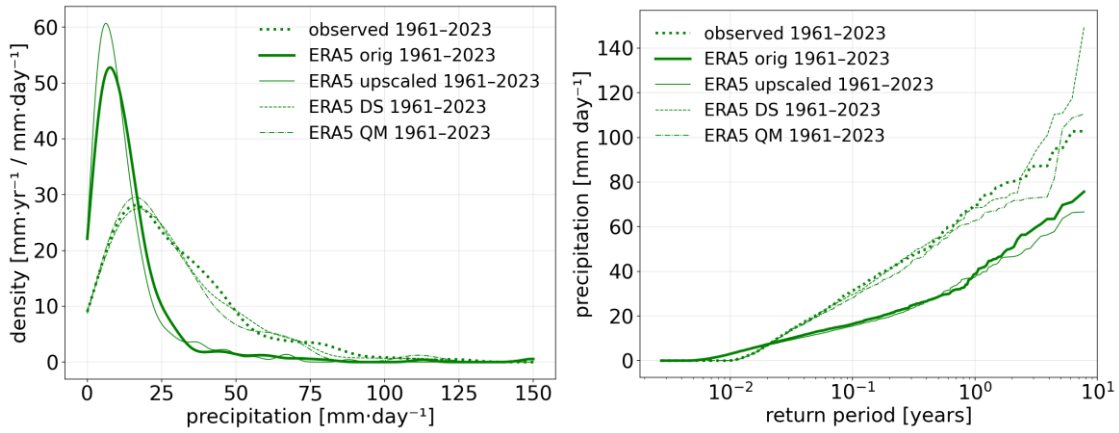
"The stochastic downscaling strategy follows the general philosophy of perfect-prognosis and analog-based approaches previously proposed in the literature (e.g. Volosciuk et al., 2017; Switanek et al., 2022). EXSoDOS does not introduce a fundamentally new class of stochastic models; instead, it adapts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

Finally, we note that the key novelty of EXSoDOS was already highlighted in the abstract:

"*Existing downscale products successfully reduce overall biases of past or future climateological variables, but the representation of variability and extreme events including their past and future shifts under climate change are still not addressed. A new stochastic model, EXSoDOS, addresses this gap by the DOwnScaling of weather EXtremes Shifts for ensemble climate projections using ground-based measurements, reanalysis, and global climate models.*"

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). Report distribution overlap, CRPS, and high-quantile loss ($p \geq 0.95$).

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:



We further extended the Perkins score table (Table 2) with additional statistics and scores (also in line with the answer to comments to referee1), also reporting the scores of quantile mapping (QM). Here is the table showing the scores for daily precipitation in Sikasso (other cases can be found in the full Table 2):

	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 (precipitation only): 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.

Given the new figures and table, we discuss the comparison with standard quantile mapping as follows:

"*Results and scores with standard quantile mapping (ERA5 QM) are added next to the full correlated sampling (ERA5 DS), see Figs. 4 and 5. While the results from the 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 we discuss in one of our replies 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 in case of daily precipitation for Sikasso. The first one is the full 63-year period (1961-2023) 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 odd years and then applied to generate time series for even years ('even models'). Conversely, 3 additional models are calibrated on even years to generate odd years ('uneven models').

In Tab. S2, 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 S2: 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 Fig. S3.

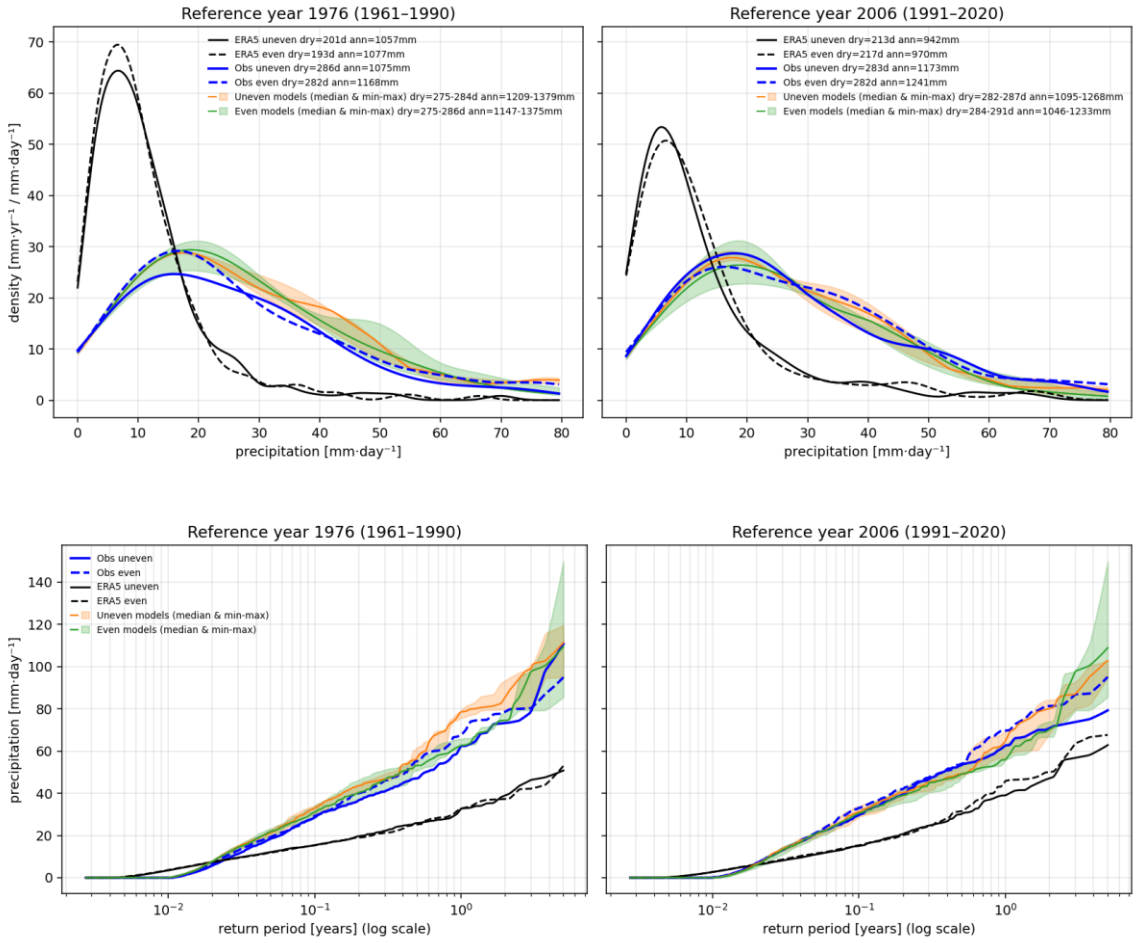


Figure S3: distributions (upper panels) and return levels (lower panels) for observations (Obs), ERA5 (ERA5), and downscaled results from models with the different calibration sets (models). We show results for even (even) and odd (uneven) years. Left panels show results for 1961-1990 and right panels show results for 1991-2020.

We find that variations in distributions of model output (even/odd; different calibration sets) are of the same order to variations in the distributions of observation samples (even vs. Odd years;1991-2020 vs 1961-1990). Differences are most pronounced in the tails—and the different calibration sets lead to a variability of comparable magnitude. Such sensitivity to calibration sets needs to be taken into account in climate model assessments.

The results above are included in the supporting information (also in line with a reply to a similar comment of referee 2): "***Text S1: Correlation stability and stationarity of the statistical relationship under climate change and its effect on model results"***

We refer to these results in the method section as follows:

"*Correlation stability and stationarity of the statistical relationship under climate change and its effect on model results are assessed for daily precipitation in Sikasso in Text S1, Tab. S2 and Fig. S3. It was found that the spread across simulations using 6 different calibration sets (even and odd years; 1961–2023, 1961–1990, 1991–2020) is of the same order as observation sampling sets (even vs odd years; 1961–1990 vs 1991–2020). The underlying correlation coefficient where also found to be stable. This indicates that statistical relationships are robust under climate change. Nevertheless, uncertainties arising from different calibration sets needs to be kept in mind in climate change assessments.*"

We also briefly mention the sensitivity tests in the conclusion section:

"… ***One should interpret distribution shifts, especially those in far tails, conditional on the uncertainty related to calibration sets and representativeness/stability of correlation between coarse-scale predictor and point-scale predictand under climate change (see sensitivity analysis in Text S1)****, methodological choices (e.g., detrending/retrending), and the realism of the shifts of the climate predictors at the coarse scale provided by the global climate models. …"*

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 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 up to 10-year return periods supported by record length and the calibration/validation split:

"*To only retain statistically relevant results for the current validation period 1961-2023, we only show return periods of up to 10 years which lead to an averaging over least 3 validation samples for that maximum period over the 63-year period, ie., 63 years divided by 10 years, and divided by two since only half of the measurements are used for the validation.*"

2. In conclusions (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). We also emphasize that users should interpret changes in the far tail as conditional on the uncertainty to calibration sets (discussed in our reply above):

"One should keep in mind the limitations of statistical downscaling when employing the model for future climate assessment. ***In this manuscript, the model is calibrated and validated with 63 year records, only providing assessments for return periods of up to 10 years. For higher return periods, one requires either substantially longer observational records or an explicit extreme-value extrapolation step requiring additional assumptions (e.g., employing Generalized Pareto Distribution or Generalized Extreme Value Distribution)****. One should interpret distribution shifts, especially those in far tails, conditional on the* ***uncertainty related to calibration sets and representativeness/stability of correlation between coarse-scale predictor and point-scale predictand under climate change (see sensitivity analysis in Text S1),*** *methodological choices (e.g., detrending/retrending), and the realism of the shifts of the climate predictors at the coarse scale provided by the global climate models."*

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. We agree, so we expanded the limitations (under sect. 4 conclusions) as follows:

"*EXSoDOS also considers only one predictor and one predictand for one location at a time, which implies several limitations. At first, the model doesn't explicitly capture its dependences on multiple local environmental (plant functional types, soil, urban environment…) and climatic parameters (e.g., boundary-layer stability or circulation indicators). As such, the complexity of physical processes may be underrepresented by the model. To perform a more in-depth analysis of changing extremes and their underlying physics, one should still rely on mechanistic high-resolution atmospheric numerical modelling. At second, the simulation of only one output variable (predictand) at a time ignores possible correlation among them, so hampers consistent representation of compound hazards (e.g., heavy drought–heat, heavy rain–wind…). At third, modelling one location at a time does not preserve spatial dependence between multiple sites (eg., a single convective rainstorm affecting multiple sites).*"

We also include perspectives (small proof-of-concept) for representing compound events and spatial dependence:

"*The stochastic model could be improved in several ways. At first, coherent time series for multiple sites and/or multiple variables can be achieved by correlating different predictors and predictands for different variables and different sites. This can by normalizing predictands and predictors, subsequently transforming them to independent variables using their correlation matrix (or Gaussian copula). After such calibration, one combines predictor variables with random sampling and transforms the variables back to the correlated space, and finally one denormalizes them again to their respective distributions. Multi-variable correlated sampling makes representation of compound hazards and spatial dependence between different sites possible. Such a strategy that introduces a correlation matrix over different predictors and predictands is conceptually similar to the multi-site approach proposed by Switanek et al., 2022.*"

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 our earlier reply above, we have included 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 and in the results above. We now would like to 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).

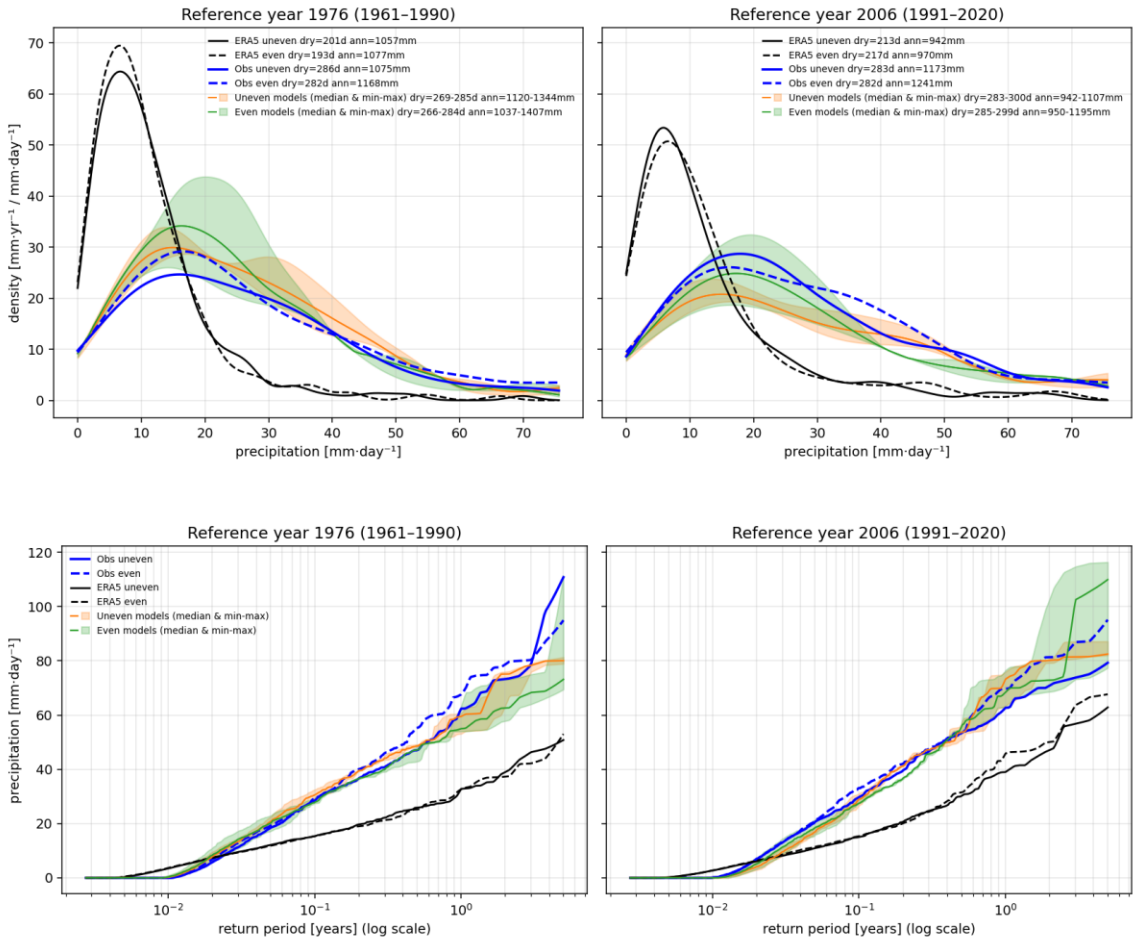


Figure S4: idem as fig. S3, but with quantile mapping instead of correlated sampling.

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 does not occur when applying correlated sampling as shown before.

Quantile mapping versus correlated sampling affects the results on the climate projections. To illustrate this, we evaluate the effect of standard quantile mapping versus correlated downscaling on future climate projections. See Fig. S5 and Tab S3 shown 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.

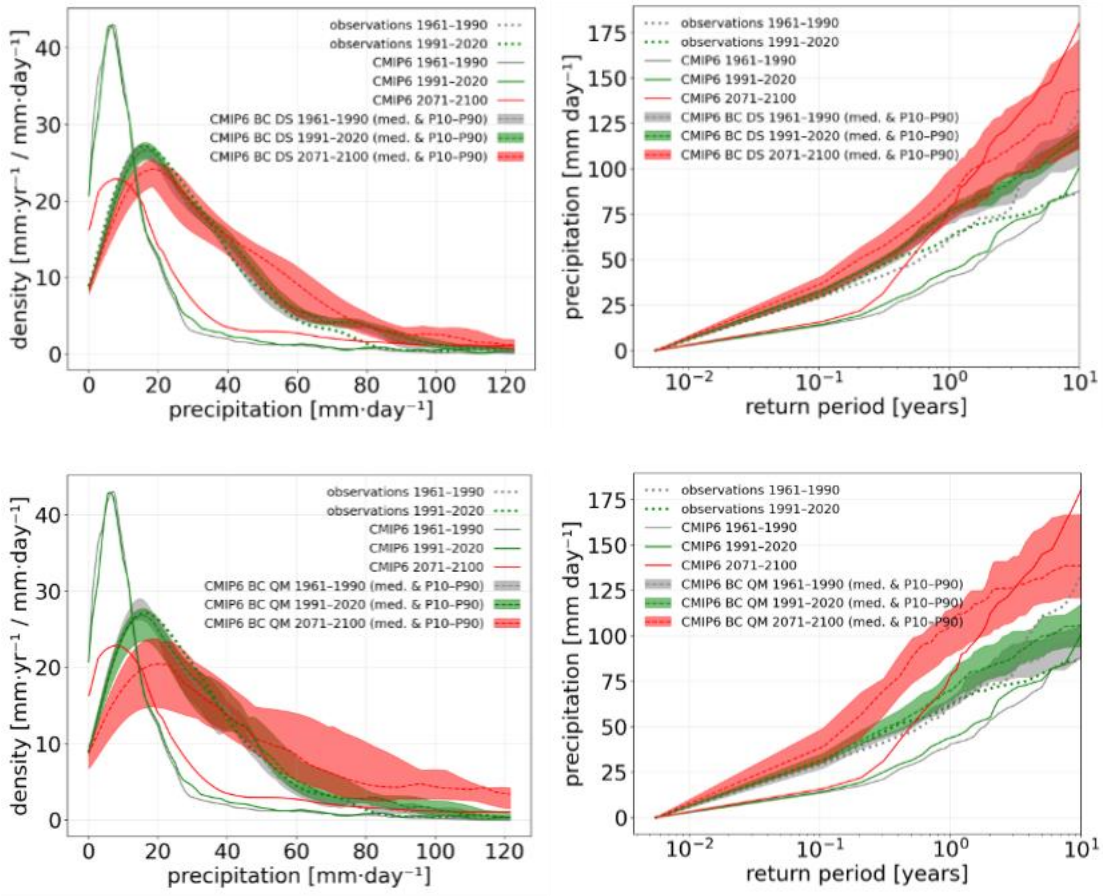


Figure S5: upper panels are the lower right panels of resp Fig. 6 and 7 (precipitation for Sikasso). The lower panels are the same but using quantile mapping for downscaling towards station level (QM).

Dataset	Window	Average	Std	P95	1y return	Dry days [d/yr]
observations	1961–1990	3.20	8.99	20.20	67.60	281.84
observations	1991–2020	3.40	9.62	22.49	69.70	281.52
CMIP6	1961–1990	2.18 (1.78–3.06)	5.14 (4.43–8.61)	9.94 (6.86–13.67)	42.39 (28.90–82.14)	229.97 (220.50–250.67)
CMIP6	1991–2020	2.35 (1.85–3.48)	5.72 (4.57–9.85)	10.86 (7.21–15.53)	47.24 (31.40–93.95)	230.83 (219.03–249.55)
CMIP6	2071–2100	2.52 (2.46–3.88)	7.89 (5.90–12.65)	11.25 (9.31–18.21)	69.75 (46.81–114.54)	235.57 (215.90–247.13)
CMIP6 BC DS	1961–1990	3.20 (3.12–3.30)	9.24 (8.82–9.30)	21.10 (20.56–22.09)	64.36 (61.90–67.14)	284.97 (282.65–285.78)
CMIP6 BC DS	1991–2020	3.40 (3.36–3.50)	9.72 (9.54–10.30)	22.54 (22.04–23.16)	69.43 (64.91–73.14)	284.48 (282.90–286.80)
CMIP6 BC DS	2071–2100	3.92 (3.28–4.26)	11.68 (10.04–12.87)	25.67 (21.93–28.59)	82.20 (70.97–90.45)	285.28 (282.83–289.37)
1961–1990	CMIP6 BC QM	2.88 (2.82–3.11)	8.19 (7.95–8.54)	18.90 (18.18–20.53)	58.56 (56.41–61.16)	283.17 (282.53–287.07)
1991–2020	CMIP6 BC QM	3.17 (3.08–3.25)	9.02 (8.51–9.94)	20.68 (19.95–20.90)	66.73 (60.75–75.50)	283.10 (281.18–287.53)
2071–2100	CMIP6 BC QM	3.69 (3.11–4.09)	12.60 (9.64–13.41)	22.52 (19.90–26.95)	96.59 (75.31–102.14)	287.18 (286.07–295.28)

Table S3: Idem as Tab. 3 in main text (only daily precipitation for Sikasso), but appending results for quantile mapping (CMIP6 BC QM) to the results with correlated sampling (CMIP6 BC DS).

We have included the above results in the supporting information as **"Text S2: comparison between quantile mapping to correlated sampling; illustration of inflation**". We also refer to these results in the main text (results) as follows:

"While the results from the 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 for

precipitation in Sikasso is also illustrated in the supporting text S2 with the help of Figs. S3, S4 and S5, and Tab. S3."

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., 2021) or CHELSA-W5E5 (Karger et al., 2023) providing 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. On the one hand, the 60 year record is required to provide sufficient data for both calibration and validation. As a 50:50 split is applied, an equivalent of 30 years of data is available for both calibration and validation, which is in line climatology assessment standards of the World Meteorological Organization (WMO, 2017). On the other hand, one can address 2 x 30 years of records to evaluate the model performance to capture past shifts in weather extremes driven by global climate models. 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 made the following improvements to the figures (see Figs. 4–7 of the revised manuscript):

- common legend for all panels, which is displayed as a separate panel in the upper right corner of the figures
- remove the statistics from the figures. These can now be found in separate tables 2 and 3 (extended with additional statistics scores for allowing for more quantitative assessment)
- explicitly mention the 10th-90th percentile envelopes.
- increased 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. We also spelled out ERA5 (EWMF ReAnalys 5) on first occurrence in abstract and main text (also for CMIP6 (Coupled Model Intercomparison Project Phase 6).

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: For the precipitation in Mali, the records have been quality filtered by MALI-METEO. For the 4 cases, we haven’t applied extra QC filter from the GHCN/global_hourly dataset. Nevertheless, we will consider the following QC filters in future applications, 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. Sensitivity to record length (30-year vs 60-year calibration for either even or uneven years) can be seen for daily precipitation by the spread results for either the even or uneven models as shown in one of our earlier replies.

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. However, long-term hourly observations and predictors are required in that case, and one may consider predictors on both daily and hourly time scales.

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 with correlation results) 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. These future perspectives have now been added to the conclusions section.

“EXSoDOS also considers only one predictor and one predictand for one location at a time, which implies several limitations. At first, the model doesn't explicitly capture its

dependences on multiple local environmental (plant functional types, soil, urban environment…) and ***climatic parameters (e.g., boundary-layer stability or circulation indicators)****. As such, complexity of physical processes may be underrepresented by the model. …*

At second, extending the predictor set further with environmental and climatic parameters (mentioned above) is a promising avenue to better represent regime-dependent land–atmosphere interactions."

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: In line with a reply of comment of referee 1, added the following information:

"EXSoDOS runs quickly and automatically. 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 quantile-delta mapping (QDM) bias-adjustment (incl. ERA5 grid upscaling) on continental grids can take hours, which is performed once per model/grid."

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 would be similar.

Additional references

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.

Brunner, A. M., Evans, L. M., Hsu, C.-Y., and Sheng, X.: Vernalization and the Chilling Requirement to Exit Bud Dormancy: Shared or Separate Regulation?, Frontiers in Plant Science, 5, https://doi.org/10.3389/fpls.2014.00732, 2014.

Pryor, S. C. and Barthelmie, R. J.: Climate Change Impacts on Wind Energy: A Review, Renewable and Sustainable Energy Reviews, 14, 430–437, https://doi.org/10.1016/j.rser.2009.07.028, 2010

Tobin, I., Jerez, S., Vautard, R., Thais, F., van Meijgaard, E., Prein, A., Déqué, M., Kotlarski, S., Maule, C. F., Nikulin, G., Noël, T., and Teichmann, C.: Climate Change Impacts on the Power Generation Potential of a European Mid-Century Wind Farms Scenario, Environmental Research Letters, 11, 034 013, https://doi.org/10.1088/1748-9326/11/3/034013, 2016.

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